



HAL
open science

Conditional Density Estimation by Penalized Likelihood Model Selection and Applications

Serge X. Cohen, Erwan Le Pennec

► **To cite this version:**

Serge X. Cohen, Erwan Le Pennec. Conditional Density Estimation by Penalized Likelihood Model Selection and Applications. [Research Report] RR-7596, INRIA Saclay, équipe SELECT; IPANEMA. 2011. inria-00575462v4

HAL Id: inria-00575462

<https://inria.hal.science/inria-00575462v4>

Submitted on 9 Dec 2011 (v4), last revised 9 Jul 2012 (v5)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Conditional Density Estimation by Penalized Likelihood Model Selection and Applications

S. X. Cohen (IPANEMA / Synchrotron Soleil)
and

E. Le Pennec (SELECT / Inria Saclay - Île de France and Université Paris Sud)

December 9, 2011

Abstract

In this technical report, we consider conditional density estimation with a maximum likelihood approach. Under weak assumptions, we obtain a theoretical bound for a Kullback-Leibler type loss for a single model maximum likelihood estimate. We use a penalized model selection technique to select a best model within a collection. We give a general condition on penalty choice that leads to oracle type inequality for the resulting estimate. This construction is applied to two examples of partition-based conditional density models, models in which the conditional density depends only in a piecewise manner from the covariate. The first example relies on classical piecewise polynomial densities while the second uses Gaussian mixtures with varying mixing proportion but same mixture components. We show how this last case is related to an unsupervised segmentation application that has been the source of our motivation to this study.

1 Introduction

Assume we observe n couples $((X_i, Y_i))_{1 \leq i \leq n}$ of random variables, we are interested in estimating the law of the second variable $Y_i \in \mathcal{Y}$ conditionally to the first one $X_i \in \mathcal{X}$. In this paper, we assume that the couples (X_i, Y_i) are independent while Y_i depends on X_i through its law. More precisely, we assume that the covariates X_i s are independent but not necessarily identically distributed. The assumption on the Y_i s are stronger: we assume that, conditionally to the X_i s, they are independents and each variable Y_i follows a law with density $s_0(\cdot|X_i)$ with respect to a common known measure $d\lambda$. Our goal is to estimate this two-variables conditional density function $s_0(\cdot|\cdot)$ from the observations.

This problem has been introduced by Rosenblatt [38] in the late 60's. He considered a stationary framework in which $s_0(y|x)$ is linked to the supposed existing densities $s_0(x)$ and $s_0(x, y)$ of respectively X_i and (X_i, Y_i) by

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

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 seems to exist before the mid 90's with a study of a spline tensor based maximum likelihood estimator proposed by Stone [40] and a bias correction of Rosenblatt's estimator due to Hyndman et al. [28].

Kernel based method have been much studied since. For instance, Fan et al. [19] and de Gooijer and Zerom [14] consider local polynomial estimator, Hall et al. [23] study a locally logistic estimator that is later extended by Hyndman and Yao [27]. In this setting, pointwise convergence properties are considered, and extensions to dependent data are often obtained. The results depend however on a critical bandwidth that should be chosen according to the regularity of the unknown conditional density. Its practical choice is rarely discussed with the notable exceptions of Bashtannyk and Hyndman [5], Fan and Yim [18] and Hall et al. [24]. Extensions to censored cases have also been discussed for instance by van Keilegom and Veraverbeke [44]. See for instance Li and Racine [32] for a comprehensive review of this topic.

In the approach of Stone [40], the conditional density is estimated through a parametrized modelization. This idea has been reused since by Györfi and Kohler [22] with an histogram based approach, by Efromovich [16, 17] with a Fourier basis, and by Brunel et al. [12] and Akakpo and Lacour [2] with piecewise polynomial representation. Those authors are able to control an integrated estimation error: with a an integrated total variation loss for the first one and a quadratic distance loss for the others. Furthermore, in the quadratic framework, they manage to construct adaptive estimators, estimators that do not require the knowledge of the regularity to be minimax optimal (up to a logarithmic factor), using respectively a blockwise attenuation principle and a model selection by penalization approach. Note that Brunel et al. [12] extend their result to censored cases while Akakpo and Lacour [2] are able to consider weakly dependent data.

In this paper, we consider a direct estimation of the conditional density function through a maximum likelihood approach. Although natural, this approach has been considered so far only by Stone [40] as mentioned before and by Blanchard et al. [10] in a classification setting with histogram type estimators. Given a set S_m of candidate functions, our estimate \hat{s}_m will be simply the maximum likelihood estimate

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

Our aim is then to analyze the finite sample performance of such an estimator in term of Kullback-Leibler type loss. As often, a trade-off between a bias term measuring the closeness of s_0 to the set S_m and a variance term depending on the complexity of the set S_m and on the sample size appears. A good set S_m will be thus one for which this trade-off leads to a small risk bound. Using a penalized model selection approach, we propose then a way to select a best model S_m^{\wedge} amongst a collection $\mathcal{S} = (S_m)_{m \in \mathcal{M}}$. For a given family of penalties $\operatorname{pen}(m)$, we define the *best* model S_m^{\wedge} as the one that minimized

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

The main result of this paper is a sufficient condition on the penalty $\operatorname{pen}(m)$ such that for any density function s_0 and any sample size n the adaptive estimate \hat{s}_m^{\wedge} performs almost as well as the best one in the family $\{\hat{s}_m\}_{m \in \mathcal{M}}$.

The very frequent use of conditional density estimation in econometrics, see Li and Racine [32] for instance, could have provided a sufficient motivation for this study. However it turns out that this work stems from a completely different subject: unsupervised hyperspectral image segmentation. Using the synchrotron beam of Soleil, the IPANEMA platform[6], for which one of the author works, is able to acquire high quality hyperspectral images, high resolution images for which a spectrum is measured at each pixel location. This provides rapidly a huge

amount of data for which an automatic processing is almost necessary. One of this processing is the segmentation of these images into homogeneous zones, so that the spectral analysis can be performed on fewer places and the geometrical structures can be exhibited. The most classical unsupervised classification rely on the density estimation of Gaussian mixture by a maximum likelihood principle. The component of the estimated mixtures will correspond to classes. We have extended this method by taking into account the localization of the pixel in the mixture weight, going thus from density estimation to conditional density estimation. As stressed by Maugis and Michel [35], understanding finely the density estimator is crucial to be able to select the right number of classes. This theoretical work has been motivated by a similar issue for the conditional density estimation case.

Section 2 Single model maximum likelihood estimate is devoted to the analysis of the maximum likelihood estimation in a single model. It starts by Section 2.1 Framework and notation in which the setting and some notation are given. The risk of the maximum likelihood in the classical case of *misspecified* parametric model is recalled in Section 2.2 Asymptotic analysis of a parametric model. Section 2.3 Jensen-Kullback-Leibler divergence and bracketing entropy provides some tools required for the extension of this analysis to more general setting presented in Section 2.4 Single model maximum likelihood estimation. We focus then in 3 Model selection and penalized maximum likelihood Model selection and penalized maximum likelihood to the multiple model case. The penalty used is described in Section 3.1 Framework while the main theorem is given in Section 3.2 A general theorem for penalized maximum likelihood conditional density estimation. Section 4 Partition-based conditional density models introduces partition-based conditional density estimator: we use model in which the conditional density depends from the covariate only in a piecewise constant manner. We study in details two instances of such model: one in which, conditionally to the covariate, the densities are piecewise polynomial for the Y variable and the other, which corresponds to our hyperspectral image segmentation motivation, in which, again conditionally to the covariate, the densities are Gaussian mixtures with the same mixture components but different mixture weights.

2 Single model maximum likelihood estimate

2.1 Framework and notation

Our statistical framework is the following: we observe n independent couples $((X_i, Y_i))_{1 \leq i \leq n} \in (\mathcal{X}, \mathcal{Y})^n$ where the X_i s are independent, but not necessarily of the 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 $d\lambda$. For any model S_m of candidate conditional densities, we estimate s_0 by the conditional density \hat{s}_m that maximize 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:

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

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

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

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 the known measure $d\lambda$, we will use the following notation

$$KL_\lambda(s, t) = KL(sd\lambda, td\lambda) = \begin{cases} \int_\Omega \frac{s}{t} \ln \frac{s}{t} td\lambda & \text{if } sd\lambda \ll td\lambda \Leftrightarrow \forall \omega \in \Omega, s(\omega) = 0 \implies t(\omega) = 0 \\ +\infty & \text{otherwise.} \end{cases}$$

Remark that, contrary to the quadratic loss, this divergence is an intrinsic quality measure: it does not depend on the reference measure $d\lambda$ but only on the probability law. As we are working with conditional densities and not with classical densities, the previous divergence should be adapted. To take into account the structure of conditional densities and the design of $(X_i)_{1 \leq i \leq n}$, we define a phantom sample $(X'_i)_{1 \leq i \leq n}$ independent of $(X_i)_{1 \leq i \leq n}$ but with the same law and 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 one 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 $KL_\lambda^{\otimes n}(s, t) = \mathbb{E} [KL_\lambda(s(\cdot|X'_i), t(\cdot|X'_i))]$.

Note also that as we are using a phantom sample, we can consider functions that depend on the observations and the design and still have an integrated type loss.

2.2 Asymptotic analysis of a parametric model

Assume that S_m is a parametric model,

$$S_m = \{s_{\theta_m}(y|x) | \theta_m \in \Theta_m \subset \mathbb{R}^{\mathcal{D}_m}\},$$

to which the true conditional density s_0 does not necessarily belong. In this case, if we let

$$\hat{\theta}_m = \operatorname{argmin}_{\theta_m \in \Theta_m} \left(\sum_{i=1}^n -\ln(s_{\theta_m}(Y_i|X_i)) \right)$$

then $\hat{s}_m = s_{\hat{\theta}_m}$. White [45] has studied this *misspecified model* setting for density estimation but its results can easily be extended to the conditional density case.

If the model is identifiable and under some (strong) regularity assumption on $\theta_m \mapsto s_{\theta_m}$, provided the $\mathcal{D}_m \times \mathcal{D}_m$ matrices $A(\theta_m)$ and $B(\theta_m)$ defined by

$$A(\theta_m)_{k,l} = \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \int \frac{\partial^2 -\log s_{\theta_m}(y|X_i)}{\partial \theta_{m,k} \partial \theta_{m,l}}(y|X_i) s_0(y|X_i) d\lambda \right]$$

$$B(\theta_m)_{k,l} = \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \int \frac{\partial \log s_{\theta_m}(y|X_i)}{\partial \theta_{m,k}} \frac{\partial \log s_{\theta_m}(y|X_i)}{\partial \theta_{m,l}} s_0(y|X_i) d\lambda \right]$$

exists, the analysis of White [45] implies that, if we let

$$\theta_m^* = \operatorname{argmin}_{\theta_m \in \Theta_m} KL_\lambda^{\otimes n}(s_0, s_{\theta_m}),$$

$\mathbb{E} [KL_\lambda^{\otimes n}(s_0, \widehat{s}_m)]$ is asymptotically equivalent to

$$KL_\lambda^{\otimes n}(s_0, s_{\theta_m^*}) + \frac{1}{2n} \operatorname{Tr}(B(\theta_m^*)A(\theta_m^*)^{-1}).$$

When s_0 belongs to the model, i.e. $s_0 = s_{\theta_m^*}$, $B(\theta_m^*) = A(\theta_m^*)$ and the previous equivalent can be rewritten as the classical parametric one

$$\min_{\theta_m} KL_\lambda^{\otimes n}(s_0, s_{\theta_m}) + \frac{1}{2n} \mathcal{D}_m.$$

This simple expression does not hold when s_0 does not belong to the parametric model as $\operatorname{Tr}(B(\theta_m^*)A(\theta_m^*)^{-1})$ cannot be simplified.

A short glimpse on the proof of the previous result shows that it depends heavily on the asymptotic normality of $\sqrt{n}(\widehat{\theta}_m - \theta_m^*)$. One may wonder if extension of this result, often called the Wilk's phenomenon [46], exists when this normality does not hold, for instance in non parametric case or when the model is not identifiable. Along this lines, Fan et al. [20] have proposed a generalization of the corresponding Chi-Square goodness-of-fit test in several setting and Boucheron and Massart [11] study the finite sample deviation of the corresponding empirical quantity in a bounded loss settings.

Our aim is to derive a non asymptotic upper bound of type

$$\mathbb{E} [KL_\lambda^{\otimes n}(s_0, \widehat{s}_m)] \leq \left(\min_{s_m \in S_m} KL_\lambda^{\otimes n}(s_0, s_m) + \frac{1}{2n} \mathcal{D}_m \right) + C_2 \frac{1}{n}$$

with as few assumptions on the set S_m as possible. Note that we only aim at having an upper bound and will not focus on the (important) question of the existence of a corresponding lower bound.

Our answer is far from definitive, the upper bound we obtained is the following weaker one

$$\mathbb{E} [JKL_{\rho,\lambda}^{\otimes n}(s_0, \widehat{s}_m)] \leq (1 + \epsilon) \left(\inf_{s_m \in S_m} KL_\lambda^{\otimes n}(s_0, s_m) + \frac{\kappa_0}{n} \mathfrak{D}_m \right) + C_2 \frac{1}{n}$$

in which the left-hand $KL_\lambda^{\otimes n}(s_0, \widehat{s}_m)$ has been replaced by a smaller divergence $JKL_{\rho,\lambda}^{\otimes n}(s_0, \widehat{s}_m)$ described below, ϵ can be chosen arbitrary small, \mathfrak{D}_m is a quantity playing the role of the dimension \mathcal{D}_m and κ_0 is a constant that depends on ϵ . This result have nevertheless the right bias/variance trade-off flavor and can be used to recover usual minimax properties of specific estimators.

2.3 Jensen-Kullback-Leibler divergence and bracketing entropy

The main visible loss is the use of a divergence smaller than the Kullback-Leibler one (but larger than the squared Hellinger distance). Namely, we use the Jensen-Kullback-Leibler divergence JKL_ρ with $\rho \in (0, 1)$ defined by

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

Note that this divergence appears explicitly with $\rho = \frac{1}{2}$ in Massart [34], but can also be found implicitly in Birgé and Massart [8] and van de Geer [42]. We use the name Jensen-Kullback-Leibler divergence in the same way Lin [33] uses the name Jensen-Shannon divergence for a sibling in its information theory work. The main tool in the proof of the previous inequality are deviation inequality for sums of random variables and their suprema. Unfortunately, this tool requires some boundedness assumptions that $\ln \frac{s_m}{s_0}$ does not satisfy. The convexification of the Jensen-Kullback-Leibler divergence is such that this problematic term is replaced by the much better behaved $\ln \frac{(1-\rho)s_0 + \rho s_m}{s_0}$.

All those divergences are nevertheless related and, denoting the squared Hellinger distance by $d^2(\text{sd}\lambda, \text{td}\lambda) = d_\lambda^2(s, t)$, we prove in Appendix

Proposition 1. *For any probability measures $\text{sd}\lambda$ and $\text{td}\lambda$ and any $\rho \in (0, 1)$*

$$C_\rho d_\lambda^2(s, t) \leq JKL_{\rho, \lambda}(s, t) \leq KL_\lambda(s, t).$$

with $C_\rho = \frac{1}{\rho} \min\left(\frac{1-\rho}{\rho}, 1\right) \left(\ln\left(1 + \frac{\rho}{1-\rho}\right) - \rho\right)$.

Furthermore, if $\text{sd}\lambda \ll \text{td}\lambda$ then

$$d_\lambda^2(s, t) \leq KL_\lambda(s, t) \leq \left(2 + \ln \left\| \frac{s}{t} \right\|_\infty\right) d_\lambda^2(s, t)$$

$$JKL_\lambda(s, t) \leq KL_\lambda(s, t) \leq JKL_\lambda(s, t) + \left\| \ln \left(\left(1 + \frac{s-t}{t}\right) \left(1 + \rho \frac{t-s}{s}\right)^{1/\rho} \right) \right\|_\infty.$$

More precisely, will appear their *tensorized* versions

$$d_\lambda^{2\otimes n}(s, t) = \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n d_\lambda^2(s(\cdot|X_i), t(\cdot|X_i)) \right] \quad \text{and} \quad JKL_{\rho, \lambda}^{\otimes n}(s, t) = \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n JKL_{\rho, \lambda}(s(\cdot|X_i), t(\cdot|X_i)) \right].$$

We focus now on the definition of \mathfrak{D}_m . It involves a bracketing entropy condition on the model S_m with respect to the Hellinger type divergence $d^{\otimes n}(s, t) = \sqrt{d^{2\otimes n}(s, t)}$. A bracket $[t^-, t^+]$ is a couple of functions such that $\forall(x, y) \in \mathcal{X} \times \mathcal{Y}, t^-(y|x) \leq t^+(y|x)$. 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 $H_{[\cdot], d^{\otimes n}}(\delta, S)$ of a set S is defined as the logarithm of the minimum number $N_{[\cdot], d^{\otimes n}}(\delta, S)$ of brackets $[t^-, t^+]$ of width $d^{\otimes n}(t^-, t^+)$ smaller than δ such that every function of S belongs to one of these brackets. \mathfrak{D}_m depends on the bracketing entropies not of the global models S_m but the ones of smaller localized sets $S_m(\tilde{s}, \sigma) = \{s_m \in S_m | d^{\otimes n}(\tilde{s}, s_m) \leq \sigma\}$. Indeed, we impose a structural assumption:

Assumption (\mathbf{H}_m) 1. *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 $s_m \in S_m$*

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

\mathfrak{D}_m is then defined as $n\sigma_m^2$ with σ_m^2 the unique root of $\frac{1}{\sigma} \phi_m(\sigma) = \sqrt{n}\sigma$. This bracketing entropy integral, often call Dudley integral, plays an important role in empirical processes theory, as stressed for instance in van der Vaart and Wellner [43]. The equation defining σ_m corresponds to a simplified 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 if the bracketing entropy scales like the dimension \mathcal{D}_m times $\ln \frac{1}{\delta}$, as it is often the case, then \mathfrak{D}_m is roughly proportional to \mathcal{D}_m . More precisely, in Appendix, we obtain

- Proposition 2.**
- if $\forall \delta \leq \sqrt{2}$, $H_{[\cdot], d^{\otimes n}}(\delta, S_m(s_m, \sigma)) \leq \mathcal{D}_m \left(C_m + \ln \frac{\sigma}{\delta} \right)$ then (H_m) is satisfied with a function ϕ_m such that $\mathfrak{D}_m = C_\star \mathcal{D}_m$,
 - if $\forall \delta \leq \sqrt{2}$, $H_{[\cdot], d^{\otimes n}}(\delta, S_m(s_m, \sigma)) \leq \mathcal{D}_m \left(C_m + \ln \frac{1}{\delta} \right)$ then (H_m) is satisfied with a function ϕ_m such that $\mathfrak{D}_m \leq \left(2C_\star + \left(\ln \frac{n}{C_\star \mathcal{D}_m} \right)_+ \right) \mathcal{D}_m$
where in both cases $C_\star = (\sqrt{C_m} + \sqrt{\pi})^2$.

2.4 Single model maximum likelihood estimation

For technical reason, we also need a separability assumption on our model:

Assumption (S_m) 1. *There exist some countable subset S'_m of S_m and a set \mathcal{Y}'_m with $\lambda(\mathcal{Y} \setminus \mathcal{Y}'_m) = 0$ such that for every $t \in S_m$, it exists some 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.*

We are now ready to state our risk bound theorem:

Theorem 1. *Assume we observe (X_i, Y_i) with unknown conditional density s_0 . Assume Assumptions (H_m) and (S_m) hold and let \hat{s}_m be a η -log-likelihood minimizer in S_m*

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

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, for $\mathfrak{D}_m = n\sigma_m^2$ with σ_m the unique root of $\frac{1}{\sigma} \phi_m(\sigma) = \sqrt{n}\sigma$, the likelihood estimate \hat{s}_m satisfies

$$\mathbb{E} \left[JKL_{\rho, \lambda}^{\otimes n}(s_0, \hat{s}_m) \right] \leq C_1 \left(\inf_{s_m \in S_m} KL_{\lambda}^{\otimes n}(s_0, s_m) + \frac{\kappa_0}{n} \mathfrak{D}_m \right) + C_2 \frac{1}{n} + \frac{\eta}{n}.$$

It should be noted that condition on the design of the X_i s appears only implicitly in the divergence $d^{\otimes n}$ used in the bracketing entropy definition. Furthermore, Assumption H_m could be further localized: it suffices to impose that the condition on the Dudley integral holds for a sequence of minimizer of $d^{2^{\otimes n}}(s_0, s_m)$.

We obtain thus a bound on the expected loss similar to the one obtained in the parametric case that holds for finite sample and that do not require the strong regularity assumptions of White [45]. In particular, we do not even require an identifiability condition in the parametric case. As often in empirical processes theory, the constant κ_0 appearing in the bound is pessimistic. Even in a very simple parametric model, the current best estimates are such that $\kappa_0 \mathfrak{D}_m$ is still much larger than the variance of Section 2.2 Asymptotic analysis of a parametric model. Numerical experiments show there is a hope that this is only a technical issue. The obtained bound quantifies however the expected classical bias-variance trade-off: a good model should be large enough so that the true conditional density is close from it but, at the same time, it should also be small so that the \mathfrak{D}_m term does not dominate.

It should be stressed that a result of similar flavor could have been obtained by the information theory technique of Barron et al. [4] and Kolarczyk et al. [31]. Indeed, if we replace the set S_m by a discretized version \mathfrak{S}_m so that

$$\inf_{s_m \in \mathfrak{S}_m} KL_{\lambda}^{\otimes n}(s_0, s_m) \leq \inf_{s_m \in S_m} KL_{\lambda}^{\otimes n}(s_0, s_m) + \frac{1}{n},$$

then, if we let $\widehat{\mathbf{s}}_m$ be a -log-likelihood minimizer in \mathfrak{S}_m ,

$$\mathbb{E} \left[\mathcal{D}_\lambda^{2\otimes n}(s_0, \widehat{\mathbf{s}}_m) \right] \leq \inf_{s_m \in \mathfrak{S}_m} KL_\lambda^{\otimes n}(s_0, s_m) + \frac{1}{n} \ln |\mathfrak{S}_m| + \frac{1}{n}$$

where $\mathcal{D}^{2\otimes n}$ is the tensorized Bhattacharyya-Renyi divergence, another divergence smaller than $KL^{\otimes n}$, $|\mathfrak{S}_m|$ is the cardinality of \mathfrak{S}_m and the expectation is taken conditionally to the covariates $(X_i)_{1 \leq i \leq n}$. As verified by Barron et al. [4] and Kolaczyk et al. [31], \mathfrak{S}_m can be chosen of cardinality of order $\ln n \mathcal{D}_m$ when the model is parametric. We obtain thus also a bound of type

$$\mathbb{E} \left[\mathcal{D}_\lambda^{2\otimes n}(s_0, \widehat{\mathbf{s}}_m) \right] \leq \inf_{s_m \in \mathfrak{S}_m} KL_\lambda^{\otimes n}(s_0, s_m) + \frac{C_1}{n} \ln n \mathcal{D}_m + \frac{1}{n}.$$

with better constants but with a different divergence. The bounds holds however only conditionally to the design, which can be an issue as soon as this design is random, and requires to compute an adapted discretization of the models.

3 Model selection and penalized maximum likelihood

3.1 Framework

A natural question is then the choice of the model. In the model selection framework, instead of a single model S_m , we assume we have at hand a collection of models $\mathcal{S} = \{S_m\}_{m \in \mathcal{M}}$. If we assume that Assumptions (H_m) and (S_m) hold for all models, then for every model S_m

$$\mathbb{E} \left[JKL_{\rho, \lambda}^{\otimes n}(s_0, \widehat{\mathbf{s}}_m) \right] \leq C_1 \left(\inf_{s_m \in \mathfrak{S}_m} KL_\lambda^{\otimes n}(s_0, s_m) + \frac{\kappa_0}{n} \mathfrak{D}_m \right) + C_2 \frac{1}{n} + \frac{\eta}{n}.$$

Obviously, one of the models minimizes the right hand side. Unfortunately, there is no way to know which one without knowing s_0 , i.e. without an oracle. Hence, we can not use this *oracle* model to estimate s_0 . We will nevertheless propose a data-driven strategy to select an estimate among the collection of estimates $\{\widehat{\mathbf{s}}_m\}_{m \in \mathcal{M}}$ according to a selection rule that performs almost as well as if we had known this *oracle*.

As always, using simply the -log-likelihood of the estimate in each model

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

as a criterion is not sufficient. It is an underestimation of the true risk of the estimate and this will lead to choose models that are too complex. By adding an adapted penalty $\text{pen}(m)$, one hope to compensate for both the *variance* term and the bias between $\sum_{i=1}^n -\ln \frac{\widehat{\mathbf{s}}_m(Y_i|X_i)}{s_0(Y_i|X_i)}$ and $\inf_{s_m \in \mathfrak{S}_m} KL_\lambda(s_0, s_m)$. For a given choice of $\text{pen}(m)$, the *best* model $S_{\widehat{m}}$ is chosen as the one whose index is an almost minimizer of the penalized η -log-likelihood :

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

The analysis of the previous section will be crucial as the intrinsic complexity \mathfrak{D}_m will appear in the penalty. It is no surprise that the complexity of the model collection itself also appears. We will require an information theory type assumption on our collection. We will assume the existence of a Kraft type inequality for the collection:

Assumption (K) 1. *There is a family $(x_m)_{m \in \mathcal{M}}$ of non-negative number such that*

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

It can be interpreted as a coding condition as stressed by Barron et al. [4] where a similar assumption is used. Remark that these coding terms satisfy a global condition, and thus that any permutation of them would also satisfy it. We should try to mitigate this arbitrariness by favoring choice of x_m for which the ratio with the intrinsic entropy term \mathfrak{D}_m is as small as possible. Indeed, as the condition on the penalty will be of the form

$$\text{pen}(m) \geq \kappa (\mathfrak{D}_m + x_m),$$

this will ensure that this lower bound is dominated by the intrinsic quantity \mathfrak{D}_m .

3.2 A general theorem for penalized maximum likelihood conditional density estimation

Our main theorem is then:

Theorem 2. *Assume we observe (X_i, Y_i) with unknown conditional density s_0 . Let $\mathcal{S} = (S_m)_{m \in \mathcal{M}}$ a at most countable model collection. Assume Assumption (K) holds while Assumptions (H_m) and (S_m) hold for every model $S_m \in \mathcal{S}$. Let \hat{s}_m be a η -log-likelihood minimizer in S_m*

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

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 $m \in \mathcal{M}$

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

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 \hat{s}_m with \hat{m} such that

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

satisfies

$$\mathbb{E} \left[JKL_{\rho, \lambda}^{\otimes n}(s_0, \hat{s}_{\hat{m}}) \right] \leq C_1 \inf_{m \in \mathcal{M}} \left(\inf_{s_m \in S_m} KL_{\lambda}^{\otimes n}(s_0, s_m) + \frac{\text{pen}(m)}{n} \right) + C_2 \frac{\Sigma}{n} + \frac{\eta + \eta'}{n}.$$

Note that, as in 2.4 Single model maximum likelihood estimation, the approach of Barron et al. [4] and Kolaczyk et al. [31] could have been used to obtain a similar result with the help of discretization.

This theorem extends Theorem 7.11 Massart [34] which handled only density estimation. As in this theorem, the cost of model selection with respect to the choice of the best single model is proved to be very mild. Indeed, let $\text{pen}(m) = \kappa(\mathfrak{D}_m + x_m)$ then one obtains

$$\mathbb{E} \left[JKL_{\rho, \lambda}^{\otimes n}(s_0, \hat{s}_{\hat{m}}) \right] \leq C_1 \inf_{m \in \mathcal{M}} \left(\inf_{s_m \in S_m} KL_{\lambda}^{\otimes n}(s_0, s_m) + \frac{\kappa}{n}(\mathfrak{D}_m + x_m) \right) + C_2 \frac{\Sigma}{n} + \frac{\eta + \eta'}{n}$$

$$\leq C_1 \frac{\kappa}{\kappa_0} \left(\max_{m \in \mathcal{M}} \frac{\mathfrak{D}_m + x_m}{\mathfrak{D}_m} \right) \inf_{m \in \mathcal{M}} \left(\inf_{s_m \in S_m} KL_{\lambda}^{\otimes n}(s_0, s_m) + \frac{\kappa_0}{n} \mathfrak{D}_m \right) + C_2 \frac{\Sigma}{n} + \frac{\eta + \eta'}{n}.$$

As soon as the term x_m is always small relatively to \mathfrak{D}_m , we obtain thus an oracle inequality that show that the penalized estimate satisfies up to a small factor the bound of Theorem 1 Single model maximum likelihood estimation for 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 very important: we do not have to know the best model within a collection to almost achieve its performance.

So far we do not have discussed the choice of the model collection, it is however critical to obtain a *good* estimator. There is unfortunately no universal choice and it should be adapted to the specific setting considered. Typically, if we consider conditional density of *regularity* indexed by a parameter α , a good collection is one such that for every parameter α there is a model which achieve a quasi optimal bias/variance trade-off. Efromovich [16, 17] considers Sobolev type regularity and use thus models generated by the first elements of Fourier basis. Brunel et al. [12] and Akakpo and Lacour [2] considers anisotropic regularity spaces for which they show that a collection of piecewise polynomial models is adapted. Although those choices are justified, in these papers, in a quadratic loss approach, they remain good choices in our maximum likelihood approach with a Kullback-Leibler type loss. Estimator associated to those collections are thus *adaptive* to the regularity: without knowing the *regularity* of the true conditional density, they select a model in which the estimate performs almost as well as in the *oracle* model, the best choice if the regularity was known. In both cases, one could prove that those estimators achieve the minimax rate for the considered classes, up to a logarithmic factor.

As in Section 2.4 Single model maximum likelihood estimation, the known estimate of constant κ_0 and even of \mathfrak{D}_m can be pessimistic. This leads to a theoretical penalty which can be too large in practice. A natural question is thus whether the constant appearing in the penalty can be estimated from the data without losing a theoretical guaranty on the performance? No definitive answer exists so far, but numerical experiment in specific case shows that the *slope heuristic* proposed by Birgé and Massart [9] may yields a solution.

The assumptions of the previous theorem are as general as possible. It is thus natural to question the existence of interesting model collections that satisfy its assumptions. We have mention so far the Fourier based collection proposed by Efromovich [17, 16] and the piecewise polynomial collection of Brunel et al. [12] and Akakpo and Lacour [2] considers anisotropic regularity. We focus on a variation of this last strategy. Motivated by an application to unsupervised image segmentation, we will consider model collection in which, in each model, the conditional densities depends on the covariate only in a piecewise constant manner. After a general introduction to these partition-based strategies, we study two cases: a classical one in which the conditional density depends in a piecewise polynomial manner of the variables and a newer one, which correspond to the unsupervised segmentation application, in which the conditional density are Gaussian mixture with common Gaussian components but mixing proportions depending on the covariate.

