Conditional Density Estimation by Penalized Likelihood Model Selection and Applications

Serge X. Cohen, Erwan Le Pennec

Conditional Density Estimation by Penalized Likelihood Model Selection

S. Cohen (IPANEMA/Soleil) and E. Le Pennec (SELECT/INRIA Saclay)

August 8, 2011

Abstract

In this paper, we consider conditional density estimation, and propose a general condition on the penalty of a penalized maximum likelihood estimate to obtain oracle type inequality with Kullback-Leibler type loss. Our aim is threefold: to extend a model selection theorem obtained by Massart for density estimation, to illustrate this theorem with families of *piecewise constant* conditional density estimator, and to provide some theoretical justification for a companion paper on unsupervised segmentation based on spatially varying Gaussian mixture estimation.

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_is are independent but not necessarily identically distributed. The assumption on the Y_is are stronger: we assume that, conditionally to the X_is, they are independents and each variable Y_i follows a law with density $s_0(\cdot|X_i)$ with respect to a common known measure $\mathrm{d}\lambda$. 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 [31] 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 [32] and a bias correction of Rosenblatt's estimator due to Hyndman et al. [23].

Kernel based method have been much studied since. For instance, Fan et al. [17] and de Gooijer and Zerom [13] consider local polynomial estimator, Hall et al. [20] study a locally logistic estimator that is later extended by Hyndman and Yao [22]. 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 exception

of Bashtannyk and Hyndman [6]. Extensions to censored cases have also been discussed for instance by van Keilegom and Veraverbeke [35].

In the approach of Stone [32], the conditional density is estimated through a parametrized modelization. This idea has been reused since by Györfi and Kohler [19] with an histogram based approach, by Efromovich [15, 16] with a Fourier basis, and by Brunel et al. [11] and Akakpo and Lacour [3] 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. [11] extend their result to censored cases while Akakpo and Lacour [3] are able to consider weakly dependent data.

In this paper, we use a model selection approach to propose a penalized maximum likelihood estimate of s_0, which has strangely enough only been considered by Stone [32] as mentioned before and by Blanchard et al. [9] in a classification setting with histogram type estimators; we derive results with a Kullback-Leibler type loss. As usual, we assume we are given a collection of models $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$, a collection of sets of candidate functions, and define for any of these models the maximum likelihood estimates

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

For a given penalty $\mathrm{pen}(m)$, the *best* model $S_{\widehat{m}}$ is the one defined by

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

The main result of this paper is a sufficient condition on the penalty $\mathrm{pen}(m)$ such that an oracle type inequality holds for the conditional density estimation error.

This theorem has been motivated by an application of Gaussian mixture modeling to unsupervised segmentation. This application as well as its implementation is the subject of a companion paper [12]. In this article, we describe the corresponding conditional density estimation setting as well as another example of *piecewise constant* conditional density estimation. For both examples, we show that the penalty can be chosen roughly proportional to the dimension of the model.

Section 2Penalized maximum likelihood for conditional density model selection begins with the description of the statistical model and of the divergence used. The main theorem, the one that deals with the choice of the penalty, is given in the same section. Its main assumption is linked to the bracketing entropy of the models, a link that we make explicit through two general lemmas that concludes this section. Sections 3.2Piecewise polynomial conditional densities estimation and 3.3Spatial Gaussian mixtures, models, bracketing entropy and penalties is devoted to exemplifications of this theorem for two *piecewise constant* conditional density estimators. For both an histogram type estimate (as well as some piecewise polynomial extension) and a spatial Gaussian mixture estimate, we make explicit the choice of the penalty $\mathrm{pen}(m)$ essentially through a fine control on the corresponding bracketing entropy.

2 Penalized maximum likelihood for conditional density model selection

2.1 Setting and maximum likelihood penalized estimate

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_is 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 $\mathrm{d}\lambda$. For any model S_m of candidate conditional densities, the maximum likelihood approach suggest to estimate s_0 by the conditional density $\widehat{s}_m$ 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:

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

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

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

In the model selection framework, instead of a single model S_m, a collection of models $\mathcal{S} = \{S_m\}_{m\in\mathcal{M}}$ is considered and the final estimates is chosen amongst the collection of estimates $\{\widehat{s}_m\}_{m\in\mathcal{M}}$ according to a selection rule. We consider here penalization based selection in which for each model a penalty $\mathrm{pen}(m)$ is chosen and the *best* model $S_{\widehat{m}}$ is chosen as the one whose index is an almost minimizer of the penalized η -log-likelihood :

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

Our goal is to give a condition on $\mathrm{pen}(m)$ that ensure that $\widehat{s}_{\widehat{m}}$ is a good estimate of s_0. We should first specify the meaning of *good* in the previous sentence.

Ideally, as we are working in a maximum likelihood approach, our result should be stated in term of the Kullback-Leibler divergence KL. As we consider that the reference measure λ is common and known, we can write

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

However, as most of the time in density estimation, our result will also involved a *smaller* divergence. This smaller divergence is often chose as the squared Hellinger distance $d^2(s\mathrm{d}\lambda, t\mathrm{d}\lambda) = d^2_\lambda(s,t)$. Here, we use an intermediate divergence: the Jensen-Kullback-Leibler divergence JKL_ρ with $\rho\in(0,1)$ defined by

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

Note that this divergence appears explicitly with $\rho=\frac{1}{2}$ in Massart [28], but can also be found implicitly in Birgé and Massart [8] and van de Geer [34]. We use the name Jensen-Kullback-Leibler divergence in the same way Lin [27] use the name Jensen-Shannon divergence for a sibling in its information theory work. All those divergences are related:

Proposition 1. *For any probability measures $s\mathrm{d}\lambda$ and $t\mathrm{d}\lambda$ 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 $C_\rho = \dfrac{1}{\rho}\min\left(\dfrac{1-\rho}{\rho}, 1\right)\left(\ln\left(1+\dfrac{\rho}{1-\rho}\right)-\rho\right)$.

Furthermore, if $s\mathrm{d}\lambda \ll t\mathrm{d}\lambda$ then $d_\lambda^2(s,t) \leq KL_\lambda(s,t) \leq \left(2+\ln\left\|\dfrac{s}{t}\right\|_\infty\right) d_\lambda^2(s,t)$.

As we are working with conditional densities and not with classical densities, the previous divergences should be adapted. We defined thus the following *tensorized* divergences that take into account the structure of conditional densities and the design of $(X_i)_{1\leq i\leq n}$:

$$KL_\lambda^{\otimes_n}(s,t) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n KL_\lambda(s(\cdot|X_i), t(\cdot|X_i))\right], \qquad d_\lambda^{2\otimes_n}(s,t) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n d_\lambda^2(s(\cdot|X_i), t(\cdot|X_i))\right]$$

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

Those divergences appear as the natural ones in this setting. Furthermore, they reduce to classical ones in specific settings:

- If the law of Y_is are independent of the X_is, that is $s(\cdot|X_i) = s(\cdot)$ and $t(\cdot|X_i) = t(\cdot)$ do not depend on X_i, these divergences reduce to respectively the classical $KL_\lambda(s,t)$, $d_{\lambda_0}^2(s,t)$ and $JKL_\lambda(s,t)$.
- If the X_is are not random but fixed, that is we consider a fixed design case, theses divergences are the classical fixed design type divergence in which there is no expectation.
- If the X_is are i.i.d., theses divergences can be rewritten as the simple ones: $KL_\lambda^{\otimes_n}(s,t) = \mathbb{E}\left[KL_\lambda(s(\cdot|X_i), t(\cdot|X_i))\right]$,

  $d_\lambda^{2\otimes_n}(s,t) = \mathbb{E}\left[d_\lambda^2(s(\cdot|X_i), t(\cdot|X_i))\right]$ and $JKL_{\rho,\lambda}^{\otimes_n}(s,t) = \mathbb{E}\left[JKL_{\rho,\lambda}(s(\cdot|X_i), t(\cdot|X_i))\right]$.

We are now almost ready to give a condition on pen(m) so that the following type of oracle inequality holds:

$$\mathbb{E}\left[JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{S_m\in\mathcal{S}}\left(\inf_{s_m\in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n}\right) + C_2\frac{1}{n}$$

with $C_1 > 1$ and C_2 some absolute constants.

2.2 A general theorem for penalized maximum likelihood conditional density estimation

The condition on pen(m) involves a bracketing entropy condition on the models 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.

The main assumption of our theorem involves the bracketing entropies not of the global models S_m but the ones of smaller localized sets $S_m(\widetilde{s}, \sigma) = \{s_m \in S_m | d^{\otimes_n}(\widetilde{s}, s_m) \leq \sigma\}$:

Assumption (H). *For every model S_m in the collection $\mathcal{S}$, 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 $s_m \in S_m$*

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

We will also assume the existence of a Kraft type inequality for the collection:

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

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

We further need a technical separability assumption:

Assumption (S). *For any model S_m in the collection $\mathcal{S}$, 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.*

Under these assumptions, we obtain:

Theorem 1. *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 Assumptions (H), (K) and (S) hold and let $\widehat{s}_m$ be a η -log-likelihood minimizer in S_m*

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

Then for any $\rho \in (0,1)$ and any $C_1 > 1$, there are two constants κ_0 and C_2 depending only on ρ and C_1 such that, as soon as for every index $m \in \mathcal{M}$

$$\mathrm{pen}(m) \geq \kappa\left(n\sigma_m^2 + x_m\right) \quad \text{with } \kappa > \kappa_0$$

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

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

satisfies

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

When considering *conditional* densities that do not depend on the covariates, this theorem reduces exactly to Theorem 7.11 of Massart [28] on density estimation. Furthermore, our proof is an adaptation of Massart's one. It should be noted that condition on the design of the X_is appears only implicitly in the divergence $d^{\otimes n}$ used in the bracketing entropy definition.

This oracle inequality may appear similar to the ones obtained by Barron et al. [5] and Kolaczyk et al. [26]; there are however some important differences. In those works, each models

considered contains a single density. In that case, the bracketing entropy is always 0 and so is $n\sigma_m^2$. Thus, only the model collection term x_m appears. The price to pay is a discretization of the parameter spaces that makes the corresponding estimator hardly implementable. They, nevertheless, obtain better constants ($\kappa = \frac{1}{2}$ as soon as $\Sigma = 1$, $C_1 = 1$ and $C_2 = 0$) but for a tensorized Bhattacharyya-Renyi divergence conditioned to the covariates $(X_i)_{1\leq i\leq n}$ instead of $JKL^{\otimes_n}$. Although the Bhattacharyya-Renyi divergence and Jensen-Kullback-Leibler one are not easily compared, one should stress that, as soon as the X_is are random, a integrated divergence is more meaningful than a conditioned one.

The terms $n\sigma_m^2$ and x_m appearing in the condition on the penalty

$$\text{pen}(m) \geq \kappa \left(n\sigma_m^2 + x_m\right) \quad \text{with } \kappa > \kappa_0$$

have quite different interpretation. The term $n\sigma_m^2$ is obtained by looking at the model S_m alone through a Dudley type integral of bracketing entropy of localized model. It can be seen an intrinsic measure of complexity. The terms x_m should be defined simultaneously for all models in the collection in order to satisfy a Kraft type inequality

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

This can be interpreted as a coding condition and thus we call x_ms coding terms. 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 $n\sigma_m^2$ is as small as possible.

The mere existence of an oracle type inequality such as the one of Theorem 1A general theorem for penalized maximum likelihood conditional density estimation does not bring much information; it only means that one can obtain a bound on the error of the estimator. A first question, as we use as family of maximum likelihood, is how this estimator compares to the best fixed one in this family? A second question is then if this best fixed one is good? Although the focus of this paper is not on these issues, we can make the following heuristic remarks. If we assume that a Wilks's phenomenon [36] holds as it is done, for instance, by Akaike [1] for AIC criterion or, better, if we are able to prove it, it is expected that the risk of a fixed model $\widehat{s}_m$ behaves asymptotically like $\inf_{s_m\in S_m} KL(s_0, s_m) + \lambda\frac{\dim(S_m)}{n}$ where $\dim(S_m)$ is the dimension of the model and $\lambda = \frac{1}{2}$ if we follow Akaike but may be larger as in Boucheron and Massart [10]. In that case, it suffices to show that the penalty $\text{pen}(m)$ can be chosen roughly proportional to this dimension to obtain that the penalized estimator has a risk of the same order than the one of the best fixed one. The performance of the best fixed model estimator depends heavily on the model collection, and thus no general result can be obtained.

2.3 Bracketing entropy, Dudley integral and complexity

Often, the main difficulty to apply Theorem 1A general theorem for penalized maximum likelihood conditional density estimation is to control the bracketing entropy term $n\sigma_m^2$. We provide here two general propositions that link such a bound to either the bracketing entropy of the local models $S_m(s_m, \sigma)$ appearing in Assumption (H) or the (larger) bracketing entropy of the larger global models S_m. As hinted in the previous paragraph, we introduce a parameter $\mathcal{D}_m$ that plays the role of a dimension. It will be either the intrinsic dimension of the model S_m or a slight upper bound of it.

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

Proposition 2. *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):

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

Furthermore the unique root σ_m of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$ satisfies

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

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

Proposition 3. *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):

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

Furthermore the unique root σ_m of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$ satisfies

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

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

3 *Piecewise constant* conditional densities

3.1 Covariate partitioning and conditional density estimation

In this section, we exemplify our main theorem for some *piecewise constant* conditional density estimate that are very much in the spirit of Donoho [14]. Following Kolaczyk et al. [26], we will only consider, as candidate estimates, conditional densities that can be written as

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

where $\mathcal{P}^{\mathcal{X}}$ is partition of $\mathcal{X}$, $\mathcal{R}_l^{\mathcal{X}}$ denotes a generic region in this partition, $\mathbf{1}$ denotes the characteristic function of a set and $s(y|\mathcal{R}_l^{\mathcal{X}})$ is a density for any $\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}$. We denote by $\|\mathcal{P}^{\mathcal{X}}\|$ the number of region in this partition.

To simplify the exposition, we will assume that $\mathcal{X}$ is $[0,1]^{d_X}$ and focus on five different hyperrectangle based collections of partitions:

- 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 $\{\mathcal{R}_l^{\mathcal{X}}\}_{1 \leq l \leq \|\mathcal{P}^{\mathcal{X}}\|}$ are thus hypercubes whose measure $|\mathcal{R}_l^{\mathcal{X}}|$ satisfies $|\mathcal{R}_l^{\mathcal{X}}| = 2^{-d_X J}$. We stop the recursion as soon as the number of step J satisfies $\frac{2^{d_X}}{n} \geq |\mathcal{R}_l^{\mathcal{X}}| \geq \frac{1}{n}$.
  - The recursive dyadic partition collection (RDP($\mathcal{X}$)) in which at each step an hypercube of measure $|\mathcal{R}_l^{\mathcal{X}}| \geq \frac{2^{d_X}}{n}$ 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 $|\mathcal{R}_l^{\mathcal{X}}| \geq \frac{2}{n}$ 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 $|\mathcal{R}_l^{\mathcal{X}}| \geq \frac{2}{n}$ 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 $\frac{1}{n}$.

We denote by $\mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}$ the corresponding partition collection where $\star(\mathcal{X})$ is either UDP($\mathcal{X}$), RDP($\mathcal{X}$), RDSP($\mathcal{X}$), RSP($\mathcal{X}$) or HRP($\mathcal{X}$). As shown by Kolaczyk and Nowak [25], Huang et al. [21] or Willet and Nowak [37], the first four partition collections, ($\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RDP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RDSP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RSP}(\mathcal{X})}$), have a tree structure that is crucial for an efficient implementation of the corresponding estimate. As, in contrast to our companion paper, we do not focus on this computational aspect in this article, we have also added the much more complex to deal with collection $\mathcal{S}_{\mathcal{P}}^{\mathrm{HRP}(\mathcal{X})}$.

