Conditional Density Estimation by Penalized Likelihood Model Selection and Applications

Serge Cohen, Erwan Le Pennec

Conditional Density Estimation by Penalized Likelihood Model Selection and Applications

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

July 9, 2012

Abstract

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

1 Introduction

Assume we observe n pairs $((X_i, Y_i))_{1\leq i\leq n}$ 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 pairs (X_i, Y_i) are independent while Y_i depends on X_i through its law. More precisely, we assume that the covariates X_i are independent but not necessarily identically distributed. Assumptions 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 [42] 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 $s_{0'}(x)$ and $s_{0''}(x, y)$. Few other references on this subject seem to exist before the mid 90's with a study of a spline tensor based maximum likelihood estimator proposed by Stone [44] and a bias correction of Rosenblatt's estimator due to Hyndman et al. [31].

Kernel based method have been much studied since. For instance, Fan et al. [22] and de Gooijer and Zerom [17] consider local polynomial estimator, Hall et al. [26] study a locally logistic estimator that is later extended by Hyndman and Yao [30]. In this setting, pointwise convergence properties are considered, and extensions to dependent data are often obtained. The results depend however on a critical bandwidth that should be chosen according to the regularity of the unknown conditional density. Its practical choice is rarely discussed with the notable exceptions of Bashtannyk and Hyndman [5], Fan and Yim [21] and Hall et al. [27]. Extensions to censored cases have also been discussed for instance by van Keilegom and Veraverbeke [48]. See for instance Li and Racine [36] for a comprehensive review of this topic.

In the approach of Stone [44], the conditional density is estimated through a parametrized modelization. This idea has been reused since by Györfi and Kohler [25] with a histogram based approach, by Efromovich [19, 20] with a Fourier basis, and by Brunel et al. [13] and Akakpo and Lacour [2] with piecewise polynomial representation. Those authors are able to control an integrated estimation error: with 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. [13] extend their result to censored cases while Akakpo and Lacour [2] are able to consider weakly dependent data.

In this paper, we consider a direct estimation of the conditional density function through a maximum likelihood approach. Although natural, this approach has been considered so far only by Stone [44] as mentioned before and by Blanchard et al. [11] in a classification setting with histogram type estimators. Assume we have a set S_m of candidate conditional densities, our estimate $\widehat{s}_m$ will be simply the maximum likelihood estimate

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

Although this estimator may look like a maximum likelihood estimator of the joint density of (X_i, Y_i), it does not generally coincide, even when the X_is are assumed to be i.i.d., with such an estimator as every function of S_m is assumed to be a conditional density and not a density. The only exceptions are when the X_is are assumed in the model to be i.i.d. uniform or non random and equal. Our aim is then to analyze the finite sample performance of such an estimator in term of Kullback-Leibler type loss. As often, a trade-off between a bias term measuring the closeness of s_0 to the set S_m and a variance term depending on the complexity of the set S_m and on the sample size appears. A good set S_m will be thus one for which this trade-off leads to a small risk bound. Using a penalized model selection approach, we propose then a way to select the best model $S_{\widehat{m}}$ among a collection $\mathcal{S} = (S_m)_{m \in \mathcal{M}}$. For a given family of penalties $\operatorname{pen}(m)$, we define the *best* model $S_{\widehat{m}}$ as the one that minimized

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

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

The very frequent use of conditional density estimation in econometrics, see Li and Racine [36] for instance, could have provided a sufficient motivation for this study. However it turns

out that this work stems from a completely different subject: unsupervised hyperspectral image segmentation. Using the synchrotron beam of Soleil, the IPANEMA platform[6], for which one of the author works, is able to acquire high quality hyperspectral images, high resolution images for which a spectrum is measured at each pixel location. This provides rapidly a huge amount of data for which an automatic processing is almost necessary. One of this processing is the segmentation of these images into homogeneous zones, so that the spectral analysis can be performed on fewer places and the geometrical structures can be exhibited. The most classical unsupervised classification method relies on the density estimation of Gaussian mixture by a maximum likelihood principle. The component of the estimated mixtures will correspond to classes. In the spirit of Kolaczyk et al. [34] and Antoniadis et al. [3], we have extended this method by taking into account the localization of the pixel in the mixture weight, going thus from density estimation to conditional density estimation. As stressed by Maugis and Michel [39], understanding finely the density estimator is crucial to be able to select the right number of classes. This theoretical work has been motivated by a similar issue for the conditional density estimation case.

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

2 Single model maximum likelihood estimate

2.1 Framework and notation

Our statistical framework is the following: we observe n independent pairs $((X_i, Y_i))_{1\leq i\leq n} \in (\mathcal{X}, \mathcal{Y})^n$ where the X_i's are independent, but not necessarily of 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, a set comprising some candidate conditional densities, we estimate s_0 by the conditional density $\widehat{s}_m$ that maximizes the likelihood (conditionally to $(X_i)_{1\leq i\leq n}$) or equivalently that minimizes the opposite of the log-likelihood, denoted -log-likelihood from now on:

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

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

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

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

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

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

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

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

- If the law of Y_i is independent of X_i, that is $s(\cdot|X_i) = s(\cdot)$ and $t(\cdot|X_i) = t(\cdot)$ do not depend on X_i, these divergences reduce to the classical $KL_\lambda(s,t)$.
- If the X_i's are not random but fixed, that is we consider a fixed design case, this divergence is the classical fixed design type divergence in which there is no expectation.
- If the X_i's are i.i.d., this divergence is nothing but $KL_\lambda^{\otimes_n}(s,t) = \mathbb{E}\left[KL_\lambda(s(\cdot|X_1), t(\cdot|X_1))\right]$.

Note that this divergence is an *integrated* divergence as it is the average over the index i of the mean with respect to the law of X_i of the divergence between the conditional densities for a given covariate value. Remark in particular that more weight is given to regions of high density of the covariates than to regions of low density and, in particular, the values of the divergence outside the supports of the X_i's are not used. In particular, if we assume that each X_i has a law with density with respect to a common finite positive measure μ and that all those densities are lower and upper bounded then all our results hold, up to modification in constants, by replacing the definition of $KL_\lambda^{\otimes_n}(s,t)$ (and their likes) by the more classical

$$KL_\lambda^{\otimes_n}(s,t) = \int_{\mathcal{X}} KL(s(\cdot|x), t(\cdot|x))\mathrm{d}\mu.$$

We stress that these types of loss is similar to the one used in the machine-learning community (see for instance Catoni [14] that has inspired our notations). Such kind of losses appears also, but less often, in regression with random design (see for instance Birgé [8]) or in other conditional density estimation studies (see for instance Brunel et al. [13] and Akakpo and Lacour [2]). When $\hat{s}$ is an estimator, or any function that depends on the observation, $KL_\lambda^{\otimes_n}(s,\hat{s})$ measures this (random) integrated divergence between s and $\hat{s}$ conditionally to the observation while $\mathbb{E}\left[KL_\lambda^{\otimes_n}(s,\hat{s})\right]$ is the average of this random quantity with respect to the observations.

2.2 Asymptotic analysis of a parametric model

Assume that S_m is a parametric model of conditional densities,

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

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

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

then $\widehat{s}_m = s_{\widehat{\theta}_m}$. White [49] has studied this *misspecified model* setting for density estimation but its results can easily been extended to the conditional density case.

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

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

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

exists, the analysis of White [49] implies that, if we let

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

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

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

When s_0 belongs to the model, i.e. $s_0 = s_{\theta_m^\star}$, $B(\theta_m^\star) = A(\theta_m^\star)$ and thus the previous asymptotic equivalent of $\mathbb{E}\left[KL_\lambda^{\otimes n}(s_0, \widehat{s}_m) \right]$ is the classical parametric one

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

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

A short glimpse on the proof of the previous result shows that it depends heavily on the asymptotic normality of $\sqrt{n}(\widehat{\theta}_m - \theta_m^\star)$. One may wonder if extension of this result, often called the Wilk's phenomenon [50], exists when this normality does not hold, for instance in non parametric case or when the model is not identifiable. Along these lines, Fan et al. [23] propose a generalization of the corresponding Chi-Square goodness-of-fit test in several settings and Boucheron and Massart [12] study the finite sample deviation of the corresponding empirical quantity in a bounded loss setting.