4 Partition-based conditional density models

4.1 Covariate partitioning and conditional density estimation

Following an idea developed by Kolaczyk et al. [31], we will partition the covariate domain and consider candidate estimates conditional densities that depends 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 Willet and Nowak [47] 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 region in this partition, the model we consider will thus be specified by a partition \mathcal{P} and a set \mathcal{G} of $\|\mathcal{P}\|$ -uples 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 will study two examples: in the first one, \mathcal{F} will indeed be a product of piecewise polynomial density sets, while in the second one \mathcal{F} will be a set of $\|\mathcal{P}\|$ -uples 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 $\hat{s}_{\mathcal{P}, \mathcal{F}}$ such that

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

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 $[0, 1]^{d_X}$ but 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 will only consider five specific hyperrectangle based collections of partitions of $[0, 1]^{d_X}$:

- Two are recursive dyadic partition collections.
 - The uniform dyadic partition collection (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 step J satisfies $\frac{2^{d_X}}{n} \geq |\mathcal{R}_l| \geq \frac{1}{n}$.
 - The recursive dyadic partition collection (RDP(\mathcal{X})) in which at each step an 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 (RDSP(\mathcal{X})) in which at each step an 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 (RSP(\mathcal{X})) in which at each step an 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 hyperrectangles partition collection (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 measure is larger than $\frac{1}{n}$.

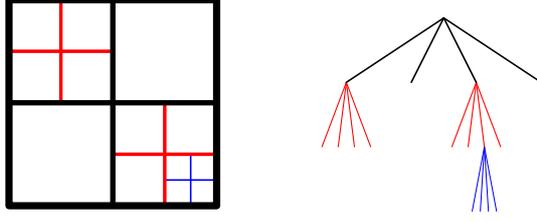


Figure 1: Example of a recursive dyadic partition with its associated dyadic tree.

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

As noticed by Kolaczyk and Nowak [30], Huang et al. [26] or Willet and Nowak [47], the first four partition collections, $(\mathcal{S}_{\mathcal{P}}^{\text{UDP}(\mathcal{X})}, \mathcal{S}_{\mathcal{P}}^{\text{RDP}(\mathcal{X})}, \mathcal{S}_{\mathcal{P}}^{\text{RDSP}(\mathcal{X})}, \mathcal{S}_{\mathcal{P}}^{\text{RSP}(\mathcal{X})})$, have a tree structure. Figure 1 Example of a recursive dyadic partition with its associated dyadic tree illustrates this structure in the case of a $\text{RDP}(\mathcal{X})$ partition. This specific structure will prove to be useful to obtain efficient numerical algorithm to perform the model selection. For sake of completeness, we have also added the much more complex to deal with collection $\mathcal{S}_{\mathcal{P}}^{\text{HRP}(\mathcal{X})}$, for which only exhaustive search algorithm exists.

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

Proposition 3. *For any of the five described partition collections $\mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}$, $\exists A_0^*, B_0^*, c_0^*$ and Σ_0 such that for all $c \geq c_0^{\star(\mathcal{X})}$:*

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

where those constant can be chosen as follows:

	$\star = \text{UDP}(\mathcal{X})$	$\star = \text{RDP}(\mathcal{X})$	$\star = \text{RDSP}(\mathcal{X})$	$\star = \text{RSP}(\mathcal{X})$	$\star = \text{HRP}(\mathcal{X})$
A_0^*	$\ln \left(\max \left(2, 1 + \frac{\ln n}{d_X \ln 2} \right) \right)$	0	0	0	0
B_0^*	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^*	0	$\frac{2^d}{2^d - 1}$	2	2	1
Σ_0	$1 + \frac{\ln n}{\dim X \ln 2}$	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 . This is useful to prove Assumption (K) for the model collections of the next sections.

In those sections, we will study the two different choices proposed above for the set \mathcal{F} . We first consider the piecewise polynomial strategy similar to the one proposed by Willet and Nowak [47] defined for $\mathcal{Y} = [0, 1]^{d_Y}$ in which the set \mathcal{F} is a product of sets. We then consider the 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.

4.2 Piecewise polynomial conditional densities estimation

In this section, we let $\mathcal{X} = [0, 1]^{d_x}$, $\mathcal{Y} = [0, 1]^{d_y}$ and λ be the Lebesgue measure dy which is a probability measure on \mathcal{Y} . Our candidate density $s(y|x \in \mathcal{R}_l)$ will then be 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 density $s(y|x \in \mathcal{R}_l)$ is a square of polynomial on each hyperrectangle $\mathcal{R}_{l,k}^y$ of the partition \mathcal{Q}_l . This differs from the choice of Willet and Nowak [47] 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 ensures the positiveness of our estimator and turns out to be 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 $\mathbf{D} = (\mathbf{D}_1, \dots, \mathbf{D}_{d_y})$, we define first for any partition $\mathcal{Q} = \{\mathcal{R}_k^y\}_{1 \leq k \leq \|\mathcal{Q}\|}$ of $\mathcal{Y} = [0, 1]^{d_y}$ the set $\mathcal{F}_{\mathcal{Q}, \mathbf{D}}$ of square of piecewise polynomial density of maximum degree \mathbf{D} defined in the partition \mathcal{Q} :

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

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

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

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

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

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

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

As shown by Willet and Nowak [47], the maximum likelihood estimate in this model can be obtained by an independent computation on each subset $\mathcal{R}_{l,k}^\times$:

$$\hat{P}_{\mathcal{R}_{l,k}^\times} = \frac{\sum_{i=1}^n \mathbf{1}_{\{(X_i, Y_i) \in \mathcal{R}_{l,k}^\times\}}}{\sum_{i=1}^n \mathbf{1}_{\{X_i \in \mathcal{R}_l\}}} \underset{P, \deg(P) \leq \mathbf{D}, \int_{\mathcal{R}_{l,k}^y} P^2(y) dy = 1}{\operatorname{argmin}} \sum_{i=1}^n \mathbf{1}_{\{(X_i, Y_i) \in \mathcal{R}_{l,k}^\times\}} \ln(P^2(Y_i)).$$

This property will be important to be able to use the efficient optimization algorithms of Willet and Nowak [47] and Huang et al. [26].

Our model collection is obtained by considering all partition \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,

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

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

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

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

Theorem 3. 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 degree for the polynomial $\mathbf{D} \in \mathbb{N}^{d_y}$.

Let

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

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 2A general theorem for penalized maximum likelihood conditional density estimation satisfies

$$\begin{aligned} \mathbb{E} \left[JKL_{\rho, \lambda}^{\otimes n}(s_0, \widehat{s}_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}) \right] &\leq C_1 \inf_{S_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}} \in \mathcal{S}} \left(\inf_{s_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}} \in S_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}} KL_{\lambda}^{\otimes n}(s_0, s_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}) + \frac{\text{pen}(\mathcal{P}^{\mathcal{Q}}, \mathbf{D})}{n} \right) \\ &\quad + C_2 \frac{1}{n} + \frac{\eta + \eta'}{n} \end{aligned}$$

as soon as

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

for

$$\tilde{\kappa} > \kappa_0 \left(C_{\star} + c_{\star} \left(A_0^{\star(\mathcal{X})} + \frac{B_0^{\star(\mathcal{X})} + A_0^{\star(\mathcal{Y})} + B_0^{\star(\mathcal{Y})}}{\prod_{d=1}^{d_y} (\mathbf{D}_d + 1)} \right) + 2 \ln n \right).$$

where κ_0 and C_2 are the constants of Theorem 2A general theorem for penalized maximum likelihood conditional density estimation that depend only on ρ and C_1 .

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

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

This additive structure of the penalty allows the use of the fast partition optimization algorithm of Donoho [15] and Huang et al. [26] as soon as the partition collection is tree structured.

In Appendix, we prove the more involved requirement

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

in which the complexity part and the coding part appear more explicitly. Using the crude bound $\|\mathcal{P}^{\mathcal{Q}}\| \geq 1$, one see that the penalty can still be chosen as a sum of penalties over each hyperrectangle. The loss with respect to the optimal penalty is of order $\kappa \log \|\mathcal{P}^{\mathcal{Q}}\| \mathcal{D}_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}$, 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 shown in Appendix. For example, the problematic term $2 \ln(n/\sqrt{\|\mathcal{P}^{\mathcal{Q}}\|})$ disappears if \mathcal{P} belongs to $\mathcal{S}_{\mathcal{P}}^{\text{UDP}(\mathcal{X})}$ while \mathcal{Q}_l is independent of \mathcal{R}_l and belongs to $\mathcal{S}_{\mathcal{P}}^{\text{UDP}(\mathcal{X})}$. Choosing the degrees \mathbf{D} of the polynomial among a family \mathcal{D}^M either globally or locally as proposed by Willet and Nowak [47] is also possible. It suffices to modify the coding part in the involved requirement accordingly: this can be achieved by replacing respectively $A_0^{\star(\mathcal{X})}$ by $A_0^{\star(\mathcal{X})} + \ln |\mathcal{D}^M|$ for the global optimization and $B_0^{\star(\mathcal{Y})}$ by $B_0^{\star(\mathcal{Y})} + \ln |\mathcal{D}^M|$ for the local optimization. The requirement can still be satisfied with minor loss with a penalty with additive structure.

Finally, reusing ideas of Willet and Nowak [47], Akakpo [1] or Akakpo and Lacour [2], one can obtain the quasi optimal minimaxity of this estimator for anisotropic Besov spaces (see for instance in Karaivanov and Petrushev [29] for a definition) whose regularity index are smaller than 1 along the axes of \mathcal{X} and smaller than $\mathbf{D} + 1$ along the axes of \mathcal{Y} .

4.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties

In this section, we consider an extension of Gaussian mixture that takes account into a 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 classifications.

In a classical Gaussian mixture model, the observations Y_i , that are assumed to belong to \mathbb{R}^p , are supposed to be obtained as the realization of a Gaussian selected from a finite set with a certain probability. Let K be the number of different Gaussians, often call the number of clusters, the density s 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 k th 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 -uple of Gaussian $(\Phi_{\theta_1}, \dots, \Phi_{\theta_K})$. Those Gaussians can share for instance the same shape, the same volume or the same diagonalization basis. The classical choice are described for instance in Biernacki et al. [7]. Using the EM algorithm, or one of its extension, one can efficiently obtained the proportions $\hat{\pi}_k$ and the Gaussian parameters $\hat{\theta}_k$ 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:

$$\hat{k}(y) = \operatorname{argmax}_{1 \leq k \leq \hat{K}} \hat{\pi}_k \Phi_{\hat{\theta}_k}(y).$$

Using our partition-based strategy, we extend this model by allowing the mixing proportion to vary according to the covariate X_i with the same mixture components. This is similar to the ideas of Kolaczyk et al. [31] and Antoniadis et al. [3], a piecewise constant component proportion, up to the choice of mixing densities. In a semantic analysis context, in which documents replaces pixels, a similar Gaussian mixture with varying weight, but without the partition structure, has been proposed by Si and Jin [39] as an extension of a general mixture based semantic analysis model introduced by Hofmann [25] under the name *Probabilistic Latent Semantic Analysis*. A similar model has also been considered in the work of Young and Hunter [48]. 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 -uples of Gaussian can be chosen in the same way than in the classical Gaussian mixture case. Using a penalized maximum likelihood strategy, a partition $\hat{\mathcal{P}}$, a number of Gaussian components \hat{K} , their parameters $\hat{\theta}_k$ and all the mixing proportions $\hat{\pi}[\hat{\mathcal{R}}_l]$ can be estimated. Each observation (x, y) can then be assigned to one the induced classes by a maximum likelihood principle:

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

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 spectrum. 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 et al. [31], we have been able to implement this penalized estimator and to test it on real datasets. Figures 2 Segmentations of a slice of a Bockay double bass showing a double layer varnish structure a) obtained with constant mixing proportions on the unit square b) obtained with piecewise constant mixing proportions on an optimal recursive dyadic partition (shown in black) Segmentations of a slice of a Bockay double bass showing a double layer varnish structure a) obtained with constant mixing proportions on the unit square b) obtained with piecewise constant mixing proportions on an optimal recursive dyadic partition (shown in black) and 3 Segmentations of painting layers. The leftmost image is a visible light microscopy image of the analyzed zone showing three layers from bottom to top (embedding resine, lead white in oil and a pictural layer containing azurite and quartz). The middle one is the segmentation obtain with constant mixing proportions on the unit square while the rightmost one is obtained with piecewise constant mixing proportions on an optimal recursive dyadic partition (shown in black) Segmentations of painting layers. The leftmost image is a visible light microscopy image of the analyzed zone showing three layers from bottom to top (embedding resine, lead white in oil and a pictural layer containing azurite and quartz). The middle one is the segmentation obtain with constant mixing proportions on the unit square while the rightmost one is obtained with piecewise constant mixing proportions on an optimal recursive dyadic partition (shown in black) illustrates this methodology. The first one corresponds to a study of a slice of a Bockay double bass. The two different varnish layers has been separated automatically. Figure 3 Segmentations of painting layers. The leftmost image is a visible light microscopy image of the analyzed zone showing three layers from bottom to top (embedding resine, lead white in oil and a pictural layer containing azurite and quartz). The middle one is the segmentation obtain with constant mixing proportions on the unit square while the rightmost one is obtained with piecewise constant mixing proportions on an optimal recursive dyadic partition (shown in black) Segmentations of painting layers. The leftmost image is a visible light microscopy image of the analyzed zone showing three layers from bottom to top (embedding resine, lead white in oil and a pictural layer containing azurite and quartz). The middle one is the segmentation obtain with constant mixing proportions on the unit square while the rightmost one is obtained with piecewise constant mixing proportions on an optimal recursive dyadic partition (shown in black) is painting layers analysis. The three layers are clearly discriminated with both the technique but the geometry is better preserved with our spatialized approach. In both cases, the use of 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.

We should now specify the models we consider. As we follow the construction of Section 4.1 Covariate partitioning and conditional density estimation, for a given segmentation \mathcal{P} , this amounts to specify the set \mathcal{F} to which belong the $\|\mathcal{P}\|$ -uples 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 , $\pi[\mathcal{R}_l]$, will be chosen freely among all vector of the $K - 1$ dimensional simplex \mathcal{S}_{K-1} :

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

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

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 along the different classes, and let E^\perp be its

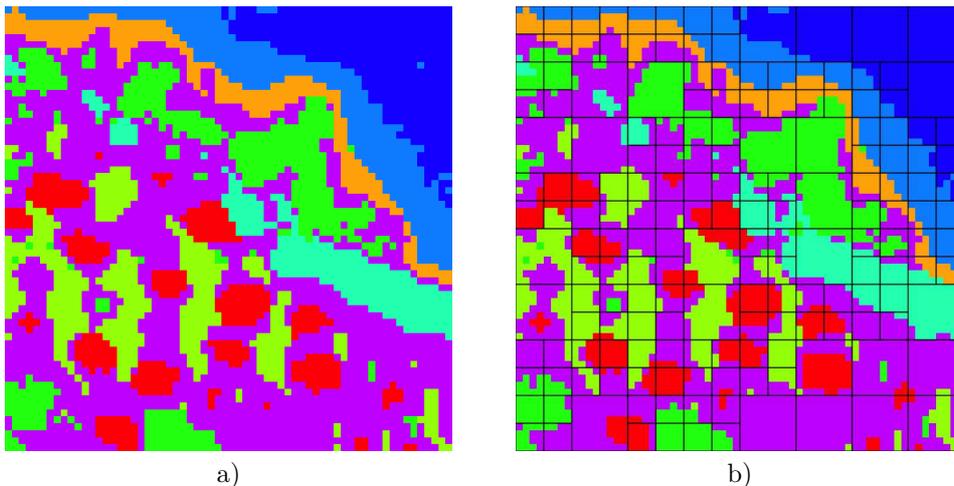


Figure 2: Segmentations of a slice of a Bockay double bass showing a double layer varnish structure a) obtained with constant mixing proportions on the unit square b) obtained with piecewise constant mixing proportions on an optimal recursive dyadic partition (shown in black).

orthogonal, in which all classes behave similarly. We will thus assume 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 depends on k while $\Phi_{\theta_{E^\perp}}$ is independent of k . Our model are then specified by the choice of a set \mathcal{G}_E^K for the K -uples $(\Phi_{\theta_{E,1}}, \dots, \Phi_{\theta_{E,K}})$ (or equivalently a set Θ_E^K for the K -uples of parameters $(\theta_{E,1}, \dots, \theta_{E,K})$) and a set \mathcal{G}_{E^\perp} for the Gaussian $\Phi_{\theta_{E^\perp}}$ (or equivalently a set Θ_{E^\perp} for its parameter θ_{E^\perp}). The resulting model is denoted $S_{\mathcal{P},K,\mathcal{G}}$

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

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

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

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

The K -uple of means $\mu = (\mu_1, \dots, \mu_K)$ is either known or unknown without any restriction. A stronger structure will be imposed on the K -uple of covariance matrices $(\Sigma_1, \dots, \Sigma_K)$. To define it, we need to introduce a decomposition of any covariance matrix Σ into $LDAD'$ where $L = |\Sigma|^{1/p_E}$ 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 will nevertheless impose a structure on the K -uple $(\Sigma_1, \dots, \Sigma_K)$ through structures

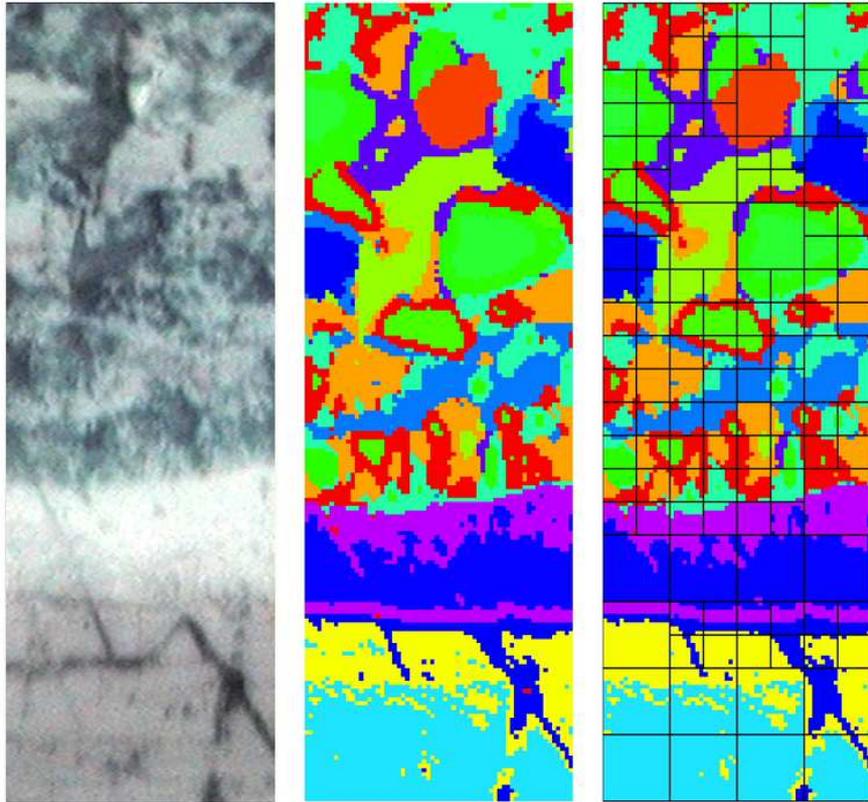


Figure 3: Segmentations of painting layers. The leftmost image is a visible light microscopy image of the analyzed zone showing three layers from bottom to top (embedding resin, lead white in oil and a pictorial layer containing azurite and quartz). The middle one is the segmentation obtained with constant mixing proportions on the unit square while the rightmost one is obtained with piecewise constant mixing proportions on an optimal recursive dyadic partition (shown in black).

on the corresponding K -uples of (L_1, \dots, L_K) , (D_1, \dots, D_K) and (A_1, \dots, A_K) . They will be either known, unknown but with a common value or unknown without any restriction. The corresponding set will be indexed by $[\mu_\star L_\star D_\star A_\star]_{p_E}^K$ 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.

In order to have a set with finite bracketing entropy, we have to 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 $\forall 1 \leq k \leq K, L_m \leq L_k \leq L_M$ for some known positive values L_m and L_M . To describe the constraints on the renormalized eigenvalue matrix A_k , we define the set $\mathcal{A}(\lambda_m, \lambda_M, p_E)$ of diagonal matrix A such that $|A| = 1$ and $\forall 1 \leq i \leq p_E, \lambda_m \leq A_{i,i} \leq \lambda_M$. Our assumption will be that all the A_k belong to $\mathcal{A}(\lambda_m, \lambda_M, p_E)$ for some known values λ_m and λ_M .

Among the $3^4 = 81$ such possible sets, six of them have been already studied by Maugis and Michel [35, 37] in their classical Gaussian mixture model analysis:

- $[\mu_0 L_K D_0 A_0]_{p_E}^K$ in which only the volume of the variance of a class is unknown. They use this model with a single class to model the non discriminant variables in E^\perp .
- $[\mu_K L_K D_0 A_K]_{p_E}^K$ in which one assumes that the unknown variances Σ_k can be diagonalized in the same known basis D_0 .
- $[\mu_K L_K D_K A_K]_{p_E}^K$ in which everything is free,
- $[\mu_K L D_0 A]_{p_E}^K$ in which the variances Σ_k are assumed to be equal and diagonalized in the known basis D_0 .
- $[\mu_K L D_0 A_K]_{p_E}^K$ in which the volumes L_k are assumed to be equal and the variance can be diagonalized in the known basis D_0
- $[\mu_K L D A]_{p_E}^K$ in which the variances Σ_k are only assumed to be equal

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 -uples of Gaussian specified by a space E and two parameter sets, a set $\Theta_{[\mu_\star L_\star D_\star A_\star]_{p_E}^K}$ of K -uples of Gaussian parameters for the differentiated space E and a set $\Theta_{[\mu_\star L_\star D_\star A_\star]_{p_{E^\perp}}^K}$ of Gaussian parameters for its orthogonal E^\perp . Those two sets are chosen among the ones described above with the same constants a , L_m , L_M , λ_m and λ_M . One verify that

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

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

Without any assumption on the design, we obtain

Theorem 4. *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 2A general theorem for penalized maximum likelihood conditional density estimation satisfies

$$\mathbb{E} \left[JKL_{\rho, \lambda}^{\otimes n}(s_0, \widehat{s}_{\mathcal{P}, K, \mathcal{G}}) \right] \leq C_1 \inf_{S_{\mathcal{P}, K, \mathcal{G}} \in \mathcal{S}} \left(\inf_{S_{\mathcal{P}, K, \mathcal{G}} \in S_{\mathcal{P}, K, \mathcal{G}}} KL_{\lambda}^{\otimes n}(s_0, S_{\mathcal{P}, K, \mathcal{G}}) + \frac{\text{pen}(\mathcal{P}, K, \mathcal{G})}{n} \right) + C_2 \frac{1}{n} + \frac{\eta + \eta'}{n}$$

as soon as

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

for

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

with $\kappa > \kappa_0$ where κ_0 and C_2 are the constants of Theorem 2A general theorem for penalized maximum likelihood conditional density estimation that depend only on ρ and C_1 and

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

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 will perform almost as well as the best possible model in term of conditional density estimation. As in the proof of Antoniadis et al. [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 a on penalty corresponding to K and the set \mathcal{G} . Indeed

$$\text{pen}(\mathcal{P}, K, \mathcal{G}) = \sum_{\mathcal{R}_l \in \mathcal{P}} \tilde{\kappa}_1 (K - 1) + \tilde{\kappa}_1 \left(\dim \left(\Theta_{[\mu_\star, L_\star, D_\star, A_\star]_{p_E}^K} \right) + \dim \left(\Theta_{[\mu_\star, L_\star, D_\star, A_\star]_{p_{E^\perp}}} \right) \right) + \tilde{\kappa}_2 \mathcal{D}_E$$

satisfies the requirement of Theorem 4Spatial Gaussian mixtures, models, bracketing entropy and penalties. This structure is the key for our numerical minimization algorithm in which one optimizes alternatively the Gaussian parameters with an EM algorithm and the partition with the same fast optimization strategy than in the previous section.

In Appendix, we obtain a weaker requirement

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

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 than in the previous paragraph.

Our theoretical result on the conditional density estimation does not guaranty good segmentation performance. If the data is 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. Our numerical experiments give nevertheless segmentations that have been considered meaningful by the physicists at IPANEMA. We think thus our analysis provides a first theoretical ground to our numerical method.

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 [9] 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.

4.4 Bracketing entropy of Gaussian families

A key ingredient in the proof of 4Spatial Gaussian mixtures, models, bracketing entropy and penaltiesSpatial Gaussian mixtures, models, bracketing entropy and penalties is a generalization of a result of Maugis and Michel [35, 36] that control the bracketing entropy the Gaussian families $\mathcal{G}_{[\cdot]_E^K}$ with respect to the d^{\max} distance defined by

$$d^{\max}((s_1, \dots, s_K), (t_1, \dots, t_K)) = \sup_{1 \leq k \leq K} d^2(s_k, t_k).$$

Here, $[(t_1^-, \dots, t_K^-), (t_1^+, \dots, t_K^+)]$ is a bracket containing (s_1, \dots, s_K) if

$$\forall 1 \leq k \leq K, \forall y \in E, \quad t_k^-(y) \leq s_k(y) \leq t_k^+(y).$$

As it can be of interest on his own, we state it here:

Proposition 4. *Let $\kappa \geq \frac{3}{4}$ and*

$$\gamma_\kappa = \min \left(\frac{3(\kappa - \frac{3}{4})}{2(1 + \frac{2\kappa}{9})(1 + \frac{2}{9})^2(1 + \frac{1}{9})}, \frac{(\kappa - \frac{1}{2})}{2(1 + \frac{2\kappa}{9})(1 + \frac{2}{9})^2} \right) \quad \beta_\kappa = \sqrt{\kappa^2 \cosh\left(\frac{2}{9\kappa}\right) + \frac{1}{4}}$$

$$\text{Assume} \quad \begin{cases} a \geq \frac{\sqrt{\gamma_\kappa}}{18\beta_\kappa} \sqrt{L_m \lambda_m \frac{\lambda_m}{\lambda_M} \frac{\delta}{p_E}} \\ \ln \left(\frac{L_M}{L_m} \right) \geq \frac{1}{20\beta_\kappa} \delta \\ \frac{\lambda_M}{\lambda_m} \ln \left(\frac{\lambda_M}{\lambda_m} \right) \geq \frac{1}{27(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa} \frac{\delta}{p_E} \end{cases}$$

Then for any $\delta \in [0, \sqrt{2}]$,

$$H_{[\cdot], d^{\max}}(\delta/9, \mathcal{G}_{[\mu_\star, L_\star, D_\star, A_\star]_{p_E}^K}) \leq \mathcal{I}_{[\mu_\star, L_\star, D_\star, A_\star]_{p_E}^K} + \mathcal{D}_{[\mu_\star, L_\star, D_\star, A_\star]_{p_E}^K} \ln \frac{1}{\delta}$$

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

$$\mathcal{I}_{[\mu_*, L_*, D_*, A_*]_{p_E}^K} = c_{\mu_*} \mathcal{I}_{\mu, p_E} + c_{L_*} \mathcal{I}_{L, p_E} + c_{D_*} \mathcal{I}_{D, p_E} + c_{A_*} \mathcal{I}_{A, p_E} \text{ with } \begin{cases} c_{\mu_0} = c_{L_0} = c_{D_0} = c_{A_0} = 0 \\ c_{\mu_K} = c_{L_K} = c_{D_K} = c_{A_K} = K \\ c_{\mu} = c_L = c_D = c_A = 1 \end{cases},$$

$$\begin{cases} \mathcal{D}_{\mu, p_E} = p_E \\ \mathcal{D}_L = 1 \\ \mathcal{D}_{D, p_E} = \frac{p_E(p_E-1)}{2} \\ \mathcal{D}_{A, p_E} = p_E - 1 \end{cases} \text{ and } \begin{cases} \mathcal{I}_{\mu, p_E} = p_E \left(\ln \left(\frac{36a\beta_\kappa p_E}{\sqrt{\gamma_\kappa L_m \lambda_m} \frac{\lambda_m}{\lambda_M}} \right) \right) \\ \mathcal{I}_{L, p_E} = \ln \left(40\beta_\kappa \ln \left(\frac{L_M}{L_m} \right) p_E \right) \\ \mathcal{I}_{D, p_E} = \frac{p_E(p_E-1)}{2} \left(\frac{\ln c}{\frac{p_E(p_E-1)}{2}} + \left(\ln \left(36(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa \frac{\lambda_M}{\lambda_m} p_E \right) \right) \right) \\ \mathcal{I}_{A, p_E} = (p_E - 1) \left(\ln \left(108(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa \frac{\lambda_M}{\lambda_m} \ln \left(\frac{\lambda_M}{\lambda_m} \right) p_E \right) \right) \end{cases}.$$

A Proofs for Section 2 Single model maximum likelihood estimate (Single model maximum likelihood estimate)

A.1 Proof of Proposition 1 Jensen-Kullback-Leibler divergence and bracketing entropy

Proof of Proposition 1 Jensen-Kullback-Leibler divergence and bracketing entropy. We first notice that, by convexity of the Kullback-Leibler divergence,

$$JKL_{\rho, \lambda}(s, t) = \frac{1}{\rho} KL_\lambda(s, (1-\rho)s + \rho t) \leq \frac{1}{\rho} ((1-\rho)KL_\lambda(s, s) + \rho KL(s, t)) = KL_\lambda(s, t).$$

Then let $d\lambda' = ((1-\rho)s + \rho t)d\lambda$, the function $u = \frac{s-t}{(1-\rho)s + \rho t}$ remains in $[-1/\rho, 1/(1-\rho)]$, and is such that $\frac{sd\lambda}{d\lambda'} = 1 + \rho u$ and $\frac{td\lambda}{d\lambda'} = 1 - (1-\rho)u$.

$$\begin{aligned} \text{Now, } JKL_\rho(sd\lambda, td\lambda) &= \frac{1}{\rho} KL(sd\lambda, (1-\rho)s + \rho td\lambda) = \frac{1}{\rho} KL((1+\rho u)d\lambda', d\lambda') \\ &= \frac{1}{\rho} KL_{\lambda'}(1+\rho u, 1) = \frac{1}{\rho} \int (1+\rho u) \ln(1+\rho u) d\lambda' \\ \text{and as } \int u d\lambda' &= 0 &= \frac{1}{\rho} \int ((1+\rho u) \ln(1+\rho u) - \rho u) d\lambda. \end{aligned}$$

$$\begin{aligned} \text{Similarly, } d^2(sd\lambda, td\lambda) &= d^2((1+\rho u)d\lambda', (1-(1-\rho)u)d\lambda') = d_{\lambda'}^2(1+\rho u, 1-(1-\rho)u) \\ &= 2 - 2 \int \sqrt{1+\rho u} \sqrt{1-(1-\rho)u} d\lambda' = 2 \int \left(1 - \sqrt{1+(2\rho-1)u - \rho(1-\rho)u^2} \right) d\lambda' \\ &= 2 \int \left(1 - \sqrt{1+(2\rho-1)u - \rho(1-\rho)u^2} + \left(\rho - \frac{1}{2}\right)u \right) d\lambda' \end{aligned}$$

Now let $\Phi(x) = (1+x) \ln(1+x) - x$, one can verify that $\Phi(x)/x^2$ is non increasing on $[-1, +\infty]$, so that $\forall u \in [-1/\rho, 1/(1-\rho)]$, $\Phi(\rho u) = \frac{\Phi(\rho u)}{\rho^2 u^2} \rho^2 u^2 \geq \frac{\Phi(\frac{\rho}{1-\rho})}{\rho^2/(1-\rho)^2} \rho^2 u^2$ so that

$$(1+\rho u) \ln(1+\rho u) - \rho u \geq \left(\left(1 + \frac{\rho}{1-\rho}\right) \ln \left(1 + \frac{\rho}{1-\rho}\right) - \frac{\rho}{1-\rho} \right) (1-\rho)^2 u^2$$

$$\geq (1 - \rho) \left(\ln \left(1 + \frac{\rho}{1 - \rho} \right) - \rho \right) u^2$$

Along the same lines, one can verify that $\forall u \in [-1/\rho, 1/(1 - \rho)]$

$$1 - \sqrt{1 + (2\rho - 1)u - \rho(1 - \rho)u^2} + \left(\rho - \frac{1}{2}\right)u \leq \frac{\max(\rho, 1 - \rho)}{2}u^2.$$

This implies thus

$$\begin{aligned} & \frac{1}{\rho} ((1 + \rho u) \ln(1 + \rho u) - \rho u) \\ & \geq \frac{1}{\rho \max(\rho, 1 - \rho)} (1 - \rho) \left(\ln \left(1 + \frac{\rho}{1 - \rho} \right) - \rho \right) 2 \left(1 - \sqrt{1 + (2\rho - 1)u - \rho(1 - \rho)u^2} + \left(\rho - \frac{1}{2}\right)u \right) \\ & \geq \frac{1}{\rho} \min\left(\frac{1 - \rho}{\rho}, 1\right) \left(\ln \left(1 + \frac{\rho}{1 - \rho} \right) - \rho \right) 2 \left(1 - \sqrt{1 + (2\rho - 1)u - \rho(1 - \rho)u^2} + \left(\rho - \frac{1}{2}\right)u \right) \end{aligned}$$

which yields the first inequality.