One of the key property of these partition collections is the existence of Kraft type inequalities:

Proposition 4. *If we define*

	$\star = \mathrm{UDP}(\mathcal{X})$	$\star = \mathrm{RDP}(\mathcal{X})$	$\star = \mathrm{RDSP}(\mathcal{X})$	$\star = \mathrm{RSP}(\mathcal{X})$	$\star = \mathrm{HRP}(\mathcal{X})$
$A_0^\star$	$\ln\left(\max\left(2, 1+\frac{\ln n}{d_X \ln 2}\right)\right)$	0	0	0	0
$B_0^\star$	0	$\ln 2$	$\lceil \ln(1+d_X)\rceil_{\ln 2}$	$\lceil \ln(1+d_X)\rceil_{\ln 2} + \lceil \ln n\rceil_{\ln 2}$	$d_X \lceil \ln n\rceil_{\ln 2}$
$C_0^\star$	$\ln\left(\max\left(2, 1+\frac{\ln n}{d_X \ln 2}\right)\right)$	$\ln 2$	$\lceil \ln(1+d_X)\rceil_{\ln 2}$	$\lceil \ln(1+d_X)\rceil_{\ln 2} + \lceil \ln n\rceil_{\ln 2}$	$d_X \lceil \ln n\rceil_{\ln 2}$
$c_0^\star$	0	$\frac{2_X^d}{2_X^d - 1}$	2	2	1
Σ_0	$1+\frac{\ln n}{\dim X \ln 2}$	2	$2(1+d_X)$	$4(1+d_X)n$	$(2n)^{d_X}$

then, for any of the five described partition collections $\mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}$, for all $c \geq c_0^{\star(\mathcal{X})}$:

$$\sum_{\mathcal{P}^{\mathcal{X}} \in \mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}} e^{-c\left(A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})} \|\mathcal{P}^{\mathcal{X}}\|\right)} \leq \Sigma_0^{\star(\mathcal{X})} e^{-cC_0^{\star(\mathcal{X})}}.$$

where $\lceil x \rceil_{\ln 2}$ is the smallest multiple of $\ln 2$ larger than x.

In the next sections, we consider two different strategies for the densities $s(y|x \in \mathcal{R}_l^{\mathcal{X}})$ in each hyperrectangle $\mathcal{R}_l^{\mathcal{X}}$ of the partition: a piecewise polynomial strategy similar to the one proposed by Willet and Nowak [37] when $\mathcal{Y} = [0,1]^{d_Y}$ and a Gaussian mixture with common mixed densities strategy that extends the setting of Maugis and Michel [29]. This last example is the main subject of our companion paper in which it is used for an unsupervised segmentation task.

3.2 Piecewise polynomial conditional densities estimation

3.2.1 Piecewise polynomial conditional densities

In this section $\mathcal{X} = [0,1]^{d_X}$, $\mathcal{Y} = [0,1]^{d_Y}$ and λ is the Lebesgue measure $\mathrm{d}y$ which is a probability measure on $\mathcal{Y}$. We should now specify the possible choices for the candidate density $s(y|x \in \mathcal{R}_l^{\mathcal{X}})$: we reuse a hyperrectangle partitioning strategy this time for $\mathcal{Y} = [0,1]^{d_Y}$ and impose that our candidate density $s(y|x \in \mathcal{R}_l^{\mathcal{X}})$ is a square of polynomial on each hyperrectangles $\mathcal{R}_{l,k}^{\mathcal{Y}}$ of the partition $\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})$. This differs from the choice of Willet and Nowak [37] in which the candidate density is simply a polynomial. The two choices coincide however when the polynomial is chosen amongst 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. [9] in which $\mathcal{Y}$ is a finite discrete set.

More precisely, for any partition $\mathcal{P}^{\mathcal{X}} = \{\mathcal{R}_l^{\mathcal{X}}\}_{1 \leq l \leq \|\mathcal{P}^{\mathcal{X}}\|}$ of $\mathcal{X} = [0,1]^{d_X}$ and any collection of partitions $\mathcal{P}^{\mathcal{Y}} = (\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}}))_{1 \leq l \leq \|\mathcal{P}^{\mathcal{X}}\|} = \left(\{\mathcal{R}_{l,k}^{\mathcal{Y}}\}_{1 \leq k \leq \|\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})\|}\right)_{1 \leq l \leq \|\mathcal{P}^{\mathcal{X}}\|}$ of $\mathcal{Y} = [0,1]^{d_Y}$, we define the partition $\mathcal{P}^{\mathcal{X},\mathcal{Y}}$ of $\mathcal{X} \times \mathcal{Y}$ as

$$\left\{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}} = \mathcal{R}_l^{\mathcal{X}} \times \mathcal{R}_{l,k}^{\mathcal{Y}} \middle| \mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}, \mathcal{R}_{l,k}^{\mathcal{Y}} \in \mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})\right\}.$$

We let $|\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}| = |\mathcal{R}_l^{\mathcal{X}}||\mathcal{R}_{l,k}^{\mathcal{Y}}|$ be the measure of the product hyperrectangle $\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}$ and denote the total number of hyperrectangles of $\mathcal{P}^{\mathcal{X},\mathcal{Y}}$ by $\|\mathcal{P}^{\mathcal{X},\mathcal{Y}}\|$ which satisfies

$$\|\mathcal{P}^{\mathcal{X},\mathcal{Y}}\| = \sum_{\mathcal{R}^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \|\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})\|.$$

We let then $S_{\mathcal{P}^{\mathcal{X}},\mathcal{Y},D_Y^{\mathrm{M}}}$ be the set of conditional densities such that

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

where $P_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}}$ is a polynomial of degree at most $D_Y^{\mathrm{M}} = (D_{Y,1}^{\mathrm{M}}, \ldots, D_{Y,d_Y}^{\mathrm{M}})$. Note that as by definition of a density $\int_{[0,1]^{d_Y}} s(y|x)\mathrm{d}y = 1$, this imposes that for any $\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}$,

$$\sum_{\mathcal{R}_{l,k}^{\mathcal{Y}} \in \mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})} \int_{\mathcal{R}_{l,k}^{\mathcal{Y}}} P^2_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}}(y)\mathrm{d}y = 1.$$

By construction,

$$\dim(S_{\mathcal{P}^{\mathcal{X}},\mathcal{Y},D_Y^{\mathrm{M}}}) = \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \left(\|\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})\| \prod_{d=1}^{d_Y} (D_{Y,d}^{\mathrm{M}} + 1) - 1 \right).$$

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

$$\mathcal{D}_{\mathcal{P}^{\mathcal{X}},\mathcal{Y},D_Y^{\mathrm{M}}} = \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \|\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})\| \prod_{k=1}^{d_Y} (D_{Y,k}^{\mathrm{M}} + 1).$$

As shown by Willet and Nowak [37], the maximum likelihood estimate in this model can be obtained by an independent computation on each subset $\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}$:

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

This property will be important for the use of the efficient optimization algorithms of Willet and Nowak [37] and Huang et al. [21].

As described in the previous section, the partition $\mathcal{P}^{\mathcal{X}} = \{\mathcal{R}_l^{\mathcal{X}}\}_{1 \leq l \leq \|\mathcal{P}^{\mathcal{X}}\|}$ will be selected amongst either $\mathrm{UDP}(\mathcal{X})$, $\mathrm{RDP}(\mathcal{X})$, $\mathrm{RDSP}(\mathcal{X})$, $\mathrm{RSP}(\mathcal{X})$ or $\mathrm{HRP}(\mathcal{X})$ collections with respect to $[0,1]^{d_X}$ while all the partitions $\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})$ are selected amongst either $\mathrm{UDP}(\mathcal{Y})$, $\mathrm{RDP}(\mathcal{Y})$, $\mathrm{RDSP}(\mathcal{Y})$, $\mathrm{RSP}(\mathcal{Y})$ or $\mathrm{HRP}(\mathcal{Y})$ collections with respect to $[0,1]^{d_Y}$.

3.2.2 Conditional density estimation theorem

We are now ready to state a theorem that shows that a penalty roughly proportional to the dimension of the model is sufficient to control the estimation error without any assumption on the design.

Theorem 2. *Fix a collection $\star(\mathcal{X})$ amongst $\mathrm{UDP}(\mathcal{X})$, $\mathrm{RDP}(\mathcal{X})$, $\mathrm{RDSP}(\mathcal{X})$, $\mathrm{RSP}(\mathcal{X})$ or $\mathrm{HRP}(\mathcal{X})$ for $\mathcal{X} = [0,1]^{d_X}$, a collection $\star(\mathcal{Y})$ amongst $\mathrm{UDP}(\mathcal{Y})$, $\mathrm{RDP}(\mathcal{Y})$, $\mathrm{RDSP}(\mathcal{Y})$, $\mathrm{RSP}(\mathcal{Y})$ or $\mathrm{HRP}(\mathcal{Y})$ and a degree for the polynomial $D_Y^{\mathrm{M}} \in \mathbb{N}^{d_Y}$.*

Let

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

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 1A general theorem for penalized maximum likelihood conditional density estimation satisfies

$$\mathbb{E}\left[JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}}})\right] \leq C_1 \inf_{S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}} \in \mathcal{S}} \left(\inf_{s_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}} \in S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}}} KL^{\otimes_n}(s_0, s_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}}) + \frac{\mathrm{pen}(\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}})}{n} \right) + C_2 \frac{1}{n} + \frac{\eta + \eta'}{n}$$

as soon as

$$\mathrm{pen}(\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}) \geq \kappa \left(\left(C_\star + 2 \ln \frac{n}{\|\mathcal{P}^{\mathcal{X},\mathcal{Y}}\|} \right) \mathcal{D}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}} + c_\star \left(A_0^{\star(\mathcal{X})} + \left(B_0^{\star(\mathcal{X})} + A_0^{\star(\mathcal{Y})} \right) \|\mathcal{P}^{\mathcal{X}}\| + B_0^{\star(\mathcal{Y})} \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}} \|\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})\| \right) \right)$$

with $\kappa > \kappa_0$ where κ_0 and C_2 are the constants of Theorem 1A general theorem for penalized maximum likelihood conditional density estimation that depend only on ρ and C_1.

Theorem 2Conditional density estimation theorem is obtained by combining

Proposition 5. *Under the assumptions of Theorem 2Conditional density estimation theorem, it exists a $D_\star$ such that for any model $S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}}$, the function*

$$\phi_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}}(\sigma) = \sigma \sqrt{\mathcal{D}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}}} \left(\sqrt{\frac{1}{2} \ln \frac{n^2}{\|\mathcal{P}^{\mathcal{X},\mathcal{Y}}\|} + D_\star} + \sqrt{\pi} \right)$$

satisfies the required property of Assumption (H).

Furthermore, $\sigma_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}}$ the unique root of $\frac{1}{\sigma}\phi_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}}(\sigma) = \sqrt{n}\sigma$ satisfies

$$n\sigma_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}} \leq \left(C_\star + \ln \frac{n^2}{\|\mathcal{P}^{\mathcal{X},\mathcal{Y}}\|} \right) \mathcal{D}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}}$$

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

and

Proposition 6. *Under the assumptions of Theorem 2Conditional density estimation theorem, for any collection $\mathcal{S}$, it exists a $c_\star > 0$ such that for*

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

Assumption (K) is satisfied with $\sum_{S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}} \in \mathcal{S}} e^{-x_{\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}}} \leq 1.$

with Theorem 1A general theorem for penalized maximum likelihood conditional density estimation.

Note that as $\|\mathcal{P}^{\mathcal{X}}\| \leq \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \|\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})\| \leq \frac{\mathcal{D}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}}{\sum_{d=1}^{d_Y} D_{Y,d}^{\mathrm{M}}}$, the condition on the penalty in Theorem 2Conditional density estimation theorem is weaker than

$$\mathrm{pen}(\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}) \geq \kappa \left(C_\star + 2 \ln \frac{n}{\|\mathcal{P}^{\mathcal{X},\mathcal{Y}}\|} + c_\star \left(A_0^{\star(\mathcal{X})} + \frac{B_0^{\star(\mathcal{X})} + A_0^{\star(\mathcal{Y})} + B_0^{\star(\mathcal{Y})}}{\sum_{d=1}^{d_Y} D_{Y,d}^{\mathrm{M}}} \right) \right) \mathcal{D}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}.$$

The lower bound on the penalty is thus roughly proportional to a multiple of the dimension of the model, the multiplicative factor being constant over n up to a logarithmic factor.

Some variations around this Theorem can be obtained from the proof in Appendix. For example, if we assume that $\mathcal{P}^{\mathcal{X}}$ belongs to $\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$ and that $\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_L)$ is independent of $\mathcal{R}_l^{\mathcal{X}}$ and belongs to $\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$, the term $\ln \frac{n^2}{\sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \|\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})\|}$ can disappear. More precisely,

$$\mathbb{E}\left[JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}}) \right] \leq C_1 \inf_{S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} \in \mathcal{M}} \left(\inf_{s_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} \in S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}} KL^{\otimes_n}(s_0, s_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}) + \frac{\mathrm{pen}(\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}})}{n} \right)$$
$$+ C_2 \frac{\ln n}{n} + \frac{\eta + \eta'}{n}$$

as soon as

$$\mathrm{pen}(\mathcal{P}^{\mathcal{X},\mathcal{Y}}, D_Y^{\mathrm{M}}) \geq \kappa_\star \left(C_\star \mathcal{D}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} + c_\star \right).$$

Choosing the degrees D_Y^{M} of the polynomial amongst a family $\mathcal{D}_Y^{\mathrm{M}}$ either globally or locally as proposed by Willet and Nowak [37] is also possible. It suffices to modify the coding part in Proposition 6Conditional density estimation theorem accordingly: this can be achieve by replacing respectively $A_0^{\star(\mathcal{X})}$ by $A_0^{\star(\mathcal{X})} + \ln |\mathcal{D}_Y^{\mathrm{M}}|$ for the global optimization and $B_0^{\star(\mathcal{Y})}$ by $B_0^{\star(\mathcal{Y})} + \ln |\mathcal{D}_Y^{\mathrm{M}}|$ for the local optimization.

Although we provide no proof of it, reusing ideas of Willet and Nowak [37], Akakpo [2] or Akakpo and Lacour [3], one could deduce from this result the quasi optimal minimaxity of this estimator for anisotropic Besov spaces (see for instance in [24] for a definition) whose regularity index are smaller than 1 along the axes of $\mathcal{X}$ and smaller than $D_Y^{\mathrm{M}} + 1$ along the axes of $\mathcal{Y}$.

We focus now on the proof of Proposition 5Conditional density estimation theorem. The key is a fine control on the bracketing entropy of localized models which allows to use Proposition 2Bracketing entropy, Dudley integral and complexity. The proof of Proposition 6Conditional density estimation theorem, which is postponed to Appendix, is obtained easily by using Proposition 4Covariate partitioning and conditional density estimation for both $\mathcal{P}^{\mathcal{X}}$ and $\mathcal{P}^{\mathcal{Y}}$.

3.2.3 $\|\cdot\|_2$ and $\|\cdot\|_\infty$ structures

The key observation here is a link between the $\|\cdot\|_2$ and the $\|\cdot\|_\infty$ structures of the square roots of the models. Indeed, following Massart [28], we define the following tensorial *norm* on functions $u(y|x)$

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

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

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

If $\sqrt{S_m}$ is a subset of a linear space $\overline{\sqrt{S_m}}$ of dimension $\mathcal{D}_m$, as it is the case in our piecewise polynomial model, and if this linear space is such that $\forall u \in \overline{\sqrt{S_m}},\ \|u\|_2^{\otimes_n} < +\infty$,

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

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

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

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

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

For $\|\cdot\|_\infty$ type norms, bracketing and bracketing entropy are closely linked. Indeed, for any u, $[u-\delta/2, u+\delta/2]$ is a δ-bracket for the $\|\cdot\|_\infty^{\otimes_n}$ norm, so that any covering of $\left\{u \in \overline{\sqrt{S_m}} \middle| \|u\|_2^{\otimes_n} \leq \sigma\right\}$ by $\|\cdot\|_\infty^{\otimes_n}$ ball of radius $\delta/2$ yields a covering by the corresponding brackets. So that one obtains finally

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

The following proposition derived from Massart [28] 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 7. *For any basis $\{\phi_k\}_{1\leq k\leq \mathcal{D}_m}$ of $\overline{\sqrt{S_m}}$ such that*

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

let

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

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

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

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

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

We can now start

Proof of Proposition 5Conditional density estimation theorem. Using a basis of Legendre polynomials, we are able to derive from Proposition 7$\|\cdot\|_2$ and $\|\cdot\|_\infty$ structures

Proposition 8. *It exists*

$$\overline{r}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} \leq \prod_{d=1}^{d_Y}\left(\sqrt{D_{Y,d}^{\mathrm{M}}+1}\sqrt{2D_{Y,d}^{\mathrm{M}}+1}\right)\sup_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}}\frac{1}{\sqrt{\|\mathcal{P}^{\mathcal{X},\mathcal{Y}}\|}\sqrt{|\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}|}}$$

so that $\forall s_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} \in S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}$,

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

with $\mathcal{C}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} = \ln\left(\kappa_\infty \overline{r}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}\right)$ *and* $\kappa_\infty \leq 2\sqrt{2\pi e}$.