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

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

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

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

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

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

2.3 Jensen-Kullback-Leibler divergence and bracketing entropy

The main visible loss is the use of a divergence smaller than the Kullback-Leibler one (but larger than the squared Hellinger distance and the squared L_1 loss whose definitions are recalled later). Namely, we use the Jensen-Kullback-Leibler divergence JKL_ρ with $\rho \in (0,1)$ defined by

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

Note that this divergence appears explicitly with $\rho=\frac{1}{2}$ in Massart [38], but can also be found implicitly in Birgé and Massart [9] and van de Geer [46]. We use the name Jensen-Kullback-Leibler divergence in the same way Lin [37] uses the name Jensen-Shannon divergence for a sibling in his information theory work. The main tools in the proof of the previous inequality are deviation inequalities for sums of random variables and their suprema. Those tools require a boundness assumption on the controlled functions that is not satisfied by the -log-likelihood differences $-\ln\frac{s_m}{s_0}$. When considering the Jensen-Kullback-Leibler divergence, those ratios are implicitly replaced by ratios $-\frac{1}{\rho}\ln\frac{(1-\rho)s_0+\rho s_m}{s_0}$ that are close to the -log-likelihood differences when the s_m are close to s_0 and always upper bounded by $-\frac{\ln(1-\rho)}{\rho}$. This divergence is smaller than the Kullback-Leibler one but larger, up to a constant factor, than the squared Hellinger one, $d_\lambda^2(s,t)=\int_\Omega|\sqrt{s}-\sqrt{t}|^2\mathrm{d}\lambda$, and the squared L_1 distance, $\|s-t\|_{\lambda,1}^2=\left(\int_\Omega|s-t|\mathrm{d}\lambda\right)^2$, as proved in Appendix

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

$$\max(C_\rho/4,\rho/2)\|s-t\|_{\lambda,1}^2 \leq JKL_{\rho,\lambda}(s,t) \leq KL_\lambda(s,t).$$

Furthermore, if $s\mathrm{d}\lambda \ll t\mathrm{d}\lambda$ then

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

while

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

More precisely, as we are in a conditional density setting, we use their *tensorized* versions

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

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

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

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

Note that the function $\sigma \mapsto \int_0^\sigma \sqrt{H_{[\cdot],d_\lambda^{\otimes_n}}(\delta, S_m)}\,\mathrm{d}\delta$ does always satisfy this assumption. $\mathfrak{D}_m$ is then defined as $n\sigma_m^2$ with σ_m^2 the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$. A good choice of ϕ_m is one which leads to a small upper bound of $\mathfrak{D}_m$. This bracketing entropy integral, often call Dudley integral, plays an important role in empirical processes theory, as stressed for instance in van der Vaart and Wellner [47] and in Kosorok [35]. The equation defining σ_m corresponds to a crude optimization of a supremum bound as shown explicitly in the proof. This definition is obviously far from being very explicit but it turns out that it can be related to an entropic dimension of the model. Recall that the classical entropy dimension of a compact set S with respect to a metric d can be defined as the smallest non negative real $\mathcal{D}$ such that there is a non negative $\mathcal{V}$ such that

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

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

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

As hinted by the notation, for parametric model, under mild assumption on the parametrization, this bracketing dimension coincides with the usual one. Under such assumption, one can prove that $\mathfrak{D}_m$ is proportional to $\mathcal{D}_m$. More precisely, working with the localized set $S_m(s,\sigma)$ instead of S_m, we obtain in Appendix, we obtain

Proposition 2. • *if $\exists \mathcal{D}_m \geq 0, \exists \mathcal{C}_m \geq 0, \forall \delta \in (0, \sqrt{2}], H_{[\cdot],d_\lambda^{\otimes_n}}(\delta, S_m) \leq \mathcal{V}_m + \mathcal{D}_m \ln\frac{1}{\delta}$ then*

- *if* $\mathcal{D}_m > 0$, *(*H_m*) holds with* $\mathfrak{D}_m \leq \left(2C_{\star,m} + 1 + \left(\ln \frac{n}{eC_{\star,m}\mathcal{D}_m}\right)_+\right)\mathcal{D}_m$ *with* $C_{\star,m} = \left(\sqrt{\frac{\mathcal{V}_m}{\mathcal{D}_m}} + \sqrt{\pi}\right)^2$,
- *if* $\mathcal{D}_m = 0$, *(*H_m*) holds with* $\phi_m(\sigma) = \sigma\sqrt{\mathcal{V}_m}$ *such that* $\mathfrak{D}_m = \mathcal{V}_m$,

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

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

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

which implies that those bounds are still useful when δ and σ are large. Assume now that all models are such that $\frac{\mathcal{V}_m}{\mathcal{D}_m} \leq \mathcal{C}$, i.e. their log-volumes $\mathcal{V}_m$ grow at most linearly with the dimension (as it is the case for instance for hypercubes with the same width). One deduces that Assumptions (H_m) hold simultaneously for every model with a common constant $C_\star = \left(\sqrt{\mathcal{C}} + \sqrt{\pi}\right)^2$. The model complexity $\mathfrak{D}_m$ can thus be chosen roughly proportional to the dimension in this case, this justifies the notation as well as our claim at the end of the previous section.

2.4 Single model maximum likelihood estimation

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

Assumption (Sep$_m$). *There exist a countable subset* S'_m *of* S_m *and a set* $\mathcal{Y}'_m$ *with* $\lambda(\mathcal{Y} \backslash \mathcal{Y}'_m) = 0$ *such that for every* $t \in S_m$, *there exists a sequence* $(t_k)_{k\geq 1}$ *of elements of* S'_m *such that for every* x *and for every* $y \in \mathcal{Y}'_m$, $\ln(t_k(y|x))$ *goes to* $\ln(t(y|x))$ *as* k *goes to infinity.*

We are now ready to state our risk bound theorem:

Theorem 1. *Assume we observe* (X_i, Y_i) *with unknown conditional density* s_0. *Assume* S_m *is a set of conditional densities for which Assumptions (*H_m*) and (Sep*$_m$*) 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, for* $\mathfrak{D}_m = n\sigma_m^2$ *with* σ_m *the unique root of* $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$, *the likelihood estimate* $\widehat{s}_m$ *satisfies*

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

This theorem holds without any assumption on the design X_i, in particular we do not assume that the covariates admit upper or lower bounded densities. The law of the design appears however in the divergence $JKL_\lambda^{\otimes_n}$ and $KL_\lambda^{\otimes_n}$ used to assess the quality of the estimate as well as in the definition of the divergence $d_\lambda^{\otimes_n}$ used to measure the bracketing entropy. By construction, those quantities however do not involve the values of the conditional densities outside the support of the X_is and put more focus on the regions of high density of covariates than the other. Note that Assumption H_m could be further localized: it suffices to impose that the condition on the Dudley integral holds for a sequence of minimizer of $d_\lambda^{2\otimes_n}(s_0, s_m)$.

We obtain thus a bound on the expected loss similar to the one obtained in the parametric case that holds for finite sample and that do not require the strong regularity assumptions of White [49]. In particular, we do not even require an identifiability condition in the parametric case. As often in empirical processes theory, the constant κ_0 appearing in the bound is pessimistic. Even in a very simple parametric model, the current best estimates are such that $\kappa_0 \mathfrak{D}_m$ is still much larger than the variance of Section 2.2. Numerical experiments show there is a hope that this is only a technical issue. The obtained bound quantifies however the expected classical bias-variance trade-off: a good model should be large enough so that the true conditional density is close from it but, at the same time, it should also be small so that the $\mathfrak{D}_m$ term does not dominate.

It should be stressed that a result of similar flavor could have been obtained by the information theory technique of Barron et al. [4] and Kolaczyk et al. [34]. Indeed, if we replace the set S_m by a *discretized* version $\mathfrak{S}_m$ so that

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

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

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

where $\mathcal{D}_\lambda^{2\otimes_n}$ is the tensorized Bhattacharyya-Renyi divergence, another divergence smaller than $KL^{\otimes_n}$, $|\mathfrak{S}_m|$ is the cardinality of $\mathfrak{S}_m$ and expectation is taken conditionally to the covariates $(X_i)_{1 \leq i \leq n}$. As verified by Barron et al. [4] and Kolaczyk et al. [34], $\mathfrak{S}_m$ can be chosen of cardinality of order $\ln n \, \mathcal{D}_m$ when the model is parametric. We obtain thus also a bound of type

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

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

3 Model selection and penalized maximum likelihood

3.1 Framework

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

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

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

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

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

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

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

The analysis of the previous section turns out to be crucial as the intrinsic complexity $\mathfrak{D}_m$ appears in the assumption on the penalty. It is no surprise that the complexity of the model collection itself also appears. We need an information theory type assumption on our collection; we assume thus 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*

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

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

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

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

3.2 A general theorem for penalized maximum likelihood conditional density estimation

Our main theorem is then:

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

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

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

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

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

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

satisfies

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

Note that, as in 2.4, the approach of of Barron et al. [4] and Kolaczyk et al. [34] could have been used to obtain a similar result with the help of discretization.

This theorem extends Theorem 7.11 Massart [38] which handles only density estimation. As in this theorem, the cost of model selection with respect to the choice of the best single model is proved to be very mild. Indeed, let $\mathrm{pen}(m) = \kappa(\mathfrak{D}_m + x_m)$ then one obtains

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

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

So far we do not have discussed the choice of the model collection, it is however critical to obtain a *good* estimator. There is unfortunately no universal choice and it should be adapted to the specific setting considered. Typically, if we consider conditional density of *regularity* indexed by a parameter α, a good collection is one such that for every parameter α there is a model which achieves a quasi optimal bias/variance trade-off. Efromovich [19, 20] considers Sobolev type regularity and use thus models generated by the first elements of Fourier basis. Brunel et al. [13] and Akakpo and Lacour [2] considers anisotropic regularity spaces for which they show that a collection of piecewise polynomial models is adapted. Although those choices are justified, in these papers, in a quadratic loss approach, they remain good choices in our maximum likelihood approach with a Kullback-Leibler type loss. Estimator associated to those collections are thus *adaptive* to the regularity: without knowing the *regularity* of the true conditional density, they select a model in which the estimate performs almost as well as in the *oracle* model, the best choice if the regularity was known. In both cases, one could prove that those estimators achieve the minimax rate for the considered classes, up to a logarithmic factor.