For the second series of inequalities,

$$d^2(sd\lambda, td\lambda) = d_{td\lambda}^2\left(\frac{s}{t}, 1\right) = \int \left(\sqrt{\frac{s}{t}} - 1 \right)^2 td\lambda,$$

while

$$KL(sd\lambda, td\lambda) = KL_{td\lambda}\left(\frac{s}{t}, 1\right) = \int \frac{s}{t} \ln \frac{s}{t} td\lambda = \int \left(\frac{s}{t} \ln \frac{s}{t} - \frac{s}{t} + 1 \right) td\lambda.$$

It turns out that $\forall x \in [0, M]$,

$$(\sqrt{x} - 1)^2 \leq x \ln x - x + 1 \leq (2 + (\ln M)_+) (\sqrt{x} - 1)^2$$

which yields the announced result. Finally, a straightforward computation yields

$$\begin{aligned} \frac{1}{\rho} \ln \frac{s}{(1 - \rho s + \rho t)} &= -\ln \left(\left(1 + \rho \frac{t - s}{s} \right)^{1/\rho} \right) \\ &= \ln \frac{s}{t} - \ln \left(\frac{s}{t} \left(1 + \rho \frac{t - s}{s} \right)^{1/\rho} \right) \end{aligned}$$

which gives the results by integration. □

A.2 Proof of Proposition 2 Jensen-Kullback-Leibler divergence and bracketing entropy

Proposition 2 Jensen-Kullback-Leibler divergence and bracketing entropy is split into two propositions: one when one controls the the bracketing entropy of the local models $S_m(s_m, \sigma)$ and another when one control the (larger) bracketing entropy of the larger global models S_m .

If one is able to bound the bracketing entropy of the local models, one can use:

Proposition 5. Assume for any $\sigma \in [0, \sqrt{2}]$ and any $\delta \in [0, \sigma]$

$$H_{[\cdot], d^{\otimes n}}(\delta, S_m(s_m, \sigma)) \leq \mathcal{D}_m \left(\mathcal{C}_m + \ln \frac{\sigma}{\delta} \right).$$

Then the function

$$\phi_m(\sigma) = \sigma \sqrt{\mathcal{D}_m} \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} \right)$$

satisfies the properties required in Assumption (H_m) :

- The function $\delta \mapsto \frac{1}{\delta} \phi_m(\delta)$ is non-increasing.
- $\int_0^\sigma \sqrt{H_{[\cdot], d^{\otimes n}}(\delta, S_m(s_m, \sigma))} d\delta \leq \phi_m(\sigma)$.

Furthermore, $\mathfrak{D}_m = n\sigma_m^2$ where σ_m is the unique root of $\frac{1}{\sigma} \phi_m(\sigma) = \sqrt{n}\sigma$ satisfies

$$\mathfrak{D}_m = \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} \right)^2 \mathcal{D}_m.$$

Otherwise, if one is only able to bound the bracketing entropy of the global model, one still has:

Proposition 6. Assume for any $\delta \in [0, \sqrt{2}]$,

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

Then the function

$$\phi_m(\sigma) = \sigma \sqrt{\mathcal{D}_m} \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \left(\sqrt{\ln \frac{1}{\sigma \wedge 1}} \right) \right).$$

satisfies the properties required in Assumption (H_m) :

- The function $\delta \mapsto \frac{1}{\delta} \phi_m(\delta)$ is non-increasing.
- $\int_0^\sigma \sqrt{H_{[\cdot], d^{\otimes n}}(\delta, S_m)} d\delta \leq \phi_m(\sigma)$.

Furthermore, $\mathfrak{D}_m = n\sigma_m^2$ where σ_m is the unique root of $\frac{1}{\sigma} \phi_m(\sigma) = \sqrt{n}\sigma$ satisfies

$$\mathfrak{D}_m \leq \left(2 \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} \right)^2 + \left(\ln \frac{n}{\left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} \right)^2 \mathcal{D}_m} \right)_+ \right) \mathcal{D}_m$$

where $(x)_+ = x$ if $x \geq 0$ and $(x)_+ = 0$ otherwise.

Proof of Proposition 5 *Proof of Proposition 2* *Jensen-Kullback-Leibler divergence and bracketing entropy.*
The function

$$\delta \mapsto \frac{1}{\delta} \phi_m(\delta) = \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} \right) \sqrt{\mathcal{D}_m}$$

is non increasing.

Now,

$$\begin{aligned} \int_0^\sigma \sqrt{H_{[\cdot], d^{\otimes n}}(\delta, S_m(s_m, \sigma))} d\delta &\leq \int_0^\sigma \sqrt{\mathcal{D}_m \left(\mathcal{C}_m + \ln \frac{\sigma}{\delta} \right)} d\delta \leq \int_0^\sigma \sqrt{\mathcal{C}_m} + \sqrt{\ln \frac{\sigma}{\delta}} d\delta \sqrt{\mathcal{D}_m} \\ &\leq \sigma \int_0^1 \sqrt{\mathcal{C}_m} + \sqrt{\ln \frac{1}{\delta}} d\delta \sqrt{\mathcal{D}_m} \end{aligned}$$

We use now

Lemma 1. For any $\sigma \in [0, 1]$, $\int_0^\sigma \sqrt{\ln \frac{1}{\delta}} d\delta \leq \sigma \left(\sqrt{\ln \frac{1}{\sigma}} + \sqrt{\pi} \right)$.

proved in Maugis and Michel [35] to obtain

$$\leq \sigma \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} \right) \sqrt{\mathcal{D}_m}$$

By definition of $\phi_m(\sigma)$:

$$\frac{1}{\sigma} \phi_m(\sigma) = \sqrt{n} \sigma \Leftrightarrow \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} \right) \sqrt{\mathcal{D}_m} = \sqrt{n} \sigma \Leftrightarrow \sigma = \frac{1}{\sqrt{n}} \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} \right) \sqrt{\mathcal{D}_m}$$

Squaring this equality and multiplying by n yields the equality of the Proposition. \square

Proof of Proposition 6 *Proof of Proposition 2* *Jensen-Kullback-Leibler divergence and bracketing entropy.*
The function

$$\delta \mapsto \frac{1}{\delta} \phi_m(\delta) = \sqrt{\mathcal{D}_m} \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \sqrt{\ln \frac{1}{\delta \wedge 1}} \right).$$

is non increasing by construction.

Now

$$\begin{aligned} \int_0^\sigma \sqrt{H_{[\cdot], d^{\otimes n}}(\delta, S_m)} d\delta &\leq \int_0^{\sigma \wedge 1} \sqrt{\mathcal{D}_m} \sqrt{\mathcal{C}_m + \ln \frac{1}{\delta}} d\delta + \int_{\sigma \wedge 1}^\sigma \sqrt{\mathcal{D}_m} \sqrt{\mathcal{C}_m} d\delta \\ &\leq \left(\sigma \sqrt{\mathcal{C}_m} + \int_0^{\sigma \wedge 1} \sqrt{\ln \frac{1}{\delta}} d\delta \right) \sqrt{\mathcal{D}_m} \end{aligned}$$

and using Lemma 1 *Proof of Proposition 2* *Jensen-Kullback-Leibler divergence and bracketing entropy*

$$\begin{aligned} \int_0^\sigma \sqrt{H_{[\cdot], d^{\otimes n}}(\delta, S_m)} d\delta &\leq \left(\sigma \sqrt{\mathcal{C}_m} + (\sigma \wedge 1) \left(\sqrt{\ln \frac{1}{\sigma \wedge 1}} + \sqrt{\pi} \right) \right) \sqrt{\mathcal{D}_m} \\ &\leq \sigma \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \sqrt{\ln \frac{1}{\sigma \wedge 1}} \right) \sqrt{\mathcal{D}_m} \end{aligned}$$

$$\frac{1}{\sigma} \phi_m(\sigma) = \sqrt{n} \sigma \Leftrightarrow \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \sqrt{\ln \frac{1}{\sigma \wedge 1}} \right) \sqrt{\mathcal{D}_m} = \sqrt{n} \sigma$$

$$\Leftrightarrow \sigma = \frac{1}{\sqrt{n}} \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \sqrt{\ln \frac{1}{\sigma \wedge 1}} \right) \sqrt{\mathcal{D}_m}$$

This implies

$$\sigma_m \geq \frac{1}{\sqrt{n}} \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} \right) \sqrt{\mathcal{D}_m}$$

which implies by plugging this bound in the initial equality

$$\begin{aligned} \sigma_m &\leq \frac{1}{\sqrt{n}} \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \sqrt{\ln \frac{\sqrt{n}}{(\sqrt{\mathcal{C}_m} + \sqrt{\pi}) \sqrt{\mathcal{D}_m} \wedge \sqrt{n}}} \right) \sqrt{\mathcal{D}_m} \\ &\leq \frac{1}{\sqrt{n}} \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \sqrt{\frac{1}{2} \left(\ln \frac{n}{(\sqrt{\mathcal{C}_m} + \sqrt{\pi})^2 \mathcal{D}_m} \right)_+} \right) \sqrt{\mathcal{D}_m} \end{aligned}$$

The bound of the Proposition is obtained by squaring this inequality, using the inequality $(\sqrt{a} + \sqrt{b})^2 \leq 2(a + b)$ and multiplying by n . \square

A.3 Proof of Theorem 1 Single model maximum likelihood estimation

Proof of Theorem 1 Single model maximum likelihood estimation. For any non random function g we define its empirical process $P_n^{\otimes n}(g)$ by

$$P_n^{\otimes n}(g) = \frac{1}{n} \sum_{i=1}^n g(X_i, Y_i)$$

and its mean $P^{\otimes n}(g)$ by

$$P^{\otimes n}(g) = \mathbb{E} [P_n^{\otimes n}(g)] = \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n g(X'_i, Y'_i) \right]$$

where (X'_i, Y'_i) is again an independent copy of (X_i, Y_i) . Note that g may depends on the covariate and thus the last term can not generally be simplified. We will denote by $\nu_n^{\otimes n}(g)$ the recentred process $P_n^{\otimes n}(g) - P^{\otimes n}(g)$.

Note that using this definition,

$$\begin{aligned} KL^{\otimes n}(s_0, t) &= P^{\otimes n} \left(-\ln \left(\frac{t}{s_0} \right) \right) \\ JKL_{\rho}^{\otimes n}(s_0, t) &= P^{\otimes n} \left(-\frac{1}{\rho} \ln \left(\frac{(1-\rho)s_0 + \rho t}{s_0} \right) \right) \end{aligned}$$

where we have dropped the indice λ to simplify the notation.

By construction, \widehat{s}_m satisfies

$$P_n^{\otimes n}(-\ln \widehat{s}_m) \leq \inf_{s_m \in S_m} P_n^{\otimes n}(-\ln s_m) + \frac{\eta}{n}$$

We let \bar{s}_m be a function such that

$$KL^{\otimes n}(s_0, \bar{s}_m) \leq \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\delta_{KL}}{n}.$$

We define then the functions $kl(\bar{s}_m)$, $kl(\hat{s}_m)$, and $jdkl(\hat{s}_m)$ by

$$kl(\bar{s}_m) = -\ln\left(\frac{\bar{s}_m}{s_0}\right) \quad kl(\hat{s}_m) = -\ln\left(\frac{\hat{s}_m}{s_0}\right) \quad jdkl(\hat{s}_m) = -\frac{1}{\rho}\ln\left(\frac{(1-\rho)s_0 + \rho\hat{s}_m}{s_0}\right)$$

By construction

$$P_n^{\otimes n}(kl(\hat{s}_m)) \leq P_n^{\otimes n}(kl(\bar{s}_m)) + \frac{\eta}{n}$$

Since, by concavity of the logarithm,

$$jdkl(\hat{s}_m) = -\frac{1}{\rho}\ln\left(\frac{(1-\rho)s_0 + \rho\hat{s}_m}{s_0}\right) \leq -\frac{1}{\rho}\left((1-\rho)\ln\frac{s_0}{s_0} + \rho\ln\frac{\hat{s}_m}{s_0}\right) = -\ln\frac{\hat{s}_m}{s_0} = kl(\hat{s}_m),$$

$$P_n^{\otimes n}(jdkl(\hat{s}_m)) \leq P_n^{\otimes n}(kl(\bar{s}_m)) + \frac{\eta}{n}$$

and thus

$$P_n^{\otimes n}(jdkl(\hat{s}_m)) - \nu_n^{\otimes n}(kl(\bar{s}_m)) \leq P_n^{\otimes n}(kl(\bar{s}_m)) - \nu_n^{\otimes n}(jdkl(\hat{s}_m)) + \frac{\eta}{n}$$

using the definition of $jdkl(\hat{s}_m)$ and of $kl(\bar{s}_m)$, we deduce

$$JKL_\rho^{\otimes n}(s_0, \hat{s}_m) - \nu_n^{\otimes n}(kl(\bar{s}_m)) \leq \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) - \nu_n^{\otimes n}(jdkl(\hat{s}_m)) + \frac{\eta}{n} + \frac{\delta_{KL}}{n}$$

where $JKL_\rho^{\otimes n}(s_0, \hat{s}_m)$ is still a random variable.

We rely now on a control on the deviation of $\nu_n^{\otimes n}(jdkl(\hat{s}_m))$ through its conditional expectation. For any random variable Z and any event A such that $\mathbb{P}\{A\} > 0$, we let $\mathbb{E}^A[Z] = \frac{E[Z\mathbf{1}_{\{A\}}]}{\mathbb{P}\{A\}}$. It is sufficient to control those quantity for all A to obtain a control of the deviation. More precisely,

Lemma 2. *Let Z be a random variable, assume it exists a non decreasing Ψ such that for all A such that $\mathbb{P}\{A\} > 0$, $\mathbb{E}^A[Z] \leq \Psi\left(\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)\right)$. then for all x $\mathbb{P}\{Z > \Psi(x)\} \leq e^{-x}$.*

Here, we can prove

Lemma 3. *There exist three absolute constants $\kappa'_0 > 4$, κ'_1 and κ'_2 such that, under Assumption (H), for all $m \in \mathcal{M}$, for every $y_m > \sigma_m$ and every event A such that $\mathbb{P}\{A\} > 0$,*

$$\mathbb{E}^A\left[\nu_n^{\otimes n}\left(\frac{-jdkl(\hat{s}_m)}{y_m^2 + \kappa'_0 d^{2\otimes n}(s_0, \hat{s}_m)}\right)\right] \leq \frac{\kappa'_1 \sigma_m}{y_m} + \kappa'_2 \frac{1}{\sqrt{ny_m^2}} \sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)} + \frac{18}{ny_m^2 \rho} \ln\left(\frac{1}{\mathbb{P}\{A\}}\right).$$

Combining Lemma 2 Proof of Theorem 1 Single model maximum likelihood estimation and Lemma 3 Proof of Theorem 1 Single model maximum likelihood estimation implies that except on a set of probability less than e^{-x} , for any $y_m > \sigma_m$,

$$\frac{-\nu_n^{\otimes n}(jdkl(\hat{s}_m))}{y_m^2 + \kappa'_0 d^{2\otimes n}(s_0, \hat{s}_m)} \leq \frac{\kappa'_1 \sigma_m}{y_m} + \kappa'_2 \sqrt{\frac{x}{ny_m^2}} + \frac{18}{\rho} \frac{x}{ny_m^2}.$$

Choosing $y_m = \theta \sqrt{\sigma_m^2 + \frac{x}{n}}$ with $\theta > 1$ to be fixed later, we deduce that, except on a set of probability less than e^{-x} ,

$$\frac{-\nu_n^{\otimes n}(jkl(\widehat{s}_m))}{y_m^2 + \kappa'_0 d^{2\otimes n}(s_0, \widehat{s}_m)} \leq \frac{\kappa'_1 + \kappa'_2}{\theta} + \frac{18}{\theta^2 \rho}$$

Thus, except on the same set, we have

$$JKL_\rho^{\otimes n}(s_0, \widehat{s}_m) - \nu_n^{\otimes n}(kl(\overline{s}_m)) \leq \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \left(\frac{\kappa'_1 + \kappa'_2}{\theta} + \frac{18}{\theta^2 \rho} \right) (y_m^2 + \kappa'_0 d^{2\otimes n}(s_0, \widehat{s}_m)) + \frac{\eta}{n} + \frac{\delta_{KL}}{n}.$$

Let $\epsilon_{\text{pen}} > 0$, we define θ_{pen} by $\left(\frac{\kappa'_1 + \kappa'_2}{\theta_{\text{pen}}} + \frac{18}{\theta_{\text{pen}}^2 \rho} \right) \kappa'_0 = C_\rho \epsilon_{\text{pen}}$ with $C_\rho = \frac{1}{\rho} \min(\frac{1-\rho}{\rho}, 1) \left(\ln \left(1 + \frac{\rho}{1-\rho} \right) - \rho \right)$ and, using $C_\rho d^{2\otimes n}(s_0, \widehat{s}_m) \leq JKL_\rho^{\otimes n}(s_0, \widehat{s}_m)$, we obtain

$$\begin{aligned} (1 - \epsilon_{\text{pen}}) JKL_\rho^{\otimes n}(s_0, \widehat{s}_m) - \nu_n^{\otimes n}(kl(\overline{s}_m)) &\leq \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{C_\rho \epsilon_{\text{pen}} y_m^2}{\kappa'_0} + \frac{\eta}{n} + \frac{\delta_{KL}}{n} \\ &\leq \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa'_0} \left(\sigma_m^2 + \frac{x}{n} \right) + \frac{\eta}{n} + \frac{\delta_{KL}}{n} \end{aligned}$$

As $\nu_n^{\otimes n}(kl(\overline{s}_m))$ is integrable (and of mean 0), we obtain by integration

$$\mathbb{E} [JKL_\rho^{\otimes n}(s_0, \widehat{s}_m)] \leq \frac{1}{1 - \epsilon_{\text{pen}}} \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{1}{1 - \epsilon_{\text{pen}}} \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa'_0} \frac{1}{n} + \frac{\eta}{n} + \frac{\delta_{KL}}{n}$$

or if we let $\kappa_0 = \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa'_0}$

$$\leq \frac{1}{1 - \epsilon_{\text{pen}}} \left(\inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \kappa_0 \sigma_m^2 \right) + \frac{\kappa_0}{1 - \epsilon_{\text{pen}}} \frac{1}{n} + \frac{\eta}{n} + \frac{\delta_{KL}}{n}.$$

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

$$\mathbb{E} [JKL_\rho^{\otimes n}(s_0, \widehat{s}_m)] \leq \frac{1}{1 - \epsilon_{\text{pen}}} \left(\inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \kappa_0 \sigma_m^2 \right) + \frac{\kappa_0}{1 - \epsilon_{\text{pen}}} \frac{1}{n} + \frac{\eta + \eta'}{n}$$

and thus $C_1 = \frac{1}{1 - \epsilon_{\text{pen}}}$ and $C_2 = \frac{\kappa_0}{1 - \epsilon_{\text{pen}}}$. \square

B Proofs for Section 3 Model selection and penalized maximum likelihood (Model selection and penalized maximum likelihood)

B.1 Proof of Theorem 2A general theorem for penalized maximum likelihood conditional density estimation

Proof of Theorem 2A general theorem for penalized maximum likelihood conditional density estimation.
For any model S_m , we let \overline{s}_m be a function such that

$$KL^{\otimes n}(s_0, \overline{s}_m) \leq \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\delta_{KL}}{n}.$$

Let $m \in \mathcal{M}$ such that $KL^{\otimes n}(s, \bar{s}_m) < +\infty$ and let

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

For every $m' \in \mathcal{M}'$,

$$P_n^{\otimes n}(kl(\hat{s}_{m'})) + \frac{\text{pen}(m')}{n} \leq P_n^{\otimes n}(kl(\hat{s}_m)) + \frac{\text{pen}(m)}{n} + \frac{\eta'}{n} \leq P_n^{\otimes n}(kl(\bar{s}_m)) + \frac{\text{pen}(m)}{n} + \frac{\eta + \eta'}{n}$$

Since, by concavity of the logarithm, $jdkl(\hat{s}_{m'}) \leq kkl(\hat{s}_{m'})$,

$$P_n^{\otimes n}(jdkl(\hat{s}_{m'})) + \frac{\text{pen}(m')}{n} \leq P_n^{\otimes n}(kkl(\bar{s}_m)) + \frac{\text{pen}(m)}{n} + \frac{\eta + \eta'}{n}$$

and thus

$$P^{\otimes n}(jdkl(\hat{s}_{m'})) - \nu_n^{\otimes n}(kkl(\bar{s}_m)) \leq P^{\otimes n}(kkl(\bar{s}_m)) + \frac{\text{pen}(m)}{n} - \nu_n^{\otimes n}(jdkl(\hat{s}_{m'})) - \frac{\text{pen}(m')}{n} + \frac{\eta + \eta'}{n}$$

using the definition of $jdkl(\hat{s}_{m'})$ and of $kkl(\bar{s}_m)$, we deduce

$$JKL_{\rho}^{\otimes n}(s_0, \hat{s}_{m'}) - \nu_n^{\otimes n}(kkl(\bar{s}_m)) \leq \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\text{pen}(m)}{n} - \nu_n^{\otimes n}(jdkl(\hat{s}_{m'})) - \frac{\text{pen}(m')}{n} + \frac{\eta + \eta'}{n} + \frac{\delta_{KL}}{n}$$

Combining again Lemma 2Proof of Theorem 1Single model maximum likelihood estimation and Lemma 3Proof of Theorem 1Single model maximum likelihood estimation, we deduce that except on a set of probability less than $e^{-x_{m'}-x}$, for any $y_{m'} > \sigma_{m'}$,

$$\frac{-\nu_n^{\otimes n}(jdkl(\hat{s}_{m'}))}{y_{m'}^2 + \kappa'_0 d^{2\otimes n}(s_0, \hat{s}_{m'})} \leq \frac{\kappa'_1 \sigma_{m'}}{y_{m'}} + \kappa'_2 \sqrt{\frac{x_{m'} + x}{ny_{m'}^2}} + \frac{18}{\rho} \frac{x_{m'} + x}{ny_{m'}^2}.$$

Choosing this time $y_{m'} = \theta \sqrt{\sigma_{m'}^2 + \frac{x_{m'} + x}{n}}$ with $\theta > 1$ to be fixed later, we deduce that, except on a set of probability less than $e^{-x_{m'}-x}$,

$$\frac{-\nu_n^{\otimes n}(jdkl(\hat{s}_{m'}))}{y_{m'}^2 + \kappa'_0 d^{2\otimes n}(s_0, \hat{s}_{m'})} \leq \frac{\kappa'_1 + \kappa'_2}{\theta} + \frac{18}{\theta^2 \rho}$$

Using the Kraft condition of Assumption (K), we deduce that if we make this choice of $y_{m'}$ for all model m' , this properties holds simultaneously for all $m' \in \mathcal{M}$ except on a set of probability less than Σe^{-x} .

Thus, except on the same set, simultaneously for all $m \in \mathcal{M}'$, we have

$$JKL_{\rho}^{\otimes n}(s_0, \hat{s}_{m'}) - \nu_n^{\otimes n}(kkl(\bar{s}_m)) \leq \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\text{pen}(m)}{n} + \left(\frac{\kappa'_1 + \kappa'_2}{\theta} + \frac{18}{\theta^2 \rho} \right) (y_{m'}^2 + \kappa'_0 d^{2\otimes n}(s_0, \hat{s}_{m'})) - \frac{\text{pen}(m')}{n} + \frac{\eta + \eta'}{n} + \frac{\delta_{KL}}{n}.$$

Let $\epsilon_{\text{pen}} > 0$, we define θ_{pen} by $\left(\frac{\kappa'_1 + \kappa'_2}{\theta_{\text{pen}}} + \frac{18}{\theta_{\text{pen}}^2 \rho} \right) \kappa'_0 = C_{\rho} \epsilon_{\text{pen}}$ with $C_{\rho} = \frac{1}{\rho} \min(\frac{1-\rho}{\rho}, 1) \left(\ln \left(1 + \frac{\rho}{1-\rho} \right) - \rho \right)$ and, using $C_{\rho} d^{2\otimes n}(s_0, \hat{s}_{m'}) \leq JKL_{\rho}^{\otimes n}(s_0, \hat{s}_{m'})$, we obtain

$$(1 - \epsilon_{\text{pen}})JKL_{\rho}^{\otimes n}(s_0, \hat{s}_{m'}) - \nu_n^{\otimes n}(kkl(\bar{s}_m)) \leq \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\text{pen}(m)}{n}$$

$$+ \frac{C_\rho \epsilon_{\text{pen}} y_{m'}^2}{\kappa'_0} - \frac{\text{pen}(m')}{n} + \frac{\eta + \eta'}{n} + \frac{\delta_{KL}}{n}.$$

We should now study $\frac{C_\rho \epsilon_{\text{pen}} y_{m'}^2}{\kappa'_0} - \frac{\text{pen}(m')}{n}$:

$$\frac{C_\rho \epsilon_{\text{pen}} y_{m'}^2}{\kappa'_0} - \frac{\text{pen}(m')}{n} = \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa'_0} \left(\sigma_m^2 + \frac{x_m + x}{n} \right) - \frac{\text{pen}(m')}{n}$$

and by construction if $\kappa_0 = \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa'_0}$

$$\frac{C_\rho \epsilon_{\text{pen}} y_{m'}^2}{\kappa'_0} - \frac{\text{pen}(m')}{n} \leq \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa'_0} \frac{x}{n} - \left(1 - \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa \kappa'_0}\right) \frac{\text{pen}(m')}{n}.$$

We deduce thus, except on a set of probability smaller than Σe^{-x} , simultaneously for any $m' \in \mathcal{M}'$

$$\begin{aligned} (1 - \epsilon_{\text{pen}}) JKL_\rho^{\otimes n}(s_0, \widehat{s}_{m'}) + \left(1 - \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa \kappa'_0}\right) \frac{\text{pen}(m')}{n} - \nu_n^{\otimes n}(kl(\overline{s}_m)) \\ \leq \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\text{pen}(m)}{n} + \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa'_0} \frac{x}{n} + \frac{\eta + \eta'}{n} + \frac{\delta_{KL}}{n} \end{aligned}$$

As $\nu_n^{\otimes n}(kl(\overline{s}_m))$ is integrable (and of mean 0), we derive that $M = \sup_{m' \in \mathcal{M}'} \frac{\text{pen}(m')}{n}$ is almost surely finite, so that as $\kappa \frac{x_{m'}}{n} \leq M$ for every $m' \in \mathcal{M}'$, one has

$$\Sigma \geq \sum_{m' \in \mathcal{M}'} e^{-x_{m'}} \geq |\mathcal{M}'| e^{-\frac{Mn}{\kappa}}$$

and thus \mathcal{M}' is almost surely finite. This implies that the some minimizer \widehat{m} of $P_n^{\otimes n}(-\ln(\widehat{s}_m)) + \frac{\text{pen}(m)}{n}$ exists.