Now

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

Remark that when $\star(\mathcal{X}) = \mathrm{UDP}(\mathcal{X})$, $\star(\mathcal{Y}) = \mathrm{UDP}(\mathcal{Y})$ and $\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})$ is independent of $\mathcal{R}_l^{\mathcal{X}}$, 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

$$D_\star = \ln\left(\kappa_\infty \prod_{k=1}^{d_Y}\left(\sqrt{D_{Y,k}^{\mathrm{M}}+1}\sqrt{2D_{Y,k}^{\mathrm{M}}+1}\right)\right)$$

we have thus $\forall s_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} \in S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}$,

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

Proposition 2Bracketing entropy, Dudley integral and complexity combined with the inequality

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

concludes the proof. □

3.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties

3.3.1 Spatial Gaussian mixture models

In this section, we assume that $\mathcal{Y} = \mathbb{R}^p$ and we model the conditional density $s(\cdot|x)$ by Gaussian mixtures with varying proportions

$$s(\cdot|x) = \sum_{k=1}^{K}\pi_k(x)\Phi_{\theta_k}(\cdot)$$

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 K the number of mixture components, μ_k the mean of the kth component, Σ_k its covariance matrix, $\theta_k = (\mu_k, \Sigma_k)$ and $\pi_k(x)$ its proportion for the value x of the covariate. As mentioned in Introduction, this work has been motivated by an application of this setting to unsupervised segmentation. In this application described with full details in our companion paper, we observe a $n = n_1 \times n_2$ hyperspectral image (at each pixel x_i of the image, a spectrum Y_i is measured) and try to infer a partition of the image into homogeneous regions.

This problem is related to the one of unsupervised classification in which one observes a collection of Y_i and tries to split them into homogeneous classes. A classical method to provide an answer to this ill posed problem is to assume that the Y_i are i.i.d. random variables with a density that is close to a mixture of K densities, to estimate the best possible mixture, and then to assign to each observation a class that correspond to one of the mixed densities by a simple maximum likelihood principle. The most classical choice for the mixed densities is Gaussian densities as described for example in Biernacki et al. [7]. The work of Maugis and Michel [29] that inspired us explains how to chose the number of classes by a penalized maximum likelihood principle. Note that this corresponds to the setting of Kolaczyk et al. [26] and Antoniadis et al. [4], a piecewise constant component proportion, up to the choice of mixing densities.

The conditional densities we consider are thus of the form

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

where K is the number of component, $\mathcal{P}^{\mathcal{X}}$ is a partition of $\mathcal{X}$, θ_k is the parameter of the kth Gaussian and $\pi = (\pi[\mathcal{R}_l^{\mathcal{X}}])_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}}$ is the set of proportions on each hyperrectangle $\mathcal{R}_l^{\mathcal{X}}$. With a slight abuse of notation, we write

$$\pi[\mathcal{R}^{\mathcal{X}}(x)] = \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \pi[\mathcal{R}_l^{\mathcal{X}}] \, \mathbf{1}_{\left\{x \in \mathcal{R}_l^{\mathcal{X}}\right\}}$$

so that the previous conditional density can be rewritten

$$s_{K,\mathcal{P}^{\mathcal{X}},\theta,\pi}(\cdot|x) = \sum_{k=1}^{K} \pi_k[\mathcal{R}^{\mathcal{X}}(x)] \, \Phi_{\theta_k}(\cdot) \, .$$

We consider then model $S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}}$ with a fixed number of class K, a fixed partition $\mathcal{P}^{\mathcal{X}}$ to be chosen withing one of the collections of Section 3.1Covariate partitioning and conditional density estimation and a given set $\mathcal{F}$ for the K-uples $(\Phi_{\theta_1}, \ldots, \Phi_{\theta_K})$ (or equivalently by a set $\Theta_{\mathcal{F}}$ for $\theta = (\theta_1, \ldots, \theta_K)$). Within this model, the free parameters are the mixing proportions $\pi[\mathcal{R}_l^{\mathcal{X}}]$ on each hyperrectangle of the partition and the parameters θ within $\Theta_{\mathcal{F}}$.

Following Maugis and Michel [29], we will assume a specific structure for the set $\mathcal{F}$ that allows variable selection. We let E be an arbitrary space of $\mathcal{Y} = \mathbb{R}^p$ and assume that

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

where $\Phi_{E,\theta_{E,k}}(y)$ depends only on the projection of y on a space E and $\Phi_{E^\perp,\theta_{E^\perp}}(y)$ depends only on the projection of y on its orthogonal $E^\perp$. As hinted by the notation, we assume that $\Phi_{E^\perp,\theta_{E^\perp}}$

is independent of k. We assume that $\Phi_{E^\perp,\theta_{E^\perp}}$ belongs to a certain set $\mathcal{F}_{E^\perp}$ (or equivalently $\theta_{E^\perp} \in \Theta_{E^\perp}$) while the K-uple $(\Phi_{E,\theta_{E,1}}, \ldots, \Phi_{E,\theta_{E,K}})$ belongs to a certain set $\mathcal{F}_{E,K}$. We denote $S_{K,\mathcal{P}^\mathcal{X},\mathcal{F}}$ the corresponding model:

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

where $\mathcal{S}_{K-1}$ is the $K-1$ dimensional simplex:

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

The spaces $\mathcal{F}_{E,K}$ and $\mathcal{F}_{E^\perp}$ are chosen amongst the *classical* Gaussian K-uples described in Biernacki et al. [7]. For a space E of dimension p_E and a fixed number K of classes, we define the application $\Psi_{E,K}$ that maps $(\theta_1, \ldots, \theta_K)$ into $(\Phi_{E,\theta_1}, \ldots, \Phi_{E,\theta_K})$, and define our sets $\mathcal{F}_{[\cdot]_E^K}$ as the image through Ψ_K of some subset $\Theta_{[\cdot]_{p_E}^K}$ of K-uples of parameters. We obtain thus

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

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

For the means, we will always assume that they satisfy $\forall k, |\mu_k| \leq a$ for a known a. We consider cases where $\mu = (\mu_1, \ldots, \mu_K)$ is either known and equal to $\mu_0 = (\mu_{0,1}, \ldots, \mu_{0,K})$ or free.

Our model will also differ by the structure imposed on the collection of covariance matrix $(\Sigma_1, \ldots, \Sigma_K)$. We decompose 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.

Let $\mathcal{A}(\lambda_m, \lambda_M, p_E)$ as the set of diagonal matrix A such that $|A| = 1$ and $\forall 1 \leq i \leq p_E, \lambda_m \leq A_{i,i} \leq \lambda_M$ and $\mathcal{S}^+(p_E)$ the set of positive definite matrix, we nevertheless have an application Γ_K that maps

$$\begin{array}{c} (\mathbb{R}^{p_E})^K \times \left(\mathbb{R}^+\right)^K \times (SO(p_E))^K \times (\mathcal{A}(0, +\infty, p_E))^K \to \left(\mathbb{R}^{p_E}, \mathcal{S}^+(p_E)\right)^K \\ ((\mu_1, \ldots, \mu_K), (L_1, \ldots, L_K), (D_1, \ldots, D_K), (A_1, \ldots, A_K)) \mapsto ((\mu_1, L_1 D_1 A_1 D_1'), \ldots, (\mu_K, L_K D_K A_K D_K')) \end{array}$$

and allows thus to define $\Theta_{\mathcal{F}_{[\cdot]_{p_E}^K}}$ through a subset of $\mathbb{R}^K \times (\mathbb{R}^+)^K \times (SO(p_E))^K \times (\mathcal{A}(0, +\infty, p_E))^K$.

We will always assume that for all $1 \leq k \leq K$, $L_m \leq L_k \leq L_M$ and $A_k \in \mathcal{A}(\lambda_m, \lambda_M, p_E)$ for some (known) positive L_m, L_M, λ_m and λ_M. The volume L (respectively the basis D and the shape A) may be either known and equal to L_0, unknown with a common value L (respectively D and A) on all classes, or unknown but free to be different on all classes with values L_k (respectively D_k and A_k).

The resulting collections of K-uples will be indexed by $[\mu_\star\, \mathrm{L}_\star\, \mathrm{D}_\star\, \mathrm{A}_\star]_{p_E}^K$ where $\star = 0$ means that the quantity is known, $\star = K$ that the quantity is unknown and possibly different for every class and its lack means that there is a common unknown value over all classes.

Amongst all possible combinations, five of them have been studied by Maugis and Michel [29]:

- $[\mu_0\, \mathrm{L}_K\, \mathrm{D}_0\, \mathrm{A}_0]_{p_E}^K$ 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 $E^\perp$.
- $[\mu_K\, \mathrm{L}_K\, \mathrm{D}_0\, \mathrm{A}_K]_{p_E}^K$ in which one assumes that the unknown variances Σ_k can be diagonalized in the same known basis D_0.
- $[\mu_K\, \mathrm{L}_K\, \mathrm{D}_K\, \mathrm{A}_K]_{p_E}^K$ in which everything is free.
- $[\mu_K\, \mathrm{L}\, \mathrm{D}_0\, \mathrm{A}]_{p_E}^K$ in which the variances Σ_k are assumed to be equal and diagonalized in the known basis D_0.
- $[\mu_K\, \mathrm{L}\, \mathrm{D}\, \mathrm{A}]_{p_E}^K$ in which the variances Σ_k are only assumed to be equal.

We consider collection $\mathcal{S}$ of models $S_{K,\mathcal{P}^\mathcal{X},\mathcal{F}}$ where the number of class K is unknown, partition $\mathcal{P}^\mathcal{X}$ is chosen amongst a given collection $\mathcal{S}_\mathcal{P}^\star$ and $\mathcal{F}$ is chosen amongst three type of sets. The space E is chosen as $\mathrm{span}\{e_i\}_{i\in I}$ where e_i is the canonical basis of $\mathbb{R}^p$ and I a subset of $\{1,\ldots,p\}$ is either known, equal to $\{1,\ldots,p_E\}$ or free; while the index $[\mu_\star\, \mathrm{L}_\star\, \mathrm{D}_\star\, \mathrm{A}_\star]$ of $\mathcal{F}_E = \mathcal{F}_{[\mu_\star\, \mathrm{L}_\star\, \mathrm{D}_\star\, \mathrm{A}_\star]_E^K}$ and $\mathcal{F}_{E^\perp} = \mathcal{F}_{[\mu_\star\, \mathrm{L}_\star\, \mathrm{D}_\star\, \mathrm{A}_\star]_{E^\perp}}$ are chosen amongst a given set, where the corresponding spaces $\mathcal{F}_E$ and $\mathcal{F}_{E^\perp}$ are obtained with the same constants a, L_m, L_M, λ_m and λ_M. Note that

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

3.3.2 Spatial Gaussian mixture density estimation theorem

Without any assumption on the design, we obtain

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

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 1A general theorem for penalized maximum likelihood conditional density estimation satisfies

$$\mathbb{E}\left[JKL_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{K,\mathcal{P}^\mathcal{X},\mathcal{F}}})\right] \le C_1 \inf_{S_{K,\mathcal{P}^\mathcal{X},\mathcal{F}}\in\mathcal{M}} \left(\inf_{s_{K,\mathcal{P}^\mathcal{X},\mathcal{F}}\in S_{K,\mathcal{P}^\mathcal{X},\mathcal{F}}} KL^{\otimes n}(s_0, s_{K,\mathcal{P}^\mathcal{X},\mathcal{F}}) + \frac{\mathrm{pen}(K,\mathcal{P}^\mathcal{X},\mathcal{F})}{n} \right) + C_2\frac{1}{n} + \frac{\eta+\eta'}{n}$$

as soon as

$$\mathrm{pen}(K,\mathcal{P}^\mathcal{X},\mathcal{F}) \ge \kappa\Bigg(\left(C_\star + \left(\ln\frac{n}{C_\star \dim(S_{K,\mathcal{P}^\mathcal{X},\mathcal{F}})} \right)_+ \right) \dim(S_{K,\mathcal{P}^\mathcal{X},\mathcal{F}}) + c_\star\left(A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})}\|\mathcal{P}^\mathcal{X}\| + (K-1) + \theta_E \right) \Bigg)$$

with $\kappa > \kappa_0$ where κ_0 and C_2 are the constants of Theorem 1A general theorem for penalized maximum likelihood conditional density estimation that depend only on ρ and C_1 and

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

Note that this result can be applied in the exact setting of Antoniadis et al. [4] and allows to obtain their results without any discretization of the mixing proportion. The optimality of this modified estimate or of the one obtained by estimating simultaneously the mixture components and their spatial proportion in their horizon models also holds. Performances are also guaranteed when the design of the X_is is random.

As $\|\mathcal{P}^{\mathcal{X}}\| \leq \dfrac{\dim(S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}})}{K-1}$ and $K-1 \leq \dfrac{\dim(S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}})}{\|\mathcal{P}^{\mathcal{X}}\|}$, the condition on the penalty term in Theorem 3Spatial Gaussian mixture density estimation theorem is weaker than

$$\mathrm{pen}(K,\mathcal{P}^{\mathcal{X}},\mathcal{F}) \geq \kappa\left(\left(C_\star + \left(\ln\frac{n}{C_\star \dim(S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}})}\right)_+ + c_\star\left(A_0^{\star(\mathcal{X})} + \frac{B_0^{\star(\mathcal{X})}}{K-1} + \frac{1}{\|\mathcal{P}^{\mathcal{X}}\|}\right)\right)\dim(S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}}) + c_\star\theta_E\right).$$

The penalty term can be chosen, up to the variable selection term θ_E, roughly proportional to the dimension of the model, with a proportionality factor constant up to a logarithmic term with n. Furthermore under the weak assumption that the means of the mixture component are unknown and possibly different, $p_E \leq \dfrac{\dim(S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}})}{K}$ so that the θ_E term can also be controlled by a *multiple* of the model dimension.

Theorem 3Spatial Gaussian mixture density estimation theorem is obtained by combining

Proposition 9. *It exists a constant C depending only on a, L_m, L_M, λ_m and λ_M such that for any model $S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}}$ of Theorem 3Spatial Gaussian mixture density estimation theorem:*

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

satisfies the property required in Assumption (H).

Furthermore, $\sigma_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}}$ the unique root of $\frac{1}{\sigma}\phi_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}}(\sigma) = \sqrt{n}\sigma$ satisfies

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

and

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

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

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

with Theorem 1A general theorem for penalized maximum likelihood conditional density estimation and defining $C_\star = \left(C+\sqrt{\pi}\right)^2$. As in the previous section, the main difficulty lies in the control of the bracketing entropy of the models. We focus thus on the proof of Proposition 9Spatial Gaussian mixture density estimation theorem and postpone the one of Proposition 10Spatial Gaussian mixture density estimation theorem to Appendix.

Proof of Proposition 9Spatial Gaussian mixture density estimation theorem. 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 $H_{[\cdot],d^{\otimes n}}(\delta, S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}})$.