As in Section 2.4, the known estimate of constant κ_0 and even of $\mathfrak{D}_m$ can be pessimistic. This leads to a theoretical penalty which can be too large in practice. A natural question is thus

whether the constant appearing in the penalty can be estimated from the data without loosing a theoretical guaranty on the performance? No definitive answer exists so far, but numerical experiment in specific case shows that the *slope heuristic* proposed by Birgé and Massart [10] may yield a solution.

The assumptions of the previous theorem are as general as possible. It is thus natural to question the existence of interesting model collections that satisfy its assumptions. We have mention so far the Fourier based collection proposed by Efromovich [20, 19] and the piecewise polynomial collection of Brunel et al. [13] and Akakpo and Lacour [2] considers anisotropic regularity. We focus on a variation of this last strategy. Motivated by an application to unsupervised image segmentation, we consider model collection in which, in each model, the conditional densities depend on the covariate only in a piecewise constant manner. After a general introduction to these partition-based strategies, we study two cases: a classical one in which the conditional density depends in a piecewise polynomial manner of the variables and a newer one, which correspond to the unsupervised segmentation application, in which the conditional densities are Gaussian mixture with common Gaussian components but mixing proportions depending on the covariate.

4 Partition-based conditional density models

4.1 Covariate partitioning and conditional density estimation

Following an idea developed by Kolaczyk et al. [34], we partition the covariate domain and consider candidate conditional density estimates that depend on the covariate only through the region it belongs. We are thus interested in conditional densities that can be written as

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

where $\mathcal{P}$ is partition of $\mathcal{X}$, $\mathcal{R}_l$ denotes a generic region in this partition, $\mathbf{1}$ denotes the characteristic function of a set and $s(y|\mathcal{R}_l)$ is a density for any $\mathcal{R}_l \in \mathcal{P}$. Note that this strategy, called as in Willett and Nowak [51] partition-based, shares a lot with the CART-type strategy proposed by Donoho [18] in an image processing setting.

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

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

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

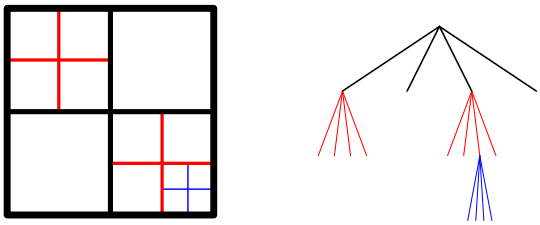


Figure 1: Example of a recursive dyadic partition with its associated dyadic tree.

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

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

As noticed by Kolaczyk and Nowak [33], Huang et al. [29] or Willett and Nowak [51], the first four partition collections, ($\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RDP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RDSP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RSP}(\mathcal{X})}$), have a tree structure. Figure 1 illustrates this structure for a RDP($\mathcal{X}$) partition. This specific structure is mainly used to obtain an efficient numerical algorithm performing the model selection. For sake of completeness, we have also added the much more complex to deal with collection $\mathcal{S}_{\mathcal{P}}^{\mathrm{HRP}(\mathcal{X})}$, for which only exhaustive search algorithms exist.

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

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

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

Those constants can be chosen as follow:

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

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

In those sections, we study the two different choices proposed above for the set $\mathcal{F}$. We first consider a piecewise polynomial strategy similar to the one proposed by Willett and Nowak [51] defined for $\mathcal{Y} = [0,1]^{d_Y}$ in which the set $\mathcal{F}$ is a product of sets. We then consider a Gaussian mixture strategy with varying mixing proportion but common mixture components that extends the work of Maugis and Michel [39] and has been the original motivation of this work. In both cases, we prove that the penalty can be chosen roughly proportional to the dimension.

4.2 Piecewise polynomial conditional density estimation

In this section, we let $\mathcal{X} = [0,1]^{d_X}$, $\mathcal{Y} = [0,1]^{d_Y}$ and λ be the Lebesgue measure $\mathrm{d}y$. Note that, in this case, λ is a probability measure on $\mathcal{Y}$. Our candidate density $s(y|x \in \mathcal{R}_l)$ is then chosen among piecewise polynomial densities. More precisely, we reuse a hyperrectangle partitioning strategy this time for $\mathcal{Y} = [0,1]^{d_Y}$ and impose that our candidate conditional density $s(y|x \in \mathcal{R}_l)$ is a square of polynomial on each hyperrectangle $\mathcal{R}_{l,k}^{\mathcal{Y}}$ of the partition $\mathcal{Q}_l$. This differs from the choice of Willett and Nowak [51] in which the candidate density is simply a polynomial. The two choices coincide however when the polynomial is chosen among the constant ones. Although our choice of using squares of polynomial is less natural, it already ensures the positiveness of the candidates so that we only have to impose that the integrals of the piecewise polynomials are equal to 1 to obtain conditional densities. It turns out to be also crucial to obtain a control of the local bracketing entropy of our models. Note that this setting differs from the one of Blanchard et al. [11] in which $\mathcal{Y}$ is a finite discrete set.

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

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

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

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

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

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

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

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

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

As shown by Willett and Nowak [51], the maximum likelihood estimate in this model can be obtained by an independent computation on each subset $\mathcal{R}_{l,k}^{\times}$:

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

This property is important to be able to use the efficient optimization algorithms of Willett and Nowak [51] and Huang et al. [29].

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

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

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

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

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

Theorem 3. *Fix a collection $\star(\mathcal{X})$ among* UDP($\mathcal{X}$), RDP($\mathcal{X}$), RDSP($\mathcal{X}$), RSP($\mathcal{X}$) *or* HRP($\mathcal{X}$) *for $\mathcal{X} = [0,1]^{d_X}$, a collection $\star(\mathcal{Y})$ among* UDP($\mathcal{Y}$), RDP($\mathcal{Y}$), RDSP($\mathcal{Y}$), RSP($\mathcal{Y}$) *or* HRP($\mathcal{Y}$) *and a maximal degree for the polynomials $\mathbf{D} \in \mathbb{N}^{d_Y}$.*

Let

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

Then there exist a $C_\star > 0$ and a $c_\star > 0$ independent of n, such that for any ρ and for any $C_1 > 1$, the penalized estimator of Theorem 2 satisfies

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

as soon as

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

for

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

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

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

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

This additive structure of the penalty allows to use the fast partition optimization algorithm of Donoho [18] and Huang et al. [29] as soon as the partition collection is tree structured.

In Appendix, we obtain a weaker requirement on the penalty

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

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

Some variations around this Theorem can be obtained through simple modifications of its proof as explained in Appendix. For example, the term $2\ln(n/\sqrt{\|\mathcal{Q}^{\mathcal{P}}\|})$ disappears if $\mathcal{P}$ belongs to $\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$ while $\mathcal{Q}_l$ is independent of $\mathcal{R}_l$ and belongs to $\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$. Choosing the degrees $\mathbf{D}$ of the polynomial among a family $\mathcal{D}^M$ either globally or locally as proposed by Willett and Nowak [51] is also possible. The constant $C_\star$ is replaced by its maximum over the family considered, while the coding part is modified by replacing respectively $A_0^{\star(\mathcal{X})}$ by $A_0^{\star(\mathcal{X})} + \ln|\mathcal{D}^M|$ for a global optimization and $B_0^{\star(\mathcal{Y})}$ by $B_0^{\star(\mathcal{Y})} + \ln|\mathcal{D}^M|$ a the local optimization. Such a penalty can be further modified into an additive one with only minor loss. Note that even if the family and its maximal degree grows with n, the *constant* $C_\star$ grows at a logarithic rate in n as long as the maximal degree grows at most polynomially with n.

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

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

as shown by Kolaczyk and Nowak [33]. We can thus reuse ideas from Willett and Nowak [51], Akakpo [1] or Akakpo and Lacour [2] to infer the quasi optimal minimaxity of this estimator for anisotropic Besov spaces (see for instance in Karaivanov and Petrushev [32] for a definition) whose regularity indices are smaller than 1 along the axes of $\mathcal{X}$ and smaller than $\mathbf{D}+1$ along the axes of $\mathcal{Y}$.