For this minimizer, one has with probability greater than $1 - \Sigma e^{-x}$,

$$\begin{aligned} (1 - \epsilon_{\text{pen}}) JKL_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{m}}) - \nu_n^{\otimes n}(kl(\overline{s}_m)) \\ \leq \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\text{pen}(m)}{n} + \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa'_0} \frac{x}{n} + \frac{\eta + \eta'}{n} + \frac{\delta_{KL}}{n} \end{aligned}$$

which yields by integration

$$\begin{aligned} \mathbb{E} [JKL_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})] &\leq \frac{1}{1 - \epsilon_{\text{pen}}} \inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{1}{1 - \epsilon_{\text{pen}}} \frac{\text{pen}(m)}{n} + \frac{1}{1 - \epsilon_{\text{pen}}} \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa'_0} \frac{\Sigma}{n} + \frac{\eta + \eta'}{n} + \frac{\delta_{KL}}{n} \\ &\leq \frac{1}{1 - \epsilon_{\text{pen}}} \left(\inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\text{pen}(m)}{n} \right) + \frac{\kappa_0}{1 - \epsilon_{\text{pen}}} \frac{\Sigma}{n} + \frac{\eta + \eta'}{n} + \frac{\delta_{KL}}{n}. \end{aligned}$$

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

$$\mathbb{E} [JKL_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})] \leq \frac{1}{1 - \epsilon_{\text{pen}}} \left(\inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\text{pen}(m)}{n} \right) + \frac{\kappa_0}{1 - \epsilon_{\text{pen}}} \frac{\Sigma}{n} + \frac{\eta + \eta'}{n}$$

and thus $C_1 = \frac{1}{1 - \epsilon_{\text{pen}}}$ and $C_2 = \frac{\kappa_0}{1 - \epsilon_{\text{pen}}}$. \square

B.2 Proof of Lemma 2 Proof of Theorem 1 Single model maximum likelihood estimation

Proof of Lemma 2 Proof of Theorem 1 Single model maximum likelihood estimation. Let $A = \{Z > \Psi(x)\}$. Either $\mathbb{P}\{A\} = 0 \leq e^{-x}$ or

$$\mathbb{E}^A[Z] \leq \Psi \left(\ln \left(\frac{1}{\mathbb{P}\{A\}} \right) \right).$$

Now in the later case,

$$\mathbb{E}^A[Z] = \frac{\mathbb{E}[Z \mathbf{1}_{\{Z > \Psi(x)\}}]}{\mathbb{P}\{Z > \Psi(x)\}} \geq \Psi(x).$$

We have thus $\Psi(x) \leq \Psi \left(\ln \left(\frac{1}{\mathbb{P}\{A\}} \right) \right)$ which implies $x \leq \ln \left(\frac{1}{\mathbb{P}\{A\}} \right)$ as Ψ is not decreasing. This last inequality yields $\mathbb{P}\{A\} \leq e^{-x}$ which concludes the proof. \square

B.3 Proof of Lemma 3 Proof of Theorem 1 Single model maximum likelihood estimation

We should now prove Lemma 3 Proof of Theorem 1 Single model maximum likelihood estimation which contains most of the differences with Massart [34]'s proof.

Proof of Lemma 3 Proof of Theorem 1 Single model maximum likelihood estimation. In this lemma, we want to control the deviation of

$$\nu_n^{\otimes n}(-jkl(\widehat{s}_m)) = \nu_n^{\otimes n} \left(\frac{1}{\rho} \ln \left(\frac{(1-\rho)s_0 + \rho\widehat{s}_m}{s_0} \right) \right).$$

Note that for any \widetilde{s}_m to be fixed later, if we let $jkl(\widetilde{s}) = -\frac{1}{\rho} \ln \left(\frac{(1-\rho)s_0 + \rho\widetilde{s}_m}{s_0} \right)$, then $-jkl(\widehat{s}_m) = -jkl(\widetilde{s}) + (-jkl(\widehat{s}_m) + jkl(\widetilde{s}))$ with

$$-jkl(\widehat{s}_m) + jkl(\widetilde{s}) = \frac{1}{\rho} \ln \left(\frac{(1-\rho)s_0 + \rho\widehat{s}_m}{(1-\rho)s_0 + \rho\widetilde{s}_m} \right)$$

To control the behavior of these quantities, we use the following key properties of Jensen-Kullback-Leibler related quantities (a rewriting of Lemma 7.26 of Massart [34])

Lemma 4. *Let P be some probability measure with density s_0 with respect to some measure λ and s, t be some non-negative and λ integrable functions, then one has for every integer $k \geq 2$*

$$P \left(\left| \ln \left(\frac{s_0 + s}{s_0 + t} \right) \right|^k \right) \leq \frac{k!}{2} \left(\frac{9 \|\sqrt{s} - \sqrt{t}\|_{\lambda,2}^2}{8} \right) 2^{k-2}$$

where $\|\cdot\|_{\lambda,2}$ is the λ - L^2 norm so that $\|\sqrt{s} - \sqrt{t}\|_{\lambda,2}^2$ is nothing but the extended Hellinger distance.

In our context this implies, conditioning first by $(X_i)_{1 \leq i \leq n}$, applying the previous inequality for each $(s_0(\cdot|X_i), s(\cdot|X_i), t(\cdot|X_i))$ and then taking the expectation, that

$$P^{\otimes n} \left(\left| \frac{1}{\rho} \ln \left(\frac{s_0 + \frac{\rho}{1-\rho}s}{s_0 + \frac{\rho}{1-\rho}t} \right) \right|^k \right) \leq \frac{k!}{2} \left(\frac{9d^{2 \otimes n}(s, t)}{8\rho(1-\rho)} \right) \left(\frac{2}{\rho} \right)^{k-2}.$$

As

Theorem 5. Assume f is a function such that

$$\begin{aligned} P^{\otimes n}(|f|^2) &\leq V \\ \forall k \geq 3, \quad P^{\otimes n}((f)_+^k) &\leq \frac{k!}{2} V b^{k-2}. \end{aligned}$$

Then for all A such that $\mathbb{P}\{A\} > 0$

$$\mathbb{E}^A(\nu_n^{\otimes n}(f)) \leq \frac{\sqrt{2V}}{\sqrt{n}} \sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right) + \frac{b}{n} \ln\left(\frac{1}{\mathbb{P}\{A\}}\right)}.$$

holds, these bounds are sufficient to obtain a Bernstein type control for $ijkl(\tilde{s})$

$$\mathbb{E}^A[-\nu_n^{\otimes n}(ijkl(\tilde{s}))] \leq \frac{3}{2\sqrt{\rho(1-\rho)}} \frac{\sqrt{d^{2\otimes n}(s_0, \tilde{s}_m)}}{\sqrt{n}} \sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right) + \frac{2}{n\rho} \ln\left(\frac{1}{\mathbb{P}\{A\}}\right)}.$$

To cope with the randomness of \hat{s}_m , we rely on the following much more involved theorem (a rewriting of Theorem 6.8 of Massart [34])

Theorem 6. Let \mathcal{G} be some countable class of real valued and measurable functions on ξ . Assume that there exist some positive numbers v and b such that for all $f \in \mathcal{G}$ and all integers $k \geq 2$

$$P^{\otimes n}(|f|^k) \leq \frac{k!}{2} V b^{k-2}$$

Assume furthermore that for any positive number δ , it exists some finite set $B(\delta)$ of brackets covering \mathcal{G} such that for any bracket $[g^-, g^+] \in B(\delta)$ and all integer $k \geq 2$

$$P^{\otimes n}(|g^+ - g^-|^k) \leq \frac{k!}{2} \delta^2 b^{k-2}$$

Let $e^{H(\delta)}$ denote the minimal cardinality of such a covering. It exists some absolute constant κ such that, for any $\epsilon \in (0, 1]$ and any measurable set A with $\mathbb{P}\{A\} > 0$, we have

$$E^A \left[\sup_{f \in \mathcal{G}} \nu_n^{\otimes n}(f) \right] \leq E + \frac{(1+6\epsilon)\sqrt{2V}}{\sqrt{n}} \sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right) + \frac{2b}{n} \ln\left(\frac{1}{\mathbb{P}\{A\}}\right)}$$

where

$$E = \frac{\kappa}{\epsilon} \frac{1}{\sqrt{n}} \int_0^{\epsilon\sqrt{V}} \sqrt{H(\delta) \wedge n\delta} + \frac{2(b+\sqrt{V})}{n} H(\sqrt{V}).$$

Furthermore $\kappa \leq 27$.

If we consider

$$\begin{aligned} \mathcal{G}_m(\tilde{s}_m, \sigma) &= \left\{ -ijkl(s_m) + jkl(\tilde{s}) = \frac{1}{\rho} \ln\left(\frac{s_0 + \frac{\rho}{1-\rho}s_m}{s_0 + \frac{\rho}{1-\rho}\tilde{s}_m}\right) \middle| s_m \in S_m, d^{2\otimes n}(\tilde{s}_m, s_m) \leq \sigma \right\} \\ &= \left\{ \frac{1}{\rho} \ln\left(\frac{s_0 + \frac{\rho}{1-\rho}s_m}{s_0 + \frac{\rho}{1-\rho}\tilde{s}_m}\right) \middle| s_m \in S_m(\tilde{s}_m, \sigma) \right\}. \end{aligned}$$

then the first assumption of Theorem 6Proof of Lemma 3Proof of Theorem 1Single model maximum likelihood estimation holds with $V = \left(\frac{3\sigma}{2\sqrt{2\rho(1-\rho)}}\right)^2$ and $b = \frac{2}{\rho}$.

We are thus focusing on

$$\begin{aligned} W_m(\tilde{s}_m, \sigma) &= \sup_{f \in \mathcal{G}_m(\tilde{s}_m, \sigma)} \nu_n^{\otimes n}(f) = \sup_{s_m \in S_m(\tilde{s}_m, \sigma)} \nu_n^{\otimes n}(-jkl(s_m) + jkl(\tilde{s})) \\ &= \sup_{s_m \in S_m(\tilde{s}_m, \sigma)} \nu_n^{\otimes n}(-jkl(s_m)) + \nu_n^{\otimes n}(jkl(\tilde{s})) \end{aligned}$$

Now if $[t^-, t^+]$ is a bracket containing s , then

$$g^- = \frac{1}{\rho} \ln \left(\frac{s_0 + \frac{\rho}{1-\rho} t^-}{s_0 + \frac{\rho}{1-\rho} \tilde{s}_m} \right) \leq \frac{1}{\rho} \ln \left(\frac{s_0 + \frac{\rho}{1-\rho} s}{s_0 + \frac{\rho}{1-\rho} \tilde{s}_m} \right) \leq \frac{1}{\rho} \ln \left(\frac{s_0 + \frac{\rho}{1-\rho} t^+}{s_0 + \frac{\rho}{1-\rho} \tilde{s}_m} \right) = g^+$$

and

$$g^+ - g^- = \frac{1}{\rho} \ln \left(\frac{s_0 + \frac{\rho}{1-\rho} t^+}{s_0 + \frac{\rho}{1-\rho} \tilde{s}_m} \right) - \frac{1}{\rho} \ln \left(\frac{s_0 + \frac{\rho}{1-\rho} t^-}{s_0 + \frac{\rho}{1-\rho} \tilde{s}_m} \right) = \frac{1}{\rho} \ln \left(\frac{s_0 + \frac{\rho}{1-\rho} t^+}{s_0 + \frac{\rho}{1-\rho} t^-} \right)$$

So that

$$P^{\otimes n}(|g^+ - g^-|^k) \leq \frac{k!}{2} \delta^2 b^{k-2}$$

as soon as $\frac{3d^{\otimes n}(t^-, t^+)}{2\sqrt{2\rho(1-\rho)}} \leq \delta$. This implies that, for any $\delta > 0$, one can construct a set of brackets satisfying the second assumption of Theorem 6 Proof of Lemma 3 Proof of Theorem 1 Single model maximum likelihood estimation from a set of brackets of $d^{\otimes n}$ width smaller than $\frac{2\sqrt{2\rho(1-\rho)}}{3} \delta$ covering $S_m(\tilde{s}_m, \sigma)$. That is

$$H(\delta) \leq H_{[\cdot], d^{\otimes n}} \left(\frac{2\sqrt{2\rho(1-\rho)}}{3} \delta, S_m(\tilde{s}_m, \sigma) \right).$$

As $\mathcal{G} \subset S_m$ satisfies (S), one can apply the theorem. We obtain for every measurable set A with $\mathbb{P}\{A\} > 0$,

$$\begin{aligned} \mathbb{E}^A [W_m(\tilde{s}_m, \sigma)] &\leq E + \frac{(1+6\epsilon)3\sigma}{2\sqrt{\rho(1-\rho)}\sqrt{n}} \sqrt{\ln \left(\frac{1}{\mathbb{P}\{A\}} \right) + \frac{4}{\rho n} \ln \left(\frac{1}{\mathbb{P}\{A\}} \right)} \\ \text{where } E &= \frac{\kappa}{\epsilon} \frac{1}{\sqrt{n}} \int_0^{\epsilon \frac{3\sigma}{2\sqrt{2\rho(1-\rho)}}} \sqrt{H_{[\cdot], d^{\otimes n}} \left(\frac{2\sqrt{2\rho(1-\rho)}}{3} \delta, S_m(\tilde{s}_m, \sigma) \right) \wedge n d \delta} \\ &\quad + \frac{2\left(\frac{2}{\rho} + \frac{3\sigma}{2\sqrt{2\rho(1-\rho)}}\right)}{n} H_{[\cdot], d^{\otimes n}} \left(\frac{2\sqrt{2\rho(1-\rho)}}{3} \frac{3\sigma}{2\sqrt{2\rho(1-\rho)}}, S_m(\tilde{s}_m, \sigma) \right) \\ &= \frac{3\kappa}{2\epsilon\sqrt{2\rho(1-\rho)}} \frac{1}{\sqrt{n}} \int_0^{\epsilon\sigma} \sqrt{H_{[\cdot], d^{\otimes n}}(\delta, S_m(\tilde{s}_m, \sigma)) \wedge n d \delta} \\ &\quad + \frac{2\left(\frac{2}{\rho} + \frac{3\sigma}{2\sqrt{2\rho(1-\rho)}}\right)}{n} H_{[\cdot], d^{\otimes n}}(\sigma, S_m(\tilde{s}_m, \sigma)) \end{aligned}$$

Choosing $\epsilon = 1$ leads to

$$\mathbb{E}^A [W_m(\tilde{s}_m, \sigma)] \leq E + \frac{21\sigma}{2\sqrt{\rho(1-\rho)}\sqrt{n}} \sqrt{\ln \left(\frac{1}{\mathbb{P}\{A\}} \right) + \frac{4}{\rho n} \ln \left(\frac{1}{\mathbb{P}\{A\}} \right)}$$

where

$$E = \frac{3\kappa}{2\sqrt{2\rho(1-\rho)}} \frac{1}{\sqrt{n}} \int_0^\sigma \sqrt{H_{[\cdot], d^{\otimes n}}(\delta, S_m(\tilde{s}_m, \sigma))} \wedge n d\delta + \frac{2\left(\frac{2}{\rho} + \frac{3\sigma}{2\sqrt{2\rho(1-\rho)}}\right)}{n} H_{[\cdot], d^{\otimes n}}(\sigma, S_m(\tilde{s}_m, \sigma))$$

By definition, if we assume $\tilde{s}_m \in S_m$, $\int_0^\sigma \sqrt{H_{[\cdot], d^{\otimes n}}(\delta, S_m(\tilde{s}_m, \sigma))} \wedge n d\delta \leq \phi_m(\sigma)$, as well as $\delta \mapsto H_{[\cdot], d^{\otimes n}}(\delta, S_m(\tilde{s}_m, \sigma))$ is non-increasing. This implies

$$H_{[\cdot], d^{\otimes n}}(\sigma, S_m(\tilde{s}_m, \sigma)) \leq \left(\frac{1}{\sigma} \int_0^\sigma \sqrt{H_{[\cdot], d^{\otimes n}}(\delta, S_m(\tilde{s}_m, \sigma))} d\delta \right)^2 \leq \frac{\phi_m^2(\sigma)}{\sigma^2}.$$

Inserting these bounds in the previous inequality yields

$$\begin{aligned} E &\leq \frac{3\kappa}{2\sqrt{2\rho(1-\rho)}} \frac{\phi_m(\sigma)}{\sqrt{n}} + \left(\frac{4}{\rho} + \frac{3\sigma}{\sqrt{2\rho(1-\rho)}} \right) \frac{\phi_m^2(\sigma)}{n\sigma^2} \\ &\leq \left(\frac{3\kappa}{2\sqrt{2\rho(1-\rho)}} + \left(\frac{4}{\rho} + \frac{3\sigma}{\sqrt{2\rho(1-\rho)}} \right) \frac{\phi_m(\sigma)}{\sqrt{n}\sigma^2} \right) \frac{\phi_m(\sigma)}{\sqrt{n}}. \end{aligned}$$

As $\delta \mapsto \delta^{-1}\phi_m(\delta)$ is also non-increasing, so is $\delta \mapsto \delta^{-2}\phi_m(\delta)$. The definition of σ_m by the equation $\frac{\phi_m(\sigma_m)}{\sqrt{n}\sigma_m^2} = 1$ is such that this is an $O\left(\frac{\phi_m(\sigma)}{\sqrt{n}}\right)$ as soon as $\sigma \geq \sigma_m$. Indeed under this assumption,

$$E \leq \left(\frac{3\kappa}{2\sqrt{2\rho(1-\rho)}} + \frac{4}{\rho} + \frac{3\sigma}{\sqrt{2\rho(1-\rho)}} \right) \frac{\phi_m(\sigma)}{\sqrt{n}}$$

and

$$\mathbb{E}^A [W_m(\tilde{s}_m, \sigma)] \leq \left(\frac{3\kappa}{2\sqrt{2\rho(1-\rho)}} + \frac{4}{\rho} + \frac{3\sigma}{\sqrt{2\rho(1-\rho)}} \right) \frac{\phi_m(\sigma)}{\sqrt{n}} + \frac{21\sigma}{2\sqrt{\rho(1-\rho)}\sqrt{n}} \sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)} + \frac{4}{\rho n} \ln\left(\frac{1}{\mathbb{P}\{A\}}\right)$$

Using now $\sigma \leq \sqrt{2}$, we let $\kappa_1'' = \left(\frac{3\kappa}{2\sqrt{2\rho(1-\rho)}} + \frac{4}{\rho} + \frac{3}{\sqrt{\rho(1-\rho)}} \right) \leq \left(\frac{81}{2\sqrt{\rho(1-\rho)}} + \frac{4}{\rho} + \frac{3}{\sqrt{\rho(1-\rho)}} \right)$ as $\kappa \leq 27$, $\kappa_2'' = \frac{21}{2\sqrt{\rho(1-\rho)}}$, we have thus obtained $\forall \sigma > \sigma_m$,

$$\mathbb{E}^A \left[\sup_{s_m \in S_m(\tilde{s}_m, \sigma)} \nu_n^{\otimes n}(-jkl(s_m) + jkl(\tilde{s})) \right] \leq \kappa_1'' \frac{\phi_m(\sigma)}{\sqrt{n}} + \frac{\kappa_2'' \sigma}{\sqrt{n}} \sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)} + \frac{4}{\rho n} \ln\left(\frac{1}{\mathbb{P}\{A\}}\right).$$

Thanks to Assumption (S), we can use the *peeling* lemma (Lemma 4.23 of [34]):

Lemma 5. *Let S be some countable set, $\tilde{s} \in S$ and $a : S \rightarrow \mathbb{R}^+$ such that $a(\tilde{s}) = \inf_{s \in S} a(s)$. Let Z be some random process indexed by S and let*

$$B(\sigma) = \{s \in S | a(s) \leq \sigma\},$$

assume that for any positive σ the non-negative random variable $\sup_{s \in B(\sigma)} (Z(s) - Z(\tilde{s}))$ has finite expectation. Then, for any function ψ on \mathbb{R}^+ such that $\psi(x)/x$ is non-increasing on \mathbb{R}^+ and

$$\mathbb{E} \left[\sup_{s \in B(\sigma)} (Z(s) - Z(\tilde{s})) \right] \leq \psi(\sigma), \quad \text{for any } \sigma \geq \sigma_* \geq 0,$$

one has for any positive number $x \geq \sigma_*$

$$\mathbb{E} \left[\sup_{s \in S} \frac{Z(s) - Z(\tilde{s})}{x^2 + a^2(s)} \right] \leq 4x^{-2} \psi(x).$$

With $S = S_m$, $\tilde{s} = \tilde{s}_m \in S_m$ to be specified with $a(s) = d^{2\otimes n}(\tilde{s}_m, s)$ and $Z(s) = -jkl(s)$.
 Provided $y_m \geq \sigma_m$, one obtains

$$\mathbb{E}^A \left[\sup_{s_m \in S_m} \nu_n^{\otimes n} \left(\frac{-jkl(s_m) + jkl(\tilde{s}_m)}{y_m^2 + d^{2\otimes n}(\tilde{s}_m, s_m)} \right) \right] \leq 4\kappa_1'' \frac{\phi_m(y_m)}{\sqrt{ny_m^2}} + \frac{4\kappa_2'' \sigma}{\sqrt{ny_m^2}} \sqrt{\ln \left(\frac{1}{\mathbb{P}\{A\}} \right)} + \frac{16}{\rho ny_m^2} \ln \left(\frac{1}{\mathbb{P}\{A\}} \right).$$

Now using again the monotonicity of $\delta \mapsto \delta^{-1} \phi_m(\delta)$ and the definition of σ_m , $\forall y_m \geq \sigma_m$,

$$\frac{\phi_m(y_m)}{\sqrt{ny_m}} \leq \frac{\phi_m(\sigma_m)}{\sqrt{n\sigma_m}} = \sigma_m$$

and therefore

$$\mathbb{E}^A \left[\sup_{s_m \in S_m} \nu_n^{\otimes n} \left(\frac{-jkl(s_m) + jkl(\tilde{s}_m)}{y_m^2 + d^{2\otimes n}(\tilde{s}_m, s_m)} \right) \right] \leq \frac{4\kappa_1'' \sigma_m}{y_m} + \frac{4\kappa_2''}{\sqrt{ny_m^2}} \sqrt{\ln \left(\frac{1}{\mathbb{P}\{A\}} \right)} + \frac{16}{\rho ny_m^2} \ln \left(\frac{1}{\mathbb{P}\{A\}} \right).$$

We can now choose \tilde{s}_m such that for every $s_m \in S_m$

$$d^{2\otimes n}(s_0, \tilde{s}_m) \leq (1 + \epsilon_d) d^{2\otimes n}(s_0, s_m)$$

so that

$$\begin{aligned} d^{2\otimes n}(\tilde{s}_m, s_m) &= P^{\otimes n} (d^2(\tilde{s}_m, s_m)) \leq P^{\otimes n} \left((d(\tilde{s}_m, s_0) + d(s_0, s_m))^2 \right) \leq 2P^{\otimes n} (d^2(\tilde{s}_m, s_0) + d^2(s_0, s_m)) \\ &\leq 2(2 + \epsilon_d) d^{2\otimes n}(s_0, s_m). \end{aligned}$$

For this choice, one obtains

$$\mathbb{E}^A \left[\sup_{s_m \in S_m} \nu_n^{\otimes n} \left(\frac{-jkl(s_m) + jkl(\tilde{s}_m)}{y_m^2 + 2(2 + \epsilon_d) d^{2\otimes n}(s_0, s_m)} \right) \right] \leq \frac{4\kappa_1'' \sigma_m}{y_m} + \frac{4\kappa_2''}{\sqrt{ny_m^2}} \sqrt{\ln \left(\frac{1}{\mathbb{P}\{A\}} \right)} + \frac{16}{\rho ny_m^2} \ln \left(\frac{1}{\mathbb{P}\{A\}} \right)$$

which implies

$$\mathbb{E}^A \left[\nu_n^{\otimes n} \left(\frac{-jkl(\hat{s}_m) + jkl(\tilde{s}_m)}{y_m^2 + 2(2 + \epsilon_d) d^{2\otimes n}(s_0, \hat{s}_m)} \right) \right] \leq \frac{4\kappa_1'' \sigma_m}{y_m} + \frac{4\kappa_2''}{\sqrt{ny_m^2}} \sqrt{\ln \left(\frac{1}{\mathbb{P}\{A\}} \right)} + \frac{16}{\rho ny_m^2} \ln \left(\frac{1}{\mathbb{P}\{A\}} \right).$$

We turn back to the control of $-\nu_n^{\otimes n}(jkl(\tilde{s}_m))$. Our Bernstein type control yields

$$\mathbb{E}^A [-\nu_n^{\otimes n}(jkl(\tilde{s}_m))] \leq \frac{3}{2\sqrt{\rho(1-\rho)}} \frac{\sqrt{d^{2\otimes n}(s_0, \tilde{s}_m)}}{\sqrt{n}} \sqrt{\ln \left(\frac{1}{\mathbb{P}\{A\}} \right)} + \frac{2}{\rho n} \ln \left(\frac{1}{\mathbb{P}\{A\}} \right)$$

or for any $y_m > 0$ and any $\kappa' > 0$:

$$\mathbb{E}^A \left[\frac{-\nu_n^{\otimes n}(jkl(\tilde{s}_m))}{y_m^2 + \kappa'^2 d^{2\otimes n}(s_0, \tilde{s}_m)} \right] \leq \frac{1}{y_m^2 + \kappa'^2 d^{2\otimes n}(s_0, \tilde{s}_m)} \left(\frac{3}{2\sqrt{\rho(1-\rho)}} \frac{\sqrt{d^{2\otimes n}(s_0, \tilde{s}_m)}}{\sqrt{n}} \sqrt{\ln \left(\frac{1}{\mathbb{P}\{A\}} \right)} + \frac{2}{\rho n} \ln \left(\frac{1}{\mathbb{P}\{A\}} \right) \right)$$

$$\leq \frac{3}{4\kappa'\sqrt{\rho(1-\rho)}} \frac{1}{\sqrt{ny_m^2}} \sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)} + \frac{2}{\rho ny_m^2} \ln\left(\frac{1}{\mathbb{P}\{A\}}\right).$$

We derive thus

$$\begin{aligned} \mathbb{E}^A \left[\nu_n^{\otimes n} \left(\frac{-jkl(\widehat{s}_m) + jkl(\widetilde{s}_m)}{y_m^2 + 2(2 + \epsilon_d)d^{2\otimes n}(s_0, \widehat{s}_m)} \right) + \frac{-\nu_n^{\otimes n}(jkl(\widetilde{s}_m))}{y_m^2 + \kappa'^2 d^{2\otimes n}(s_0, \widetilde{s}_m)} \right] \\ \leq \frac{4\kappa_1''\sigma_m}{y_m} + \left(4\kappa_2'' + \frac{3}{4\kappa'\sqrt{\rho(1-\rho)}} \right) \frac{1}{\sqrt{ny_m^2}} \sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)} + \frac{18}{\rho ny_m^2} \ln\left(\frac{1}{\mathbb{P}\{A\}}\right) \end{aligned}$$

Let κ'_d such that $\kappa'_d{}^2 = 2(2 + \epsilon_d)/(1 + \epsilon_d)$, using $d^{2\otimes n}(s_0, \widehat{s}_m) \geq d^{2\otimes n}(s_0, \widetilde{s}_m)/(1 + \epsilon_d)$, we have

$$\nu_n^{\otimes n} \left(\frac{-jkl(\widehat{s}_m) + jkl(\widetilde{s}_m)}{y_m^2 + 2(2 + \epsilon_d)d^{2\otimes n}(s_0, \widehat{s}_m)} \right) + \frac{-\nu_n^{\otimes n}(jkl(\widetilde{s}_m))}{y_m^2 + \kappa_d'^2 d^{2\otimes n}(s_0, \widetilde{s}_m)} \geq \nu_n^{\otimes n} \left(\frac{-jkl(\widehat{s}_m)}{y_m^2 + 2(2 + \epsilon_d)d^{2\otimes n}(s_0, \widehat{s}_m)} \right)$$

and thus

$$\mathbb{E}^A \left[\nu_n^{\otimes n} \left(\frac{-jkl(\widehat{s}_m)}{y_m^2 + \kappa_0' d^{2\otimes n}(s_0, \widehat{s}_m)} \right) \right] \leq \frac{\kappa_1'\sigma_m}{y_m} + \kappa_2' \frac{1}{\sqrt{ny_m^2}} \sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)} + \frac{18}{ny_m^2\rho} \ln\left(\frac{1}{\mathbb{P}\{A\}}\right).$$

where $\kappa_0' = 2(2 + \epsilon_d)/(1 + \epsilon_d)$, $\kappa_1' = 4\kappa_1''$ and $\kappa_2' = 4\kappa_2'' + 3/(4\sqrt{\rho(1-\rho)}\kappa_d')$. \square

B.4 Behavior of the constants of Theorem 1 Single model maximum likelihood estimation and Theorem 2 A general theorem for penalized maximum likelihood conditional density estimation

We explain now the behavior of the constants κ_0 and C_2 with respect to C_1 and ρ . As shown in the proof, if we let $\epsilon_{\text{pen}} = 1 - \frac{1}{C_1}$ then $C_1 = \frac{1}{1-\epsilon_{\text{pen}}}$ and $C_2 = \frac{\kappa_0}{1-\epsilon_{\text{pen}}} = \kappa_0 C_1$ so that it suffices to study the behavior of κ_0 .