More precisely, we will derive an upper bound of an upper bound of the bracketing entropy $H_{[\cdot],d^{\otimes n}}(\delta, S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}})$ that is independent of the distribution law of $(X_i)_{1\leq i\leq n}$: the bracketing entropy with a sup norm Hellinger distance $d^{\sup} = \sqrt{d^{2\sup}}$, $H_{[\cdot],d^{\sup}}(\delta, S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}})$, where $d^{2\sup}$ is defined by

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

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

We prove in Appendix

Proposition 11. *It exists a constant C depending only on a, L_m, L_M, λ_m and λ_M such that for any model $S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}}$ of Theorem 3Spatial Gaussian mixture density estimation theorem:*

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

which combined with Proposition 3Bracketing entropy, Dudley integral and complexity leads immediately to Proposition 9Spatial Gaussian mixture density estimation theorem. □

3.3.3 Bracketing entropy of Gaussian families

A key step in the proof of Proposition 11Spatial Gaussian mixture density estimation theorem is a generalization of a result of Maugis and Michel [29, 30] that control the bracketing entropy the Gaussian families $\mathcal{F}_{[\cdot]_E^K}$ with respect to the $d^{\max}$ distance defined by

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

Here, $[(t_1^-,\ldots,t_K^-),(t_1^+,\ldots,t_K^+)]$ is a bracket containing $(s_1,\ldots,s_K)$ 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 12. *Let $\kappa \geq \frac{3}{4}$ and $\gamma_\kappa = \min\left(\frac{\kappa-\frac{3}{4}}{2(1+\frac{2\kappa}{9})(1+\frac{2}{9})(1+\frac{1}{9})}, \frac{3(\kappa-\frac{1}{2})}{2(1+\frac{2\kappa}{9})(1+\frac{2}{9})^3}\right)$.*

$$\textit{Assume} \qquad \begin{cases} a \geq \frac{\sqrt{\gamma_\kappa}}{18\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}}\sqrt{L_m\lambda_m\frac{\lambda_m}{\lambda_M}}\frac{\delta}{p_E} \\ \ln\left(\frac{L_M}{L_m}\right) \geq \frac{1}{20\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}}\delta \\ \frac{\lambda_M}{\lambda_m}\ln\left(\frac{\lambda_M}{\lambda_m}\right) \geq \frac{1}{27(1+\frac{2}{9})(\sqrt{2}+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}}\frac{\delta}{p_E} \end{cases}$$

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

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

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

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

$$\begin{cases} \mathcal{D}_{\mu,p_E} = p_E \\ \mathcal{D}_{\mathrm{L}} = 1 \\ \mathcal{D}_{\mathrm{D},p_E} = \frac{p_E(p_E-1)}{2} \\ \mathcal{D}_{\mathrm{A},p_E} = p_E - 1 \end{cases} \quad \textit{and} \quad \begin{cases} \mathcal{I}_{\mu,p_E} = p_E\left(\ln\left(\frac{36a\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}p_E}{\sqrt{\gamma_\kappa L_m\lambda_m\frac{\lambda_m}{\lambda_M}}}\right)\right) \\ \mathcal{I}_{\mathrm{L},p_E} = \ln\left(40\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}\ln\left(\frac{L_M}{L_m}\right)p_E\right) \\ \mathcal{I}_{\mathrm{D},p_E} = \frac{p_E(p_E-1)}{2}\left(\frac{\ln c}{\frac{p_E(p_E-1)}{2}} + \left(\ln\left(36(1+\frac{2}{9})(\sqrt{2}+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}\frac{\lambda_M}{\lambda_m}p_E\right)\right)\right) \\ \mathcal{I}_{\mathrm{A},p_E} = (p_E-1)\left(\ln\left(108(1+\frac{2}{9})(\sqrt{2}+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}\frac{\lambda_M}{\lambda_m}\ln\left(\frac{\lambda_M}{\lambda_m}\right)p_E\right)\right) \end{cases}$$

A Proofs for Section 2Penalized maximum likelihood for conditional density model selection (Penalized maximum likelihood for conditional density model selection)

A.1 Proof of Proposition 1Setting and maximum likelihood penalized estimate

Proof of Proposition 1Setting and maximum likelihood penalized estimate. We first notice that, by convexity of the Kullback-Leibler divergence,

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

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

Now,
$$\begin{aligned} JKL_\rho(s\mathrm{d}\lambda, t\mathrm{d}\lambda) &= \frac{1}{\rho}KL(s\mathrm{d}\lambda,(1-\rho)s+\rho t\mathrm{d}\lambda) = \frac{1}{\rho}KL((1+\rho u)\mathrm{d}\lambda',\mathrm{d}\lambda') \\ &= \frac{1}{\rho}KL_{\lambda'}(1+\rho u,1) = \frac{1}{\rho}\int(1+\rho u)\ln(1+\rho u)\mathrm{d}\lambda' \end{aligned}$$

and as $\int u\mathrm{d}\lambda' = 0$
$$= \frac{1}{\rho}\int\left((1+\rho u)\ln(1+\rho u)-\rho u\right)\mathrm{d}\lambda.$$

Similarly,
$$\begin{aligned} d^2(s\mathrm{d}\lambda, t\mathrm{d}\lambda) &= d^2((1+\rho u)\mathrm{d}\lambda', (1-(1-\rho)u)\mathrm{d}\lambda') = d^2_{\lambda'}(1+\rho u, 1-(1-\rho)u) \\ &= 2-2\int\sqrt{1+\rho u}\sqrt{1-(1-\rho)u}\mathrm{d}\lambda' = 2\int\left(1-\sqrt{1+(2\rho-1)u-\rho(1-\rho)u^2}\right)\mathrm{d}\lambda' \\ &= 2\int\left(1-\sqrt{1+(2\rho-1)u-\rho(1-\rho)u^2}+(\rho-\frac{1}{2})u\right)\mathrm{d}\lambda' \end{aligned}$$

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 $\forall u\in[-1/\rho,1/(1-\rho)]$, $\Phi(\rho u) = \frac{\Phi(\rho u)}{\rho^2u^2}\rho^2u^2 \geq \frac{\Phi(\frac{\rho}{1-\rho})}{\rho^2/(1-\rho)^2}\rho^2u^2$ so that

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

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

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

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

This implies thus

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

which yields the first inequality.

For the second series of inequalities,

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

while

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

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. □

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

Proof of Theorem 1A general theorem for penalized maximum likelihood conditional density estimation. Let g be a non random function, 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

$$P^{\otimes_n}(g) = \mathbb{E}\left[P_n^{\otimes_n}(g)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n g(X_i, Y_i)\right].$$

Note that g may depends on the covariate X_i and thus the last term can not generally be simplified. We will denote by $\nu_n^{\otimes_n}(g)$ the recentred process $P_n^{\otimes_n}(g) - P^{\otimes_n}(g)$.

For any model S_m, one assumes the existence of two functions $\widehat{s}_m$ and $\overline{s}_m$, such that

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

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

We define then the functions $kl(\overline{s}_m)$, $kl(\widehat{s}_m)$, and $jkl(\widehat{s}_m)$ by

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

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

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

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

$$P_n^{\otimes_n}(kl(\widehat{s}_{m'})) + \frac{\mathrm{pen}(m')}{n} \leq P_n^{\otimes_n}(kl(\widehat{s}_m)) + \frac{\mathrm{pen}(m)}{n} + \frac{\eta'}{n} \leq P_n^{\otimes_n}(kl(\overline{s}_m)) + \frac{\mathrm{pen}(m)}{n} + \frac{\eta + \eta'}{n}$$

Since, by concavity of the logarithm,

$$jkl(\widehat{s}_{m'}) = -\frac{1}{\rho}\ln\left(\frac{(1-\rho)s_0 + \rho\widehat{s}_{m'}}{s_0}\right) \leq -\frac{1}{\rho}\left((1-\rho)\ln\frac{s_0}{s_0} + \rho\ln\frac{\widehat{s}_{m'}}{s_0}\right) = -\ln\frac{\widehat{s}_{m'}}{s_0} = kl(\widehat{s}_{m'}),$$

$$P_n^{\otimes_n}(jkl(\widehat{s}_{m'})) + \frac{\mathrm{pen}(m')}{n} \leq P_n^{\otimes_n}(kl(\overline{s}_m)) + \frac{\mathrm{pen}(m)}{n} + \frac{\eta + \eta'}{n}$$

and thus

$$P^{\otimes_n}(jkl(\widehat{s}_{m'})) - \nu_n^{\otimes_n}(kl(\overline{s}_m)) \leq P^{\otimes_n}(kl(\overline{s}_m)) + \frac{\mathrm{pen}(m)}{n} - \nu_n^{\otimes_n}(jkl(\widehat{s}_{m'})) - \frac{\mathrm{pen}(m')}{n} + \frac{\eta + \eta'}{n}$$

using the definition of $jkl(\widehat{s}_{m'})$ and of $kl(\overline{s}_m)$, we deduce

$$JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{m'}) - \nu_n^{\otimes_n}(kl(\overline{s}_m)) \leq \inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} - \nu_n^{\otimes_n}(jkl(\widehat{s}_{m'})) - \frac{\mathrm{pen}(m')}{n} + \frac{\eta + \eta'}{n} + \frac{\delta_{KL}}{n}$$

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

Lemma 1. *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 2. *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 $\mathbb{P}\{A\} > 0$,*

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

Combining Lemma 1Proof of Theorem 1A general theorem for penalized maximum likelihood conditional density estimation and Lemma 2Proof of Theorem 1A general theorem for penalized maximum likelihood conditional density estimation implies that except on a set of probability less than $e^{-x_{m'}-x}$, for any $y_{m'} > \sigma_{m'}$,

$$\frac{-\nu_n^{\otimes_n}(jkl(\widehat{s}_{m'}))}{y_{m'}^2 + \kappa_0' d^{2\otimes_n}(s_0, \widehat{s}_{m'})} \leq \frac{\kappa_1' \sigma_{m'}}{y_{m'}} + \kappa_2' \sqrt{\frac{x_{m'}+x}{n y_{m'}^2}} + \frac{18}{\rho} \frac{x_{m'}+x}{n y_{m'}^2}.$$

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

$$\frac{-\nu_n^{\otimes_n}(jkl(\widehat{s}_{m'}))}{y_{m'}^2 + \kappa_0' d^{2\otimes_n}(s_0, \widehat{s}_{m'})} \leq \frac{\kappa_1' + \kappa_2'}{\theta} + \frac{18}{\theta^2 \rho}$$

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

$$JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{m'}) - \nu_n^{\otimes_n}(kl(\overline{s}_m)) \leq \inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} + \left(\frac{\kappa_1' + \kappa_2'}{\theta} + \frac{18}{\theta^2 \rho}\right)\left(y_{m'}^2 + \kappa_0' d^{2\otimes_n}(s_0, \widehat{s}_{m'})\right) - \frac{\mathrm{pen}(m')}{n} + \frac{\eta + \eta'}{n} + \frac{\delta_{KL}}{n}.$$

Let $\epsilon_{\mathrm{pen}} > 0$, we define θ_{pen} by $\left(\frac{\kappa_1'+\kappa_2'}{\theta_{\mathrm{pen}}} + \frac{18}{\theta_{\mathrm{pen}}^2 \rho}\right)\kappa_0' = C_\rho\, \epsilon_{\mathrm{pen}}$ 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

$$(1-\epsilon_{\mathrm{pen}}) JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{m'}) - \nu_n^{\otimes_n}(kl(\overline{s}_m)) \leq \inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} + \frac{C_\rho \epsilon_{\mathrm{pen}} y_{m'}^2}{\kappa_0'} - \frac{\mathrm{pen}(m')}{n} + \frac{\eta + \eta'}{n} + \frac{\delta_{KL}}{n}.$$

We should now study $\frac{C_\rho \epsilon_{\mathrm{pen}} y_{m'}^2}{\kappa_0'} - \frac{\mathrm{pen}(m')}{n}$:

$$\frac{C_\rho \epsilon_{\mathrm{pen}} y_{m'}^2}{\kappa_0'} - \frac{\mathrm{pen}(m')}{n} = \frac{C_\rho \epsilon_{\mathrm{pen}} \theta_{\mathrm{pen}}^2}{\kappa_0'}\left(\sigma_m^2 + \frac{x_m + x}{n}\right) - \frac{\mathrm{pen}(m')}{n}$$

and by construction if, in the constraint defining the penalties, $\kappa_0 \geq \frac{C_\rho \epsilon_{\mathrm{pen}} \theta_{\mathrm{pen}}^2}{\kappa_0'}$

$$\frac{C_\rho \epsilon_{\mathrm{pen}} y_{m'}^2}{\kappa_0'} - \frac{\mathrm{pen}(m')}{n} \leq \frac{C_\rho \epsilon_{\mathrm{pen}} \theta_{\mathrm{pen}}^2}{\kappa_0'} \frac{x}{n} - (1 - \frac{C_\rho \epsilon_{\mathrm{pen}} \theta_{\mathrm{pen}}^2}{\kappa \kappa_0'}) \frac{\mathrm{pen}(m')}{n}.$$

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

$$(1-\epsilon_{\mathrm{pen}}) JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{m'}) + (1 - \frac{C_\rho \epsilon_{\mathrm{pen}} \theta_{\mathrm{pen}}^2}{\kappa \kappa_0'}) \frac{\mathrm{pen}(m')}{n} - \nu_n^{\otimes_n}(kl(\overline{s}_m))$$

$$\leq \inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} + \frac{C_\rho \epsilon_{\mathrm{pen}} \theta^2_{\mathrm{pen}}}{\kappa'_0} \frac{x}{n} + \frac{\eta + \eta'}{n} + \frac{\delta_{KL}}{n}$$

As $\nu_n^{\otimes_n}(kl(\overline{s}_m))$ is integrable (and of mean 0), we derive that $M = \sup_{m' \in \mathcal{M}'} \frac{\mathrm{pen}(m')}{n}$ 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 $\hat{m}$ of $P_n^{\otimes_n}(-\ln(\widehat{s}_m)) + \frac{\mathrm{pen}(m)}{n}$ exists.

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

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

which yields by integration

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

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

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

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

Proof of Lemma 1Proof of Theorem 1A general theorem for penalized maximum likelihood conditional density estimo
Let $A = \{Z > \Psi(x)\}$. Either $P\{A\} = 0 \leq e^{-x}$ or

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

Now in the later case,

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

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. □

A.3 Proof of Lemma 2Proof of Theorem 1A general theorem for penalized maximum likelihood conditional density estimation

We should now prove Lemma 2Proof of Theorem 1A general theorem for penalized maximum likelihood conditional density estimation which contains most of the differences with Massart [28]'s proof.

Proof of Lemma 2Proof of Theorem 1A general theorem for penalized maximum likelihood conditional density estimo
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 $\widetilde{s}$ to be fixed later, if we let $jkl(\widetilde{s}) = -\frac{1}{\rho}\ln\left(\frac{(1-\rho)s_0+\rho\widetilde{s}}{s_0}\right)$, then $-jkl(\widehat{s}_m) = -jkl(\widetilde{s}) + (-jkl(\widehat{s}_m) + jkl(\widetilde{s}))$ with

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

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 [28])

Lemma 3. *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 4. *Assume f is a function such that*

$$P^{\otimes_n}\left(|f|^2\right) \leq V$$
$$\forall k \geq 3, \quad P^{\otimes_n}\left((f)_+^k\right) \leq \frac{k!}{2}Vb^{k-2}.$$

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 $jkl(\widetilde{s})$

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

To cope with the randomness of $\widehat{s}_m$, we rely on the following much more involved theorem (a rewriting of Theorem 6.8 of Massart [28])

Theorem 5. *Let $\mathcal{F}$ 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 $f \in \mathcal{F}$ 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 $\mathcal{F}$ 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

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

where

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

Furthermore $\kappa \leq 27$.

If we consider

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

then the first assumption of Theorem 5Proof of Lemma 2Proof of Theorem 1A general theorem for penalized maximum likelihood conditional density 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

$$\begin{aligned}W_m(\widetilde{s}, \sigma) &= \sup_{f \in \mathcal{F}_m(\widetilde{s},\sigma)} \nu_n^{\otimes_n}(f) = \sup_{s_m \in S_m(\widetilde{s},\sigma)} \nu_n^{\otimes_n}\left(-jkl(s_m) + jkl(\widetilde{s})\right) \\ &= \sup_{s_m \in S_m(\widetilde{s},\sigma)} \nu_n^{\otimes_n}\left(-jkl(s_m)\right) + \nu_n^{\otimes_n}\left(jkl(\widetilde{s})\right)\end{aligned}$$

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

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

and

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

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 5Proof of Lemma 2Proof of Theorem 1A general theorem for penalized maximum likelihood conditional density estimation from a set of brackets of $d^{\otimes_n}$ width smaller than $\frac{2\sqrt{2\rho(1-\rho)}}{3}\delta$ covering $S_m(\widetilde{s},\sigma)$. That is

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

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

$$\mathbb{E}^A\left[W_m(\widetilde{s},\sigma)\right] \leq E + \frac{(1+6\epsilon)3\sigma}{2\sqrt{\rho(1-\rho)}\sqrt{n}}\sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)} + \frac{4}{\rho n}\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)$$

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

Choosing $\epsilon = 1$ leads to

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

where

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

By definition, $\int_0^{\sigma}\sqrt{H_{[\cdot],d^{\otimes_n}}\left(\delta, S_m(\widetilde{s}_m,\sigma)\right)\wedge n}\mathrm{d}\delta \leq \phi_m(\sigma)$, as well as $\delta \mapsto H_{[\cdot],d^{\otimes_n}}\left(\delta, S_m(\widetilde{s}_m,\sigma)\right)$ is non-increasing. This implies

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

Inserting these bounds in the previous inequality yields

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

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

As $\delta \mapsto \delta^{-1}\phi_m(\delta)$ is also non-increasing, so is $\delta \mapsto \delta^{-2}\phi_m(\delta)$. Now by definition, $\frac{\phi_m(\sigma_m)}{\sqrt{n}\sigma_m^2} = 1$. This implies thus, as soon as $\sigma \geq \sigma_m$

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

and

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

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 $\forall \sigma > \sigma_m$,

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

Thanks to Assumption (S), we can use the *pealing* lemma (Lemma 4.23 of [28]):

Lemma 4. *Let S be some countable set, $\widetilde{s} \in S$ and $a : S \to \mathbb{R}^+$ such that $a(\widetilde{s}) = \inf_{s \in S} a(s)$. 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 $\sup_{s \in B(\sigma)} (Z(s) - Z(\widetilde{s}))$ has finite expectation. Then, for any function ψ on $\mathbb{R}^+$ such that $\psi(x)/x$ is non-increasing on $\mathbb{R}^+$ and

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

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

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

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

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

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

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

and therefore

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

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

$$d^{2\otimes_n}(s_0,\widetilde{s}_m)\leq(1+\epsilon_d)d^{2\otimes_n}(s_0,s_m)$$

so that

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

For this choice, one obtains

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

which implies

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

We turn back to the control of $-\nu_n^{\otimes_n}(jkl(\widetilde{s}_m))$. Our Berstein type control yields

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

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

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

We derive thus

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

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

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

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 $\kappa_2'=4\kappa_2''+3/(4\sqrt{\rho(1-\rho)}\kappa_d')$. □