4.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties

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

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

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

where

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

with μ_k the mean of the kth component, Σ_k its covariance matrix, $\theta_k = (\mu_k, \Sigma_k)$ and π_k its mixing proportion. A model $S_{K,\mathcal{G}}$ is obtained by specifying the number of component K as well as a set $\mathcal{G}$ to which should belong the K-tuple of Gaussian $(\Phi_{\theta_1}, \ldots, \Phi_{\theta_K})$. Those Gaussians can share for instance the same shape, the same volume or the same diagonalization basis. The classical choices are described for instance in Biernacki et al. [7]. Using the EM algorithm, or one of its extension, one can efficiently obtain the proportions $\widehat{\pi}_k$ and the Gaussian parameters $\widehat{\theta}_k$ of the maximum likelihood estimate within such a model. Using tools also derived from Massart [38], Maugis and Michel [39] show how to choose the number of classes by a penalized maximum likelihood principle. These Gaussian mixture models are often used in unsupervised classification application: one observes a collection of Y_i and tries to split them into homogeneous classes. Those classes are chosen as the Gaussian components of an estimated Gaussian mixture close to the density of the observations. Each observation can then be assigned to a class by a simple maximum likelihood principle:

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

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

To overcome this issue, Kolaczyk et al. [34] and Antoniadis et al. [3] propose to use mixture model in which the mixing proportions depend on the covariate X_i while the mixture components remain constant. We propose to estimate simultaneously those mixing proportions and the mixture components with our partition-based strategy. In a semantic analysis context, in which documents replace pixels, a similar Gaussian mixture with varying weight, but without the

partition structure, has been proposed by Si and Jin [43] as an extension of a general mixture based semantic analysis model introduced by Hofmann [28] under the name *Probabilistic Latent Semantic Analysis*. A similar model has also been considered in the work of Young and Hunter [52]. In our approach, for a given partition $\mathcal{P}$, the conditional density $s(\cdot|x)$ are modeled as

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

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

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

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

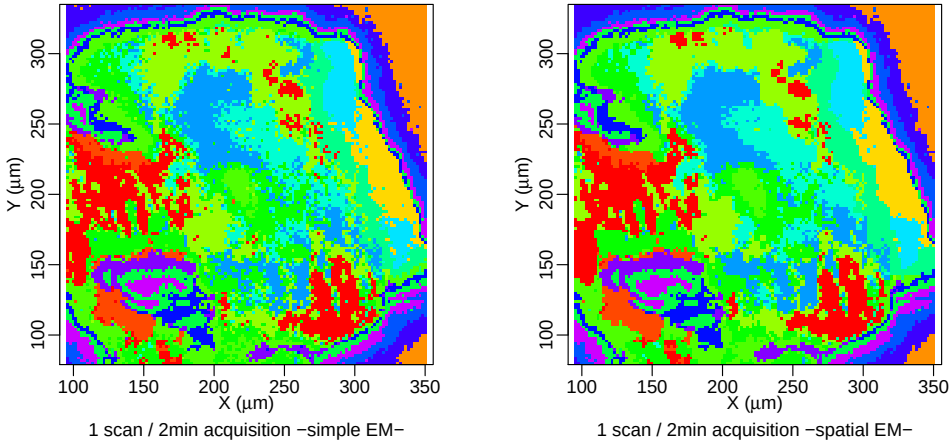
$$\widehat{k}(x, y) = \underset{1 \le k \le \widehat{K}}{\operatorname{argmax}}\, \widehat{\pi}_k[\widehat{\mathcal{R}_l}(x)] \Phi_{\widehat{\theta}_k}(y).$$

This is the strategy we have used at IPANEMA [6] to segment, in an unsupervised manner, hyperspectral images. In these images, a spectrum Y_i, with around 1000 frequency bands, is measured at each pixel location X_i and our aim was to derive a partition in *homogeneous* regions without any human intervention. This is a precious help for users of this imaging technique as this allows to focus the study on a few representative spectrums. Combining the classical EM strategy for the Gaussian parameter estimation (see for instance Biernacki et al. [7]) and dynamic programming strategies for the partition, as described for instance by Kolaczyk et al. [34], we have been able to implement this penalized estimator and to test it on real datasets.

Figure 2 illustrates this methodology. The studied sample is a thin cross-section of maple with a single layer of hide glue on top of it, prepared recently using materials and processes from the Cité de la Musique, using materials of the same type and quality that is used for lutherie. This sample is to serve as reference material to study the spectral variation of the hide glue at the various steps of the process. We present here the result for a low signal to noise ratio acquisition requiring only two minutes of scan. Using piecewise constant mixing proportions instead of constant mixing proportions leads to a better geometry of the segmentation, with less isolated points and more structured boundaries. As described in a more applied study [16], this methodology permits to work with a much lower signal to noise ratio and thus allows to reduce significantly the acquisition time.

We should now specify the models we consider. As we follow the construction of Section 4.1, for a given segmentation $\mathcal{P}$, this amounts to specify the set $\mathcal{F}$ to which belong the $\|\mathcal{P}\|$-tuples of densities $(s(y|\mathcal{R}_l))_{\mathcal{R}_l \in \mathcal{P}}$. As described above, we assume that $s(y|\mathcal{R}_l) = \sum_{k=1}^{K} \pi_k[\mathcal{R}_l] \Phi_{\theta_k}(y)$. The mixing proportions within the region $\mathcal{R}_l$, $\pi[\mathcal{R}_l]$, are chosen freely among all vectors of the $K-1$ dimensional simplex $\mathcal{S}_{K-1}$:

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



(a) (b)

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

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

To allow variable selection, we follow Maugis and Michel [39] and let E be an arbitrary subspace of $\mathcal{Y} = \mathbb{R}^p$, that is expressed differently for the different classes, and let $E^\perp$ be its orthogonal, in which all classes behave similarly. We assume thus that

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

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

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

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

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

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

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

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

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

Among the $3^4 = 81$ such possible sets, six of them have been already studied by Maugis and Michel [39, 41] in their classical Gaussian mixture model analysis:

- $[\mu_0 \, \mathrm{L}_K \, \mathrm{D}_0 \, \mathrm{A}_0]_{p_E}^K$ 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.
- $[\mu_K \, \mathrm{L} \, \mathrm{D}_0 \, \mathrm{A}_K]_{p_E}^K$ in which the volumes L_k are assumed to be equal and the variance can be diagonalized in the known basis D_0
- $[\mu_K \, \mathrm{L} \, \mathrm{D} \, \mathrm{A}]_{p_E}^K$ in which the variances Σ_k are only assumed to be equal

All these cases, as well as the others, are covered by our analysis with a single proof.

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

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

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

Without any assumptions on the design, we obtain

Theorem 4. *Assume the collection $\mathcal{S}$ is one of the collections of the previous paragraph.*

Then, there exist a $C_\star > \pi$ and a $c_\star > 0$, such that, for any ρ and for any $C_1 > 1$, the penalized estimator of Theorem 2 satisfies

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

as soon as

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

for

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

with $\kappa > \kappa_0$ where κ_0 and C_2 are the constants of Theorem 2 that depend only on ρ and C_1 and

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

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

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

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

satisfies the requirement of Theorem 4. This structure is the key for our numerical minimization algorithm in which one optimizes alternately the Gaussian parameters with an EM algorithm and the partition with the same fast optimization strategy as in the previous section.

In Appendix, we obtain a weaker requirement

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

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

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

Two major questions remain nevertheless open. Can we calibrate the penalty (choosing the constants) in a datadriven way while guaranteeing the theoretical performance in this specific setting? Can we derive a non asymptotic classification result from this conditional density result? The *slope heuristic*, proposed by Birgé and Massart [10], we have used in our numerical experiments, seems a promising direction. Deriving a theoretical justification in this conditional estimation setting would be much better. Linking the non asymptotic estimation behavior to a non asymptotic classification behavior appears even more challenging.

4.4 Bracketing entropy of Gaussian families

A key ingredient in the proof of 4 is a generalization of a result of Maugis and Michel [39, 40] controlling the bracketing entropy the Gaussian families $\mathcal{G}_{[\cdot]_E^K}$ with respect to the $d_\lambda^{\max}$ distance defined by

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

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

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

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

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

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

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

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

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

where c_S is an universal constant.

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

A.1 Proof of Proposition 1

Proof of Proposition 1. 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 $\dfrac{t\mathrm{d}\lambda}{\mathrm{d}\lambda'} = 1-(1-\rho)u$.

$$\begin{aligned}
\text{Now,} \qquad JKL_\rho(s\mathrm{d}\lambda, t\mathrm{d}\lambda) &= \frac{1}{\rho} KL(s\mathrm{d}\lambda, (1-\rho)s + \rho t\mathrm{d}\lambda) = \frac{1}{\rho} KL((1+\rho u)\mathrm{d}\lambda', \mathrm{d}\lambda') \\
&= \frac{1}{\rho} KL_{\lambda'}(1+\rho u, 1) = \frac{1}{\rho}\int (1+\rho u)\ln(1+\rho u)\mathrm{d}\lambda' \\
\text{and as } \int u\mathrm{d}\lambda' = 0 \qquad &= \frac{1}{\rho}\int \left((1+\rho u)\ln(1+\rho u) - \rho u\right)\mathrm{d}\lambda.
\end{aligned}$$

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

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

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

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.