Now κ_0 is defined as equal to $\frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa_0'}$ with θ_{pen} the root of $\left(\frac{\kappa_1' + \kappa_2'}{\theta_{\text{pen}}} + \frac{18}{\theta_{\text{pen}}^2 \rho} \right) \kappa_0' = C_\rho \epsilon_{\text{pen}}$ where we use the constants appearing in Lemma 3 Proof of Theorem 1 Single model maximum likelihood estimation. This implies

$$\kappa_0 = \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa_0'} = \theta_{\text{pen}}^2 \left(\frac{\kappa_1' + \kappa_2'}{\theta_{\text{pen}}} + \frac{18}{\theta_{\text{pen}}^2 \rho} \right) = \theta_{\text{pen}} (\kappa_1' + \kappa_2') + \frac{18}{\rho}.$$

Solving the implied quadratic equation $\theta_{\text{pen}} (\kappa_1' + \kappa_2') + \frac{18}{\rho} = \theta_{\text{pen}}^2 \frac{C_\rho \epsilon_{\text{pen}}}{\kappa_0'}$ yields

$$\theta_{\text{pen}} = \frac{\kappa_0' (\kappa_1' + \kappa_2') \left(\sqrt{1 + \frac{72 C_\rho \epsilon_{\text{pen}}}{\rho \kappa_0' (\kappa_1' + \kappa_2')^2}} + 1 \right)}{2 C_\rho \epsilon_{\text{pen}}}$$

and thus

$$\kappa_0 = \frac{\kappa_0' (\kappa_1' + \kappa_2')^2 \left(\sqrt{1 + \frac{72 C_\rho \epsilon_{\text{pen}}}{\rho \kappa_0' (\kappa_1' + \kappa_2')^2}} + 1 \right)}{2 C_\rho \epsilon_{\text{pen}}} + \frac{18}{\rho}$$

Now

$$\kappa_1' = 4\kappa_1'' = 4 \left(\frac{3\kappa}{2\sqrt{2\rho(1-\rho)}} + \frac{4}{\rho} + \frac{3}{\sqrt{\rho(1-\rho)}} \right) = \frac{1}{\sqrt{\rho(1-\rho)}} \left(3\kappa\sqrt{2} + 12 + 16\sqrt{\frac{1-\rho}{\rho}} \right)$$

and using that for any $\epsilon > 0$, once ϵ_d is small enough, $2 > \kappa'_d \geq 2(1 - \epsilon)$

$$\kappa'_2 = 4\kappa''_2 + \frac{3}{4\sqrt{\rho(1-\rho)}\kappa'_d} \leq \frac{42}{\sqrt{\rho(1-\rho)}} + \frac{3}{8\sqrt{\rho(1-\rho)}(1-\epsilon)} = \frac{1}{\sqrt{\rho(1-\rho)}} \left(42 + \frac{3}{8(1-\epsilon)} \right)$$

so that

$$(\kappa'_1 + \kappa'_2)^2 \leq \frac{1}{\rho(1-\rho)} \left(3\kappa\sqrt{2} + 54 + \frac{3}{8(1-\epsilon)} + 16\sqrt{\frac{1-\rho}{\rho}} \right)^2.$$

Now using $4 < \kappa'_0 \leq 4(1 + \epsilon)$

$$\begin{aligned} \kappa_0 &\leq \frac{4(1 + \epsilon) \left(3\kappa\sqrt{2} + 54 + \frac{3}{8(1-\epsilon)} + 16\sqrt{\frac{1-\rho}{\rho}} \right)^2 \left(\sqrt{1 + \frac{72C_\rho\epsilon_{\text{pen}}}{\rho\kappa'_0(\kappa'_1 + \kappa'_2)^2}} + 1 \right)}{2\rho(1-\rho)C_\rho\epsilon_{\text{pen}}} + \frac{18}{\rho} \\ &\leq \frac{1}{C_\rho\rho(1-\rho)\epsilon_{\text{pen}}} \\ &\quad \times \left(2(1 + \epsilon) \left(3\kappa\sqrt{2} + 54 + \frac{3}{8(1-\epsilon)} + 16\sqrt{\frac{1-\rho}{\rho}} \right)^2 \left(\sqrt{1 + \frac{72C_\rho\epsilon_{\text{pen}}}{\rho\kappa'_0(\kappa'_1 + \kappa'_2)^2}} + 1 \right) + 18C_\rho(1-\rho)\epsilon_{\text{pen}} \right) \end{aligned}$$

This implies that κ_0 scales when ρ is close to 1 proportionally to

$$\frac{1}{C_\rho\rho(1-\rho)\epsilon_{\text{pen}}} = \frac{\rho}{(1-\rho)^2 \left(\ln \left(1 + \frac{\rho}{1-\rho} \right) - \rho \right) \epsilon_{\text{pen}}}$$

and thus explodes when ρ goes to 1 as well as when ϵ_{pen} goes to 0.

Note that, as it is almost always the case in density estimation, these constants are rather large, mostly because of the crude constant appearing in Theorem 6Proof of Lemma 3Proof of Theorem 1Single model maximum likelihood estimation. Indeed if we denote $\sigma_{\mathcal{M}}$ the supremum over all models of the collection the right hand side of the previous bound on κ_0 can already be replaced by

$$\begin{aligned} &\frac{1}{C_\rho\rho(1-\rho)\epsilon_{\text{pen}}} \\ &\quad \times \left(2(1 + \epsilon) \left(3\kappa\sqrt{2} + 42 + 6\sqrt{2}\sigma_{\mathcal{M}} + \frac{3}{8(1-\epsilon)} + 16\sqrt{\frac{1-\rho}{\rho}} \right)^2 \left(\sqrt{1 + \frac{72C_\rho\epsilon_{\text{pen}}}{\rho\kappa'_0(\kappa'_1 + \kappa'_2)^2}} + 1 \right) + 18C_\rho(1-\rho)\epsilon_{\text{pen}} \right). \end{aligned}$$

which is much smaller than the previous quantity as soon as $\sigma_{\mathcal{M}}$ is much smaller than $\sqrt{2}$, which can be ensured in the models of Section 4Partition-based conditional density models provided we limit their maximum dimension well below n , for instance to $n/\ln^2(n)$.

C Proof for Section 4.1Covariate partitioning and conditional density estimation (Covariate partitioning and conditional density estimation)

Proof of Proposition 3Covariate partitioning and conditional density estimation. We start by the UDP case, as we stop as soon as $\frac{2^d}{n} > 2^{-d_X J} \leq \frac{1}{n}$, $J \leq \frac{\ln n}{d_X \ln 2}$ and thus there is at most $1 + \frac{\ln n}{d_X \ln 2}$ different partitions in the collection, which allows to prove the proposition in this case.

The proof for the RDP, RSDP and RSP cases are handled simultaneously. Indeed all these partition collection are recursive partition collection and thus corresponds to tree structures. More precisely, any RDP can be represented by a 2_X^d -ary tree in which a node has a value 0 if it has no child or the value 1 otherwise. Along the same lines, any RSDP (respectively RSP) can be represented by a dyadic tree in which a node has the value 0 if it has no child or the one plus the number of the dimension of the split (respectively one plus the number of the dimension and the position of the split). Such a tree can be encoded by the ordered list of the values of his nodes. The total length of the code us thus given by the number of nodes $N(\mathcal{P})$ time the encoding cost (respectively $\lceil \frac{\ln 2}{\ln 2} \rceil$ bits, $\lceil \frac{\ln(1+d_X)}{\ln 2} \rceil$ bits and $\lceil \frac{\ln(1+d_X)}{\ln 2} \rceil + \lceil \frac{\ln n}{\ln 2} \rceil$). As this code is decodable, it satisfies the Kraft inequality and thus, using the definition of $B_0^{*(\mathcal{X})}$,

$$\sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{*(\mathcal{X})}} 2^{-N(\mathcal{P}) \frac{B_0^{*(\mathcal{X})}}{\ln 2}} \leq 1 \Leftrightarrow \sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{*(\mathcal{X})}} e^{-N(\mathcal{P}) B_0^{*(\mathcal{X})}} \leq 1.$$

It turns out that the number of nodes $N(\mathcal{P})$ can be computed from the number of hyperrectangles of the partition $\|\mathcal{P}\|$, which is also the number of leaves in the tree. Indeed, each inner node has exactly 2_X^d children in the RDP case and only 2 in the RSDP and RSP case, while, in all cases, every node but the root has a single parent. Let $d = d_X + d_Y$ in the RDP case and $d = 1$ in the RSDP and RSP case, we have then $2^d(N(\mathcal{P}) - \|\mathcal{P}\|) = N(\mathcal{P}) - 1$ and thus

$$N(\mathcal{P}) = \frac{2^d \|\mathcal{P}\| - 1}{2^d - 1} = \frac{2^d}{2^d - 1} \|\mathcal{P}\| + \left(1 - \frac{2^d}{2^d - 1}\right) = c_0^{*(\mathcal{X})} \|\mathcal{P}\| + (1 - c_0^{*(\mathcal{X})})$$

with $c_0^{*(\mathcal{X})}$ as defined in the proposition. Plugging this in the Kraft inequality leads to

$$\sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{*(\mathcal{X})}} e^{-c_0^{*(\mathcal{X})} B_0^{*(\mathcal{X})} \|\mathcal{P}\| + B_0^{*(\mathcal{X})} (c_0^{*(\mathcal{X})} - 1)} \leq 1 \Leftrightarrow \sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{*(\mathcal{X})}} e^{-c_0^{*(\mathcal{X})} B_0^{*(\mathcal{X})} \|\mathcal{P}\|} \leq e^{B_0^{*(\mathcal{X})} (1 - c_0^{*(\mathcal{X})})}.$$

Let now $c \geq c_0^{*(\mathcal{X})}$,

$$\sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{*(\mathcal{X})}} e^{-c B_0^{*(\mathcal{X})} \|\mathcal{P}\|} \leq \sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{*(\mathcal{X})}} e^{-(c - c_0^{*(\mathcal{X})}) B_0^{*(\mathcal{X})} \|\mathcal{P}\|} e^{-c_0^{*(\mathcal{X})} B_0^{*(\mathcal{X})} \|\mathcal{P}\|}$$

and as $\|\mathcal{P}\| \geq 1$

$$\begin{aligned} &\leq e^{-(c - c_0^{*(\mathcal{X})}) B_0^{*(\mathcal{X})}} \sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{*(\mathcal{X})}} e^{-c_0^{*(\mathcal{X})} B_0^{*(\mathcal{X})} \|\mathcal{P}\|} \\ &\leq e^{-(c - c_0^{*(\mathcal{X})}) B_0^{*(\mathcal{X})}} e^{(1 - c_0^{*(\mathcal{X})}) B_0^{*(\mathcal{X})}} = e^{B_0^{*(\mathcal{X})}} e^{-c B_0^{*(\mathcal{X})}} = \sum_0^{*(\mathcal{X})} e^{-c C_0^{*(\mathcal{X})}} \end{aligned}$$

which concludes this three cases.

For the HRP cases, it is sufficient to give the uppermost coordinate of the hyperrectangles ordered in a uniquely decodable way based on the following observation: assume we have a current list of hyperrectangles, the complementary of the union of these hyperrectangles is either empty if the list contains all the hyperrectangles of the partition or contains a lowermost point that is the lowermost corner of a unique hyperrectangle. Furthermore, this hyperrectangle is completely specified by its uppermost corner coordinates. Starting with an empty list, an HRP

partition can thus be entirely specified by the list of uppermost corner coordinates obtained through this scheme.

This leads to a code with $\|\mathcal{P}\| \times d_X \left\lceil \frac{\ln n}{\ln 2} \right\rceil$ bits for each partition that satisfies the Kraft inequality

$$\sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{\text{HRP}(\mathcal{X})}} 2^{-\|\mathcal{P}\| \frac{B_0^{\text{HRP}(\mathcal{X})}}{\ln 2}} \leq 1 \Leftrightarrow \sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{\text{HRP}(\mathcal{X})}} e^{-c_0^{\text{HRP}(\mathcal{X})} B_0^{\text{HRP}(\mathcal{X})} \|\mathcal{P}\|} \leq 1$$

Now for any $c \geq c_0^{\text{HRP}(\mathcal{X})}$,

$$\begin{aligned} \sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{\text{HRP}(\mathcal{X})}} e^{-c B_0^{\text{HRP}(\mathcal{X})} \|\mathcal{P}\|} &= \sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{\text{HRP}(\mathcal{X})}} e^{-(c-c_0^{\text{HRP}(\mathcal{X})}) B_0^{\text{HRP}(\mathcal{X})} \|\mathcal{P}\|} e^{-c_0^{\text{HRP}(\mathcal{X})} B_0^{\text{HRP}(\mathcal{X})} \|\mathcal{P}\|} \\ &\leq e^{-(c-c_0^{\text{HRP}(\mathcal{X})}) B_0^{\text{HRP}(\mathcal{X})}} \sum_{\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{\text{HRP}(\mathcal{X})}} e^{-c_0^{\text{HRP}(\mathcal{X})} B_0^{\text{HRP}(\mathcal{X})} \|\mathcal{P}\|} \\ &\leq e^{-(c-c_0^{\text{HRP}(\mathcal{X})}) B_0^{\text{HRP}(\mathcal{X})}} = e^{B_0^{\text{HRP}(\mathcal{X})}} e^{-c B_0^{\text{HRP}(\mathcal{X})}} = \sum_0^{\text{HRP}(\mathcal{X})} e^{-c B_0^{\text{HRP}(\mathcal{X})}} \end{aligned}$$

which concludes the proof. \square

D Proof for Section 4.2 Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation)

Theorem 3 Piecewise polynomial conditional densities estimation is obtained by proving that Assumption $(H_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}})$ and $(S_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}})$ hold for any model $S_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}$ while Assumption (K) holds for any model collection. Theorem 3 Piecewise polynomial conditional densities estimation is then a consequence of Theorem 2A general theorem for penalized maximum likelihood conditional density estimation.

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

Proposition 7. *Under the assumptions of Theorem 3 Piecewise polynomial conditional densities estimation, it exists a D_* such that for any model $S_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}$, the function*

$$\phi_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}(\sigma) = \sigma \sqrt{\mathcal{D}_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}} \left(\sqrt{\frac{1}{2} \ln \frac{n^2}{\|\mathcal{P}^{\mathcal{Q}}\|} + D_*} + \sqrt{\pi} \right)$$

satisfies the properties of Assumption $(H_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}})$.

Furthermore, $\mathfrak{D}_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}} = n \sigma_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}^2$ where $\sigma_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}$ is the unique root of $\frac{1}{\sigma} \phi_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}(\sigma) = \sqrt{n}$ satisfies

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

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

This is deduced from

Proposition 8. $\forall S_{\mathcal{P}\mathcal{Q},\mathbf{D}}, \forall s_{\mathcal{P}\mathcal{Q},\mathbf{D}} \in S_{\mathcal{P}\mathcal{Q},\mathbf{D}},$

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

which is later proved and Proposition 2 Jensen-Kullback-Leibler divergence and bracketing entropy combined with the inequality

$$\left(\sqrt{\frac{1}{2} \ln \frac{n^2}{\sum_{\mathcal{R}_l \in \mathcal{P}} \|\mathcal{Q}_l\|} + D_\star} + \sqrt{\pi} \right)^2 \leq \ln \frac{n^2}{\|\mathcal{P}\mathcal{Q}\|} + 2D_\star + 2\pi.$$

The proof of Kraft type assumption,

Proposition 9. *Under the assumptions of Theorem 3 Piecewise polynomial conditional densities estimation, for any collection \mathcal{S} , it exists a $c_\star > 0$ such that for*

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

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

is obtained by using Proposition 3 Covariate partitioning and conditional density estimation for both \mathcal{P} and \mathcal{Q} . The complete proof is postponed after the one Proposition 8 Proof for Section 4.2 Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation).

D.1 Proof of Proposition 8 Proof for Section 4.2 Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation)

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

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

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

A, the reference measure is the Lebesgue measure, $\|u\|_\infty^{2\otimes n} \geq \|u\|_2^{2\otimes n}$. By definition $d^{\otimes n}(s, t) = \|\sqrt{s} - \sqrt{t}\|_2^{\otimes n}$ and thus for any model S_m and any function $s_m \in S_m$

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

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

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

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

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

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

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

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

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

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

Proposition 10. *For any basis $\{\phi_k\}_{1 \leq k \leq \mathcal{D}_m}$ of $\sqrt{S_m}$ such that*

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

let

$$\bar{\tau}_m(\{\phi_k\}) = \sup_{\sum_{k=1}^{\mathcal{D}_m} \beta_k \phi_k \in \sqrt{S_m} \setminus \{0\}} \frac{1}{\sqrt{\mathcal{D}_m}} \frac{\left\| \sum_{k=1}^{\mathcal{D}_m} \beta_k \phi_k \right\|_{\infty}^{\otimes n}}{\|\beta\|_{\infty}}.$$

and let $\bar{\tau}_m$ be the infimum over all suitable bases.

Then $\bar{\tau}_m \geq 1$ and

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

with $\mathcal{C}_m = \ln(\kappa_{\infty} \bar{\tau}_m)$ 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 10 Proof of Proposition 8 Proof for Section 4.2 Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation)

Proposition 11. *It exists*

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

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

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

with $\mathcal{C}_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}} = \ln(\kappa_{\infty} \bar{\tau}_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}})$ and $\kappa_{\infty} \leq 2\sqrt{2\pi e}$.

Now one verify that

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

Remark that when $\star(\mathcal{X}) = \text{UDP}(\mathcal{X})$, $\star(\mathcal{Y}) = \text{UDP}(\mathcal{Y})$ and \mathcal{Q}_l is independent of \mathcal{R}_l , all the hyperrectangles have the same size 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 8 Proof for Section 4.2 Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation) as $\forall s_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}} \in S_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}$,

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

D.2 Proofs of Proposition 10 Proof of Proposition 8 Proof for Section 4.2 Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation) and Proposition 11 Proof of Proposition 8 Proof for Section 4.2 Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation)

Proof of Proposition 10 Proof of Proposition 8 Proof for Section 4.2 Piecewise polynomial conditional densities estimation
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 $\bar{r}_m(\phi) \geq 1$.

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

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

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

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

The definition of \bar{r}_m implies then that

$$\begin{aligned} \left\| \sum_{k=1}^{\mathcal{D}_m} \beta_k \phi_k - \sum_{k=1}^{\mathcal{D}_m} \beta'_k \phi_k \right\|_{\infty}^{\otimes n} &\leq \bar{r}_m(\phi) \sqrt{\mathcal{D}_m} \|\beta - \beta'\|_{\infty} \\ &\leq \frac{\delta}{2}. \end{aligned}$$

The set $\left\{ \sum_{k=1}^{\mathcal{D}_m} \beta_k \phi_k \mid \beta \in \mathcal{G}_m(\delta, \sigma) \right\}$ is thus a $\frac{\delta}{2}$ covering of $\left\{ u \in \sqrt{\mathcal{S}_m} \mid \|u\|_2^{\otimes n} \leq \sigma \right\}$ 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 $\frac{\delta}{\sqrt{\mathcal{D}_m} \bar{r}_m(\phi)}$ 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 $\|\beta' - \beta\|_{\infty} \leq \frac{\delta}{\sqrt{\mathcal{D}_m} \bar{r}_m(\phi)}$. As $\|\beta' - \beta\|_2 \leq \sqrt{\mathcal{D}_m} \|\beta' - \beta\|_{\infty}$, this implies $\|\beta\|_2 \leq \sigma + \frac{\delta}{\bar{r}_m(\phi)}$. We deduce then

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

and thus

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

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

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

which concludes the proof. \square

Instead of Proposition 11 Proof of Proposition 8 Proof for Section 4.2 Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation), we prove an extended version of it in which the degree of the conditional densities may depends on the hyperrectangle. More precisely, we reuse the partition $\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{*(\mathcal{X})}$ and the partitions $\mathcal{Q}_l \in \mathcal{S}_{\mathcal{P}}^{*(\mathcal{Y})}$ for $\mathcal{R}_l \in \mathcal{P}$ and define now the model $S_{\mathcal{P}\mathcal{Q}, \mathbf{D}}$ as the set of conditional densities such that

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

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

By construction,

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

The corresponding linear space $\overline{\sqrt{S_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}}}$ is

$$\left\{ \sum_{\mathcal{R}_{l,k}^{\times} \in \mathcal{P}^{\mathcal{Q}}} P_{\mathcal{R}_{l,k}^{\times}}(y) \mathbf{1}_{\{(x,y) \in \mathcal{R}_{l,k}^{\times}\}} \mid \deg(P_{\mathcal{R}_{l,k}^{\times}}) \leq \mathbf{D}(\mathcal{R}_{l,k}^{\times}) \right\}.$$

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

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

Note that the space $S_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}$ 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 12. *It exists*

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

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

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

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

Proposition 11 Proof of Proposition 8 Proof for Section 4.2 Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation) is deduced from this proposition with the help of the simple upper bound

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

To be able to choose the polynomial degree, it is sufficient to upper bound the ratio

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

over all choices of degrees to obtain Assumption $H_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}$ for both the global choice and the local choice. 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 $A_0^{*(\mathcal{X})}$ by $A_0^{*(\mathcal{X})} + \ln |\mathcal{D}^M|$ for the global optimization and $B_0^{*(\mathcal{Y})}$ by $B_0^{*(\mathcal{Y})} + \ln |\mathcal{D}^M|$ for the local optimization, where $|\mathcal{D}^M|$ denotes the size of the family of possible degrees turns out to be sufficient.

Proof of Proposition 12 Proofs of Proposition 10 Proof of Proposition 8 Proof for Section 4.2 Piecewise polynomial conditional densities estimation Let L_D be the one dimensional Legendre polynomial of degree D and $G_D = \sqrt{2D+1}L_D$ its rescaled version, we recall that

$$\forall D \in \mathbb{N}, \quad \|G_D\|_{\infty} = \sqrt{2D+1} \quad \text{and} \quad \forall (D, D') \in \mathbb{N}^2, \quad \int G_D(t) G_{D'}(t) dt = \delta_{D, D'}$$

Let $D \in \mathbb{N}^{d_Y}$, we define G_D as the polynomial

$$G_{D_1, \dots, D_{d_Y}}(y) = G_{D_1}(y_1) \times \dots \times G_{D_{d_Y}}(y_{d_Y}),$$

by construction

$$\forall D \in \mathbb{N}^{d_Y}, \quad \|G_D\|_\infty = \prod_{1 \leq d \leq d_Y} \sqrt{2D_d + 1}$$

and

$$\forall (D, D') \in \mathbb{N}^{d_Y \times 2}, \quad \int_{y \in [0, 1]^{d_Y}} G_D(y) G_{D'}(y) dy = \delta_{D, D'}.$$

Now for any hyperrectangle $\mathcal{R}_{l,k}^\times$, we define $G_D^{\mathcal{R}_{l,k}^\times}(x, y) = \frac{1}{\sqrt{|\mathcal{R}_{l,k}^\times|}} G_D(T^{\mathcal{R}_{l,k}^\times}(y)) \mathbf{1}_{\{(x,y) \in \mathcal{R}_{l,k}^\times\}}(x)$ where $T^{\mathcal{R}_{l,k}^\times}$ is the affine transform that maps $\mathcal{R}_{l,k}^\times$ into $[0, 1]^{d_Y}$ so that

$$\forall \mathcal{R}_{l,k}^\times \in \mathcal{P}^{\mathcal{Q}}, \forall D \in \mathbb{N}^{d_Y}, \quad \|G_D^{\mathcal{R}_{l,k}^\times}\|_\infty = \frac{1}{\sqrt{|\mathcal{R}_{l,k}^\times|}} \prod_{1 \leq d \leq d_Y} \sqrt{2D_d + 1}$$

and

$$\begin{aligned} \forall (\mathcal{R}_{l,k}^\times, \mathcal{R}_{l',k'}^\times) \in (\mathcal{P}^{\mathcal{Q}})^2, \forall (D, D') \in \mathbb{N}^{d_Y \times 2}, \\ \int_{x \in [0, 1]^{d_X}} \int_{y \in [0, 1]^{d_Y}} G_D^{\mathcal{R}_{l,k}^\times}(x, y) G_{D'}^{\mathcal{R}_{l',k'}^\times}(x, y) dy dx = \delta_{\mathcal{R}_{l,k}^\times, \mathcal{R}_{l',k'}^\times} \delta_{D, D'}. \end{aligned}$$

Using the piecewise structure, one deduces

$$\begin{aligned} & \mathbb{E} \left[\left\| \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{P}^{\mathcal{Q}}} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \beta_D^{\mathcal{R}_{l,k}^\times} G_D^{\mathcal{R}_{l,k}^\times}(X_i, \cdot) \right\|_2^2 \right] \\ &= \mathbb{E} \left[\sum_{\mathcal{R}_l \in \mathcal{P}} \frac{\mathbf{1}_{\{X_1 \in \mathcal{R}_l\}}}{|\mathcal{R}_l|} \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{Q}_l} \int_{(x,y) \in \mathcal{R}_{l,k}^\times} \left| \sum_{D \leq \mathbf{D}(\mathcal{R}_l, \mathcal{R}_{l,k}^\times)} \beta_D^{\mathcal{R}_l, \mathcal{R}_{l,k}^\times} G_D^{\mathcal{R}_l, \mathcal{R}_{l,k}^\times}(x, y) \right|^2 dy dx \right] \\ &= \mathbb{E} \left[\sum_{\mathcal{R}_l \in \mathcal{P}} \frac{\mathbf{1}_{\{X_1 \in \mathcal{R}_l\}}}{|\mathcal{R}_l|} \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{Q}_l} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \left| \beta_D^{\mathcal{R}_{l,k}^\times} \right|^2 \right] \\ &= \sum_{\mathcal{R}_l \in \mathcal{P}} \frac{\mathbb{P}\{X_1 \in \mathcal{R}_l\}}{|\mathcal{R}_l|} \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{Q}_l} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \left| \beta_D^{\mathcal{R}_{l,k}^\times} \right|^2. \end{aligned}$$

The space $\sqrt{S_{\mathcal{P}^{\mathcal{Q}}, \mathbf{D}}}$ is spanned by

$$\left\{ G_D^{\mathcal{R}_{l,k}^\times} \mid \mathcal{R}_{l,k}^\times \in \mathcal{P}^{\mathcal{Q}}, D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times) \right\}$$

but also by the rescaled $\phi_D^{\mathcal{R}_{l,k}^\times} = \frac{1}{\sqrt{\mu_X(\mathcal{R}_l)}} G_D^{\mathcal{R}_{l,k}^\times}$ where $\mu_X(\mathcal{R}_l) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{P}\{X_i \in \mathcal{R}_l\}}{|\mathcal{R}_l|}$. For these functions, one has

$$\begin{aligned}
& \left\| \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{P}^{\mathcal{Q}}} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \beta_D^{\mathcal{R}_{l,k}^\times} \phi_D^{\mathcal{R}_{l,k}^\times} \right\|_2^{2 \otimes n} \\
&= \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \left\| \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{P}^{\mathcal{Q}}} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \beta_D^{\mathcal{R}_{l,k}^\times} \phi_D^{\mathcal{R}_{l,k}^\times}(X_i, \cdot) \right\|_2^2 \right] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{P}^{\mathcal{Q}}} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \frac{\beta_D^{\mathcal{R}_{l,k}^\times}}{\sqrt{\mu_X(\mathcal{R}_l)}} G_D^{\mathcal{R}_{l,k}^\times}(X_i, \cdot) \right\|_2^2 \right] \\
&= \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{P}^{\mathcal{Q}}} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \left| \beta_D^{\mathcal{R}_{l,k}^\times} \right|^2 = \left\| \beta_D^{\mathcal{R}_{l,k}^\times} \right\|_2^2.
\end{aligned}$$

For the $\|\cdot\|_\infty$ type norm, we have

$$\begin{aligned}
& \left\| \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{P}^{\mathcal{Q}}} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \beta_D^{\mathcal{R}_{l,k}^\times} \phi_D^{\mathcal{R}_{l,k}^\times} \right\|_\infty^{2 \otimes n} \\
&= \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \left\| \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{P}^{\mathcal{Q}}} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \beta_D^{\mathcal{R}_{l,k}^\times} \phi_D^{\mathcal{R}_{l,k}^\times}(X_i, \cdot) \right\|_\infty^2 \right] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{P}^{\mathcal{Q}}} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \beta_D^{\mathcal{R}_{l,k}^\times} \phi_D^{\mathcal{R}_{l,k}^\times}(X_i, \cdot) \right\|_\infty^2 \right] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\sum_{\mathcal{R}_l \in \mathcal{P}} \mathbf{1}_{\{X_i \in \mathcal{R}_l\}} \sup_{\mathcal{R}_{l,k}^y \in \mathcal{Q}_l} \left\| \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \beta_D^{\mathcal{R}_{l,k}^\times} \phi_D^{\mathcal{R}_{l,k}^\times}(X_i, \cdot) \right\|_\infty^2 \right] \\
&\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\sum_{\mathcal{R}_l \in \mathcal{P}} \mathbf{1}_{\{X_i \in \mathcal{R}_l\}} \sup_{x \in \mathcal{R}_l} \sup_{\mathcal{R}_{l,k}^y \in \mathcal{Q}_l} \left(\sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \left| \beta_D^{\mathcal{R}_{l,k}^\times} \right| \left\| \phi_D^{\mathcal{R}_{l,k}^\times}(x, \cdot) \right\|_\infty \right)^2 \right] \\
&\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\sum_{\mathcal{R}_l \in \mathcal{P}} \mathbf{1}_{\{X_i \in \mathcal{R}_l\}} \sup_{\mathcal{R}_{l,k}^y \in \mathcal{Q}_l} \frac{1}{\mu_X(\mathcal{R}_l) |\mathcal{R}_{l,k}^\times|} \left(\sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \|G_D\|_\infty \right)^2 \left\| \beta_D^{\mathcal{R}_{l,k}^\times} \right\|_\infty^2 \right] \\
&\leq \sum_{\mathcal{R}_l \in \mathcal{P}} |\mathcal{R}_l| \sup_{\mathcal{R}_{l,k}^y \in \mathcal{Q}_l} \frac{1}{|\mathcal{R}_{l,k}^\times|} \left(\sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \|G_D\|_\infty \right)^2 \left\| \beta_D^{\mathcal{R}_{l,k}^\times} \right\|_\infty^2.
\end{aligned}$$

Now

$$\begin{aligned} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \|G_D\|_\infty &= \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \prod_{d=1}^{d_Y} \|G_{D_d}\|_\infty = \prod_{d=1}^{d_Y} \left(\sum_{D_d \leq \mathbf{D}_d(\mathcal{R}_{l,k}^\times)} \|G_{D_d}\|_\infty \right) \\ &= \prod_{d=1}^{d_Y} \left(\sum_{D_d \leq \mathbf{D}_d(\mathcal{R}_{l,k}^\times)} \sqrt{2D_d + 1} \right) \leq \sup_{\mathcal{R}_{l',k'}^\times \in \mathcal{P}^\mathcal{Q}} \prod_{d=1}^{d_Y} \left(\sum_{D_d \leq \mathbf{D}_d(\mathcal{R}_{l',k'}^\times)} \sqrt{2D_d + 1} \right) \end{aligned}$$

while

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

This implies

$$\begin{aligned} &\frac{\left\| \sum_{\mathcal{R}_{l,k}^\times \in \mathcal{P}^\mathcal{Q}} \sum_{D \leq \mathbf{D}(\mathcal{R}_{l,k}^\times)} \beta_D^{\mathcal{R}_{l,k}^\times} \phi_D^{\mathcal{R}_{l,k}^\times} \right\|_\infty^{2\otimes n}}{\mathcal{D}_{\mathcal{P}^\mathcal{Q}, \mathbf{D}} \left\| \beta_D^{\mathcal{R}_{l,k}^\times} \right\|_\infty^2} \\ &\leq \frac{\left(\sup_{\mathcal{R}_{l',k'}^\times \in \mathcal{P}^\mathcal{Q}} \prod_{d=1}^{d_Y} \left(\sum_{D_d \leq \mathbf{D}_d(\mathcal{R}_{l',k'}^\times)} \sqrt{2D_d + 1} \right) \right)^2}{\inf_{\mathcal{R}_{l',k'}^\times \in \mathcal{P}^\mathcal{Q}} \prod_{d=1}^{d_Y} (\mathbf{D}_d(\mathcal{R}_{l',k'}^\times) + 1)} \sum_{\mathcal{R}_l \in \mathcal{P}} |\mathcal{R}_l| \sup_{\mathcal{R}_{l,k}^y \in \mathcal{Q}_l} \frac{1}{\|\mathcal{P}^\mathcal{Q}\| \|\mathcal{R}_{l,k}^\times\|} \\ &\leq \left(\frac{\sup_{\mathcal{R}_{l',k'}^\times \in \mathcal{P}^\mathcal{Q}} \prod_{d=1}^{d_Y} \left(\sum_{D_d \leq \mathbf{D}_d(\mathcal{R}_{l',k'}^\times)} \sqrt{2D_d + 1} \right)}{\inf_{\mathcal{R}_{l',k'}^\times \in \mathcal{P}^\mathcal{Q}} \prod_{d=1}^{d_Y} \sqrt{\mathbf{D}_d(\mathcal{R}_{l',k'}^\times) + 1}} \sum_{\mathcal{R}_l \in \mathcal{P}} |\mathcal{R}_l| \sup_{\mathcal{R}_{l,k}^y \in \mathcal{Q}_l} \frac{1}{\sqrt{\|\mathcal{P}^\mathcal{Q}\|} \sqrt{|\mathcal{R}_{l,k}^\times|}} \right)^2 \\ &\leq \left(\frac{\sup_{\mathcal{R}_{l',k'}^\times \in \mathcal{P}^\mathcal{Q}} \prod_{d=1}^{d_Y} \left(\sum_{D_d \leq \mathbf{D}_d(\mathcal{R}_{l',k'}^\times)} \sqrt{2D_d + 1} \right)}{\inf_{\mathcal{R}_{l',k'}^\times \in \mathcal{P}^\mathcal{Q}} \prod_{d=1}^{d_Y} \sqrt{\mathbf{D}_d(\mathcal{R}_{l',k'}^\times) + 1}} \sup_{\mathcal{R}_{l,k}^\times \in \mathcal{P}^\mathcal{Q}} \frac{1}{\sqrt{\|\mathcal{P}^\mathcal{Q}\|} \sqrt{|\mathcal{R}_{l,k}^\times|}} \right)^2. \end{aligned}$$