A.4 Behavior of the constants of Theorem 1A 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 2Proof of Theorem 1A general theorem for penalized maximum likelihood conditional density estimation. This implies

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

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

$$\left(\kappa_1' + \kappa_2'\right)^2 \leq \frac{1}{\rho(1-\rho)} \left(3\kappa\sqrt{2} + 54 + \frac{3}{8(1-\epsilon)} + 16\sqrt{\frac{1-\rho}{\rho}} \right)^2.$$

Now using $4 < \kappa_0' \leq 4(1+\epsilon)$

$$\begin{aligned}
\kappa_0 &\leq \frac{4(1+\epsilon) \left(3\kappa\sqrt{2} + 54 + \frac{3}{8(1-\epsilon)} + 16\sqrt{\frac{1-\rho}{\rho}} \right)^2 \left(\sqrt{1 + \frac{72 C_\rho \epsilon_{\text{pen}}}{\rho \kappa_0' (\kappa_1' + \kappa_2')^2}} + 1 \right)}{2\rho(1-\rho) C_\rho \epsilon_{\text{pen}}} + \frac{18}{\rho} \\
&\leq \frac{1}{C_\rho \rho (1-\rho) \epsilon_{\text{pen}}} \\
&\quad \times \left(2(1+\epsilon) \left(3\kappa\sqrt{2} + 54 + \frac{3}{8(1-\epsilon)} + 16\sqrt{\frac{1-\rho}{\rho}} \right)^2 \left(\sqrt{1 + \frac{72 C_\rho \epsilon_{\text{pen}}}{\rho \kappa_0' (\kappa_1' + \kappa_2')^2}} + 1 \right) + 18 C_\rho (1-\rho) \epsilon_{\text{pen}} \right)
\end{aligned}$$

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 5Proof of Lemma 2Proof of Theorem 1A general theorem for penalized maximum likelihood conditional density 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

$$\frac{1}{C_\rho \rho(1-\rho)\epsilon_{\text{pen}}} \times \left(2(1+\epsilon)\left(3\kappa\sqrt{2}+42+6\sqrt{2}\sigma_{\mathcal{M}}+\frac{3}{8(1-\epsilon)}+16\sqrt{\frac{1-\rho}{\rho}}\right)^2\left(\sqrt{1+\frac{72C_\rho\epsilon_{\text{pen}}}{\rho\kappa_0'(\kappa_1'+\kappa_2')^2}}+1\right)+18C_\rho(1-\rho)\epsilon_{\text{pen}}\right).$$

which is much smaller than the previous quantity as soon as $\sigma_{\mathcal{M}}$ is much smaller than $\sqrt{2}$, which can be ensure in the models of Section 3.3Spatial Gaussian mixtures, models, bracketing entropy and penalties provided we limit their maximum dimension well below n, for instance to $n/\ln^2(n)$.

B Proof for Section 2.3Bracketing entropy, Dudley integral and complexity (Bracketing entropy, Dudley integral and complexity)

Proof of Proposition 2Bracketing entropy, Dudley integral and complexity. 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,

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

We use now

Lemma 5. *For any $\sigma \in [0,1]$, $\int_0^\sigma \sqrt{\ln\frac{1}{\delta}}\,\mathrm{d}\delta \le \sigma\left(\sqrt{\ln\frac{1}{\sigma}}+\sqrt{\pi}\right)$.*

proved in Maugis and Michel [29] to obtain

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

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

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

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

Proof of Proposition 3Bracketing entropy, Dudley integral and complexity. 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

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

and using Lemma 5Proof for Section 2.3Bracketing entropy, Dudley integral and complexity (Bracketing entropy, Dudley integral and complexity)

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

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

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

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

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. □

C Proof for Section 3.1Covariate partitioning and conditional density estimation (Covariate partitioning and conditional density estimation)

Proof of Proposition 4Covariate partitioning and conditional density estimation. We start by the UDP case, as we stop as soon as $\frac{2_X^d}{n} > 2^{-d_X J} \leq \frac{1}{n}$, $J \leq \frac{\ln n}{d_X \ln 2}$ 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^d_X-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 $N(\mathcal{P}^{\mathcal{X}})$ time the encoding cost (respectively $\lceil \frac{\ln 2}{\ln 2} \rceil$ bits, $\left\lceil \frac{\ln(1+d_X)}{\ln 2} \right\rceil$ bits and $\left\lceil \frac{\ln(1+d_X)}{\ln 2} \right\rceil + \lceil \frac{\ln n}{\ln 2} \rceil$). As this code is decodable, it satisfies the Kraft inequality and thus, using the definition of $B_0^{\star(\mathcal{X})}$,

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

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

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

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

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

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

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

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

$$\begin{aligned} &\leq e^{-(c - c_0^{\star(\mathcal{X})}) B_0^{\star(\mathcal{X})}} \sum_{\mathcal{P}^{\mathcal{X}} \in \mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}} e^{-c_0^{\star(\mathcal{X})} B_0^{\star(\mathcal{X})} \|\mathcal{P}^{\mathcal{X}}\|} \\ &\leq e^{-(c - c_0^{\star(\mathcal{X})}) B_0^{\star(\mathcal{X})}} e^{(1 - c_0^{\star(\mathcal{X})}) B_0^{\star(\mathcal{X})}} = e^{B_0^{\star(\mathcal{X})}} e^{-c B_0^{\star(\mathcal{X})}} = \Sigma_0^{\star(\mathcal{X})} e^{-c C_0^{\star(\mathcal{X})}} \end{aligned}$$

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 $\|\mathcal{P}^{\mathcal{X}}\| \times d_X \left\lceil \frac{\ln n}{\ln 2} \right\rceil$ bits for each partition that satisfies the Kraft inequality

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

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

$$\begin{aligned}
\sum_{\mathcal{P}^{\mathcal{X}} \in \mathcal{S}_{\mathcal{P}}^{\mathrm{HRP}(\mathcal{X})}} e^{-cB_0^{\mathrm{HRP}(\mathcal{X})} \|\mathcal{P}^{\mathcal{X}}\|} &= \sum_{\mathcal{P}^{\mathcal{X}} \in \mathcal{S}_{\mathcal{P}}^{\mathrm{HRP}(\mathcal{X})}} e^{-(c-c_0^{\mathrm{HRP}(\mathcal{X})}) B_0^{\mathrm{HRP}(\mathcal{X})} \|\mathcal{P}^{\mathcal{X}}\|} e^{-c_0^{\mathrm{HRP}(\mathcal{X})} B_0^{\mathrm{HRP}(\mathcal{X})} \|\mathcal{P}^{\mathcal{X}}\|} \\
&\leq e^{-(c-c_0^{\mathrm{HRP}(\mathcal{X})}) B_0^{\mathrm{HRP}(\mathcal{X})}} \sum_{\mathcal{P}^{\mathcal{X}} \in \mathcal{S}_{\mathcal{P}}^{\mathrm{HRP}(\mathcal{X})}} e^{-c_0^{\mathrm{HRP}(\mathcal{X})} B_0^{\mathrm{HRP}(\mathcal{X})} \|\mathcal{P}^{\mathcal{X}}\|} \\
&\leq e^{-(c-c_0^{\mathrm{HRP}(\mathcal{X})}) B_0^{\mathrm{HRP}(\mathcal{X})}} = e^{B_0^{\mathrm{HRP}(\mathcal{X})}} e^{-cB_0^{\mathrm{HRP}(\mathcal{X})}} = \Sigma_0^{\mathrm{HRP}(\mathcal{X})} e^{-cC_0^{\mathrm{HRP}(\mathcal{X})}}
\end{aligned}$$

which concludes the proof.

□

D Proof for Section 3.2Piecewise polynomial conditional densities estimation (Piecewise polynomial conditional densities estimation)

D.1 Model coding

Proof of Proposition 6Conditional density estimation theorem. By construction

$$\begin{aligned}
\sum_{S_{\mathcal{P}^{\mathcal{X}},\mathcal{Y},D_Y^{\mathrm{M}}} \in \mathcal{S}} e^{-x_{\mathcal{P}^{\mathcal{X}},\mathcal{Y},D_Y^{\mathrm{M}}}} &= \sum_{\mathcal{P}^{\mathcal{X}} \in \mathcal{S}_{\mathcal{P}}^{\star}(\mathcal{X})} \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \sum_{\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}}) \in \mathcal{S}_{\mathcal{P}}^{\star}(\mathcal{Y})} e^{-x_{\mathcal{P}^{\mathcal{X}},\mathcal{Y},D_Y^{\mathrm{M}}}} \\
&= \sum_{\mathcal{P}^{\mathcal{X}} \in \mathcal{S}_{\mathcal{P}}^{\star}(\mathcal{X})} \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \sum_{\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}}) \in \mathcal{S}_{\mathcal{P}}^{\star}(\mathcal{Y})} e^{-c_\star \left(A_0^{\star(\mathcal{X})} + \left(B_0^{\star(\mathcal{X})} + A_0^{\star(\mathcal{Y})} \right) \|\mathcal{P}^{\mathcal{X}}\| + B_0^{\star(\mathcal{Y})} \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \|\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})\| \right)} \\
&= \sum_{\mathcal{P}^{\mathcal{X}} \in \mathcal{S}_{\mathcal{P}}^{\star}(\mathcal{X})} e^{-c_\star \left(A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})} |\mathcal{P}| \right)} \prod_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \left(\sum_{\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}}) \in \mathcal{S}_{\mathcal{P}}^{\star}(\mathcal{Y})} e^{-c_\star \left(A_0^{\star(\mathcal{Y})} + B_0^{\star(\mathcal{Y})} \|\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})\| \right)} \right)
\end{aligned}$$

By Proposition 4Covariate partitioning and conditional density estimation, one can find $c_\star \geq \max(1, c_0^{\star(\mathcal{X})}, c_0^{\star(\mathcal{Y})})$ such that

$$\sum_{\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}}) \in \mathcal{S}_{\mathcal{P}}^{\star}(\mathcal{Y})} e^{-c_\star \left(A_0^{\star(\mathcal{Y})} + B_0^{\star(\mathcal{Y})} \|\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})\| \right)} \leq 1$$

and

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

Plugging these bounds in the previous equality yields

$$\sum_{S_{\mathcal{P}^{\mathcal{X}},\mathcal{Y},D_Y^{\mathrm{M}}}\in\mathcal{S}} e^{-x_{\mathcal{P}^{\mathcal{X}},\mathcal{Y},D_Y^{\mathrm{M}}}} \leq \sum_{\mathcal{P}^{\mathcal{X}}\in\mathcal{S}_{\mathcal{P}}^{\star}(\mathcal{X})} e^{-c_\star\left(A_0^{\star(\mathcal{X})}+B_0^{\star(\mathcal{X})}\|\mathcal{P}^{\mathcal{X}}\|\right)} \leq 1.$$

The proposition holds with the modified weights for polynomial as

$$\sum_{D_Y^{\mathrm{M}}\in\mathcal{D}_Y^{\mathrm{M}}} e^{-c_\star \ln|\mathcal{D}_Y^{\mathrm{M}}|} = |\mathcal{D}_Y^{\mathrm{M}}|^{1-c_\star} \leq 1$$

as soon as $c_\star \geq 1$. □

D.2 Entropy

Proof of Proposition 7 $\|\cdot\|_2$ and $\|\cdot\|_\infty$ structures. Let $(\phi_k)_{1\leq k\leq\mathcal{D}_m}$ be a basis of $\overline{\sqrt{S_m}}$ satisfying

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

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

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

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

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

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

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

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

The definition of $\overline{r}_m$ implies then that

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

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

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

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

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

and thus

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

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

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

which concludes the proof. □

Instead of Proposition 8 $\|\cdot\|_2$ and $\|\cdot\|_\infty$ structures, 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 $\mathcal{P}^{\mathcal{X}} \in \mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}$ and the partitions $\mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}}) \in \mathcal{S}_{\mathcal{P}}^{\star(\mathcal{Y})}$ for $\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}$ and define now the model $S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}$ as the set of conditional densities such that

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

where $P_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}}$ is a polynomial of degree at most $D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}) = \left(D_{Y,1}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}), \ldots, D_{Y,d_Y}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})\right)$. By construction,

$$\dim(S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}) = \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \left(\left(\sum_{\mathcal{R}_{l,k}^{\mathcal{Y}} \in \mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})} \prod_{d=1}^{d_Y} \left(D_{Y,d}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}) + 1\right)\right) - 1\right).$$

The corresponding linear space $\overline{\sqrt{S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}}}$ is

$$\left\{ \sum_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} P_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}}(y) \mathbf{1}_{\left\{(x,y)\in\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}\right\}} \middle| \deg\left(P_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}}\right) \leq D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}) \right\}$$

of dimension

$$\mathcal{D}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} = \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \sum_{\mathcal{R}_{l,k}^{\mathcal{Y}} \in \mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})} \prod_{d=1}^{d_Y} \left(D_{Y,d}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}) + 1\right) = \sum_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y} \left(D_{Y,d}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}) + 1\right)$$

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

Proposition 13. *It exists*

$$\overline{r}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} \leq \frac{\sup_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y} \left(\sum_{D_{Y,d} \leq D_{Y,d}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \sqrt{2D_{Y,d}+1}\right)}{\inf_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y} \sqrt{D_{Y,d}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}) + 1}} \sup_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} \frac{1}{\sqrt{\|\mathcal{P}^{\mathcal{X}}\|}\sqrt{|\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}|}}$$

such that $\forall s_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} \in S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}$,

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

with $\mathcal{C}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} = \ln\left(\kappa_\infty \overline{r}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}\right)$ *and* $\kappa_\infty \leq 2\sqrt{2\pi e}$.

Proposition 8$\|\cdot\|_2$ and $\|\cdot\|_\infty$ structures is deduced from this proposition with the help of the simple upper bound

$$\sum_{D_{Y,d}\leq D_{Y,d}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \sqrt{2D_{Y,d}+1} \leq (D_{Y,d}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})+1)\sqrt{2D_{Y,d}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})+1}.$$

For the polynomial degree choice, it is sufficient to upper bound the ratio

$$\frac{\sup_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}}\prod_{d=1}^{d_Y}\left(\sum_{D_{Y,d}\leq D_{Y,d}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})}\sqrt{2D_{Y,d}+1}\right)}{\inf_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}}\prod_{d=1}^{d_Y}\sqrt{D_{Y,d}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})+1}}$$

over all choices of degrees and to apply the same reasoning as in the proof of Proposition 9Spatial Gaussian mixture density estimation theorem.