Recall now that $KL(s\mathrm{d}\lambda, t\mathrm{d}\lambda) \geq \frac{1}{2}\|s-t\|^2_{\lambda,1}$ so that

$$\begin{aligned} JKL_\rho(s\mathrm{d}\lambda, t\mathrm{d}\lambda) &= \frac{1}{\rho} KL(s\mathrm{d}\lambda, (1-\rho)s + \rho t \mathrm{d}\lambda) \\ &\geq \frac{1}{2\rho}\|s - ((1-\rho)s + \rho t)\|^2_{\lambda,1} \\ &\geq \frac{1}{2\rho}\|\rho(s-t)\|^2_{\lambda,1} \\ &\geq \frac{\rho}{2}\|s-t\|^2_{\lambda,1}. \end{aligned}$$

Combining this result with $d^2_\lambda(s,t) \geq \frac{1}{4}\|s-t\|^2_{\lambda,1}$ allows to conclude.

For the third 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 result.

For the last bound, we use an idea of Kolaczyk and Nowak [33]:

$$\begin{aligned} KL(s\mathrm{d}\lambda, t\mathrm{d}\lambda) &= \int s\ln\left(\frac{s}{t}\right) d\lambda \\ &= KL(s\mathrm{d}\lambda, t\mathrm{d}\lambda) \qquad\qquad = \int \left(t - s + s\ln\left(\frac{s}{t}\right)\right) d\lambda \end{aligned}$$

and as $\log x \leq x - 1$

$$\leq \int \left(t - s + s\left(\frac{s}{t} - 1\right)\right) d\lambda$$

and assuming that t does not vanish

$$\begin{aligned} &\leq \int \frac{1}{t}\left(t^2 - 2st + s^2\right)) d\lambda \\ &\leq \left\|\frac{1}{t}\right\|_\infty \|t-s\|^2_{\lambda,2} \end{aligned}$$

□

A.2 Proof of Proposition 2

For sake of simplicity, we remove from now on the subscript reference to the common measure λ from all notations.

Proposition 2 is split into three propositions: Proposition 5 handles the cases of bracketing dimension 0, Proposition 6 applies when one control the bracketing entropy of the models S_m while 7 applies using bounds on the bracketing entropy of the local models $S_m(s_m, \sigma)$. Recall that Assumption (H_m)is the existence of a non-decreasing function ϕ_m such that $\delta \mapsto \frac{1}{\delta}\phi_m(\delta)$ is non-increasing and $\int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}(\delta, S_m(s_m,\sigma))}\,\mathrm{d}\delta \leq \phi_m(\sigma)$. The complexity term $\mathfrak{D}_m$ is then defined by $n\sigma_m^2$ where σ_m is the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$.

For the case of bracketing dimension 0, it suffices to show that the property holds for the local models as $H_{[\cdot],d^{\otimes n}}(\delta, S_m(s_m,\sigma)) \leq H_{[\cdot],d^{\otimes n}}(\delta, S_m)$.

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

$$H_{[\cdot],d^{\otimes n}}(\delta, S_m(s_m,\sigma)) \leq \mathcal{V}_m$$

then the function

$$\phi_m(\sigma) = \sigma\sqrt{\mathcal{V}_m}$$

satisfies the properties required in Assumption (H_m).

Furthermore, $\mathfrak{D}_m = \mathcal{V}_m$.

Proof. One check easily that ϕ_m is non-decreasing while $\delta \mapsto \frac{1}{\delta}\phi_m(\delta) = \sqrt{\mathcal{V}_m}$ is constant and thus non-increasing.

Using the assumption on the entropy,

$$\begin{aligned}\int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}(\delta, S_m(s_m,\sigma))}\,\mathrm{d}\delta &= \int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}(\delta \wedge \sqrt{2}, S_m(s_m, \sigma\wedge\sqrt{2}))}\,\mathrm{d}\delta \\ &\leq \int_0^\sigma \sqrt{\mathcal{V}_m}\,\mathrm{d}\delta \\ &\leq \sigma\sqrt{\mathcal{V}_m} = \phi_m(\sigma).\end{aligned}$$

Finally, the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$ is $\sigma_m = \sqrt{\frac{\mathcal{V}_m}{n}}$ which implies $\mathfrak{D}_m = n\sigma_m^2 = \mathcal{V}_m$. □

If one is only able to bound the bracketing entropy of the global model, one has:

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

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

Then the function

$$\phi_m(\sigma) = \sigma\sqrt{\mathcal{D}_m}\left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \sqrt{\ln\frac{1}{\sigma\wedge e^{-1/2}}}\right).$$

satisfies the properties required in Assumption (H_m).

Furthermore, $\mathfrak{D}$ satisfies

$$\mathfrak{D}_m \leq \left(2\left(\sqrt{\mathcal{C}_m} + \sqrt{\pi}\right)^2 + 1 + \left(\ln \frac{n}{e\left(\sqrt{\mathcal{C}_m} + \sqrt{\pi}\right)^2 \mathcal{D}_m} \right)_+ \right) \mathcal{D}_m$$

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

Proof. When $\sigma \geq e^{-1/2}$,

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

which is non-decreasing and such that $\delta \mapsto \frac{1}{\delta}\phi_m(\delta) = \sqrt{\mathcal{D}_m}\left(\sqrt{\mathcal{C}_m} + \sqrt{\pi}\right)$ is constant and thus non-increasing.

When $\sigma \leq e^{-1/2}$,

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

and thus

$$\begin{aligned}
\phi_m'(\sigma) &= \sqrt{\mathcal{D}_m}\left(\left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \sqrt{\ln\frac{1}{\sigma}}\right) - \frac{1}{2\sqrt{\ln\frac{1}{\sigma}}}\right) \\
&= \sqrt{\mathcal{D}_m}\left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \frac{1}{\sqrt{\ln\frac{1}{\sigma}}}\left(\ln\frac{1}{\sigma} - 1/2\right)\right) \geq 0
\end{aligned}$$

as $\ln\frac{1}{\sigma} - 1/2 \geq 0$ when $\sigma \leq e^{-1/2}$. ϕ_m is thus non-decreasing. The function

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

is strictly decreasing and thus non-increasing.

Now

$$\begin{aligned}
\int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}(\delta, S_m)}\,\mathrm{d}\delta &= \int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}(\delta \wedge \sqrt{2}, S_m)}\,\mathrm{d}\delta \\
&\leq \int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}(\delta \wedge e^{-1/2}, S_m)}\,\mathrm{d}\delta \\
&\leq \int_0^{\sigma\wedge e^{-1/2}} \sqrt{\mathcal{D}_m}\sqrt{\mathcal{C}_m + \ln\frac{1}{\delta}}\,\mathrm{d}\delta + \int_{\sigma\wedge e^{-1/2}}^\sigma \sqrt{\mathcal{D}_m}\sqrt{\mathcal{C}_m + \ln\frac{1}{e^{-1/2}}}\,\mathrm{d}\delta \\
&\leq \left(\sigma\sqrt{\mathcal{C}_m} + \int_0^{\sigma\wedge e^{-1/2}} \sqrt{\ln\frac{1}{\delta}}\,\mathrm{d}\delta + \int_{\sigma\wedge e^{-1/2}}^\sigma \sqrt{\ln\frac{1}{e^{-1/2}}}\,\mathrm{d}\delta\right)\sqrt{\mathcal{D}_m}.
\end{aligned}$$

We now rely on

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

proved in Maugis and Michel [39] to deduce

$$\begin{aligned}\int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}(\delta, S_m)}\,\mathrm{d}\delta &\leq \left(\sigma\sqrt{\mathcal{C}_m} + (\sigma \wedge e^{-1/2})\left(\sqrt{\ln\frac{1}{\sigma\wedge e^{-1/2}}} + \sqrt{\pi}\right)\right.\\ &\qquad \left. + (\sigma - \sigma\wedge e^{-1/2})_+\sqrt{\ln\frac{1}{e^{-1/2}}}\right)\sqrt{\mathcal{D}_m}\\ &\leq \sigma\left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \sqrt{\ln\frac{1}{\sigma\wedge e^{-1/2}}}\right)\sqrt{\mathcal{D}_m}\end{aligned}$$

$$\begin{aligned}&\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma \quad \Leftrightarrow \quad \left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \sqrt{\ln\frac{1}{\sigma\wedge e^{-1/2}}}\right)\sqrt{\mathcal{D}_m} = \sqrt{n}\sigma\\ \Leftrightarrow\quad &\sigma = \frac{1}{\sqrt{n}}\left(\sqrt{\mathcal{C}_m} + \sqrt{\pi} + \sqrt{\ln\frac{1}{\sigma\wedge e^{-1/2}}}\right)\sqrt{\mathcal{D}_m}\end{aligned}$$

This implies

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

which implies by inserting this bound in the initial equality

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

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

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

Proposition 7. *Assume for any $\sigma \in (0, \sqrt{2}]$ and any $\delta \in (0, \sigma]$*

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

Then the function

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

satisfies the properties required in Assumption (H_m).