The proposition is then obtained by a simple application of Proposition 10 Proof of Proposition 8 Proof for Section 4.2 Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation). \square

D.3 Proof of Proposition 9 Proof for Section 4.2 Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation)

Proof. By construction

$$\sum_{S_{\mathcal{P}^\mathcal{Q}, \mathbf{D}} \in \mathcal{S}} e^{-x_{\mathcal{P}^\mathcal{Q}, \mathbf{D}}} = \sum_{\mathcal{P} \in \mathcal{S}_p^{+(x)}} \sum_{\mathcal{R}_l \in \mathcal{P}} \sum_{\mathcal{Q}_l \in \mathcal{S}_p^{+(y)}} e^{-x_{\mathcal{P}^\mathcal{Q}, \mathbf{D}}}$$

$$\begin{aligned}
&= \sum_{\mathcal{P} \in \mathcal{S}_p^{*(\mathcal{X})}} \sum_{\mathcal{R}_l \in \mathcal{P}} \sum_{\mathcal{Q}_l \in \mathcal{S}_p^{*(\mathcal{Y})}} e^{-c_* \left(A_0^{*(\mathcal{X})} + (B_0^{*(\mathcal{X})} + A_0^{*(\mathcal{Y})}) \|\mathcal{P}\| + B_0^{*(\mathcal{Y})} \sum_{\mathcal{R}_l \in \mathcal{P}} \|\mathcal{Q}_l\| \right)} \\
&= \sum_{\mathcal{P} \in \mathcal{S}_p^{*(\mathcal{X})}} e^{-c_* \left(A_0^{*(\mathcal{X})} + B_0^{*(\mathcal{X})} \|\mathcal{P}\| \right)} \prod_{\mathcal{R}_l \in \mathcal{P}} \left(\sum_{\mathcal{Q}_l \in \mathcal{S}_p^{*(\mathcal{Y})}} e^{-c_* \left(A_0^{*(\mathcal{Y})} + B_0^{*(\mathcal{Y})} \|\mathcal{Q}_l\| \right)} \right)
\end{aligned}$$

By Proposition 3 Covariate partitioning and conditional density estimation, one can find $c_* \geq \max(1, c_0^{*(\mathcal{X})}, c_0^{*(\mathcal{Y})})$ such that

$$\sum_{\mathcal{Q}_l \in \mathcal{S}_p^{*(\mathcal{Y})}} e^{-c_* \left(A_0^{*(\mathcal{Y})} + B_0^{*(\mathcal{Y})} \|\mathcal{Q}_l\| \right)} \leq 1$$

and

$$\sum_{\mathcal{P} \in \mathcal{S}_p^{*(\mathcal{X})}} e^{-c_* \left(A_0^{*(\mathcal{X})} + B_0^{*(\mathcal{X})} \|\mathcal{P}\| \right)} \leq 1.$$

Plugging these bounds in the previous equality yields

$$\sum_{S_{\mathcal{P}, \mathcal{Q}, \mathbf{D}} \in \mathcal{S}} e^{-x_{\mathcal{P}, \mathcal{Q}, \mathbf{D}}} \leq \sum_{\mathcal{P} \in \mathcal{S}_p^{*(\mathcal{X})}} e^{-c_* \left(A_0^{*(\mathcal{X})} + B_0^{*(\mathcal{X})} \|\mathcal{P}\| \right)} \leq 1.$$

The proposition holds with the modified weights for polynomial as

$$\sum_{\mathbf{D} \in \mathcal{D}^M} e^{-c_* \ln |\mathcal{D}^M|} = |\mathcal{D}^M|^{1-c_*} \leq 1$$

as soon as $c_* \geq 1$. □

E Proofs for Section 4.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties (Spatial Gaussian mixtures, models, bracketing entropy and penalties)

As in the piecewise polynomial density case, Theorem 4 Spatial Gaussian mixtures, models, bracketing entropy and penalties is obtained by showing that Assumptions $(H_{\mathcal{P}, K, \mathcal{G}})$, $(S_{\mathcal{P}, K, \mathcal{G}})$ and (K) hold for any collection.

Again, one easily verifies that Assumption $(S_{\mathcal{P}, K, \mathcal{G}})$ holds. For the complexity assumption, one has

Proposition 13. *It exists a constant C depending only on a , L_m , L_M , λ_m and λ_M such that for any model $S_{\mathcal{P}, K, \mathcal{G}}$ of Theorem 4 Spatial Gaussian mixtures, models, bracketing entropy and penalties:*

$$\phi_{\mathcal{P}, K, \mathcal{G}}(\sigma) = \sigma \left(\sqrt{C} + \sqrt{\pi} + \sqrt{\ln \frac{1}{\sigma \wedge 1}} \right) \sqrt{\dim(S_{\mathcal{P}, K, \mathcal{G}})}$$

satisfies the properties required in Assumption $(H_{\mathcal{P}, K, \mathcal{G}})$.

Furthermore, $\mathfrak{D}_{\mathcal{P},K,\mathcal{G}} = n\sigma_{\mathcal{P},K,\mathcal{G}}^2$ where $\sigma_{\mathcal{P},K,\mathcal{G}}$ is the unique root of $\frac{1}{\sigma}\phi_{\mathcal{P},K,\mathcal{G}}(\sigma) = \sqrt{n}\sigma$ satisfies

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

For the Kraft assumption, one can verify that

Proposition 14. *For any collections \mathcal{S} of Theorem 4 Spatial Gaussian mixtures, models, bracketing entropy and penalties, 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 $\sum_{S_{\mathcal{P},K,\mathcal{G}} \in \mathcal{S}} e^{-x_{\mathcal{P},K,\mathcal{G}}} \leq 1$.

As for the piecewise polynomial case section, the main difficulty lies in the control of the bracketing entropy of the models. A proof of Proposition 14 Proofs for Section 4.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties (Spatial Gaussian mixtures, models, bracketing entropy and penalties) can be found in our technical report [13].

We focus thus on the proof of Proposition 13 Proofs for Section 4.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties (Spatial Gaussian mixtures, models, bracketing entropy and penalties). Due to the complex structure of the spatial mixture, we did not succeed in bounding the bracketing entropy of local model. We derive only some upper bound on the bracketing entropy $H_{[\cdot],d^{\otimes n}}(\delta, S_{\mathcal{P},K,\mathcal{G}})$. More precisely, we will derive an upper bound of an upper bound of the bracketing entropy $H_{[\cdot],d^{\otimes n}}(\delta, S_{\mathcal{P},K,\mathcal{G}})$ that is independent of the distribution law of $(X_i)_{1 \leq i \leq n}$: the bracketing entropy with a sup norm Hellinger distance $d^{\text{sup}} = \sqrt{d^{2\text{sup}}}$, $H_{[\cdot],d^{\text{sup}}}(\delta, S_{\mathcal{P},K,\mathcal{G}})$, where $d^{2\text{sup}}$ is defined by

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

Obviously $d^{2\text{sup}} \geq d^{2\otimes n}$ and thus $H_{[\cdot],d^{\text{sup}}}(\delta, S_{\mathcal{P},K,\mathcal{G}}) \geq H_{[\cdot],d^{\otimes n}}(\delta, S_{\mathcal{P},K,\mathcal{G}})$, while this quantity is independent of the design.

As

Proposition 15. *It exists a constant C depending only on a , L_m , L_M , λ_m and λ_M such that for any model $S_{\mathcal{P},K,\mathcal{G}}$ of Theorem 4 Spatial Gaussian mixtures, models, bracketing entropy and penalties:*

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

Proposition 13 Proofs for Section 4.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties (Spatial Gaussian mixtures, models, bracketing entropy and penalties) is a direct consequence of Proposition 2 Jensen-Kullback-Leibler divergence and bracketing entropy.

E.1 Model coding

Proof of Proposition 14 Proofs for Section 4.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties (S
This proposition is a simple combination of Theorem 3 Piecewise polynomial conditional densities estimation, of classical Kraft type inequalities for order selection and variable selection (see for instance in the book of Massart [34]):

Lemma 6. • For the selection of the model order K , let $x_K = (K - 1)$, for $c > 0$

$$\sum_{K \geq 1} e^{-cx_K} = \frac{1}{1 - e^{-c}}$$

- For the ordered variable selection case, $E = \text{span}\{e_i\}_{i \in I}$ with $I = \{1, \dots, p_E\}$, let $\theta_E = p_E$, for $c > 0$

$$\sum_E e^{-c\theta_E} = \frac{1}{e^c - 1} \leq 1.$$

- For the non ordered variable selection case, $E = \text{span}\{e_i\}_{i \in I}$ with $I \subset \{1, \dots, p\}$, let $\theta_E = \left(1 + \theta + \ln \frac{p}{p_E}\right) p_E$, for $c \geq 1$,

$$\sum_E e^{-c\theta_E} = \frac{e^{-(c-1)(1+\theta)}}{1 - e^{-\theta}}.$$

and on a crude bound on the number of different models indexed by $[\mu_\star L_\star D_\star A_\star]^K$ and $[\mu_\star L_\star D_\star A_\star]$. Using that there is at most $3 \times 3 \times 3 \times 3$ different type of models $[\mu_\star L_\star D_\star A_\star]^K$ and $2 \times 2 \times 2 \times 2$ different type of models $[\mu_\star L_\star D_\star A_\star]$, and $3^4 \times 2^4 = 1296$, we obtain

$$\begin{aligned} \sum_{S_K, \mathcal{P}, \mathcal{G} \in \mathcal{S}} e^{-x_{K, \mathcal{P}, \mathcal{G}}} &= \sum_{K \in \mathbb{N}^*} \sum_{\mathcal{P} \in \mathcal{S}_p^*} \sum_E \sum_{[\mu_\star L_\star D_\star A_\star]^K} \sum_{[\mu_\star L_\star D_\star A_\star]} e^{-c_\star (A_0^{\star(x)} + B_0^{\star(x)} \|\mathcal{P}\| + (K-1) + \mathcal{D}_E)} \\ &= \left(\sum_{K \in \mathbb{N}^*} e^{-c_\star(K-1)} \right) \left(\sum_{\mathcal{P} \in \mathcal{S}_p^*} e^{-c_\star (A_0^{\star(x)} + B_0^{\star(x)} \|\mathcal{P}\|)} \right) \\ &\quad \times \left(\sum_E e^{-c_\star \mathcal{D}_E} \right) \left(\sup_{K \in \mathbb{N}^*} \sum_{[\mu_\star L_\star D_\star A_\star]^K} \sum_{[\mu_\star L_\star D_\star A_\star]} \right) \\ &\leq 1296 \frac{1}{1 - e^{-c_\star}} \Sigma_0^* e^{-c_\star C_0^*} \begin{cases} 1 & \text{if } E \text{ is known,} \\ \frac{1}{e^{c_\star} - 1} & \text{if } E \text{ is chosen amongst} \\ & \text{spaces spanned by the first} \\ & \text{coordinates,} \\ 2e^{-(c_\star-1)(1+\ln 2)} & \text{if } E \text{ is free.} \end{cases} \end{aligned}$$

Choosing c_\star slightly larger than $\max(1, c_0^*)$ yields the result. \square

E.2 Entropy of spatial mixtures

Proof of Proposition 15 Proofs for Section 4.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties (S) While we use the 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

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

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

A key tool is the following Lemma that allows to decompose the entropy in three parts:

Lemma 7. For any $\delta \in (0, \sqrt{2}]$,

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

We bound those bracketing entropies with the help of two lemmas.

Lemma 8. For any $\delta \in [0, \sqrt{2}]$,

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

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

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

proved in Genovese and Wasserman [21] implies the existence of a universal constant $\mathcal{C}_{\mathcal{S}}$ such that

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

while Proposition 4 Bracketing entropy of Gaussian families handles the bracketing entropy of Gaussian K -uples collection. It implies the existence of two constants $\mathcal{C}_{[\star]^*}$ and $\mathcal{C}_{[\star]}$ depending only on a, L_m, L_M, λ_m and λ_M such that

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

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

As $\dim(\mathcal{S}_{K, \mathcal{P}, \mathcal{G}}) = \|\mathcal{P}\|(K-1) + \dim(\mathcal{G}_E^K) + \dim(\mathcal{G}_{E^\perp})$, we obtain Proposition 15 Proofs for Section 4.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties (Spatial Gaussian mixtures, models, bracketing entropy and penalties) with $C = \max(\mathcal{C}_{\mathcal{S}}, \mathcal{C}_{[\star]^*}, \mathcal{C}_{[\star]})$. \square

E.3 Entropy of Gaussian families

Instead of Proposition 4 Bracketing entropy of Gaussian families, we prove the slightly stronger

Proposition 16. Let $\kappa \geq \frac{3}{4}$ and

$$\gamma_\kappa = \min \left(\frac{3(\kappa - \frac{3}{4})}{2(1 + \frac{2\kappa}{9})(1 + \frac{2}{9})^2(1 + \frac{1}{9})}, \frac{(\kappa - \frac{1}{2})}{2(1 + \frac{2\kappa}{9})(1 + \frac{2}{9})^2} \right) \quad \beta_\kappa = \sqrt{\kappa^2 \cosh\left(\frac{2}{9\kappa}\right) + \frac{1}{4}}$$

$$\text{Assume } \begin{cases} a \geq \frac{\sqrt{\gamma_\kappa}}{18\beta_\kappa} \sqrt{L_m \lambda_m \frac{\lambda_M}{\lambda_m} \frac{\delta}{p_E}} \\ \ln \left(\frac{L_M}{L_m} \right) \geq \frac{1}{20\beta_\kappa} \delta \\ \frac{\lambda_M}{\lambda_m} \ln \left(\frac{\lambda_M}{\lambda_m} \right) \geq \frac{1}{27(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa} \frac{\delta}{p_E} \end{cases}$$

Then for any $\delta \in [0, \sqrt{2}]$,

$$H_{[\cdot], d^{\text{max}}}(\delta/9, \mathcal{G}_{[\mu_\star, L_\star, D_\star, A_\star]_E^K}) \leq \mathcal{I}_{[\mu_\star, L_\star, D_\star, A_\star]_{p_E}^K} + \mathcal{D}_{[\mu_\star, L_\star, D_\star, A_\star]_{p_E}^K} \ln \frac{1}{\delta}$$

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

$$\mathcal{I}_{[\mu_*, L_*, D_*, A_*]_{p_E}^K} = c_{\mu_*} \mathcal{I}_{\mu, p_E} + c_{L_*} \mathcal{I}_{L, p_E} + c_{D_*} \mathcal{I}_{D, p_E} + c_{A_*} \mathcal{I}_{A, p_E} \text{ with } \begin{cases} c_{\mu_0} = c_{L_0} = c_{D_0} = c_{A_0} = 0 \\ c_{\mu_K} = c_{L_K} = c_{D_K} = c_{A_K} = K \\ c_{\mu} = c_L = c_D = c_A = 1 \end{cases},$$

$$\begin{cases} \mathcal{D}_{\mu, p_E} = p_E \\ \mathcal{D}_L = 1 \\ \mathcal{D}_{D, p_E} = \frac{p_E(p_E-1)}{2} \\ \mathcal{D}_{A, p_E} = p_E - 1 \end{cases} \text{ and } \begin{cases} \mathcal{I}_{\mu, p_E} = p_E \left(\ln \left(\frac{36a\beta_\kappa p_E}{\sqrt{\gamma_\kappa L_m \lambda_m \frac{\lambda_m}{\lambda_M}}} \right) \right) \\ \mathcal{I}_{L, p_E} = \ln \left(40\beta_\kappa \ln \left(\frac{L_M}{L_m} \right) p_E \right) \\ \mathcal{I}_{D, p_E} = \frac{p_E(p_E-1)}{2} \left(\frac{\ln c}{\frac{p_E(p_E-1)}{2}} + \left(\ln \left(36(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa \frac{\lambda_M}{\lambda_m} p_E \right) \right) \right) \\ \mathcal{I}_{A, p_E} = (p_E - 1) \left(\ln \left(108(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa \frac{\lambda_M}{\lambda_m} \ln \left(\frac{\lambda_M}{\lambda_m} \right) p_E \right) \right) \end{cases}.$$

Furthermore, for any $p_E \leq p$

$$\begin{aligned} \mathcal{I}_{\mu, p_E} &\leq \mathcal{C}_{\mu, p} \mathcal{D}_{\mu, p_E} \\ \mathcal{I}_{L, p_E} &\leq \mathcal{C}_{L, p} \mathcal{D}_{L, p_E} \\ \mathcal{I}_{D, p_E} &\leq \mathcal{C}_{D, p} \mathcal{D}_{D, p_E} \\ \mathcal{I}_{A, p_E} &\leq \mathcal{C}_{A, p} \mathcal{D}_{A, p_E} \end{aligned}$$

with

$$\begin{aligned} \mathcal{C}_{\mu, p} &= \ln \left(\frac{36a\beta_\kappa p}{\sqrt{\gamma_\kappa L_m \lambda_m \frac{\lambda_m}{\lambda_M}}} \right) \\ \mathcal{C}_{L, p} &= \ln \left(40\beta_\kappa \ln \left(\frac{L_M}{L_m} \right) p \right) \\ \mathcal{C}_{D, p} &= \left(\ln c + \left(\ln \left(36(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa \frac{\lambda_M}{\lambda_m} p \right) \right) \right) \\ \mathcal{C}_{A, p} &= \ln \left(108(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa \frac{\lambda_M}{\lambda_m} \ln \left(\frac{\lambda_M}{\lambda_m} \right) p \right) \end{aligned}$$

and, uniformly over K ,

$$\begin{aligned} \mathcal{I}_{[\mu_*, L_*, D_*, A_*]_{p_E}^K} &\leq \max_{\mu'_*, L'_*, D'_*, A'_*, K'} \left(\mathcal{C}_{\mu, p} \frac{c_{\mu'_*} K'}{c_{\mu'_*} K' + c_{L'_*} + c_{D'_*} \frac{K'(K'-1)}{2} + c_{A'_*} (K' - 1)} \right. \\ &\quad + \mathcal{C}_{L, p} \frac{c_{L'_*}}{c_{\mu'_*} K' + c_{L'_*} + c_{D'_*} \frac{K'(K'-1)}{2} + c_{A'_*} (K' - 1)} \\ &\quad + \mathcal{C}_{D, p} \frac{c_{D'_*} \frac{K'(K'-1)}{2}}{c_{\mu'_*} K' + c_{L'_*} + c_{D'_*} \frac{K'(K'-1)}{2} + c_{A'_*} (K' - 1)} \\ &\quad \left. + \mathcal{C}_{A, p} \frac{c_{A'_*} (K' - 1)}{c_{\mu'_*} K' + c_{L'_*} + c_{D'_*} \frac{K'(K'-1)}{2} + c_{A'_*} (K' - 1)} \right) \mathcal{D}_{[\mu_*, L_*, D_*, A_*]_{p_E}^K} \\ &\leq \max(\mathcal{C}_{\mu, p}, \mathcal{C}_{L, p}, \mathcal{C}_{D, p}, \mathcal{C}_{A, p}) \mathcal{D}_{[\mu_*, L_*, D_*, A_*]_{p_E}^K} \end{aligned}$$

where the max is taken over all the Gaussian set type and all number of classes considered.

Proof of Proposition 16 Entropy of Gaussian families. We consider all models $\mathcal{G}_{[\mu, L, A, D, \delta]_E^K}$ at once by a “tensorial” construction of the $\delta/9$ bracket collection.

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

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

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

- for any δ_L , the grid $\mathcal{G}_L(L_m, L_M, \delta_L)$ of $[L_m, L_M]$:

$$\mathcal{G}_L(L_m, L_M, \delta_L) = \{L_m(1 + \delta_L)^g \mid g \in \mathbb{N}, L_m(1 + \delta_L)^g \leq L_M\}$$

- For any δ_D , the grid $\mathcal{G}_D(p_E, \delta_D)$ 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 [41]).

- for any δ_A , the grid $\mathcal{G}_A(\lambda_m, \lambda_M, p_E, \delta_A)$ of $\mathcal{A}(\lambda_m, \lambda_M(1 + \delta_A), p_E)$:

$$\mathcal{G}_A(\lambda_m, \lambda_M, p_E, \delta_A) = \{A \in \mathcal{A}(\lambda_m, \lambda_M(1 + \delta_A), p_E) \mid \forall 1 \leq i < p_E, \exists g_i \in \mathbb{N}, A_i = \lambda_m(1 + \delta_A)^{g_i}\}.$$

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 $[L_m, L_M]$, there is a $\tilde{L} \in \mathcal{G}_L(L_m, L_M, \delta_L)$ such that $(1 + \delta_L)^{-1} L_{j_L} < L \leq L_{j_L}$ while

$$|\mathcal{G}_L(L_m, L_M, \delta_L)| \leq 1 + \frac{\ln\left(\frac{L_M}{L_m}\right)}{\ln(1 + \delta_L)}.$$

If we further assume that $\delta_L \leq \frac{1}{9}$ then $\ln(1 + \delta_L) \geq \frac{9}{10}\delta_L$ and

$$|\mathcal{G}_L(L_m, L_M, \delta_L)| \leq 1 + \frac{10 \ln\left(\frac{L_M}{L_m}\right)}{9\delta_L} \leq \max\left(2, \frac{20 \ln\left(\frac{L_M}{L_m}\right)}{9\delta_L}\right).$$

By definition on a δ_D -net, for any $D \in SO(p_E)$ there is a $\tilde{D} \in \mathcal{G}_D(p_E, \delta_D)$ such that

$$\forall x, \|(\tilde{D} - D)x\|_2 \leq \delta_D \|x\|_2.$$

As proved by Szarek [41], it exists a universal constant c_S such that, as soon as $\delta_D \leq 1$

$$|\mathcal{G}_D(p_E, \delta_D)| \leq c \left(\frac{1}{\delta_D}\right)^{\frac{p_E(p_E-1)}{2}}$$

where $\frac{p_E(p_E-1)}{2}$ is the intrinsic dimension of $SO(p_E)$.

The structure of the grid $\mathcal{G}_A(\lambda_m, \lambda_M, p_E, \delta_A)$ is more complex. Although, looking at the condition on the $p_E - 1$ first diagonal values, we have

$$|\mathcal{G}_A(\lambda_m, \lambda_M, p_E, \delta_A)| \leq \left(2 + \frac{\ln\left(\frac{\lambda_M}{\lambda_m}\right)}{\ln(1 + \delta_A)} \right)^{p_E - 1}$$

where $p_E - 1$ is the intrinsic dimension of $\mathcal{A}(\lambda_m, \lambda_M, p_E)$. If we further assume that $\delta_A \leq \frac{1}{36}$ then $\ln(1 + \delta_A) \geq \frac{36}{37}\delta_A$ and thus

$$|\mathcal{G}_A(\lambda_m, \lambda_M, p_E, \delta_A)| \leq \left(2 + \frac{37 \ln\left(\frac{\lambda_M}{\lambda_m}\right)}{36\delta_A} \right)^{p_E - 1} \leq \max \left(4^{p_E - 1}, \left(\frac{74 \ln\left(\frac{\lambda_M}{\lambda_m}\right)}{36\delta_A} \right)^{p_E - 1} \right).$$

A key non trivial property of this grid is the following approximation property:

Lemma 9. *For $A \in \mathcal{A}(\lambda_m, \lambda_M, p_E)$ there is $\tilde{A} \in \mathcal{G}_A(\lambda_m, \lambda_M, p_E, \delta_A)$ such that*

$$|\tilde{A}_{i,i}^{-1} - A_{i,i}^{-1}| \leq \delta_A \lambda_m^{-1}.$$

Define $c_{\mu_0} = c_{L_0} = c_{D_0} = c_{A_0} = 0, c_{\mu_K} = c_{L_K} = c_{D_K} = c_{A_K} = K, c_\mu = c_L = c_D = c_A = 1$. Let f_{K, μ_*, p_E} be the application from $(\mathbb{R}^{p_E})^{c_{\mu_*}}$ to \mathbb{R}^K defined by

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

and f_{K, L_*} (respectively f_{K, D_*} and f_{K, A_*}) be the similar application from $(\mathbb{R}^+)^{c_{L_*}}$ into $(\mathbb{R}^+)^K$ (respectively from $(SO(p_E))^{c_{D_*}}$ into $(SO(p_E))^K$ and from $(\mathcal{A}(0, +\infty, p_E))^{c_{A_*}}$ into $(\mathcal{A}(0, +\infty, p_E))^K$).

By definition, the image of

$$([-a, a]^{p_E})^{c_{\mu_*}} \times ([L_m, L_M])^{c_{L_*}} \times (SO(p_E))^{c_{D_*}} \times (\mathcal{A}(\lambda_m, \lambda_M, p_E))^{c_{A_*}}$$

by $(f_{K, \mu_*} \otimes f_{L_*} \otimes f_{K, D_*} \otimes f_{K, A_*})$ is, up to reordering, the set of parameters of all the K -uples of Gaussian densities of type $[\mu_* L_*, D_*, A_*]^K$.

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

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

We prove that, for the γ_κ and β_κ defined in Proposition 16 Entropy of Gaussian families and any $\kappa \geq \frac{3}{4}$, the choice

$$\begin{cases} \delta_\mu = \frac{\sqrt{\gamma_\kappa L_m \lambda_m \frac{\lambda_m}{\lambda_M}}}{9\beta_\kappa} \frac{\delta}{p_E} \\ \delta_L = \frac{1}{18\beta_\kappa} \frac{\delta}{p_E} \\ \delta_D = \delta_A = \frac{1}{36(1+\frac{2}{9})(\sqrt{2}+1)\beta_\kappa} \frac{\lambda_m}{\lambda_M} \frac{\delta}{p_E} \\ \delta_\Sigma = \frac{1}{9\beta_\kappa} \frac{\delta}{p_E} \end{cases}$$

is such the image of

$$(\mathcal{G}_\mu(a, p_E, \delta_\mu))^{c_{\mu^*}} \times (\mathcal{G}_L(L_m, L_M, \delta_L))^{c_{L^*}} \times (\mathcal{G}_D(p_E, \delta_D))^{c_{D^*}} \times (\mathcal{G}_A(\lambda_m, \lambda_M, p_E, \delta_A))^{c_{A^*}}$$

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

Indeed, we have

Lemma 10. *Let $\kappa \geq \frac{3}{4}$, $\gamma_\kappa = \min\left(\frac{3(\kappa - \frac{3}{4})}{2(1 + \frac{2\kappa}{9})(1 + \frac{2}{9})^2(1 + \frac{1}{9})}, \frac{(\kappa - \frac{1}{2})}{2(1 + \frac{2\kappa}{9})(1 + \frac{2}{9})^2}\right)$ and $\beta_\kappa = \sqrt{\kappa^2 \cosh(\frac{2}{9\kappa}) + \frac{1}{4}}$. For any $0 < \delta \leq \sqrt{2}$, any $p_E \geq 1$ and any $\delta_\Sigma \leq \frac{1}{9\beta_\kappa} \frac{\delta}{p_E}$,*

Let $(\tilde{\mu}, \tilde{L}, \tilde{A}, \tilde{D}) \in [-a, a]^{p_E} \times [L_m, L_M] \times \mathcal{A}(\lambda_m, +\infty) \times SO(p_E)$, define $\tilde{\Sigma} = \tilde{L}\tilde{D}\tilde{A}\tilde{D}'$,

$$t^-(x) = (1 + \kappa\delta_\Sigma)^{-p_E} \Phi_{\tilde{\mu}, (1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(x) \quad \text{and} \quad t^+(x) = (1 + \kappa\delta_\Sigma)^{p_E} \Phi_{\tilde{\mu}, (1+\delta_\Sigma)\tilde{\Sigma}}(x).$$

then $[t^-, t^+]$ is an $\delta/9$ Hellinger bracket.

Furthermore, let $(\mu, L, A, D) \in [-a, a]^{p_E} \times [L_m, L_M] \times \mathcal{A}(\lambda_m, \lambda_M) \times SO(p_E)$ and define $\Sigma = LDAD'$. If

$$\begin{cases} \|\mu - \tilde{\mu}\|^2 \leq p_E \gamma_\kappa L_m \lambda_m \frac{\lambda_m}{\lambda_M} \delta_\Sigma^2 \\ (1 + \frac{\delta_\Sigma}{2})^{-1} \tilde{L} \leq L \leq \tilde{L} \\ \forall 1 \leq i \leq p_E, \quad |A_{i,i}^{-1} - \tilde{A}_{i,i}^{-1}| \leq \frac{1}{4(1 + \frac{2}{9})(\sqrt{2}+1)} \frac{1}{\lambda_M} \delta_\Sigma \\ \forall x \in \mathbb{R}^{p_E}, \quad \|Dx - \tilde{D}x\| \leq \frac{1}{4(1 + \frac{2}{9})(\sqrt{2}+1)} \frac{\lambda_m}{\lambda_M} \delta_\Sigma \|x\| \end{cases}$$

then $t^-(x) \leq \Phi_{\mu, \Sigma}(x) \leq t^+(x)$.

as proved later. By definition of d^{\max} , this implies that our choice of $\delta_\mu, \delta_L, \delta_D, \delta_A$ and δ_Σ is such that every K -uple of couples of the collections is a $\delta/9$ -bracket and they cover the whole set.