Proof of Proposition 13Entropy. 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

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

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

$$G_{D_{Y,1},\ldots,D_{Y,d_Y}}(y) = G_{D_{Y,1}}(y_1)\times\cdots\times G_{D_{Y,d_Y}}(y_{d_Y}),$$

by construction

$$\forall D_Y\in\mathbb{N}^{d_Y},\quad \|G_{D_Y}\|_\infty = \prod_{1\leq d\leq d_Y}\sqrt{2D_{Y,k}+1}$$

and

$$\forall (D_Y,D_Y')\in\mathbb{N}^{d_Y\times 2},\quad \int_{y\in[0,1]_Y^d} G_{D_Y}(y)G_{D_Y'}(y)\mathrm{d}y = \delta_{D_Y,D_Y'}.$$

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

$$\forall \mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}},\forall D_Y\in\mathbb{N}^{d_Y},\quad \|G_D^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}}\|_\infty = \frac{1}{\sqrt{|\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}|}}\prod_{1\leq d\leq d_Y}\sqrt{2D_{Y,k}+1}$$

and

$\forall (\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}},\mathcal{R}_{l',k'}^{\mathcal{X},\mathcal{Y}})\in(\mathcal{P}^{\mathcal{X},\mathcal{Y}})^2,\forall (D_Y,D_Y')\in\mathbb{N}^{d_Y\times 2}$,

$$\int_{x\in[0,1]^d_X}\int_{y\in[0,1]^d_Y} G_D^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}(x,y)G_{D'}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'}}(x,y)\mathrm{d}y\mathrm{d}x = \delta_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k},\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'}}\delta_{D_Y,D'_Y}.$$

Using the piecewise structure, one deduces

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}}\sum_{D_Y\le D_Y^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k})}\beta_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}G_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}(X_i,\cdot)\right\|_2^2\right]\\
&\quad=\mathbb{E}\left[\sum_{\mathcal{R}^{\mathcal{X}}_l\in\mathcal{P}^{\mathcal{X}}}\frac{\mathbf{1}_{\{X_1\in\mathcal{R}^{\mathcal{X}}_l\}}}{|\mathcal{R}^{\mathcal{X}}_l|}\sum_{\mathcal{R}^{\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{Y}}(\mathcal{R}^{\mathcal{X}}_l)}\int_{(x,y)\in\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}\left|\sum_{D_Y\le D_Y^{\mathrm{M}}(\mathcal{R}^{\mathcal{X}}_l,\mathcal{R}^{\mathcal{Y}}_{l,k})}\beta_{D_Y}^{\mathcal{R}^{\mathcal{X}}_l,\mathcal{R}^{\mathcal{Y}}_{l,k}}G_{D_Y}^{\mathcal{R}^{\mathcal{X}}_l,\mathcal{R}^{\mathcal{Y}}_{l,k}}(x,y)\right|^2\mathrm{d}y\mathrm{d}x\right]\\
&\quad=\mathbb{E}\left[\sum_{\mathcal{R}^{\mathcal{X}}_l\in\mathcal{P}^{\mathcal{X}}}\frac{\mathbf{1}_{\{X_1\in\mathcal{R}^{\mathcal{X}}_l\}}}{|\mathcal{R}^{\mathcal{X}}_l|}\sum_{\mathcal{R}^{\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{Y}}(\mathcal{R}^{\mathcal{X}}_l)}\sum_{D_Y\le D_Y^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k})}\left|\beta_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}\right|^2\right]\\
&\quad=\sum_{\mathcal{R}^{\mathcal{X}}_l\in\mathcal{P}^{\mathcal{X}}}\frac{\mathbb{P}\{X_i\in\mathcal{R}^{\mathcal{X}}_l\}}{|\mathcal{R}^{\mathcal{X}}_l|}\sum_{\mathcal{R}^{\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{Y}}(\mathcal{R}^{\mathcal{X}}_l)}\sum_{D_Y\le D_Y^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k})}\left|\beta_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}\right|^2.
\end{aligned}$$

The space $\overline{\sqrt{S_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}}}}$ is spanned by

$$\left\{G_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}\middle|\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y\le D_Y^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k})\right\}$$

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

$$\begin{aligned}
&\left\|\sum_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}}\sum_{D_Y\le D_Y^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k})}\beta_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}\phi_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}\right\|_2^{2\otimes_n}\\
&\quad=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|\sum_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}}\sum_{D_Y\le D_Y^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k})}\beta_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}\phi_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}(X_i,\cdot)\right\|_2^2\right]\\
&\quad=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\sum_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}}\sum_{D_Y\le D_Y^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k})}\frac{\beta_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}}{\sqrt{\mu_X(\mathcal{R}^{\mathcal{X}}_l)}}G_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}(X_i,\cdot)\right\|_2^2\right]\\
&\quad=\frac{1}{n}\sum_{i=1}^n\sum_{\mathcal{R}^{\mathcal{X}}_l\in\mathcal{P}^{\mathcal{X}}}\frac{\mathbb{P}\{X_i\in\mathcal{R}^{\mathcal{X}}_l\}}{|\mathcal{R}^{\mathcal{X}}_l|}\sum_{\mathcal{R}^{\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{Y}}(\mathcal{R}^{\mathcal{X}}_l)}\sum_{D_Y\le D_Y^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k})}\left|\frac{\beta_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}}{\sqrt{\mu(\mathcal{R}^{\mathcal{X}}_l)}}\right|^2\\
&\quad=\sum_{\mathcal{R}^{\mathcal{X}}_l\in\mathcal{P}^{\mathcal{X}}}\mu_X(\mathcal{R}^{\mathcal{X}}_l)\sum_{\mathcal{R}^{\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{Y}}(\mathcal{R}^{\mathcal{X}}_l)}\sum_{D_Y\le D_Y^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k})}\left|\frac{\beta_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}}{\sqrt{\mu(\mathcal{R}^{\mathcal{X}}_l)}}\right|^2\\
&\quad=\sum_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}}\sum_{D_Y\le D_Y^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k})}\left|\beta_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}\right|^2=\left\|\beta_{D_Y}^{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}}\right\|_2^2.
\end{aligned}$$

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

$$
\begin{aligned}
&\left\| \sum_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} \sum_{D_Y \le D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \beta_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}} \phi_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}} \right\|_\infty^{2\otimes_n} \\
&= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \left\| \sum_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} \sum_{D_Y \le D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \beta_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}} \phi_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}}(X_i,\cdot) \right\|_\infty^2 \right] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\left\| \sum_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} \sum_{D_Y \le D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \beta_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}} \phi_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}}(X_i,\cdot) \right\|_\infty^2 \right] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \mathbf{1}_{\{X_i \in \mathcal{R}_l^{\mathcal{X}}\}} \sup_{\mathcal{R}_{l,k}^{\mathcal{Y}} \in \mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})} \left\| \sum_{D_Y \le D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \beta_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}} \phi_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}}(X_i,\cdot) \right\|_\infty^2 \right] \\
&\le \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \mathbf{1}_{\{X_i \in \mathcal{R}_l^{\mathcal{X}}\}} \sup_{x \in \mathcal{R}_l^{\mathcal{X}}} \sup_{\mathcal{R}_{l,k}^{\mathcal{Y}} \in \mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})} \left(\sum_{D_Y \le D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \left| \beta_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}} \right| \left\| \phi_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}}(x,\cdot) \right\|_\infty \right)^2 \right] \\
&\le \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \mathbf{1}_{\{X_i \in \mathcal{R}_l^{\mathcal{X}}\}} \sup_{\mathcal{R}_{l,k}^{\mathcal{Y}} \in \mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})} \frac{1}{\mu_X(\mathcal{R}_l^{\mathcal{X}}) |\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}|} \left(\sum_{D_Y \le D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \|G_{D_Y}\|_\infty \right)^2 \left\| \beta_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}} \right\|_\infty^2 \right] \\
&\le \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} |\mathcal{R}_l^{\mathcal{X}}| \sup_{\mathcal{R}_{l,k}^{\mathcal{Y}} \in \mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})} \frac{1}{|\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}|} \left(\sum_{D_Y \le D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \|G_{D_Y}\|_\infty \right)^2 \left\| \beta_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}} \right\|_\infty^2 .
\end{aligned}
$$

Now

$$
\begin{aligned}
\sum_{D_Y \le D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \|G_{D_Y}\|_\infty &= \sum_{D_Y \le D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \prod_{d=1}^{d_Y} \|G_{D_{Y,d}}\|_\infty = \prod_{d=1}^{d_Y} \left(\sum_{D_{Y,d} \le D^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})_{Y,d}} \|G_{D_{Y,d}}\|_\infty \right) \\
&= \prod_{d=1}^{d_Y} \left(\sum_{D_d \le D^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})_{Y,d}} \sqrt{2D_{Y,d}+1} \right) \le \sup_{\mathcal{R}_{l',k'}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y} \left(\sum_{D_{Y,d} \le D^{\mathrm{M}}(\mathcal{R}_{l',k'}^{\mathcal{X},\mathcal{Y}})_{Y,d}} \sqrt{2D_{Y,d}+1} \right)
\end{aligned}
$$

while

$$
\mathcal{D}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} \ge \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \sum_{\mathcal{R}_{l,k}^{\mathcal{Y}} \in \mathcal{P}^{\mathcal{Y}}(\mathcal{R}_l^{\mathcal{X}})} \inf_{\mathcal{R}_{l',k'}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y} \left(D_{Y,k}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}) + 1 \right) \ge \left(\inf_{\mathcal{R}_{l',k'}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y} \left(D_{Y,k}^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}) + 1 \right) \right) \|\mathcal{P}^{\mathcal{X},\mathcal{Y}}\|.
$$

This implies

$$
\frac{\left\| \sum_{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}} \in \mathcal{P}^{\mathcal{X},\mathcal{Y}}} \sum_{D_Y \le D_Y^{\mathrm{M}}(\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}})} \beta_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}} \phi_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}} \right\|_\infty^{2\otimes_n}}{\mathcal{D}_{\mathcal{P}^{\mathcal{X},\mathcal{Y}},D_Y^{\mathrm{M}}} \left\| \beta_{D_Y}^{\mathcal{R}_{l,k}^{\mathcal{X},\mathcal{Y}}} \right\|_\infty^2}
$$

$$\leq \frac{\left(\sup_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y}\left(\sum_{D_{Y,d}\leq D^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'})_{Y,d}}\sqrt{2D_{Y,d}+1}\right)\right)^2}{\inf_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y}\left(D^{\mathrm{M}}_{Y,k}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'})+1\right)} \sum_{\mathcal{R}^{\mathcal{X}}_l\in\mathcal{P}^{\mathcal{X}}} |\mathcal{R}^{\mathcal{X}}_l| \sup_{\mathcal{R}^{\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{Y}}(\mathcal{R}^{\mathcal{X}}_l)} \frac{1}{\|\mathcal{P}^{\mathcal{X},\mathcal{Y}}\||\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}|}$$

$$\leq \left(\frac{\sup_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y}\left(\sum_{D_{Y,d}\leq D^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'})_{Y,d}}\sqrt{2D_{Y,d}+1}\right)}{\inf_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y}\sqrt{D^{\mathrm{M}}_{Y,k}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'})+1}} \sum_{\mathcal{R}^{\mathcal{X}}_l\in\mathcal{P}^{\mathcal{X}}} |\mathcal{R}^{\mathcal{X}}_l| \sup_{\mathcal{R}^{\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{Y}}(\mathcal{R}^{\mathcal{X}}_l)} \frac{1}{\sqrt{\|\mathcal{P}^{\mathcal{X},\mathcal{Y}}\|}\sqrt{|\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}|}}\right)^2$$

$$\leq \left(\frac{\sup_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y}\left(\sum_{D_{Y,d}\leq D^{\mathrm{M}}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'})_{Y,d}}\sqrt{2D_{Y,d}+1}\right)}{\inf_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}} \prod_{d=1}^{d_Y}\sqrt{D^{\mathrm{M}}_{Y,k}(\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l',k'})+1}} \sup_{\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}\in\mathcal{P}^{\mathcal{X},\mathcal{Y}}} \frac{1}{\sqrt{\|\mathcal{P}^{\mathcal{X},\mathcal{Y}}\|}\sqrt{|\mathcal{R}^{\mathcal{X},\mathcal{Y}}_{l,k}|}}\right)^2.$$

The proposition is then obtained by a simple application of Proposition 7$\|\cdot\|_2$ and $\|\cdot\|_\infty$ structures. □

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

E.1 Model coding

Proof of Proposition 10Spatial Gaussian mixture density estimation theorem. This proposition is a simple combination of Theorem 2Conditional density estimation theorem, of classical Kraft type inequalities for order selection and variable selection (see for instance in the book of Massart [28]):

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, $E=\mathrm{span}\{e_i\}_{i\in I}$ with $I=\{1,\ldots,p_E\}$, 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, $E=\mathrm{span}\{e_i\}_{i\in I}$ with $I\subset\{1,\ldots,p\}$, 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 $[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]^K$ and $[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]$. Using that there is at most $3\times3\times3\times3$ different type of models $[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]^K$ and $2\times2\times2\times2$ different type of models $[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]$, and $3^4\times2^4=1296$, we obtain

$$\sum_{S_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}}\in\mathcal{S}} e^{-x_{K,\mathcal{P}^{\mathcal{X}},\mathcal{F}}} = \sum_{K\in\mathbb{N}^*}\sum_{\mathcal{P}^{\mathcal{X}}\in\mathcal{S}^\star_{\mathcal{P}}}\sum_{E}\sum_{[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]^K}\sum_{[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]} e^{-c_\star\left(A_0^\star(\mathcal{X})+B_0^\star(\mathcal{X})\|\mathcal{P}^{\mathcal{X}}\|+(K-1)+\theta_E\right)}$$

$$= \left(\sum_{K \in \mathbb{N}^*} e^{-c_\star (K-1)} \right) \left(\sum_{\mathcal{P}^{\mathcal{X}} \in \mathcal{S}^\star_{\mathcal{P}}} e^{-c_\star \left(A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})} \| \mathcal{P}^{\mathcal{X}} \| \right)} \right)$$
$$\times \left(\sum_E e^{-c_\star \theta_E} \right) \left(\sup_{K \in \mathbb{N}^*} \sum_{[\mu_\star \, \mathrm{L}_\star \, \mathrm{D}_\star \, \mathrm{A}_\star]^K} \sum_{[\mu_\star \, \mathrm{L}_\star \, \mathrm{D}_\star \, \mathrm{A}_\star]} \right)$$
$$\leq 1296 \frac{1}{1 - e^{-c_\star}} \Sigma_0^\star e^{-c_\star C_0^\star} \begin{cases} 1 & \text{if } E \text{ is known,} \\ \frac{1}{e^c_\star - 1} & \text{if } E \text{ is chosen amongst spaces spanned by the first coordinates,} \\ 2e^{-(c_\star - 1)(1 + \ln 2)} & \text{if } E \text{ is free.} \end{cases}$$

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

E.2 Entropy of spatial mixtures

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

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

We say that $[(s_1, \ldots, s_K), (t_1, \ldots, t_K)]$ is a bracket of $\mathcal{F}_{E,K}$ 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 $\delta \in (0, \sqrt{2}]$,*

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

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

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

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

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

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

proved in Genovese and Wasserman [18] 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 14Entropy of Gaussian families (an extended version of Proposition 12Bracketing entropy of Gaussian families) handles the bracketing entropy of Gaussian K-uples collection. It implies the existence of two constants $\mathcal{C}_{[\star]^\star}$ and $\mathcal{C}_{[\star]}$ depending only on a, L_m, L_M, λ_m and λ_M such that

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

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

As $\dim(S_{K,\mathcal{P},\mathcal{F}}) = \|\mathcal{P}^{\mathcal{X}}\|(K-1) + \dim(\mathcal{F}_{E,K}) + \dim(\mathcal{F}_{E^\perp})$, we obtain Proposition 11Spatial Gaussian mixture density estimation theorem with $C = \max(\mathcal{C}_{\mathcal{S}}, \mathcal{C}_{[\star]^\star}, \mathcal{C}_{[\star]})$. □

E.3 Entropy of Gaussian families

Instead of Proposition 12Bracketing entropy of Gaussian families, we prove this extended version

Proposition 14. *Let* $\kappa \geq \frac{3}{4}$ *and* $\gamma_\kappa = \min\left(\frac{\kappa - \frac{3}{4}}{2(1+\frac{2\kappa}{9})(1+\frac{2}{9})(1+\frac{1}{9})}, \frac{3(\kappa-\frac{1}{2})}{2(1+\frac{2\kappa}{9})(1+\frac{2}{9})^3} \right)$.

$$\textit{Assume} \qquad \begin{cases} a \geq \frac{\sqrt{\gamma_\kappa}}{18\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}} \sqrt{L_m \lambda_m \frac{\lambda_m}{\lambda_M}} \frac{\delta}{p_E} \\ \ln\left(\frac{L_M}{L_m}\right) \geq \frac{1}{20\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}} \delta \\ \frac{\lambda_M}{\lambda_m} \ln\left(\frac{\lambda_M}{\lambda_m}\right) \geq \frac{1}{27(1+\frac{2}{9})(\sqrt{2}+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}} \frac{\delta}{p_E} \end{cases}$$

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

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

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

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

$$\begin{cases} \mathcal{D}_{\mu,p_E} = p_E \\ \mathcal{D}_{\mathrm{L}} = 1 \\ \mathcal{D}_{\mathrm{D},p_E} = \frac{p_E(p_E-1)}{2} \\ \mathcal{D}_{\mathrm{A},p_E} = p_E - 1 \end{cases} \quad \textit{and} \quad \begin{cases} \mathcal{I}_{\mu,p_E} = p_E\left(\ln\left(\frac{36a\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}p_E}{\sqrt{\gamma_\kappa L_m \lambda_m \frac{\lambda_m}{\lambda_M}}}\right)\right) \\ \mathcal{I}_{\mathrm{L},p_E} = \ln\left(40\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}\ln\left(\frac{L_M}{L_m}\right)p_E\right) \\ \mathcal{I}_{\mathrm{D},p_E} = \frac{p_E(p_E-1)}{2}\left(\frac{\ln c}{\frac{p_E(p_E-1)}{2}} + \left(\ln\left(36(1+\frac{2}{9})(\sqrt{2}+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}\frac{\lambda_M}{\lambda_m}p_E\right)\right)\right) \\ \mathcal{I}_{\mathrm{A},p_E} = (p_E-1)\left(\ln\left(108(1+\frac{2}{9})(\sqrt{2}+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}\frac{\lambda_M}{\lambda_m}\ln\left(\frac{\lambda_M}{\lambda_m}\right)p_E\right)\right) \end{cases}$$