Furthermore, $\mathfrak{D}_m$ satisfies

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

Proof of Proposition 7. By construction, the function ϕ_m is non decreasing while 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,

$$\begin{aligned}
\int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}(\delta, S_m(s_m,\sigma))}\,\mathrm{d}\delta &\leq \int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}(\delta\wedge\sqrt{2}, S_m(s_m,\sigma\wedge\sqrt{2}))} \\
&\leq \int_0^\sigma \sqrt{\mathcal{D}_m\left(\mathcal{C}_m + \ln\frac{\sigma\wedge\sqrt{2}}{\delta\wedge\sqrt{2}}\right)}\,\mathrm{d}\delta \\
&\leq \int_0^\sigma \left(\sqrt{\mathcal{C}_m} + \sqrt{\ln\frac{\sigma}{\delta}}\right)\mathrm{d}\delta\sqrt{\mathcal{D}_m} \\
&\leq \sigma\int_0^1 \left(\sqrt{\mathcal{C}_m} + \sqrt{\ln\frac{1}{\delta}}\right)\mathrm{d}\delta\sqrt{\mathcal{D}_m}
\end{aligned}$$

We now use Lemma 1 to obtain

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

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

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

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

A.3 Proof of Theorem 1

Proof of Theorem 1. For any function g, which may depend on the observed (X_i, Y_i), 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

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

where (X_i', Y_i') is an independent copy of (X_i, Y_i). Note that when g depends on the (X_i, Y_i), $P^{\otimes n}(g)$ is a random variable. Let $\nu_n^{\otimes n}(g)$ denote the recentred process $P_n^{\otimes n}(g) - P^{\otimes n}(g)$.

Using this definition,

$$KL^{\otimes n}(s_0, t) = P^{\otimes n}\left(-\ln\left(\frac{t}{s_0}\right)\right) \quad \text{and} \quad JKL_\rho^{\otimes n}(s_0, t) = P^{\otimes n}\left(-\frac{1}{\rho}\ln\left(\frac{(1-\rho)s_0 + \rho t}{s_0}\right)\right).$$

By construction, $\widehat{s}_m$ satisfies

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

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

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

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

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

By construction

$$P_n^{\otimes_n}(kl(\widehat{s}_m)) \leq P_n^{\otimes_n}(kl(\overline{s}_m)) + \frac{\eta}{n}$$

Since, by concavity of the logarithm,

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

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

and thus

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

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

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

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

We now rely 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 $\mathbb{E}^A[Z] = \frac{E[Z\mathbf{1}_{\{A\}}]}{\mathbb{P}\{A\}}$. It is sufficient to control those quantities for all A to obtain a control of the deviation. More precisely,

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

Here, we can prove

Lemma 3. *There exist three absolute constants $\kappa_0' > 4$, κ_1' and κ_2' such that, under Assumption (H), for all $m \in \mathcal{M}$, for every $y_m > \sigma_m$ and every event A such that $\mathbb{P}\{A\} > 0$,*

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

Combining Lemma 2 and Lemma 3 implies that, except on a set of probability less than e^{-x}, for any $y_m > \sigma_m$,

$$\frac{-\nu_n^{\otimes_n}(jkl(\widehat{s}_m))}{y_m^2 + \kappa_0' d^{2\otimes_n}(s_0, \widehat{s}_m)} \leq \frac{\kappa_1'\sigma_m}{y_m} + \kappa_2'\sqrt{\frac{x}{ny_m^2}} + \frac{18}{\rho}\frac{x}{ny_m^2}.$$

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

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

Thus, except on the same set,

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

Let $\epsilon_{\text{pen}} > 0$, we define θ_{pen} by $\left(\frac{\kappa_1' + \kappa_2'}{\theta_{\text{pen}}} + \frac{18}{\theta_{\text{pen}}^2\rho}\right)\kappa_0' = C_\rho\,\epsilon_{\text{pen}}$ with C_ρ defined in Proposition 1 and as $\widehat{s}_m$ is a conditional density $C_\rho\, d^{2\otimes_n}(s_0, \widehat{s}_m) \leq JKL_\rho^{\otimes_n}(s_0, \widehat{s}_m)$. Thus, we obtain

$$\begin{aligned}(1-\epsilon_{\text{pen}})JKL_\rho^{\otimes_n}(s_0, \widehat{s}_m) - \nu_n^{\otimes_n}(kl(\overline{s}_m)) &\leq \inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{C_\rho \epsilon_{\text{pen}} y_m^2}{\kappa_0'} + \frac{\eta}{n} + \frac{\delta_{KL}}{n} \\ &\leq \inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa_0'}\left(\sigma_m^2 + \frac{x}{n}\right) \\ &\quad + \frac{\eta}{n} + \frac{\delta_{KL}}{n}\end{aligned}$$

Let $\kappa_0 = \frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa_0'}$, we obtain that, with probability smaller than e^{-x},

$$\begin{aligned}JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}}) &> \frac{1}{1-\epsilon_{\text{pen}}}\left(\inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \kappa_0\sigma_m^2\right) + \frac{\eta}{n} + \frac{\delta_{KL}}{n} \\ &\quad + \frac{1}{1-\epsilon_{\text{pen}}}\nu_n^{\otimes_n}(kl(\overline{s}_m)) + \frac{\kappa_0}{1-\epsilon_{\text{pen}}}\frac{x}{n}\end{aligned}$$

which can be rewritten as, with probability smaller than e^{-x},

$$\begin{aligned}JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}}) - \left(\frac{1}{1-\epsilon_{\text{pen}}}\left(\inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \kappa_0\sigma_m^2\right) + \frac{\eta}{n} + \frac{\delta_{KL}}{n}\right) \\ + \frac{1}{1-\epsilon_{\text{pen}}}\nu_n^{\otimes_n}(kl(\overline{s}_m)) > \frac{\kappa_0}{1-\epsilon_{\text{pen}}}\frac{x}{n}\end{aligned}$$

For any non negative random variable Z and any $a > 0$, $\mathbb{E}[Z] = a\int_{z\geq 0}\mathbb{P}\{Z > az\}dz$ so

$$\begin{aligned}\mathbb{E}\left[JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}}) - \left(\frac{1}{1-\epsilon_{\text{pen}}}\left(\inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \kappa_0\sigma_m^2\right) + \frac{\eta}{n} + \frac{\delta_{KL}}{n}\right)\right] \\ + \mathbb{E}\left[\frac{1}{1-\epsilon_{\text{pen}}}\nu_n^{\otimes_n}(kl(\overline{s}_m))\right] \leq \frac{1}{1-\epsilon_{\text{pen}}}\kappa_0\frac{1}{n}\end{aligned}$$

As by construction $\nu_n^{\otimes_n}(kl(\overline{s}_m))$ is integrable and $\mathbb{E}[\nu_n^{\otimes_n}(kl(\overline{s}_m))] = 0$), we derive

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

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

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

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

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

B.1 Proof of Theorem 2

Proof of Theorem 2. For any model S_m, we let $\overline{s}_m$ be a function such that

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

Let $m \in \mathcal{M}$ such that $KL^{\otimes_n}(s, \overline{s}_m) < +\infty$ 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, $jkl(\widehat{s}_{m'}) \leq kl(\widehat{s}_{m'})$,

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

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

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

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

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

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

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

Using the Kraft condition of Assumption (K), we deduce that if we make this choice of $y_{m'}$ for all models m', this properties hold 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 $m \in \mathcal{M}'$,

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

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_ρ defined in Proposition 1 and, as $\widehat{s}_{m'}$ is a conditional density $C_\rho\, d^{2\otimes_n}(s_0, \widehat{s}_{m'}) \leq JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{m'})$, 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 we let $\kappa_0 = \frac{C_\rho \epsilon_{\mathrm{pen}} \theta_{\mathrm{pen}}^2}{\kappa_0'}$

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

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

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

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 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 $1 - \Sigma e^{-x}$,

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

which yields by the same integration technique that in the proof of the previous theorem

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

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

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

which is sligthly stronger than the result stated in the theorem with $C_1=\frac{1}{1-\epsilon_{\text{pen}}}$ and $C_2=\frac{\kappa_0}{1-\epsilon_{\text{pen}}}$ as the penalty of the select model appears in the right-hand side with a positive weight. □

B.2 Proof of Lemma 2

Proof of Lemma 2. 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,

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

Hence $\Psi(x)\leq\Psi\left(\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)\right)$ 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. □

B.3 Proof of Lemma 3

We should now prove Lemma 3 which contains most of the differences with Massart [38]'s proof.

Proof of Lemma 3. In this lemma, we want to control the deviation of

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

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

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

To control the behavior of these quantities, we use the following key properties of Jensen-Kullback-Leibler related quantities (a rewriting of Lemma 7.26 of Massart [38])

Lemma 4. *Let P be a probability measure with density s_0 with respect to a 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 this lemma, $P(g)$ stands for $\int g s_0 \mathrm{d}\lambda$ i.e. the expectation with respect to the probability $s_0\mathrm{d}\lambda$. 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

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

We now use

Theorem 5. *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 $\mathbb{P}\{A\} > 0$

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

These bounds are sufficient to obtain a Bernstein type control for $jkl(\widetilde{s})$

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

To cope with the randomness of $\widehat{s}_m$, we rely on the following much more involved theorem (a rewriting of Theorem 6.8 of Massart [38])

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

$$P^{\otimes_n}(|f|^k) \leq \frac{k!}{2}Vb^{k-2}$$

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

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

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

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

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

Furthermore $\kappa \leq 27$.