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

$$\begin{aligned} & \left(\max \left(2^{p_E}, \left(\frac{4a}{\frac{\sqrt{\gamma_\kappa L_m \lambda_m \frac{\lambda_m}{\lambda_M} \delta}}{9\beta_\kappa} p_E}} \right)^{p_E} \right) \right)^{c_{\mu^*}} \times \left(\max \left(2, \frac{20 \ln \left(\frac{L_M}{L_m} \right)}{9 \frac{1}{18\beta_\kappa} \frac{\delta}{p_E}} \right) \right)^{c_{L^*}} \\ & \times \left(c \left(\frac{1}{\frac{36(1 + \frac{2}{9})(\sqrt{2}+1)\beta_\kappa}{\lambda_M} \frac{\delta}{p_E}} \right)^{\frac{p_E(p_E-1)}{2}} \right)^{c_{D^*}} \\ & \times \left(\max \left(4^{p_E-1}, \left(\frac{74 \ln \left(\frac{\lambda_M}{\lambda_m} \right)}{36 \frac{1}{36(1 + \frac{2}{9})(\sqrt{2}+1)\beta_\kappa} \frac{\lambda_m}{\lambda_M} \frac{\delta}{p_E}} \right)^{p_E-1} \right) \right)^{c_{A^*}} \\ & \left(\max \left(2^{p_E}, \left(\frac{36a\beta_\kappa p_E}{\sqrt{\gamma_\kappa L_m \lambda_m \frac{\lambda_m}{\lambda_M} \delta}} \right)^{p_E} \right) \right)^{c_{\mu^*}} \times \left(\max \left(2, \frac{40\beta_\kappa \ln \left(\frac{L_M}{L_m} \right) p_E}{\delta} \right) \right)^{c_{L^*}} \\ & \times \left(c \left(\frac{36(1 + \frac{2}{9})(\sqrt{2}+1)\beta_\kappa \frac{\lambda_M}{\lambda_m} p_E}{\delta} \right)^{\frac{p_E(p_E-1)}{2}} \right)^{c_{D^*}} \end{aligned}$$

$$\times \left(\max \left(4^{p_E-1}, \left(\frac{108(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa \frac{\lambda_M}{\lambda_m} \ln \left(\frac{\lambda_M}{\lambda_m} \right) p_E}{\delta} \right)^{p_E-1} \right) \right)^{c_{A^*}}$$

So that under the mild assumptions that

$$\begin{cases} a \geq \frac{\sqrt{\gamma_\kappa}}{18\beta_\kappa} \sqrt{L_m \lambda_m \frac{\lambda_m}{\lambda_M} \frac{\delta}{p_E}} \\ \ln \left(\frac{L_M}{L_m} \right) \geq \frac{1}{20\beta_\kappa} \delta \\ \frac{\lambda_M}{\lambda_m} \ln \left(\frac{\lambda_M}{\lambda_m} \right) \geq \frac{1}{27(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa} \frac{\delta}{p_E} \end{cases},$$

$$\begin{aligned} & H_{[\cdot], d^{\max}}(\delta/9, \mathcal{G}_{[\mu^*, L^*, D^*, A^*]_E^K}) \\ & \leq c_{\mu^*} p_E \left(\ln \left(\frac{36a\beta_\kappa p_E}{\sqrt{\gamma_\kappa L_m \lambda_m \frac{\lambda_m}{\lambda_M}}} \right) + \ln \frac{1}{\delta} \right) \\ & \quad + c_{L^*} \left(\ln \left(40\beta_\kappa \ln \left(\frac{L_M}{L_m} \right) p_E \right) + \ln \frac{1}{\delta} \right) \\ & \quad + c_{D^*} \frac{p_E(p_E - 1)}{2} \left(\frac{\ln c}{\frac{p_E(p_E-1)}{2}} + \left(\ln \left(36(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa \frac{\lambda_M}{\lambda_m} p_E \right) + \ln \frac{1}{\delta} \right) \right) \\ & \quad + c_{A^*} (p_E - 1) \left(\ln \left(108(1 + \frac{2}{9})(\sqrt{2} + 1)\beta_\kappa \frac{\lambda_M}{\lambda_m} \ln \left(\frac{\lambda_M}{\lambda_m} \right) p_E \right) + \ln \frac{1}{\delta} \right) \end{aligned}$$

which concludes the proof. \square

E.4 Entropy of spatial mixtures (Lemmas)

Proof of Lemma 7 Entropy of spatial mixtures. This is a variation around the proof of Genovese and Wasserman [21].

Let $\{[\pi_1^-, \pi_1^+], \dots, [\pi_{N_{S_{K-1}}}^-, \pi_{N_{S_{K-1}}}^+]\}$ be a minimal covering of $\delta/3$ Hellinger bracket of the simplex \mathcal{S}_{K-1} . Let

$$\left\{ \left[(t_{E,1,1}^-, \dots, t_{E,K,1}^-), (t_{E,1,1}^+, \dots, t_{E,K,1}^+) \right], \dots, \left[(t_{E,1,N_{E,K}}^-, \dots, t_{E,K,N_{E,K}}^-), (t_{E,1,N_{E,K}}^+, \dots, t_{E,K,N_{E,K}}^+) \right] \right\}$$

be a minimal covering of $\delta/9$ sup norm Hellinger bracket of $\mathcal{G}_{E,K}$ and $\{[t_{E^\perp,1}^-, t_{E^\perp,1}^+], \dots, [t_{E^\perp,N_\perp}^-, t_{E^\perp,N_\perp}^+]\}$

be a minimal covering of $\delta/9$ Hellinger bracket of \mathcal{G}_{E^\perp} . By definition, $\ln N_{S_{K-1}} = H_{[\cdot], d}(\delta/3, \mathcal{S}_{K-1})$, $\ln N_{E,K} = H_{[\cdot], d^{\max}}(\delta/9, \mathcal{G}_{E,K})$ and $\ln N_{E^\perp} = H_{[\cdot], d}(\delta/9, \mathcal{G}_{E^\perp})$.

By construction,

$$\left\{ \left[\sum_{\mathcal{R}_l \in \mathcal{P}} \left(\sum_{k=1}^K \pi_{i[\mathcal{R}_l], k}^- t_{E, k, j}^-(y) t_{E^\perp, j_{E^\perp}}^-(y) \right) \mathbf{1}_{\{x \in \mathcal{R}_l\}}, \sum_{\mathcal{R}_l \in \mathcal{P}} \left(\sum_{k=1}^K \pi_{i[\mathcal{R}_l], k}^+ t_{E, k, j}^+(y) t_{E^\perp, l}^+(y) \right) \mathbf{1}_{\{x \in \mathcal{R}_l\}} \right] \right\}$$

$$1 \leq i[\mathcal{R}_l] \leq N_{S_{K-1}}, 1 \leq j \leq N_{E,K}, 1 \leq l \leq N_{E^\perp}$$

is a covering of the model $S_{K, \mathcal{P}, \mathcal{G}}$ of cardinality $\exp(|\mathcal{P}| H_{[\cdot], d}(\delta/3, \mathcal{S}_{K-1}) + H_{[\cdot], d^{\max}}(\delta/9, \mathcal{G}_{E,K}) + H_{[\cdot], d}(\delta/9, \mathcal{G}_{E^\perp}))$.

It remains thus only to prove that each bracket is of sup norm Hellinger d^{sup} width smaller than δ .

Using

Lemma 11. For any δ Hellinger brackets $[t^-(x), t^+(x)]$, if for any x $[u^-(x, y), u^+(x, y)]$ is an δ bracket and $\delta \leq \sqrt{2}/3$, then $[t^-(x) u^-(x, y), t^+(x) u^+(x, y)]$ is a 3δ Hellinger bracket.

we obtain immediately that

$$d^2 \left(t_{E,k,j_E}^- (\cdot) t_{E^\perp,l}^- (\cdot), t_{E,k,j_E}^+ (\cdot) t_{E^\perp,l}^+ (\cdot) \right) \leq 9(\delta/9)^2 = (\delta/3)^2.$$

We denote $[t_{k,j,l}^{--}, t_{k,j,l}^{++}]$ the corresponding $\delta/3$ Hellinger bracket.

By definition,

$$\begin{aligned} & d^{2 \sup} \left(\sum_{\mathcal{R}_i \in \mathcal{P}} \left(\sum_{k=1}^K \pi_{i[\mathcal{R}_i],k}^- t_{k,j,l}^{--} (y) \right) \mathbf{1}_{\{x \in \mathcal{R}_i\}}, \sum_{\mathcal{R}_i \in \mathcal{P}} \left(\sum_{k=1}^K \pi_{i[\mathcal{R}_i],k}^+ t_{k,j,l}^{++} (y) \right) \mathbf{1}_{\{x \in \mathcal{R}_i\}} \right) \\ &= \sup_{\mathcal{R}_i \in \mathcal{P}} d^2 \left(\sum_{k=1}^K \pi_{i[\mathcal{R}_i],k}^- t_{k,j,l}^{--}, \sum_{k=1}^K \pi_{i[\mathcal{R}_i],k}^+ t_{k,j,l}^{++} \right) \\ &\leq \sup_{i,j,l} d^2 \left(\sum_{k=1}^K \pi_{i,k}^- t_{k,j,l}^{--}, \sum_{k=1}^K \pi_{i,k}^+ t_{k,j,l}^{++} \right) \end{aligned}$$

Seeing $\pi_{i,k} g_{k,j,l}(y)$ as a function of k and y , we can use

Lemma 12. For any brackets $[t^-(x), t^+(x)]$ and if for any x $[u^-(x, y), u^+(x, y)]$ is a bracket then

$$d_y^2 \left(\int_x t^-(x) u^-(x, y) d\lambda_x(x), \int_x t^+(x) u^+(x, y) d\lambda_x(x) \right) \leq d_{x,y}^2 (t^-(x) u^-(x, y), t^+(x) u^+(x, y))$$

to obtain

$$\begin{aligned} & d^{2 \sup} \left(\sum_{\mathcal{R}_i \in \mathcal{P}} \left(\sum_{k=1}^K \pi_{i[\mathcal{R}_i],k}^- t_{k,j,l}^{--} (y) \right) \mathbf{1}_{\{x \in \mathcal{R}_i\}}, \sum_{\mathcal{R}_i \in \mathcal{P}} \left(\sum_{k=1}^K \pi_{i[\mathcal{R}_i],k}^+ t_{k,j,l}^{++} (y) \right) \mathbf{1}_{\{x \in \mathcal{R}_i\}} \right) \\ &\leq \sup_{i,j,l} d_{k,y}^2 \left(\pi_{i,k}^- t_{k,j,l}^{--}(y), \pi_{i,k}^+ t_{k,j,l}^{++}(y) \right) \end{aligned}$$

and then using again Lemma 11 Entropy of spatial mixtures (Lemmas)

$$\leq 9(\delta/3)^2 = \delta^2.$$

□

Proof of Lemma 11 Entropy of spatial mixtures (Lemmas).

$$\begin{aligned} & d^2(t^-(x) u^-(x, y), t^+(x) u^+(x, y)) \\ &= \iint \left(\sqrt{t^+(x) u^+(x, y)} - \sqrt{t^-(x) u^-(x, y)} \right)^2 d\lambda_x(x) d\lambda_y(y) \\ &= \iint \left(\sqrt{t^+(x)} \left(\sqrt{u^+(x, y)} - \sqrt{u^-(x, y)} \right) + \left(\sqrt{t^+(x)} - \sqrt{t^-(x)} \right) \sqrt{u^-(x, y)} \right)^2 d\lambda_x(x) d\lambda_y(y) \\ &= \iint \left(t^+(x) \left(\sqrt{u^+(x, y)} - \sqrt{u^-(x, y)} \right)^2 + \left(\sqrt{t^+(x)} - \sqrt{t^-(x)} \right)^2 u^-(x, y) \right) \end{aligned}$$

$$\begin{aligned}
& + 2\sqrt{t^+(x)} \left(\sqrt{t^+(x)} - \sqrt{t^-(x)} \right) \sqrt{u^-(x,y)} \left(\sqrt{u^+(x,y)} - \sqrt{u^-(x,y)} \right) \Big) d\lambda_x(x) d\lambda_y(y) \\
= & \int t^+(x) d^2(u^-(x,y), u^+(x,y)) d\lambda_x(x) + d^2(t^-(x), t^+(x)) \sup_x \int u^-(x,y) d\lambda_y(y) \\
& + 2 \int \sqrt{t^+(x)} \left(\sqrt{t^+(x)} - \sqrt{t^-(x)} \right) \int \sqrt{u^-(x,y)} \left(\sqrt{u^+(x,y)} - \sqrt{u^-(x,y)} \right) d\lambda_y(y) d\lambda_x(x) \\
\leq & \left(\sqrt{\int t^+(x) d\lambda_x(x)} \sup_x d(u^-(x,y), u^+(x,y)) + d(t^-(x), t^+(x)) \sup_x \sqrt{\int u^-(x,y) d\lambda_y(y)} \right)^2.
\end{aligned}$$

Using

Lemma 13. For any δ -Hellinger bracket $[t^-, t^+]$, $\int t^- d\lambda \leq 1$ and $\int t^+ d\lambda \leq 1 + 2(\sqrt{2} + \sqrt{3})\delta$.

we deduce using $\delta \leq \sqrt{2}/3$

$$\begin{aligned}
d^2(t^-(x) u^-(x,y), t^+(x) u^+(x,y)) & \leq \left(\sqrt{1 + 2(\sqrt{2} + \sqrt{3})\delta} + 1 \right)^2 \delta^2 \\
& \leq 9\delta^2
\end{aligned}$$

□

Proof of Lemma 12 Entropy of spatial mixtures (Lemmas).

$$\begin{aligned}
& d_y^2 \left(\int_x t^-(x) u^-(x,y) d\lambda_x(x), \int_x t^+(x) u^+(x,y) d\lambda_x(x) \right) \\
= & \int_y \left(\sqrt{\int_x t^+(x) u^+(x,y) d\lambda_x(x)} - \sqrt{\int_x t^-(x) u^-(x,y) d\lambda_x(x)} \right)^2 d\lambda_y(y) \\
= & \int_y \int_x t^+(x) u^+(x,y) d\lambda_x(x) d\lambda_y(y) + \int_y \int_x t^-(x) u^-(x,y) d\lambda_x(x) d\lambda_y(y) \\
& - 2 \int_y \sqrt{\int_x t^+(x) u^+(x,y) d\lambda_x(x)} \sqrt{\int_x t^-(x) u^-(x,y) d\lambda_x(x)} d\lambda_y(y) \\
\leq & \int_y \int_x t^+(x) u^+(x,y) d\lambda_x(x) d\lambda_y(y) + \int_y \int_x t^-(x) u^-(x,y) d\lambda_x(x) d\lambda_y(y) \\
& - 2 \int_y \int_x \sqrt{t^+(x) u^+(x,y)} \sqrt{t^-(x) u^-(x,y)} d\lambda_x(x) d\lambda_y(y) \\
\leq & d_{x,y}^2(t^-(x) u^-(x,y), t^+(x) u^+(x,y))
\end{aligned}$$

□

Proof of Lemma 13 Entropy of spatial mixtures (Lemmas). The first point is straightforward as t^- is upper-bounded by a density.

For the second point,

$$\begin{aligned}
\int t^+ d\lambda & = \int (t^+ - t^-) d\lambda + \int t^- d\lambda \leq \int (\sqrt{t^+} - \sqrt{t^-}) (\sqrt{t^+} + \sqrt{t^-}) d\lambda + 1 \\
& \leq 2 \int (\sqrt{t^+} - \sqrt{t^-}) \sqrt{t^+} d\lambda + 1 \leq 2 \left(\int (\sqrt{t^+} - \sqrt{t^-})^2 d\lambda \right)^{1/2} \left(\int t^+ d\lambda \right)^{1/2} + 1
\end{aligned}$$

$$\int t^+ d\lambda \leq 2\delta \left(\int t^+ d\lambda \right)^{1/2} + 1$$

Solving the corresponding inequality and using $\delta \leq \sqrt{2}$ yields

$$\int t^+ d\lambda \leq \left(\delta + \sqrt{1 + \delta^2} \right)^2 \leq 1 + 2 \left(\delta + \sqrt{1 + \delta^2} \right) \delta \leq 1 + 2 \left(\sqrt{2} + \sqrt{3} \right) \delta$$

□

E.5 Entropy of Gaussian families (Lemma)

Proof of Lemma 9 Entropy of Gaussian families. We define first \tilde{g}_i as the set of integers such that

$$\forall 1 \leq i < p_E, \lambda_m(1 + \delta_A)^{\tilde{g}_i} \leq A_{i,i} < \lambda_m(1 + \delta_A)^{\tilde{g}_i+1}.$$

By construction $\tilde{g}_i \in \mathbb{N}$ and $\lambda_m(1 + \delta_A)^{\tilde{g}_i} \leq \lambda_M$. Now as $A_{p_E, p_E} = \frac{1}{\prod_{i=1}^{p_E-1} A_{i,i}}$,

$$\frac{1}{\prod_{i=1}^{p_E-1} \lambda_m(1 + \delta_A)^{\tilde{g}_i+1}} = \frac{(1 + \delta_A)^{-(p_E-1)}}{\prod_{i=1}^{p_E-1} \lambda_m(1 + \delta_A)^{\tilde{g}_i}} < A_{p_E, p_E} \leq \frac{1}{\prod_{i=1}^{p_E-1} \lambda_m(1 + \delta_A)^{\tilde{g}_i}}.$$

There is thus an integer d between 0 and $p_E - 2$ such that

$$\frac{(1 + \delta_A)^{-d-1}}{\prod_{i=1}^{p_E-1} \lambda_m(1 + \delta_A)^{\tilde{g}_i}} < A_{p_E, p_E} \leq \frac{(1 + \delta_A)^{-d}}{\prod_{i=1}^{p_E-1} \lambda_m(1 + \delta_A)^{\tilde{g}_i}}.$$

Defined then $g_i = \tilde{g}_i + 1$ if $i \leq d$ and $g_i = \tilde{g}_i$ otherwise, then

$$\forall 1 \leq i < p_E, \lambda_m(1 + \delta_A)^{g_i-1} \leq A_{i,i} < \lambda_m(1 + \delta_A)^{g_i}$$

which implies $\lambda_m(1 + \delta_A)^{g_i} \leq (1 + \delta_A)\lambda_M$. Now

$$\frac{1}{\prod_{i=1}^{p_E-1} \lambda_m(1 + \delta_A)^{g_i}} = \frac{(1 + \delta_A)^{-d}}{\prod_{i=1}^{p_E-1} \lambda_m(1 + \delta_A)^{\tilde{g}_i}}$$

and thus

$$\frac{(1 + \delta_A)^{-1}}{\prod_{i=1}^{p_E-1} \lambda_m(1 + \delta_A)^{g_i}} < A_{p_E, p_E} \leq \frac{1}{\prod_{i=1}^{p_E-1} \lambda_m(1 + \delta_A)^{g_i}}$$

which implies

$$\lambda_m \leq \frac{1}{\prod_{i=1}^{p_E-1} \lambda_m(1 + \delta_A)^{g_i}} \leq (1 + \delta_A)\lambda_M.$$

Thus the diagonal matrix \tilde{A} defined by

$$\forall 1 \leq i \leq p_E - 1, \tilde{A}_{i,i} = \lambda_m(1 + \delta_A)^{g_i}$$

and $\tilde{A}_{p_E, p_E} = \frac{1}{\prod_{i=1}^{p_E-1} \tilde{A}_{i,i}}$ belongs to $\mathcal{G}_A(\lambda_m, \lambda_M, p_E, \delta_A)$. Furthermore, we can write for any $1 \leq i \leq p_E - 1$

$$\tilde{A}_{i,i}(1 + \delta_A)^{-1} \leq A_{i,i} < \tilde{A}_{i,i}(1 + \delta_A)$$

which implies

$$\tilde{A}_{i,i}^{-1}(1 + \delta_A)^{-1} < A_{i,i}^{-1} < \tilde{A}_{i,i}^{-1}(1 + \delta_A)$$

and thus

$$\begin{aligned} |A_{i,i}^{-1} - \tilde{A}_{i,i}^{-1}| &\leq \tilde{A}_{i,i}^{-1} \max(1 + \delta_A - 1, 1 - (1 + \delta_A)^{-1}) = \tilde{A}_{i,i}^{-1} \max\left(\delta_A, \frac{\delta_A}{1 + \delta_A}\right) \\ &\leq \lambda_m^{-1} \delta_A. \end{aligned}$$

Along the same lines,

$$(1 + \delta_A)^{-1} \tilde{A}_{p_E, p_E} \leq A_{p_E, p_E} \leq \tilde{A}_{p_E, p_E}$$

thus

$$\tilde{A}_{p_E, p_E}^{-1} \leq A_{p_E, p_E}^{-1} \leq (1 + \delta_A) \tilde{A}_{p_E, p_E}^{-1}$$

and

$$|\tilde{A}_{p_E, p_E}^{-1} - A_{p_E, p_E}^{-1}| \leq \tilde{A}_{p_E, p_E}^{-1} \delta_A \leq \lambda_m^{-1} \delta_A.$$

□

Proof of Lemma 10 Entropy of Gaussian families. We prove first that $[t^-, t^+]$ is an $\delta/9$ Hellinger bracket. As $(1 + \delta_\Sigma) \tilde{\Sigma}^{-1} - (1 + \delta_\Sigma)^{-1} \tilde{\Sigma}^{-1} = ((1 + \delta_\Sigma) - (1 + \delta_\Sigma)^{-1}) \tilde{\Sigma}^{-1}$ is a positive definite matrix, one can apply

Lemma 14. *Let $\Phi_{(\mu_1, \Sigma_1)}$ and $\Phi_{(\mu_2, \Sigma_2)}$ be two Gaussian densities with full rank covariance matrix in dimension p_E such that $\Sigma_1^{-1} - \Sigma_2^{-1}$ is a positive definite matrix, for any $x \in \mathbb{R}^{p_E}$*

$$\frac{\Phi_{(\mu_1, \Sigma_1)}(x)}{\Phi_{(\mu_2, \Sigma_2)}(x)} \leq \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}} \exp\left(\frac{1}{2} (\mu_1 - \mu_2)' (\Sigma_2 - \Sigma_1)^{-1} (\mu_1 - \mu_2)\right).$$

proved by Maugis and Michel [35]. This yields using eventually $\kappa \geq \frac{1}{2}$

$$\begin{aligned} \frac{t^-(x)}{t^+(x)} &= \frac{(1 + \kappa \delta_\Sigma)^{-p_E} \Phi_{\mu, (1 + \delta_\Sigma)^{-1} \tilde{\Sigma}}(x)}{(1 + \kappa \delta_\Sigma)^{p_E} \Phi_{\mu, (1 + \delta_\Sigma) \tilde{\Sigma}}(x)} \leq \frac{1}{(1 + \kappa \delta_\Sigma)^{2p_E}} \sqrt{\frac{(1 + \delta_\Sigma)^{p_E}}{(1 + \delta_\Sigma)^{-p_E}}} \leq \frac{(1 + \delta_\Sigma)^{p_E}}{(1 + \kappa \delta_\Sigma)^{2p_E}} \\ &\leq \left(\frac{1 + \delta_\Sigma}{(1 + \kappa \delta_\Sigma)^2}\right)^{p_E} \leq \left(\frac{1 + \delta_\Sigma}{1 + 2\kappa \delta_\Sigma + \kappa^2 \delta_\Sigma^2}\right)^{p_E} \leq 1 \end{aligned}$$

Concerning the Hellinger width,

$$\begin{aligned} d^2(t^-, t^+) &= \int t^-(x) dx + \int t^+(x) dx - 2 \int \sqrt{t^-(x)} \sqrt{t^+(x)} dx \\ &= (1 + \kappa \delta_\Sigma)^{-p_E} + (1 + \kappa \delta_\Sigma)^{p_E} \\ &\quad - 2(1 + \kappa \delta_\Sigma)^{-p_E/2} (1 + \kappa \delta_\Sigma)^{p_E/2} \int \sqrt{\Phi_{\mu, (1 + \delta_\Sigma)^{-1} \tilde{\Sigma}}(x)} \sqrt{\Phi_{\mu, (1 + \delta_\Sigma) \tilde{\Sigma}}(x)} dx \\ &= (1 + \kappa \delta_\Sigma)^{-p_E} + (1 + \kappa \delta_\Sigma)^{p_E} - \left(2 - d^2\left(\Phi_{\mu, (1 + \delta_\Sigma)^{-1} \tilde{\Sigma}}(x), \Phi_{\mu, (1 + \delta_\Sigma) \tilde{\Sigma}}(x)\right)\right). \end{aligned}$$

Using

Lemma 15. Let $\Phi_{(\mu_1, \Sigma_1)}$ and $\Phi_{(\mu_2, \Sigma_2)}$ be two Gaussian densities with full rank covariance matrix in dimension p_E ,

$$d^2(\Phi_{(\mu_1, \Sigma_1)}, \Phi_{(\mu_2, \Sigma_2)}) = 2 \left(1 - 2^{p_E/2} |\Sigma_1 \Sigma_2|^{-1/4} |\Sigma_1^{-1} + \Sigma_2^{-1}|^{-1/2} \exp \left(-\frac{1}{4} (\mu_1 - \mu_2)' (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2) \right) \right).$$

also proved in [35], we derive

$$\begin{aligned} d^2(t^-, t^+) &= \int t^-(x) dx + \int t^+(x) dx - 2 \int \sqrt{t^-(x)} \sqrt{t^+(x)} dx \\ &= (1 + \kappa \delta_\Sigma)^{-p_E} + (1 + \kappa \delta_\Sigma)^{p_E} - 2 \cdot 2^{p_E/2} \left((1 + \delta_\Sigma) + (1 + \delta_\Sigma)^{-1} \right)^{-p_E/2} \\ &= 2 - 2 \cdot 2^{p_E/2} \left((1 + \delta_\Sigma) + (1 + \delta_\Sigma)^{-1} \right)^{-p_E/2} + (1 + \kappa \delta_\Sigma)^{-p_E} + (1 + \kappa \delta_\Sigma)^{p_E} - 2 \end{aligned}$$

So that denoting

Lemma 16. For any $0 < \delta \leq \sqrt{2}$ and any $p_E \geq 1$, let $\kappa \geq \frac{1}{2}$ and $\beta_\kappa = \sqrt{\kappa^2 \cosh(\frac{2\kappa}{9}) + \frac{1}{4}}$, if $\delta_\Sigma \leq \frac{1}{9\beta_\kappa} \frac{\delta}{p_E}$, then

$$\delta_\Sigma \leq \frac{1}{p_E} \frac{2}{9} \leq \frac{2}{9}.$$

and

Lemma 17. For any $d \in \mathbb{N}$, for any $\delta_\Sigma > 0$,

$$2 - 2 \cdot 2^{d/2} \left((1 + \delta_\Sigma) + (1 + \delta_\Sigma)^{-1} \right)^{-d/2} \leq \frac{d^2 \delta_\Sigma^2}{4}.$$

Furthermore, if $d\delta_\Sigma \leq c$, then

$$(1 + \kappa \delta_\Sigma)^d + (1 + \kappa \delta_\Sigma)^{-d} - 2 \leq \kappa^2 \cosh(\kappa c) d^2 \delta_\Sigma^2.$$

with $c = \frac{2}{9}$ yields

$$d^2(t^-, t^+) \leq \left(\kappa^2 \cosh(\frac{2\kappa}{9}) + \frac{1}{4} \right) p_E^2 \delta_\Sigma^2 \leq \left(\frac{\delta}{9} \right)^2$$

as $\delta_\Sigma \leq \frac{1}{9\beta_\kappa} \frac{\delta}{p_E}$

We focus now on the proof that $t^-(x) \leq \Phi_{\mu, \Sigma}(x) \leq t^+(x)$. As

Lemma 18. Under the Assumptions of Lemma 10 Entropy of Gaussian families, $(1 + \delta_\Sigma) \tilde{\Sigma}^{-1} - \Sigma^{-1}$ and $\Sigma^{-1} - (1 + \delta_\Sigma) \tilde{\Sigma}^{-1}$ are positive definite and satisfies

$$\begin{aligned} \forall x \in \mathbb{R}^{p_E}, x' \left((1 + \delta_\Sigma) \tilde{\Sigma}^{-1} - \Sigma^{-1} \right) x &\geq \frac{1}{4} \tilde{L}^{-1} \frac{1}{\lambda_M} \delta_\Sigma \|x\|^2 \\ \forall x \in \mathbb{R}^{p_E}, x' \left(\Sigma^{-1} - (1 + \delta_\Sigma) \tilde{\Sigma}^{-1} \right) x &\geq \frac{3}{4(1 + \frac{2}{9})} \tilde{L}^{-1} \frac{1}{\lambda_M} \delta_\Sigma \|x\|^2 \end{aligned}$$

we can apply Lemma 14 Entropy of Gaussian families (Lemma) on Gaussian density ratio to both

$$\frac{\Phi_{\mu, \Sigma}(x)}{(1 + \kappa \delta_\Sigma)^{p_E} \Phi_{\tilde{\mu}, (1 + \delta_\Sigma) \tilde{\Sigma}}(x)} \quad \text{and} \quad \frac{(1 + \kappa \delta_\Sigma)^{-p_E} \Phi_{\tilde{\mu}, (1 + \delta_\Sigma)^{-1} \tilde{\Sigma}}(x)}{\Phi_{\mu, \Sigma}(x)}$$

in order to prove that they are smaller than 1.