Furthermore, for any $p_E \leq p$

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

with

$$\mathcal{C}_{\mu,p} = \ln\left(\frac{36a\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}p}{\sqrt{\gamma_\kappa L_m \lambda_m \frac{\lambda_m}{\lambda_M}}}\right)$$

$$\mathcal{C}_{\mathrm{L},p} = \ln\left(40\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}\ln\left(\frac{L_M}{L_m}\right)p\right)$$
$$\mathcal{C}_{\mathrm{D},p} = \left(\ln c + \left(\ln\left(36(1+\frac{2}{9})(\sqrt{2}+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}\frac{\lambda_M}{\lambda_m}p\right)\right)\right)$$
$$\mathcal{C}_{\mathrm{A},p} = \ln\left(108(1+\frac{2}{9})(\sqrt{2}+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}\frac{\lambda_M}{\lambda_m}\ln\left(\frac{\lambda_M}{\lambda_m}\right)p\right)$$

and, uniformly over K,

$$\begin{aligned}\mathcal{I}_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_{p_E}^K} \leq \max_{\mu'_\star,\mathrm{L}_\star,\mathrm{D}'_\star,\mathrm{A}'_\star,K'} \Bigg(&\mathcal{C}_{\mu,p}\frac{c_{\mu'_\star}K'}{c_{\mu'_\star}K'+c_{\mathrm{L}'_\star}+c_{\mathrm{D}'_\star}\frac{K'(K'-1)}{2}+c_{\mathrm{A}'_\star}(K'-1)}\\ &+\mathcal{C}_{\mathrm{L},p}\frac{c_{\mathrm{L}'_\star}}{c_{\mu'_\star}K'+c_{\mathrm{L}'_\star}+c_{\mathrm{D}'_\star}\frac{K'(K'-1)}{2}+c_{\mathrm{A}'_\star}(K'-1)}\\ &+\mathcal{C}_{\mathrm{D},p}\frac{c_{\mathrm{D}'_\star}\frac{K'(K'-1)}{2}}{c_{\mu'_\star}K'+c_{\mathrm{L}'_\star}+c_{\mathrm{D}'_\star}\frac{K'(K'-1)}{2}+c_{\mathrm{A}'_\star}(K'-1)}\\ &+\mathcal{C}_{\mathrm{A},p}\frac{c_{\mathrm{A}'_\star}(K'-1)}{c_{\mu'_\star}K'+c_{\mathrm{L}'_\star}+c_{\mathrm{D}'_\star}\frac{K'(K'-1)}{2}+c_{\mathrm{A}'_\star}(K'-1)}\Bigg)\mathcal{D}_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_{p_E}^K}\\ \leq \max(\mathcal{C}_{\mu,p},\mathcal{C}_{L,p},\mathcal{C}_{D,p},\mathcal{C}_{A,p})&\mathcal{D}_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_{p_E}^K}\end{aligned}$$

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

Proof of Proposition 14Entropy of Gaussian families. We consider all models $\mathcal{F}_{[\mu_\star\,\mathrm{L}_\star\,\mathrm{A}_\star\,\mathrm{D}_\star]_E^K}$ at once by a "tensorial" construction of the $\delta/9$ bracket collection.

We define first a set of grids

- for any δ_μ, the grid $\mathcal{G}_\mu(a,p_E,\delta_\mu)$ of $[-a,a]^{p_E}$:
$$\mathcal{G}_\mu(a,p_E,\delta_\mu) = \left\{g\delta_\mu \middle| g\in\mathbb{Z}^{p_E}, \|g\|_\infty \leq \frac{a}{\delta_\mu}\right\}$$
- for any δ_L, the grid $\mathcal{G}_\mathrm{L}(L_m,L_M,\delta_\mathrm{L})$ of $[L_m,L_M]$:
$$\mathcal{G}_\mathrm{L}(L_m,L_M,\delta_\mathrm{L}) = \{L_m(1+\delta_\mathrm{L})^g | g\in\mathbb{N}, L_m(1+\delta_\mathrm{L})^g \leq L_M\}$$
- For any δ_D, the grid $\mathcal{G}_\mathrm{D}(p_E,\delta_\mathrm{D})$ of $SO(p_E)$ made of the elements of a δ_D-net with respect to the $\|\cdot\|_2$ operator norm (as described by Szarek [33]).
- for any δ_A, the grid $\mathcal{G}_\mathrm{A}(\lambda_m,\lambda_M,p_E,\delta_\mathrm{A})$ of $\mathcal{A}(\lambda_m,\lambda_M(1+\delta_\mathrm{A}),p_E)$:
$$\mathcal{G}_\mathrm{A}(\lambda_m,\lambda_M,p_E,\delta_\mathrm{A}) = \{A\in\mathcal{A}(\lambda_m,\lambda_M(1+\delta_\mathrm{A}),p_E) | \forall 1\leq i<p_E, \exists g_i\in\mathbb{N}, A_i = \lambda_m(1+\delta_\mathrm{A})^{g_i}\}.$$

Obviously, for any $\mu\in[-a,a]$, there is a $\tilde{\mu}\in\mathcal{G}_\mu(a,p_E,\delta_\mu)$ such that
$$\|\tilde{\mu}-\mu\|^2 \leq p_E\delta_\mu^2$$
while
$$|\mathcal{G}_\mu(a,p_E,\delta_\mu)| \leq \left(1+2\frac{a}{\delta_\mu}\right)^{p_E} \leq \max\left(2^{p_E},\left(\frac{4a}{\delta_\mu}\right)^{p_E}\right).$$

In the same fashion, for any L in $[L_m, L_M]$, there is a $\tilde{L} \in \mathcal{G}_{\mathrm{L}}(L_m, L_M, \delta_{\mathrm{L}})$ such that $(1+\delta_{\mathrm{L}})^{-1} L_{j_L} < L \leq L_{j_L}$ while

$$|\mathcal{G}_{\mathrm{L}}(L_m, L_M, \delta_{\mathrm{L}})| \leq 1 + \frac{\ln\left(\frac{L_M}{L_m}\right)}{\ln(1+\delta_{\mathrm{L}})}.$$

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

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

By definition on a δ_{D}-net, for any $D \in SO(p_E)$ there is a $\tilde{D} \in \mathcal{G}_{\mathrm{D}}(p_E, \delta_{\mathrm{D}})$ such that

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

As proved by Szarek [33], it exists a universal constant c_S such that, as soon as $\delta_{\mathrm{D}} \leq 1$

$$|\mathcal{G}_{\mathrm{D}}(p_E, \delta_{\mathrm{D}})| \leq c\left(\frac{1}{\delta_{\mathrm{D}}}\right)^{\frac{p_E(p_E-1)}{2}}$$

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

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

$$|\mathcal{G}_{\mathrm{A}}(\lambda_m, \lambda_M, p_E, \delta_{\mathrm{A}})| \leq \left(2 + \frac{\ln\left(\frac{\lambda_M}{\lambda_m}\right)}{\ln(1+\delta_{\mathrm{A}})}\right)^{p_E-1}$$

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

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

Now we use

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

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

Let

$$\begin{cases} c_{\mu_0} = c_{\mathrm{L}_0} = c_{\mathrm{D}_0} = c_{\mathrm{A}_0} = 0 \\ c_{\mu_K} = c_{\mathrm{L}_K} = c_{\mathrm{D}_K} = c_{\mathrm{A}_K} = K \\ c_{\mu} = c_{\mathrm{L}} = c_{\mathrm{D}} = c_{\mathrm{A}} = 1 \end{cases}$$

Define $f_{K,\mu_\star,p_E}$ as the application from $(\mathbb{R}^{p_E})^{c_{\mu_\star}}$ to $\mathbb{R}^K$ defined by

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

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

By definition, the image of

$$([-a,a]^{p_E})^{c_{\mu_\star}} \times ([L_m,L_M])^{c_{\mathrm{L}_\star}} \times (SO(p_E))^{c_{\mathrm{D}_\star}} \times (\mathcal{A}(\lambda_m,\lambda_M,p_E))^{c_{\mathrm{A}_\star}}$$

by $\Psi_K \circ \Gamma_K \circ \left(f_{K,\mu_\star,p_E} \otimes f_{L_{K,\cdot},p_E} \otimes f_{K,\mathrm{D}_\star,p_E} \otimes f_{K,\mathrm{A}_\star}\right)$ is the set $\mathcal{F}_{[\mu_\star\ \mathrm{L}_\star\ \mathrm{D}_\star\ \mathrm{A}_\star]^K}$ of all the K-uples of Gaussian densities of type $[\mu_\star\,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]^K$.

We define now for any δ_Σ the application B_{K,δ_Σ} that maps $((\mu_1,\Sigma_1),\ldots,(\mu_K,\Sigma_K))$ into the K-uple of couples

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

Lemma 10. *For any mean μ and any full rank covariance matrix Σ in a space of dimension p_E, for any $0<\delta\le\sqrt{2}$, let $\kappa\ge\frac{1}{2}$ and $\delta_\Sigma\le\frac{1}{9\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}}\frac{\delta}{p_E}$, let*

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

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

shows that, as soon as $\Sigma_1,\ldots,\Sigma_K$ are full rank, for any $\kappa\ge\frac{1}{2}$, if $\delta_\Sigma=\frac{1}{9\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}}\frac{\delta}{p_E}$, then

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

is a $\delta/9$-bracket for the $d^{\max}$ norm.

We rely now on the much more involved

Lemma 11. *Let $\kappa\ge\frac{3}{4}$, $\gamma_\kappa=\min\left(\frac{\kappa-\frac{3}{4}}{2(1+\frac{2\kappa}{9})(1+\frac{2}{9})(1+\frac{1}{9})},\frac{3(\kappa-\frac{1}{2})}{2(1+\frac{2\kappa}{9})(1+\frac{2}{9})^3}\right)$. For any $0<\delta\le\sqrt{2}$, any $p_E\ge1$ and any $\delta_\Sigma\le\frac{1}{9\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}}\frac{\delta}{p_E}$,*

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

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

If

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

then $[t^-,t^+]$ is an $\delta/9$ Hellinger bracket such that $t^-(x)\le\Phi_{\mu,\Sigma}(x)\le t^+(x)$.

and on the definition of $d^{\max}$ to obtain that as soon as $\kappa>\frac{3}{4}$ the choice

$$\begin{cases}\delta_\mu=\sqrt{\gamma_\kappa L_m\lambda_m\frac{\lambda_m}{\lambda_M}}\delta_\Sigma=\frac{\sqrt{\gamma_\kappa L_m\lambda_m\frac{\lambda_m}{\lambda_M}}}{9\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}}\frac{\delta}{p_E}\\ \delta_{\mathrm{L}}=\frac{1}{2}\delta_\Sigma=\frac{1}{18\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}}\frac{\delta}{p_E}\\ \delta_{\mathrm{D}}=\delta_{\mathrm{A}}=\frac{1}{4(1+\frac{2}{9})(\sqrt{2}+1)}\frac{\lambda_m}{\lambda_M}\delta_\Sigma=\frac{1}{36(1+\frac{2}{9})(\sqrt{2}+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac{1}{4}}}\frac{\lambda_m}{\lambda_M}\frac{\delta}{p_E}\end{cases}$$

is such that the image of

$$(\mathcal{G}_\mu(a,p_E,\delta_\mu))^{c_{\mu_\star}}\times(\mathcal{G}_{\mathrm{L}}(L_m,L_M,\delta_{\mathrm{L}}))^{c_{\mathrm{L}_\star}}\times(\mathcal{G}_{\mathrm{D}}(p_E,\delta_{\mathrm{D}}))^{c_{\mathrm{D}_\star}}\times(\mathcal{G}_{\mathrm{A}}(\lambda_m,\lambda_M,p_E,\delta_{\mathrm{A}}))^{c_{\mathrm{A}_\star}}$$

by $B_{K,\delta}\circ\Gamma_K\circ\left(f_{K,\mu_\star,p_E}\otimes f_{L_{K,\cdot},p_E}\otimes f_{K,\mathrm{D}_\star,p_E}\otimes f_{K,\mathrm{A}_\star}\right)$ is a $\delta/9$-bracket covering of $\mathcal{F}_{[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]_E^K}$ for the $d^{\max}$ norm.

Its cardinality is bounded by

$$\begin{aligned}
&\left(\max\left(2^{p_E},\left(\frac{4a}{\frac{\sqrt{\gamma_\kappa L_m\lambda_m\frac{\lambda_m}{\lambda_M}}}{9\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}}\frac{\delta}{p_E}}\right)^{p_E}\right)\right)^{c_{\mu_\star}}\times\left(\max\left(2,\frac{20\ln\left(\frac{L_M}{L_m}\right)}{9\frac{1}{18\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}}\frac{\delta}{p_E}}\right)\right)^{c_{\mathrm{L}_\star}}\\
&\quad\times\left(c\left(\frac{1}{\frac{1}{36(1+\frac29)(\sqrt2+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}}\frac{\lambda_m}{\lambda_M}\frac{\delta}{p_E}}\right)^{\frac{p_E(p_E-1)}{2}}\right)^{c_{\mathrm{D}_\star}}\\
&\quad\times\left(\max\left(4^{p_E-1},\left(\frac{74\ln\left(\frac{\lambda_M}{\lambda_m}\right)}{36\frac{1}{36(1+\frac29)(\sqrt2+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}}\frac{\lambda_m}{\lambda_M}\frac{\delta}{p_E}}\right)^{p_E-1}\right)\right)^{c_{\mathrm{A}_\star}}\\
&=\left(\max\left(2^{p_E},\left(\frac{36a\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}p_E}{\sqrt{\gamma_\kappa L_m\lambda_m\frac{\lambda_m}{\lambda_M}}\delta}\right)^{p_E}\right)\right)^{c_{\mu_\star}}\times\left(\max\left(2,\frac{40\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}\ln\left(\frac{L_M}{L_m}\right)p_E}{\delta}\right)\right)^{c_{\mathrm{L}_\star}}\\
&\quad\times\left(c\left(\frac{36(1+\frac29)(\sqrt2+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}\frac{\lambda_M}{\lambda_m}p_E}{\delta}\right)^{\frac{p_E(p_E-1)}{2}}\right)^{c_{\mathrm{D}_\star}}\\
&\quad\times\left(\max\left(4^{p_E-1},\left(\frac{108(1+\frac29)(\sqrt2+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}\frac{\lambda_M}{\lambda_m}\ln\left(\frac{\lambda_M}{\lambda_m}\right)p_E}{\delta}\right)^{p_E-1}\right)\right)^{c_{\mathrm{A}_\star}}
\end{aligned}$$

So that under the mild assumptions that

$$\begin{cases}a\geq\frac{\sqrt{\gamma_\kappa}}{18\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}}\sqrt{L_m\lambda_m\frac{\lambda_m}{\lambda_M}}\frac{\delta}{p_E}\\ \ln\left(\frac{L_M}{L_m}\right)\geq\frac{1}{20\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}}\delta\\ \frac{\lambda_M}{\lambda_m}\ln\left(\frac{\lambda_M}{\lambda_m}\right)\geq\frac{1}{27(1+\frac29)(\sqrt2+1)\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}}\frac{\delta}{p_E}\end{cases},$$

$$\begin{aligned}
&H_{[\cdot],d^{\max}}(\delta/9,\mathcal{F}_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_E^K})\\
&\quad\leq c_{\mu_\star}p_E\left(\ln\left(\frac{36a\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}p_E}{\sqrt{\gamma_\kappa L_m\lambda_m\frac{\lambda_m}{\lambda_M}}}\right)+\ln\frac{1}{\delta}\right)\\
&\qquad+c_{\mathrm{L}_\star}\left(\ln\left(40\sqrt{\kappa^2\cosh(\frac{2\kappa}{9})+\frac14}\ln\left(\frac{L_M}{L_m}\right)p_E\right)+\ln\frac{1}{\delta}\right)
\end{aligned}$$

$$+ c_{\mathrm{D}_\star} \frac{p_E(p_E-1)}{2} \left(\frac{\ln c}{\frac{p_E(p_E-1)}{2}} + \left(\ln \left(36(1+\frac{2}{9})(\sqrt{2}+1) \sqrt{\kappa^2 \cosh(\frac{2\kappa}{9}) + \frac{1}{4}} \frac{\lambda_M}{\lambda_m} p_E \right) + \ln \frac{1}{\delta} \right) \right)$$

$$+ c_{\mathrm{A}_\star} (p_E - 1) \left(\ln \left(108(1+\frac{2}{9})(\sqrt{2}+1) \sqrt{\kappa^2 \cosh(\frac{2\kappa}{9}) + \frac{1}{4}} \frac{\lambda_M}{\lambda_m} \ln \left(\frac{\lambda_M}{\lambda_m} \right) p_E \right) + \ln \frac{1}{\delta} \right)$$

which proves Proposition 12Bracketing entropy of Gaussian families. □

E.4 Entropy of spatial mixtures (Lemmas)

Proof of Lemma 7Entropy of spatial mixtures. This a variation around the proof of Genovese and Wasserman [18].