If we consider

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

then the first assumption of Theorem 6 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

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

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

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

and

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

So that

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

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

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

Theorem 6 can not be used directly with the set $\mathcal{G}_m(\widetilde{s}_m,\sigma)$ as it is not necessarily countable. However, Assumption (Sep_m) implies the existence of a countable family S'_m such that

$$\mathcal{G}'_m(\widetilde{s}_m,\sigma) = \left\{ -jkl(s_m)+jkl(\widetilde{s}) = \frac{1}{\rho}\ln\left(\frac{s_0+\frac{\rho}{1-\rho}s_m}{s_0+\frac{\rho}{1-\rho}\widetilde{s}_m}\right)\middle| s_m\in S'_m, d^{2\otimes_n}(\widetilde{s}_m,s_m)\leq\sigma\right\}$$

is countable, and thus for which the conclusion of Theorem 6 holds, while $\sup_{\mathcal{G}'_m(\widetilde{s}_m,\sigma)} \nu_n^{\otimes_n}(f) = \sup_{\mathcal{G}_m(\widetilde{s}_m,\sigma)} \nu_n^{\otimes_n}(f)$ with probability 1. We deduce thus that for every measurable set A with $\mathbb{P}\{A\} > 0$,

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

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

Choosing $\epsilon = 1$ leads to

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

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 Assumption (H_m), if we assume $\widetilde{s}_m \in S_m$, $\int_0^{\sigma}\sqrt{H_{[\cdot],d^{\otimes n}}\left(\delta, S_m(\widetilde{s}_m,\sigma)\right) \wedge n}\mathrm{d}\delta \leq \phi_m(\sigma)$, 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

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

As $\delta \mapsto \delta^{-1}\phi_m(\delta)$ is also non-increasing, so is $\delta \mapsto \delta^{-2}\phi_m(\delta)$. The definition of σ_m can be rewritten as the equation $\frac{\phi_m(\sigma_m)}{\sqrt{n}\sigma_m^2} = 1$. The right-hand side of the previous inequality is thus an $O(\frac{\phi_m(\sigma)}{\sqrt{n}})$ as soos as $\sigma \geq \sigma_m$. Indeed under this assumption,

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

and

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

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 $\kappa \leq 27$, $\kappa_2'' = \frac{21}{2\sqrt{\rho(1-\rho)}}$, so that $\forall \sigma > \sigma_m$,

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

Thanks to Assumption (Sep_m), we can use the *pealing* lemma (Lemma 4.23 of [38]):

Lemma 5. *Let S be a 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 a 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

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

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

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

For this choice, one obtains

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

which implies

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

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 $\kappa' > 0$:

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

We derive thus

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

Let κ_d' such that ${\kappa_d'}^2 = 2(2+\epsilon_d)/(1+\epsilon_d)$, using $d^{2\otimes_n}(s_0,\widehat{s}_m) \geq d^{2\otimes_n}(s_0,\widetilde{s}_m)/(1+\epsilon_d)$,

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

and thus

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

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

B.4 Behavior of the constants of Theorem 1 and Theorem 2

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

Now κ_0 is defined as equal to $\frac{C_\rho \epsilon_{\text{pen}} \theta_{\text{pen}}^2}{\kappa_0'}$ with θ_{pen} the root of $\left(\frac{\kappa_1'+\kappa_2'}{\theta_{\text{pen}}} + \frac{18}{\theta_{\text{pen}}^2 \rho}\right) \kappa_0' = C_\rho \, \epsilon_{\text{pen}}$ where we use the constants appearing in Lemma 3. 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 6. Indeed let $\sigma_{\mathcal{M}}$ denote 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 ensured in the models of Section 4 provided we limit their maximum dimension well below n, for instance to $n/\ln^2(n)$.

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

Proof of Proposition 3. 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.

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

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

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

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

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

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

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

$$\sum_{\mathcal{P}\in\mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}} e^{-cB_0^{\star(\mathcal{X})}\|\mathcal{P}\|} \leq \sum_{\mathcal{P}\in\mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}} e^{-(c-c_0^{\star(\mathcal{X})})B_0^{\star(\mathcal{X})}\|\mathcal{P}\|} e^{-c_0^{\star(\mathcal{X})}B_0^{\star(\mathcal{X})}\|\mathcal{P}\|}$$

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

$$\begin{aligned}
&\leq e^{-(c-c_0^{\star(\mathcal{X})})B_0^{\star(\mathcal{X})}} \sum_{\mathcal{P}\in\mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}} e^{-c_0^{\star(\mathcal{X})}B_0^{\star(\mathcal{X})}\|\mathcal{P}\|} \\
&\leq e^{-(c-c_0^{\star(\mathcal{X})})B_0^{\star(\mathcal{X})}} e^{(1-c_0^{\star(\mathcal{X})})B_0^{\star(\mathcal{X})}} = e^{B_0^{\star(\mathcal{X})}} e^{-cB_0^{\star(\mathcal{X})}} = \Sigma_0^{\star(\mathcal{X})} e^{-cC_0^{\star(\mathcal{X})}}
\end{aligned}$$

which concludes these three cases.

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

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

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

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

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

It is then only a matter of calculation to check that if c is larger than 1 in the UDP and RDP cases and larger than $2\ln 2$ in the other cases then all these sums can be bounded by 1.

□

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

Theorem 3 is obtained by proving that Assumption ($\mathrm{H}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$) and ($\mathrm{S}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$) hold for any model $S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ while Assumption (K) holds for any model collection. Theorem 3 is then a consequence of Theorem 2.

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

Proposition 8. *Under the assumptions of Theorem 3, there exists a $D_\star$ such that for any model $S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ Assumption ($H_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$) is satisfied with a function ϕ such that*

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

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

The proof relies on the combination of Proposition 2 and

Proposition 9. $\forall S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}, \forall s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \in S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$,

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

Remark that we also use the inequality

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

By using Proposition 3 for both $\mathcal{P}$ and $\mathcal{Q}$, we obtain the Kraft type assumption:

Proposition 10. *Under the assumptions of Theorem 3, for any collection $\mathcal{S}$, there exists a $c_\star > 0$ such that for*

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

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

The complete proof is postponed after the one Proposition 9.

D.1 Proof of Proposition 9

For sake of simplicity, we remove from now on the subscript reference to the common measure λ from all notations.We rely on a link between $\|\cdot\|_2$ and $\|\cdot\|_\infty$ structures of the square roots of the models and a relationship between bracketing entropy and metric entropy for $\|\cdot\|_\infty$ norms.

Following Massart [38], we define the following tensorial *norm* on functions $u(y|x)$

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

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

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

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

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

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

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

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

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

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

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

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

The following proposition, proved in next section, is similar to a proposition of Massart [38]. It provides a bound for this last entropy term under an assumption on a link between $\|\cdot\|_\infty^{2\otimes n}$ and $\|\cdot\|_2^{2\otimes n}$ structures:

Proposition 11. *For any basis $\{\phi_k\}_{1\leq k\leq \mathcal{D}_m}$ of $\overline{\sqrt{S_m}}$ such that*

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

let

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

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

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

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

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

In our setting, using a basis of Legendre polynomials, we are able to derive from Proposition 11

Proposition 12. *For any model of Section 4.2,*

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

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

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

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

One easily verifies that

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

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

Let

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

we have slightly more than Proposition 9 as $\forall s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \in S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$,

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

D.2 Proofs of Proposition 11 and Proposition 12

Proof of Proposition 11. 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

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

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

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

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

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

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

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

and thus

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

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

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

which concludes the proof. □

Instead of Proposition 12, by mimicking a proof of Massart [38], we prove an extended version of it in which the degree of the conditional densities may depend on the hyperrectangle. More precisely, we reuse the partition $\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{\star}(\mathcal{X})$ and the partitions $\mathcal{Q}_l \in \mathcal{S}_{\mathcal{P}}^{\star}(\mathcal{Y})$ for $\mathcal{R}_l \in \mathcal{P}$ and define now the model $S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ as the set of conditional densities such that

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

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

By construction,

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

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

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

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

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

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

Proposition 13. *There exists*

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

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

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

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

Proposition 12 is deduced from this proposition with the help of the simple upper bound

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

As

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

once a maximal degree is chosen along each axis, the equivalent of constant $C_\star$ of 3 depends only on this maximal degrees. Assumption $\mathrm{H}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ holds then, with the same constants, simultaneaously for all models of both global choice and local choice strategies. Obtaining the Kraft type assumption, Assumption (K) is only a matter of taking into account the augmentation of the number of models within the collection. Replacing respectively $A_0^{\star(\mathcal{X})}$ by $A_0^{\star(\mathcal{X})} + \ln |\mathcal{D}^M|$ for global optimization and $B_0^{\star(\mathcal{Y})}$ by $B_0^{\star(\mathcal{Y})} + \ln |\mathcal{D}^M|$ for local optimization, where $|\mathcal{D}^M|$ denotes the size of the family of possible degrees, turns out to be sufficient as mentioned earlier.