For the first one, using

$$\begin{aligned} \frac{\Phi_{\mu,\Sigma}(x)}{(1+\kappa\delta_\Sigma)^{p_E}\Phi_{\tilde{\mu},(1+\delta_\Sigma)\tilde{\Sigma}}(x)} &\leq (1+\kappa\delta_\Sigma)^{-p_E} \left(\sqrt{\frac{|(1+\delta_\Sigma)\tilde{\Sigma}|}{|\Sigma|}} \exp\left(\frac{1}{2}(\mu-\tilde{\mu})'((1+\delta_\Sigma)\tilde{\Sigma}-\Sigma)^{-1}(\mu-\tilde{\mu})\right) \right) \\ &\leq \frac{(1+\delta_\Sigma)^{p_E/2}}{(1+\kappa\delta_\Sigma)^{p_E}} \left(\sqrt{\frac{|\tilde{\Sigma}|}{|\Sigma|}} \exp\left(\frac{1}{2}(\mu-\tilde{\mu})'((1+\delta_\Sigma)\tilde{\Sigma}-\Sigma)^{-1}(\mu-\tilde{\mu})\right) \right). \end{aligned}$$

Now

$$\begin{aligned} ((1+\delta_\Sigma)\tilde{\Sigma}-\Sigma)^{-1} &= ((1+\delta_\Sigma)\tilde{\Sigma}(\Sigma^{-1}-(1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1})\Sigma)^{-1} \\ &= (1+\delta_\Sigma)^{-1}\Sigma^{-1}(\Sigma^{-1}-(1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1})^{-1}\tilde{\Sigma}^{-1} \end{aligned}$$

and thus

$$\begin{aligned} (\mu-\tilde{\mu})'((1+\delta_\Sigma)\tilde{\Sigma}-\Sigma)^{-1}(\mu-\tilde{\mu}) &\leq (1+\delta_\Sigma)^{-1}L_m^{-1}\lambda_m^{-1}\frac{4(1+\frac{2}{9})}{3}\tilde{L}\lambda_M\delta_\Sigma^{-1}\tilde{L}^{-1}\lambda_m^{-1}\|\mu-\tilde{\mu}\|^2 \\ &\leq \frac{4(1+\frac{2}{9})}{3}(1+\delta_\Sigma)^{-1}\delta_\Sigma^{-1}L_m^{-1}\lambda_m^{-1}\frac{\lambda_M}{\lambda_m}p_E\gamma_\kappa L_m\lambda_m\frac{\lambda_m}{\lambda_M}\delta_\Sigma^2 \\ &\leq \frac{4(1+\frac{2}{9})}{3}\gamma_\kappa(1+\delta_\Sigma)^{-1}p_E\delta_\Sigma \end{aligned}$$

Now as by construction,

$$\frac{|\tilde{\Sigma}|}{|\Sigma|} \leq (1+\frac{1}{2}\delta)^{p_E},$$

one obtains

$$\begin{aligned} \frac{\Phi_{\mu,\Sigma}}{(1+\kappa\delta)^{p_E}\Phi_{\tilde{\mu},(1+\delta)\tilde{\Sigma}}} &\leq \frac{(1+\delta_\Sigma)^{p_E/2}}{(1+\kappa\delta_\Sigma)^{p_E}}(1+\frac{1}{2}\delta_\Sigma)^{p_E/2} \exp\left(\frac{1}{2}\frac{4(1+\frac{2}{9})}{3}\gamma_\kappa(1+\delta_\Sigma)^{-1}p_E\delta_\Sigma\right) \\ &\leq \left(\frac{\sqrt{1+\delta_\Sigma}\sqrt{1+\frac{1}{2}\delta_\Sigma}}{1+\kappa\delta_\Sigma} \exp\left(\frac{2(1+\frac{2}{9})}{3}\gamma_\kappa(1+\delta_\Sigma)^{-1}\delta_\Sigma\right)\right)^{p_E}. \end{aligned}$$

It is thus sufficient to prove that

$$\frac{\sqrt{1+\delta_\Sigma}\sqrt{1+\frac{1}{2}\delta_\Sigma}}{1+\kappa\delta_\Sigma} \exp\left(\frac{2(1+\frac{2}{9})}{3}\gamma_\kappa(1+\delta_\Sigma)^{-1}\delta_\Sigma\right) \leq 1$$

or equivalently

$$\frac{2(1+\frac{2}{9})}{3}\gamma_\kappa(1+\delta_\Sigma)^{-1}\delta_\Sigma \leq \ln\left(\frac{1+\kappa\delta_\Sigma}{\sqrt{1+\delta_\Sigma}\sqrt{1+\frac{1}{2}\delta_\Sigma}}\right).$$

Now let

$$f_1(\delta_\Sigma) = \ln\left(\frac{1+\kappa\delta_\Sigma}{\sqrt{1+\delta_\Sigma}\sqrt{1+\frac{1}{2}\delta_\Sigma}}\right) = \ln(1+\kappa\delta_\Sigma) - \frac{1}{2}\ln(1+\delta_\Sigma) - \frac{1}{2}\ln(1+\frac{1}{2}\delta_\Sigma)$$

$$f_1'(\delta_\Sigma) = \frac{\kappa}{1 + \kappa\delta_\Sigma} - \frac{\frac{1}{2}}{1 + \delta_\Sigma} - \frac{\frac{1}{4}}{1 + \frac{1}{2}\delta_\Sigma} = \frac{\frac{3}{4}(\kappa - \frac{2}{3})\delta_\Sigma + \kappa - \frac{3}{4}}{(1 + \kappa\delta_\Sigma)(1 + \delta_\Sigma)(1 + \frac{1}{2}\delta_\Sigma)}$$

and thus provided $\kappa > \frac{3}{4}$, as $\delta_\Sigma \leq \frac{2}{9}$

$$f_1'(\delta_\Sigma) > \frac{\kappa - \frac{3}{4}}{(1 + \frac{2\kappa}{9})(1 + \frac{2}{9})(1 + \frac{1}{9})}$$

Finally, as $f_1(0) = 0$, one deduces

$$f_1(\delta_\Sigma) > \frac{\kappa - \frac{3}{4}}{(1 + \frac{2\kappa}{9})(1 + \frac{2}{9})(1 + \frac{1}{9})}\delta_\Sigma \geq \frac{2(1 + \frac{2}{9})}{3}\gamma_k\delta_\Sigma \geq \frac{2(1 + \frac{2}{9})}{3}\gamma_k(1 + \delta_\Sigma)^{-1}\delta_\Sigma$$

which implies thus

$$\frac{\Phi_{\mu,\Sigma}(x)}{(1 + \kappa\delta_\Sigma)^{p_E}\Phi_{\tilde{\mu},(1+\delta_\Sigma)\Sigma}(x)} \leq 1$$

or $\Phi_{\mu,\Sigma}(x) \leq t^+(x)$.

The second case is handled in the same way.

$$\begin{aligned} & \frac{(1 + \kappa\delta_\Sigma)^{-p_E}\Phi_{\tilde{\mu},(1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(x)}{\Phi_{\mu,\Sigma}(x)} \\ & \leq (1 + \kappa\delta_\Sigma)^{-p_E} \left(\sqrt{\frac{|\Sigma|}{|(1 + \delta_\Sigma)^{-1}\tilde{\Sigma}|}} \exp\left(\frac{1}{2}(\mu - \tilde{\mu})'(\Sigma - (1 + \delta_\Sigma)^{-1}\tilde{\Sigma})^{-1}(\mu - \tilde{\mu})\right) \right) \\ & \leq \frac{(1 + \delta_\Sigma)^{p_E/2}}{(1 + \kappa\delta_\Sigma)^{p_E}} \exp\left(\frac{1}{2}(\mu - \tilde{\mu})'(\Sigma - (1 + \delta_\Sigma)^{-1}\tilde{\Sigma})^{-1}(\mu - \tilde{\mu})\right) \end{aligned}$$

Now as

$$\begin{aligned} (\Sigma - (1 + \delta_\Sigma)^{-1}\tilde{\Sigma})^{-1} &= (\Sigma((1 + \delta_\Sigma)\tilde{\Sigma}^{-1} - \Sigma^{-1})(1 + \delta_\Sigma)^{-1}\tilde{\Sigma})^{-1} \\ &= (1 + \delta_\Sigma)\tilde{\Sigma}^{-1}((1 + \delta_\Sigma)\tilde{\Sigma}^{-1} - \Sigma^{-1})^{-1}\Sigma^{-1} \end{aligned}$$

and thus

$$\begin{aligned} (\mu - \tilde{\mu})'(\Sigma - (1 + \delta_\Sigma)^{-1}\tilde{\Sigma})^{-1}(\mu - \tilde{\mu}) &\leq (1 + \delta_\Sigma)\tilde{L}^{-1}\lambda_m^{-1}4\tilde{L}\lambda_M\delta_\Sigma^{-1}L_m^{-1}\lambda_m^{-1}\|\mu - \tilde{\mu}\|^2 \\ &\leq (1 + \delta_\Sigma)L_m^{-1}\lambda_m^{-1}4\frac{\lambda_M}{\lambda_m}\delta_\Sigma^{-1}p_E\gamma_\kappa L_m\lambda_m\frac{\lambda_m}{\lambda_M}\delta_\Sigma^2 \\ &\leq 4p_E\gamma_\kappa(1 + \delta_\Sigma)\delta_\Sigma \end{aligned}$$

one deduces

$$\begin{aligned} \frac{\Phi_{\mu,\Sigma}}{(1 + \kappa\delta_\Sigma)^{p_E}\Phi_{\tilde{\mu},(1+\delta_\Sigma)\tilde{\Sigma}}} &\leq \frac{(1 + \delta_\Sigma)^{p_E/2}}{(1 + \kappa\delta_\Sigma)^{p_E}} \exp\left(\frac{1}{2}4p_E\gamma_\kappa(1 + \delta_\Sigma)\delta_\Sigma\right) \\ &\leq \left(\frac{\sqrt{1 + \delta_\Sigma}}{1 + \kappa\delta_\Sigma} \exp(2\gamma_\kappa(1 + \delta_\Sigma)\delta_\Sigma)\right)^{p_E}. \end{aligned}$$

All we need to prove is thus

$$\frac{\sqrt{1+\delta_\Sigma}}{1+\kappa\delta_\Sigma} \exp(2\gamma_\kappa(1+\delta_\Sigma)\delta_\Sigma) \leq 1$$

or equivalently

$$2\gamma_\kappa(1+\delta_\Sigma)\delta_\Sigma \leq \ln\left(\frac{1+\kappa\delta_\Sigma}{\sqrt{1+\delta_\Sigma}}\right).$$

Let

$$\begin{aligned} f_2(\delta_\Sigma) &= \ln\left(\frac{1+\kappa\delta_\Sigma}{\sqrt{1+\delta_\Sigma}}\right) = \ln(1+\kappa\delta_\Sigma) - \frac{1}{2}\ln(1+\delta_\Sigma) \\ f_2'(\delta_\Sigma) &= \frac{\kappa}{1+\kappa\delta_\Sigma} - \frac{\frac{1}{2}}{1+\delta_\Sigma} = \frac{\frac{\kappa}{2}\delta_\Sigma + \kappa - \frac{1}{2}}{(1+\kappa\delta_\Sigma)(1+\delta_\Sigma)} \end{aligned}$$

and thus provided $\kappa > \frac{1}{2}$, as $\delta_\Sigma \leq \frac{2}{9}$

$$f_2'(\delta_\Sigma) > \frac{\kappa - \frac{1}{2}}{(1 + \frac{2\kappa}{9})(1 + \frac{2}{9})}$$

Finally, as $f_2(0) = 0$, one deduces

$$f_2(\delta_\Sigma) > \frac{\kappa - \frac{1}{2}}{(1 + \frac{2\kappa}{9})(1 + \frac{2}{9})}\delta_\Sigma \geq 2\gamma_\kappa(1 + \frac{2}{9})\delta_\Sigma \geq 2\gamma_\kappa(1 + \delta_\Sigma)\delta_\Sigma$$

which implies

$$\frac{(1 + \kappa\delta_\Sigma)^{-p_E} \Phi_{\tilde{\mu}, (1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(x)}{\Phi_{\mu, \Sigma}(x)} \leq 1$$

or equivalently $t^-(x) \leq \Phi_{\mu, \Sigma}(x)$. □

Proof of Lemma 16 Entropy of Gaussian families (Lemma). A straightforward computation yields

$$\delta_\Sigma \leq \frac{1}{9\beta_\kappa} \frac{\delta}{p_E} \leq \frac{1}{p_E} \frac{\sqrt{2}}{9\sqrt{\frac{1}{4} + \frac{1}{4}}} = \frac{1}{p_E} \frac{2}{9} \leq \frac{2}{9}$$

□

Proof of Lemma 17 Entropy of Gaussian families (Lemma).

$$\begin{aligned} 2 - 2 \cdot 2^{d/2} \left((1 + \delta_\Sigma) + (1 + \delta_\Sigma)^{-1} \right)^{-d/2} &= 2 \left(1 - \left(\frac{e^{\ln(1+\delta_\Sigma)} + e^{-\ln(1+\delta_\Sigma)}}{2} \right)^{-d/2} \right) \\ &= 2 \left(1 - (\cosh(\ln(1 + \delta_\Sigma)))^{-d/2} \right) \\ &= 2f(\ln(1 + \delta_\Sigma)) \end{aligned}$$

where $f(x) = 1 - \cosh(x)^{-d/2}$. Studying this function yields

$$\begin{aligned} f'(x) &= -\frac{d}{2} \sinh(x) \cosh(x)^{-d/2-1} \\ f''(x) &= -\frac{d}{2} \cosh(x)^{-d/2} + \frac{d}{2} \left(\frac{d}{2} + 1 \right) \sinh(x)^2 \cosh(x)^{-d/2-2} \\ &= \frac{d}{2} \left(\left(\frac{d}{2} + 1 \right) \left(\frac{\sinh(x)}{\cosh(x)} \right)^2 - 1 \right) \cosh(x)^{-d/2} \end{aligned}$$

for any $x \geq 0$, as $\sinh(x) \leq \cosh(x)$ and $\cosh(x) \geq 1$, we have thus

$$f''(x) \leq \left(\frac{d}{2} \right)^2.$$

Now as $f(0) = 0$ and $f'(0) = 0$, this implies for any $x \geq 0$

$$f(x) \leq \frac{1}{4} d^2 \frac{x^2}{2}.$$

We deduce thus that

$$2 - 2 \cdot 2^{d/2} \left((1 + \delta_\Sigma) + (1 + \delta_\Sigma)^{-1} \right)^{-d/2} \leq \frac{1}{4} d^2 (\ln(1 + \delta_\Sigma))^2$$

and using $\ln(1 + \delta_\Sigma) \leq \delta_\Sigma$

$$2 - 2 \cdot 2^{d/2} \left((1 + \delta_\Sigma) + (1 + \delta_\Sigma)^{-1} \right)^{-d/2} \leq \frac{1}{4} d^2 \delta_\Sigma^2.$$

Now,

$$(1 + \kappa \delta_\Sigma)^d + (1 + \kappa \delta_\Sigma)^{-d} - 2 = 2(\cosh(d \ln(1 + \kappa \delta_\Sigma)) - 1) = 2g(d \ln(1 + \kappa \delta_\Sigma))$$

with $g(x) = \cosh(x) - 1$. Studying this function yields

$$g'(x) = \sinh(x) \quad \text{and} \quad g''(x) = \cosh(x)$$

and thus, as $g(0) = 0$ and $g'(0) = 0$, for any $0 \leq x \leq c$

$$g(x) \leq \cosh(c) \frac{x^2}{2}.$$

As $\ln(1 + \kappa \delta_\Sigma) \leq \kappa \delta_\Sigma$, $d \delta_\Sigma \leq c$ implies $d \ln(1 + \kappa \delta_\Sigma) \leq \kappa c$, we obtain thus

$$(1 + \kappa \delta_\Sigma)^d + (1 + \kappa \delta_\Sigma)^{-d} - 2 \leq \cosh(\kappa c) d^2 (\ln(1 + \kappa \delta_\Sigma))^2 \leq \kappa^2 \cosh(\kappa c) d^2 \delta_\Sigma^2.$$

□

Proof of Lemma 18 Entropy of Gaussian families (Lemma). We deduce this result from the slightly more general:

Lemma 19. Let $\delta_\Sigma \leq \frac{2}{9}$.

Let $(L, A, D) \in [L_m, L_M] \times \mathcal{A}(\lambda_m, \lambda_M) \times SO(p_E)$ and $(\tilde{L}, \tilde{A}, \tilde{D}) \in [L_m, L_M] \times \mathcal{A}(\lambda_m, +\infty) \times SO(p_E)$, define $\Sigma = LDAD'$ and $\tilde{\Sigma} = \tilde{L}\tilde{D}\tilde{A}\tilde{D}'$.

If

$$\begin{cases} (1 + \delta_L)^{-1} \tilde{L} \leq L \leq \tilde{L} \\ \forall 1 \leq i \leq p_E, \quad |A_{i,i}^{-1} - \tilde{A}_{i,i}^{-1}| \leq \delta_A \lambda_m^{-1} \\ \forall x \in \mathbb{R}^{p_E}, \quad \|Dx - \tilde{D}x\| \leq \delta_D \delta_\Sigma \|x\| \end{cases}$$

then $(1 + \delta_\Sigma) \tilde{\Sigma}^{-1} - \Sigma^{-1}$ and $\Sigma^{-1} - (1 + \delta_\Sigma) \tilde{\Sigma}^{-1}$ satisfies

$$\begin{aligned} \forall x \in \mathbb{R}^{p_E}, x' ((1 + \delta_\Sigma) \tilde{\Sigma}^{-1} - \Sigma^{-1}) x &\geq \tilde{L}^{-1} \left((\delta_\Sigma - \delta_L) \lambda_M^{-1} - (1 + \delta_\Sigma) \lambda_m^{-1} (\sqrt{2} \delta_D + \delta_A) \right) \|x\|^2 \\ \forall x \in \mathbb{R}^{p_E}, x' (\Sigma^{-1} - (1 + \delta_\Sigma) \tilde{\Sigma}^{-1}) x &\geq \frac{\tilde{L}^{-1}}{1 + \delta_\Sigma} \left(\delta_\Sigma \lambda_M^{-1} - \lambda_m^{-1} (\sqrt{2} \delta_D + \delta_A) \right) \|x\|^2 \end{aligned}$$

Indeed Lemma 16 Entropy of Gaussian families (Lemma) ensures that $\delta_\Sigma \leq \frac{2}{9}$ and if we let $\delta_L = \frac{1}{2} \delta_\Sigma$ and $\delta_A = \delta_D = \frac{1}{4(1 + \frac{2}{9})(\sqrt{2} + 1)} \frac{\lambda_m}{\lambda_M} \delta_\Sigma$, the bounds of the previous Lemma becomes $\forall x \in \mathbb{R}^{p_E}$,

$$\begin{aligned} x' ((1 + \delta_\Sigma) \tilde{\Sigma}^{-1} - \Sigma^{-1}) x &\geq \tilde{L}^{-1} \left((\delta_\Sigma - \delta_L) \lambda_M^{-1} - (1 + \delta_\Sigma) \lambda_m^{-1} (\sqrt{2} \delta_D + \delta_A) \right) \|x\|^2 \\ &\geq \tilde{L}^{-1} \left(\left(\delta_\Sigma - \frac{1}{2} \delta_\Sigma \right) \lambda_M^{-1} - (1 + \delta_\Sigma) \lambda_m^{-1} (\sqrt{2} + 1) \frac{1}{4(1 + \frac{2}{9})(\sqrt{2} + 1)} \frac{\lambda_m}{\lambda_M} \delta_\Sigma \right) \|x\|^2 \\ &\geq \frac{1}{4} \tilde{L}^{-1} \frac{1}{\lambda_M} \delta_\Sigma \|x\|^2 \end{aligned}$$

while $\forall x \in \mathbb{R}^{p_E}$,

$$\begin{aligned} x' (\Sigma^{-1} - (1 + \delta_\Sigma) \tilde{\Sigma}^{-1}) x &\geq \frac{\tilde{L}^{-1}}{1 + \delta_\Sigma} \left(\delta_\Sigma \lambda_M^{-1} - \lambda_m^{-1} (\sqrt{2} \delta_D + \delta_A) \right) \|x\|^2 \\ &\geq \frac{\tilde{L}^{-1}}{1 + \delta_\Sigma} \left(\delta_\Sigma \lambda_M^{-1} - \lambda_m^{-1} (\sqrt{2} + 1) \frac{1}{4(1 + \frac{2}{9})(\sqrt{2} + 1)} \frac{\lambda_m}{\lambda_M} \delta_\Sigma \right) \|x\|^2 \\ &\geq \frac{3}{4(1 + \frac{2}{9})} \tilde{L}^{-1} \frac{1}{\lambda_M} \delta_\Sigma \|x\|^2. \end{aligned}$$

□

Proof of Lemma 19 Entropy of Gaussian families (Lemma). By definition,

$$\begin{aligned} x' ((1 + \delta_\Sigma) \tilde{\Sigma}^{-1} - \Sigma^{-1}) x &= (1 + \delta_\Sigma) \tilde{L}^{-1} \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |\tilde{D}'_i x|^2 - L^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 \\ &= (1 + \delta_\Sigma) \tilde{L}^{-1} \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |\tilde{D}'_i x|^2 - (1 + \delta_\Sigma) \tilde{L}^{-1} \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |D'_i x|^2 \\ &\quad + (1 + \delta_\Sigma) \tilde{L}^{-1} \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |D'_i x|^2 - (1 + \delta_\Sigma) \tilde{L}^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 \end{aligned}$$

$$+ (1 + \delta_\Sigma) \tilde{L}^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 - L^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2$$

Along the same lines,

$$\begin{aligned} x' (\Sigma^{-1} - (1 + \delta_\Sigma)^{-1} \tilde{\Sigma}^{-1}) x &= L^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 - (1 + \delta_\Sigma)^{-1} \tilde{L}^{-1} \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |\tilde{D}'_i x|^2 \\ &= L^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 - (1 + \delta_\Sigma)^{-1} \tilde{L}^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 \\ &\quad + (1 + \delta_\Sigma)^{-1} \tilde{L}^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 - (1 + \delta_\Sigma)^{-1} \tilde{L}^{-1} \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |\tilde{D}'_i x|^2 \\ &\quad + (1 + \delta_\Sigma)^{-1} \tilde{L}^{-1} \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |\tilde{D}'_i x|^2 - (1 + \delta_\Sigma)^{-1} \tilde{L}^{-1} \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |D'_{j_D, i} x|^2 \end{aligned}$$

Now

$$\begin{aligned} \left| \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |\tilde{D}'_i x|^2 - \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |D'_i x|^2 \right| &\leq \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} \left| |\tilde{D}'_i x|^2 - |D'_i x|^2 \right| \\ &\leq \lambda_m^{-1} \sum_{i=1}^{p_E} \left| |\tilde{D}'_i x|^2 - |D'_i x|^2 \right| \\ &\leq \lambda_m^{-1} \sum_{i=1}^{p_E} \left| |\tilde{D}'_i x| - |D'_i x| \right| \left(|\tilde{D}'_i x| + |D'_i x| \right) \\ &\leq \lambda_m^{-1} \left(\sum_{i=1}^{p_E} |(\tilde{D}_i - D_i)' x|^2 \right)^{1/2} \left(\sum_{i=1}^{p_E} |(\tilde{D}_i + D_i)' x|^2 \right)^{1/2} \\ &\leq \lambda_m^{-1} \delta_D \|x\| \sqrt{2} \|x\| = \lambda_m^{-1} \sqrt{2} \delta_D \|x\|^2. \end{aligned}$$

Furthermore,

$$\begin{aligned} \left| \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |D'_i x|^2 - \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 \right| &\leq \sum_{i=1}^{p_E} |\tilde{A}_{i,i}^{-1} - A_{i,i}^{-1}| |D'_i x|^2 \\ &\leq \delta_A \lambda_m^{-1} \sum_{i=1}^{p_E} |D'_i x|^2 = \delta_A \lambda_m^{-1} \|x\|^2. \end{aligned}$$

We notice then that

$$\begin{aligned} (1 + \delta_\Sigma) \tilde{L}^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 - L^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 &= ((1 + \delta_\Sigma) \tilde{L}^{-1} - L^{-1}) \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 \\ &\geq (\delta_\Sigma - \delta_L) \tilde{L}^{-1} \lambda_M^{-1} \|x\|^2 \end{aligned}$$

while

$$L^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 - (1 + \delta_\Sigma)^{-1} \tilde{L}^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2 = (L^{-1} - (1 + \delta_\Sigma)^{-1} \tilde{L}^{-1}) \sum_{i=1}^{p_E} A_{i,i}^{-1} |D'_i x|^2$$

$$\begin{aligned}
&\geq (1 - (1 + \delta_\Sigma)^{-1}) L_{j,L}^{-1} \lambda_M^{-1} \|x\|^2 \\
&\geq \frac{\delta_\Sigma}{1 + \delta_\Sigma} \lambda_M^{-1} \|x\|^2
\end{aligned}$$

We deduce thus that

$$\begin{aligned}
x' ((1 + \delta_\Sigma) \tilde{\Sigma}^{-1} - \Sigma^{-1}) x &\geq (\delta_\Sigma - \delta_L) \tilde{L}^{-1} \lambda_M^{-1} \|x\|^2 - (1 + \delta_\Sigma) \tilde{L}^{-1} \lambda_m^{-1} (\sqrt{2} \delta_D + 2\delta_A) \|x\|^2 \\
&\geq \tilde{L}^{-1} \left((\delta_\Sigma - \delta_L) \lambda_M^{-1} - (1 + \delta_\Sigma) \lambda_m^{-1} (\sqrt{2} \delta_D + \delta_A) \right) \|x\|^2
\end{aligned}$$

and

$$\begin{aligned}
x' (\Sigma^{-1} - (1 + \delta_\Sigma)^{-1} \tilde{\Sigma}^{-1}) x &\geq \frac{\delta_\Sigma}{1 + \delta_\Sigma} \lambda_M^{-1} \|x\|^2 - (1 + \delta_\Sigma)^{-1} \tilde{L}^{-1} \lambda_m^{-1} (\sqrt{2} \delta_D + \delta_A) \|x\|^2 \\
&\geq \frac{\tilde{L}^{-1}}{1 + \delta_\Sigma} \left(\delta_\Sigma \lambda_M^{-1} - \lambda_m^{-1} (\sqrt{2} \delta_D + \delta_A) \right) \|x\|^2
\end{aligned}$$

□

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. *J. Comput. Graph. Statist.*, 18(1):216–237, 2008.
- [4] A. Barron, C. Huang, J. Li, and X. Luo. *MDL Principle, Penalized Likelihood, and Statistical Risk*, chapter in Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday. Tampere University Press, 2008.
- [5] D. Bashtannyk and R. Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3):279 – 298, 2001.
- [6] L. Bertrand, M.-A. Languille, S. Cohen, L. Robinet, C. Gervais, S. Leroy, D. Bernard, E. Le Pennec, W. Josse, J. Doucet, and S. Schöder. European research platform IPANEMA at the SOLEIL synchrotron for ancient and historical materials. *Journal of Synchrotron Radiation*, 18(5), September 2011.
- [7] Ch. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. Model-based cluster and discriminant analysis with the MIXMOD software. *Comput. Statist. Data Anal.*, 51(2):587–600, 2006.
- [8] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.
- [9] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73, 2007.

- [10] G. Blanchard, C. Schäfer, Y. Rozenholc, and K.R. Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2):209–241, 2007.
- [11] S. Boucheron and P. Massart. A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, pages 1–29, 2010.
- [12] E. Brunel, F. Comte, and C. Lacour. Adaptive estimation of the conditional density in presence of censoring. *Sankhyā*, 69(4):734–763, 2007.
- [13] S. Cohen and E. Le Pennec. Conditional density estimation by penalized likelihood model selection and applications. ArXiv 1103.2021, 2011.
- [14] J. de Gooijer and D. Zerom. On conditional density estimation. *Statist. Neerlandica*, 57(2):159–176, 2003.
- [15] D. Donoho. CART and best-ortho-basis: a connection. *Ann. Statist.*, 25(5):1870–1911, 1997.
- [16] S. Efromovich. Conditional density estimation in a regression setting. *Ann. Statist.*, 35(6):2504–2535, 2007.
- [17] S. Efromovich. Oracle inequality for conditional density estimation and an actuarial example. *Annals of the Institute of Statistical Mathematics*, 62:249–275, 2010.
- [18] J. Fan and T. Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.
- [19] J. Fan, Q. Yao, and H. Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996.
- [20] J. Fan, C. Zhang, and J. Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *Ann. Stat.*, 29(1):153–193, 2001.
- [21] Ch. Genovese and L. Wasserman. Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.*, 28(4):1105–1127, 2000.
- [22] L. Györfi and M. Kohler. Nonparametric estimation of conditional distributions. *IEEE Trans. Information Theory*, 53:1872–1879, 2007.
- [23] P. Hall, R. Wolff, and Q. Yao. Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.*, 94:154–163, 1999.
- [24] P. Hall, J. Racine, and Q. Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.
- [25] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence*, 1999.
- [26] Y. Huang, I. Pollak, M. Do, and C. Bouman. Fast search for best representations in multitree dictionaries. *IEEE Transactions on Image Processing*, 15(7):1779–1793, 07 2006.
- [27] R. Hyndman and Q. Yao. Nonparametric estimation and symmetry tests for conditional density functions. *Journal of nonparametric statistics*, 14(3):259–278, 2002.
- [28] R. Hyndman, D. Bashtannyk, and G. Grunwald. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5:315–336, 1996.

- [29] B. Karaivanov and P. Petrushev. Nonlinear piecewise polynomial approximation beyond besov spaces. *Applied and Computational Harmonic Analysis*, 15(3):177 – 223, 2003.
- [30] E. Kolaczyk and R. Nowak. Multiscale generalised linear models for nonparametric function estimation. *Biometrika*, 92(1):119–133, 2005.
- [31] E. Kolaczyk, J. Ju, and S. Gopal. Multiscale, multigranular statistical image segmentation. *J. Amer. Statist. Assoc.*, 100(472):1358–1369, 2005.
- [32] Q. Li and J. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, 2007.
- [33] J. Lin. Divergence measures based on the Shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145 –151, 01 1991.
- [34] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
- [35] C. Maugis and B. Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: P & S*, 2010. To appear.
- [36] C. Maugis and B. Michel. Erratum on "a non asymptotic penalized criterion for Gaussian mixture model selection". Available on their webpage, 2010.
- [37] C. Maugis and B. Michel. Data-driven penalty calibration: a case study for Gaussian mixture model selection. *ESAIM: P & S*, 2011. To appear.
- [38] M. Rosenblatt. Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pages 25–31. Academic Press, 1969.
- [39] 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*, pages 218–252, 2005.
- [40] Ch. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, 22(1):118–171, 1994.
- [41] S. Szarek. Metric entropy of homogeneous spaces, (gdansk, 1997), 395–410, banach center publ. 43, polish acad. sci., warsaw, 1998. *Quantum Probability*, pages 395–410, 1998.
- [42] S. van de Geer. The method of sieves and minimum contrast estimators. *Math. Methods Statist.*, 4:20–38, 1995.
- [43] A. van der Vaart and J. Wellner. *Weak Convergence*. Springer, 1996.
- [44] I. van Keilegom and N. Veraverbeke. Density and hazard estimation in censored regression models. *Bernoulli*, 8(5):607–625, 2002.
- [45] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1): 1–25, 1992.
- [46] S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9:60–62, 1938.

- [47] R. Willet and R. Nowak. Multiscale poisson intensity and density estimation. *IEEE Transactions on Information Theory*, 53(9):3171–3187, 2007.
- [48] D. Young and D. Hunter. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, 54(10):2253–2266, 2010.