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

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

be a minimal covering of $\delta/9$ sup norm Hellinger bracket of $\mathcal{F}_{E,K}$ and $\{ \left[t_{E^\perp,1}^-, t_{E^\perp,1}^+ \right], \ldots, \left[t_{E_\perp,N_\perp}^-, t_{E_\perp,N_{E_\perp}}^+ \right] \}$ be a minimal covering of $\delta/9$ Hellinger bracket of $\mathcal{F}_{E^\perp}$. By definition, $\ln N_{\mathcal{S}_{K-1}} = H_{[\cdot],d}(\delta/3, \mathcal{S}_{K-1})$, $\ln N_{E,K} = H_{[\cdot],d^{\max}}(\delta/9, \mathcal{F}_{E,K})$ and $\ln N_{E^\perp} = H_{[\cdot],d}(\delta/9, \mathcal{F}_{E^\perp})$.

By construction,

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

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

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

Using

Lemma 12. *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 $\left[t_{k,j,l}^{--}, t_{k,j,l}^{++} \right]$ the corresponding $\delta/3$ Hellinger bracket.

By definition,

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

$$\leq \sup_{i,j,l} d^2 \left(\sum_{k=1}^{K} \pi^{-}_{i,k}\, t^{--}_{k,j,l}, \sum_{k=1}^{K} \pi^{+}_{i,k}\, t^{++}_{k,j,l} \right)$$

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

Lemma 13. *For any brackets $[t^-(x), t^+(x)]$ and if for any x $[u^-(x,y), u^+(x,y)]$ is a bracket then*

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

to obtain

$$d^{2\sup} \left(\sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \left(\sum_{k=1}^{K} \pi^{-}_{i[\mathcal{R}_l^{\mathcal{X}}],k}\, t^{--}_{k,j,l}\,(y) \right) \mathbf{1}_{\{x \in \mathcal{R}_l^{\mathcal{X}}\}}, \sum_{\mathcal{R}_l^{\mathcal{X}} \in \mathcal{P}^{\mathcal{X}}} \left(\sum_{k=1}^{K} \pi^{+}_{i[\mathcal{R}_l^{\mathcal{X}}],k}\, t^{++}_{k,j,l}\,(y) \right) \mathbf{1}_{\{x \in \mathcal{R}_l^{\mathcal{X}}\}} \right)$$
$$\leq \sup_{i,j,l} d_{k,y}^2 \left(\pi^{-}_{i,k}\, t^{--}_{k,j,l}(y), \pi^{+}_{i,k}\, t^{++}_{k,j,l}(y) \right)$$

and then using again Lemma 12Entropy of spatial mixtures (Lemmas)

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

□

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

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

Using

Lemma 14. *For any δ Hellinger bracket $[t^-, t^+]$, $\int t^- \mathrm{d}\lambda \leq 1$ and $\int t^+ \mathrm{d}\lambda \leq 1 + 2(\sqrt{2} + \sqrt{3})\delta$.*

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

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

$$\leq 9\delta^2$$

□

Proof of Lemma 13Entropy of spatial mixtures (Lemmas).

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

□

Proof of Lemma 14Entropy of spatial mixtures (Lemmas). The first point is straightforward as t^- is upper-bounded by a density.

For the second point,

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

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

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

□

E.5 Entropy of Gaussian families (Lemma)

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

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

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

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

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

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

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

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

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

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

and thus

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

which implies

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

Thus the diagonal matrix $\tilde{A}$ defined by

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

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

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

which implies

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

and thus

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

Along the same lines,

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

thus

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

and

$$\left|\tilde{A}^{-1}_{p_E,p_E} - A^{-1}_{p_E,p_E}\right| \leq \tilde{A}^{-1}_{p_E,p_E}\delta_{\mathrm{A}} \leq \lambda_m^{-1}\delta_{\mathrm{A}}.$$

□

Proof of Lemma 10Entropy of Gaussian families. As $(1+\delta_\Sigma)\Sigma^{-1} - (1+\delta_\Sigma)^{-1}\Sigma^{-1} = \left((1+\delta_\Sigma) - (1+\delta_\Sigma)^{-1}\right)\Sigma^{-1}$ is a positive definite matrix, one can thus apply

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 such that $\Sigma_1^{-1} - \Sigma_2^{-1}$ is a positive definite matrix, for any $x \in \mathbb{R}^{p_E}$*

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

proved by Maugis and Michel [29]. This yields using eventually $\kappa \geq \frac{1}{2}$

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

Concerning the Hellinger width,

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

Using

Lemma 16. *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,*

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

also proved in [29], we derive

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

So that combining

Lemma 17. *For any $0 < \delta \leq \sqrt{2}$ and any $p_E \geq 1$, let $\kappa \geq \frac{1}{2}$ and $\delta_\Sigma \leq \frac{1}{9\sqrt{\kappa^2 \cosh(\frac{2\kappa}{9}) + \frac{1}{4}}} \frac{\delta}{p_E}$, then*

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

and

Lemma 18. *For any $d \in \mathbb{N}$, for any $\delta_\Sigma > 0$,*

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

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

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

□

Proof of Lemma 17Entropy of Gaussian families (Lemma). A straightforward computation yields

$$\delta_\Sigma \leq \frac{1}{9\sqrt{\kappa^2 \cosh(\frac{2\kappa}{9}) + \frac{1}{4}}} \frac{\delta}{p_E} \leq \frac{1}{p_E} \frac{\sqrt{2}}{9\sqrt{\frac{1}{4} + \frac{1}{4}}} = \frac{1}{p_E} \frac{2}{9} \leq \frac{2}{9}$$

□

Proof of Lemma 18Entropy of Gaussian families (Lemma).

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

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 $\cosh(x) >= 1$, 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

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

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

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

Now,

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

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

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

□

Proof of Lemma 11Entropy of Gaussian families. As $\tilde{\Sigma}$ is a full rank matrix by construction, Lemma 10Entropy of Gaussian families on "Gaussian" brackets applies and $[t^-, t^+]$ is an $\delta/9$ Hellinger bracket. Using

Lemma 19. *Under the Assumptions of Lemma 11Entropy 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*

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

we can apply Lemma 15Entropy 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

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

Now

$$\left((1+\delta_\Sigma)\tilde{\Sigma}-\Sigma\right)^{-1} = \left((1+\delta_\Sigma)\tilde{\Sigma}\left(\Sigma^{-1}-(1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1}\right)\Sigma\right)^{-1} = (1+\delta_\Sigma)^{-1}\Sigma^{-1}\left(\Sigma^{-1}-(1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1}\right)^{-1}\tilde{\Sigma}^{-1}$$

and thus

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

Now as by construction,

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

one obtains

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

It is thus sufficient to prove that

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

or equivalently

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

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

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

which implies thus

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

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

The second case is handled in the same way.

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

Now as

$$\left(\Sigma-(1+\delta_\Sigma)^{-1}\tilde{\Sigma}\right)^{-1} = \left(\Sigma\left((1+\delta_\Sigma)\tilde{\Sigma}^{-1}-\Sigma^{-1}\right)(1+\delta_\Sigma)^{-1}\tilde{\Sigma}\right)^{-1} = (1+\delta_\Sigma)\tilde{\Sigma}^{-1}\left((1+\delta_\Sigma)\tilde{\Sigma}^{-1}-\Sigma^{-1}\right)^{-1}\Sigma^{-1}$$

and thus

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

one deduces

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

All we need to prove is thus

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

or equivalently

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

Let

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

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

$$f_2(\delta_\Sigma) > \frac{\kappa - \frac{1}{2}}{(1+\frac{2\kappa}{9})(1+\frac{2}{9})}\delta_\Sigma \geq \frac{2(1+\frac{2}{9})}{3}\gamma_\kappa(1+\frac{2}{9})\delta_\Sigma \geq \frac{2(1+\frac{2}{9})}{3}\gamma_\kappa(1+\delta_\Sigma)\delta_\Sigma$$

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 19Entropy of Gaussian families (Lemma). We deduce this result from the slightly more general:

Lemma 20. *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

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

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

$$\forall x \in \mathbb{R}^{p_E}, x'\left((1+\delta_\Sigma)\tilde{\Sigma}^{-1} - \Sigma^1\right)x \geq \tilde{L}^{-1}\left((\delta_\Sigma - \delta_{\mathrm{L}})\lambda_M^{-1} - (1+\delta_\Sigma)\lambda_m^{-1}\left(\sqrt{2}\delta_{\mathrm{D}} + \delta_{\mathrm{A}}\right)\right)\|x\|^2$$
$$\forall x \in \mathbb{R}^{p_E}, x'\left(\Sigma^{-1} - (1+\delta_\Sigma)^{-1}\right)x \geq \frac{\tilde{L}^{-1}}{1+\delta_\Sigma}\left(\delta_\Sigma\lambda_M^{-1} - \lambda_m^{-1}\left(\sqrt{2}\delta_{\mathrm{D}} + \delta_{\mathrm{A}}\right)\right)\|x\|^2$$

Indeed Lemma 17Entropy of Gaussian families (Lemma) ensures that $\delta_\Sigma \leq \frac{2}{9}$ and if we let $\delta_{\mathrm{L}} = \frac{1}{2}\delta_\Sigma$ and $\delta_{\mathrm{A}} = \delta_{\mathrm{D}} = \frac{1}{4(1+\frac{2}{9})(\sqrt{2}+1)}\frac{\lambda_m}{\lambda_M}\delta_\Sigma$, the bounds of the previous Lemma becomes

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

$$\geq \frac{1}{4}\tilde{L}^{-1}\frac{1}{\lambda_M}\delta_\Sigma\|x\|^2$$

while

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

□

Proof of Lemma 20Entropy of Gaussian families (Lemma). By definition,

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

Along the same lines,

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

Now

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

$$\leq \lambda_m^{-1} \left(\sum_{i=1}^{p_E} \left| (\tilde{D}_i - D_i)' x \right|^2 \right)^{1/2} \left(\sum_{i=1}^{p_E} \left| (\tilde{D}_i + D_i)' x \right|^2 \right)^{1/2}$$
$$\leq \lambda_m^{-1} \delta_{\mathrm{D}} \|x\| \sqrt{2} \|x\| = \lambda_m^{-1} \sqrt{2} \delta_{\mathrm{D}} \|x\|^2.$$

Furthermore,

$$\left| \sum_{i=1}^{p_E} \tilde{A}_{i,i}^{-1} |D_i' x|^2 - \sum_{i=1}^{p_E} A_{i,i}^{-1} |D_i' x|^2 \right| \leq \sum_{i=1}^{p_E} \left| \tilde{A}_{i,i}^{-1} - A_{i,i}^{-1} \right| |D_i' x|^2$$
$$\leq \delta_{\mathrm{A}} \lambda_m^{-1} \sum_{i=1}^{p_E} |D_i' x|^2 = \delta_{\mathrm{A}} \lambda_m^{-1} \|x\|^2.$$

We notice then that

$$(1+\delta_\Sigma) \tilde{L}^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D_i' x|^2 - L^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D_i' x|^2 = \left((1+\delta_\Sigma) \tilde{L}^{-1} - L^{-1} \right) \sum_{i=1}^{p_E} A_{i,i}^{-1} |D_i' x|^2$$
$$\geq (\delta_\Sigma - \delta_{\mathrm{L}}) \tilde{L}^{-1} \lambda_M^{-1} \|x\|^2$$

while

$$L^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D_i' x|^2 - (1+\delta_\Sigma)^{-1} \tilde{L}^{-1} \sum_{i=1}^{p_E} A_{i,i}^{-1} |D_i' x|^2 = \left(L^{-1} - (1+\delta_\Sigma)^{-1} \tilde{L}^{-1} \right) \sum_{i=1}^{p_E} A_{i,i}^{-1} |D_i' x|^2$$
$$\geq \left(1 - (1+\delta_\Sigma)^{-1} \right) L_{j,L}^{-1} \lambda_M^{-1} \|x\|^2$$
$$\geq \frac{\delta_\Sigma}{1+\delta_\Sigma} \lambda_M^{-1} \|x\|^2$$

We deduce thus that

$$x' \left((1+\delta_\Sigma) \tilde{\Sigma}^{-1} - \Sigma^{-1} \right) x \geq (\delta_\Sigma - \delta_{\mathrm{L}}) \tilde{L}^{-1} \lambda_M^{-1} \|x\|^2 - (1+\delta_\Sigma) \tilde{L}^{-1} \lambda_m^{-1} \left(\sqrt{2} \delta_{\mathrm{D}} + 2\delta_{\mathrm{A}} \right) \|x\|^2$$
$$\geq \tilde{L}^{-1} \left((\delta_\Sigma - \delta_{\mathrm{L}}) \lambda_M^{-1} - (1+\delta_\Sigma) \lambda_m^{-1} \left(\sqrt{2} \delta_{\mathrm{D}} + \delta_{\mathrm{A}} \right) \right) \|x\|^2$$

and

$$x' \left(\Sigma^{-1} - (1+\delta_\Sigma)^{-1} \tilde{\Sigma}^{-1} \right) x \geq \frac{\delta_\Sigma}{1+\delta_\Sigma} \lambda_M^{-1} \|x\|^2 - (1+\delta_\Sigma)^{-1} \tilde{L}^{-1} \lambda_m^{-1} \left(\sqrt{2} \delta_{\mathrm{D}} + \delta_{\mathrm{A}} \right) \|x\|^2$$
$$\geq \frac{\tilde{L}^{-1}}{1+\delta_\Sigma} \left(\delta_\Sigma \lambda_M^{-1} - \lambda_m^{-1} \left(\sqrt{2} \delta_{\mathrm{D}} + \delta_{\mathrm{A}} \right) \right) \|x\|^2$$

□

References

[1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716 – 723, 1974.

[2] N. Akakpo. Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. Submitted, 2010.

[3] N. Akakpo and C. Lacour. Inhomogeneous and anisotropic conditional density estimation from dependent data. Submitted, 2011.

[4] 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.

[5] 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.

[6] D. Bashtannyk and R. Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3):279 – 298, 2001.

[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] G. Blanchard, C. Schäfer, Y. Rozenholc, and K.R. Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2):209–241, 2007.

[10] S. Boucheron and P. Massart. A high-dimensional wilks phenomenon. *Probability Theory and Related Fields*, pages 1–29, 2010.

[11] E. Brunel, F. Comte, and C. Lacour. Adaptive estimation of the conditional density in presence of censoring. *Sankhyā*, 69(4):734–763, 2007.

[12] S. Cohen and E. Le Pennec. Unsupervised hyperspectral image segmentation with spatial Gaussian mixture. In Preparation, 2011.

[13] J. de Gooijer and D. Zerom. On conditional density estimation. *Statist. Neerlandica*, 57(2): 159–176, 2003.

[14] D. Donoho. CART and best-ortho-basis: a connection. *Ann. Statist.*, 25(5):1870–1911, 1997.

[15] S. Efromovich. Conditional density estimation in a regression setting. *Ann. Statist.*, 35(6): 2504–2535, 2007.

[16] S. Efromovich. Oracle inequality for conditional density estimation and an actuarial example. *Annals of the Institute of Statistical Mathematics*, 62:249–275, 2010.

[17] J. Fan, Q. Yao, and H. Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996.

[18] Ch. Genovese and L. Wasserman. Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.*, 28(4):1105–1127, 2000.

[19] L. Györfi and M. Kohler. Nonparametric estimation of conditional distributions. *IEEE Trans. Information Theory*, 53:1872–1879, 2007.

[20] P. Hall, R. Wolff, and Q. Yao. Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.*, 94:154–163, 1999.

[21] 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.

[22] R. Hyndman and Q. Yao. Nonparametric estimation and symmetry tests for conditional density functions. *Journal of nonparametric statistics*, 14(3):259–278, 2002.

[23] R. Hyndman, D. Bashtannyk, and G. Grunwald. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5:315–336, 1996.

[24] B. Karaivanov and P. Petrushev. Nonlinear piecewise polynomial approximation beyond besov spaces. *Applied and Computational Harmonic Analysis*, 15(3):177 – 223, 2003.

[25] E. Kolaczyk and R. Nowak. Multiscale generalised linear models for nonparametric function estimation. *Biometrika*, 92(1):119–133, 2005.

[26] E. Kolaczyk, J. Ju, and S. Gopal. Multiscale, multigranular statistical image segmentation. *J. Amer. Statist. Assoc.*, 100(472):1358–1369, 2005.

[27] J. Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145 –151, 01 1991.

[28] 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.

[29] C. Maugis and B. Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: P & S*, 2009. To appear.

[30] C. Maugis and B. Michel. Erratum on "a non asymptotic penalized criterion for Gaussian mixture model selection". Available on their webpage, 2010.

[31] 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.

[32] Ch. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, 22(1):118–171, 1994.

[33] 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.

[34] S. van de Geer. The method of sieves and minimum contrast estimators. *Math. Methods Statist.*, 4:20–38, 1995.

[35] I. van Keilegom and N. Veraverbeke. Density and hazard estimation in censored regression models. *Bernoulli*, 8(5):607–625, 2002.

[36] S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9:60–62, 1938.

[37] R. Willet and R. Nowak. Multiscale poisson intensity and density estimation. *IEEE Transactions on Information Theory*, 53(9):3171–3187, 2007.