Proof of Proposition 13. Let L_D be the one dimensional Legendre polynomial of degree D on $[0,1]$ and $G_D = \sqrt{2D+1} L_D$ its rescaled version, we recall that, by definition,

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

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

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

by construction

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

and

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

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

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

and

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

Using the piecewise structure, one deduces

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

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

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

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

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

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

For $\|\cdot\|_\infty$ type norm,

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

Now

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

while

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

This implies

$$\frac{\left\| \sum_{\mathcal{R}^{\times}_{l,k} \in \mathcal{Q}^{\mathcal{P}}} \sum_{D \leq \mathbf{D}(\mathcal{R}^{\times}_{l,k})} \beta_D^{\mathcal{R}^{\times}_{l,k}} \phi_D^{\mathcal{R}^{\times}_{l,k}} \right\|_{\infty}^{2 \otimes_n}}{\mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \left\| \beta_D^{\mathcal{R}^{\times}_{l,k}} \right\|_{\infty}^2}$$

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

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

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

Proposition is then obtained by a simple application of Proposition 11. □

D.3 Proof of Proposition 10

Proof. By construction

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

By Proposition 3, one can find $c_\star \geq \max(1, c_0^{\star(\mathcal{X})}, c_0^{\star(\mathcal{Y})})$ such that

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

and

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

Plugging these bounds in the previous equality yields

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

Proposition holds with the modified weights for polynomial as

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

as soon as $c_\star \geq 1$. □

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

As in the piecewise polynomial density case, Theorem 4 is obtained by showing that Assumptions ($\mathrm{H}_{\mathcal{P},K,\mathcal{G}}$), ($\mathrm{S}_{\mathcal{P},K,\mathcal{G}}$) and (K) hold for any collection.

Again, one easily verifies that Assumption ($\mathrm{S}_{\mathcal{P},K,\mathcal{G}}$) holds. For the complexity assumption, combining 2 with a bound on the bracketing entropy of the models of type

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

one obtains

Proposition 14. *There exists a constant C depending only on a, L_-, L_+, λ_- and λ_+ such that for any model $S_{\mathcal{P},K,\mathcal{G}}$ of Theorem 4 Assumption ($H_{\mathcal{P},K,\mathcal{G}}$) is satisfied with a function ϕ such that*

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

For the Kraft assumption, one can verify that

Proposition 15. *For any collections $\mathcal{S}$ of Theorem 4, there is a $c_\star$ such that for the choice*

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

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

As for the piecewise polynomial case section, the main difficulty lies in controlling the bracketing entropy of the models. A proof of Proposition 15 can be found in our technical report [15].

We focus thus on the proof of Proposition 14. Due to the complex structure of spatial mixture, we did not manage to bound the bracketing entropy of local model. We derive only an upper bound of the bracketing entropy $H_{[\cdot],d^{\otimes n}}(\delta, S_{\mathcal{P},K,\mathcal{G}})$, but one that is independent of the distribution law of $(X_i)_{1\leq i\leq n}$: the bracketing entropy with a sup norm Hellinger distance $d^{\sup} = \sqrt{d^{2\sup}}$, $H_{[\cdot],d^{\sup}}(\delta, S_{\mathcal{P},K,\mathcal{G}})$, where $d^{2\sup}$ is defined by

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

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

Proposition 14 is a direct consequence of Proposition 2 and

Proposition 16. *There exists a constant C depending only on a, L_-, L_+, λ_- and λ_+ such that for any model $S_{\mathcal{P},K,\mathcal{G}}$ of Theorem 4:*

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

E.1 Model coding

Proof of Proposition 15. This proposition is a simple combination of Theorem 3, crude bounds on the number of different models indexed by $[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]^K$ and $[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]$ and of classical Kraft type inequalities for order selection and variable selection (see for instance in the book of Massart [38]):

Lemma 6.

- *For the selection of 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 $I = \{1,\dots,p_E\}$, let $\theta_E = p_E$, for $c > 0$*

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

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

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

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 $3^4 \times 2^4 = 1296$, we obtain

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

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

E.2 Entropy of spatial mixtures

Proof of Proposition 16. While we use classical Hellinger distance to measure the complexity of the simplex $\mathcal{S}_{K-1}$ and the set $\mathcal{G}_{E^\perp}$, we use a sup norm Hellinger distance on $\mathcal{G}_E^K$ defined by

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

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

Using a similar proof than Genovese and Wasserman [24], we decompose the entropy in three parts with:

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

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

We bound those bracketing entropies with the help of two results. We first use a Lemma proved in Genovese and Wasserman [24] that implies the existence of a universal constant $\mathcal{C}_{\mathcal{S}}$ such that

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

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

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

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

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

We then rely on Proposition 4 to handle the bracketing entropy of Gaussian K-tuples collection. It implies the existence of two constants $\mathcal{C}_{[\star]^\star}$ and $\mathcal{C}_{[\star]}$ depending only on a, L_-, L_+, λ_- and λ_+ such that

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

As $\dim(S_{K,\mathcal{P},\mathcal{G}})=\|\mathcal{P}\|(K-1)+\dim(\mathcal{G}_E^K)+\dim(\mathcal{G}_{E^\perp})$, we obtain Proposition 16 with $C=\max(\mathcal{C}_{\mathcal{S}},\mathcal{C}_{[\star]^\star},\mathcal{C}_{[\star]})$. □

E.3 Entropy of Gaussian families

Instead of Proposition 4, we prove the slightly stronger

Proposition 17. *Let $\kappa\ge\frac{3}{4}$ and*

$$\gamma_\kappa=\min\left(\frac{3(\kappa-\frac{3}{4})}{2(1+\frac{\kappa}{6})(1+\frac{1}{6})(1+\frac{1}{12})},\frac{(\kappa-\frac{1}{2})}{2(1+\frac{\kappa}{6})(1+\frac{1}{6})}\right)\qquad\beta_\kappa=\sqrt{\kappa^2\cosh(\frac{\kappa}{6})+\frac{1}{2}}$$

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

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

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

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

$$\begin{cases} \mathcal{D}_{\mu,p_E} = p_E \\ \mathcal{D}_\mathrm{L} = 1 \\ \mathcal{D}_{\mathrm{D},p_E} = \frac{p_E(p_E-1)}{2} \\ \mathcal{D}_{\mathrm{A},p_E} = p_E - 1 \end{cases} \quad \textit{and} \quad \begin{cases} \mathcal{V}_{\mu,p_E} = p_E \ln\left(1 + \frac{18\beta_\kappa a p_E}{\sqrt{\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}}\right) \\ \mathcal{V}_{\mathrm{L},p_E} = \ln\left(1 + \frac{39}{2}\beta_\kappa \ln\left(\frac{L_+}{L_-}\right) p_E\right) \\ \mathcal{V}_{\mathrm{D},p_E} = \frac{p_E(p_E-1)}{2}\left(\frac{2\ln c_S}{p_E(p_E-1)} + \left(\ln\left(126\beta_\kappa \frac{\lambda_+}{\lambda_-} p_E\right)\right)\right) \\ \mathcal{V}_{\mathrm{A},p_E} = (p_E-1)\ln\left(2 + \frac{255}{2}\beta_\kappa \frac{\lambda_+}{\lambda_-} \ln\left(\frac{\lambda_+}{\lambda_-}\right) p_E\right) \end{cases}$$

where c_S *is a universal constant.*

Furthermore, for any $p_E \leq p$

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

with

$$\begin{aligned} \mathcal{C}_{\mu,p} &= \ln\left(1 + \frac{18\beta_\kappa a p}{\sqrt{\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}}\right) \\ \mathcal{C}_{\mathrm{L},p} &= \ln\left(1 + \frac{39}{2}\beta_\kappa \ln\left(\frac{L_+}{L_-}\right) p\right) \\ \mathcal{C}_{\mathrm{D},p} &= \left(2\ln c_S + \left(\ln\left(126\beta_\kappa \frac{\lambda_+}{\lambda_-} p\right)\right)\right) \\ \mathcal{C}_{\mathrm{A},p} &= \ln\left(2 + \frac{255}{2}\beta_\kappa \frac{\lambda_+}{\lambda_-} \ln\left(\frac{\lambda_+}{\lambda_-}\right) p\right) \end{aligned}$$

and, uniformly over K*,*

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

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

Proposition 4 is obtained by setting $\kappa = 1$ and using the crude bounds $1/9 \leq \gamma_\kappa \leq 1/4$, $1 \leq \beta_\kappa \leq 2$.

Proof of Proposition 17. We consider all models $\mathcal{G}_{[\mu_\star \, \mathrm{L}_\star \, \mathrm{A}_\star \, \mathrm{D}_\star]_E^K}$ at once by a "tensorial" construction of a suitable $\delta/9$ bracket collection.

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

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

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

- For any δ_L, the grid $\mathcal{G}_\mathrm{L}(L_-, L_+, \delta_\mathrm{L})$ of $[L_-, L_+]$:

$$\mathcal{G}_\mathrm{L}(L_-, L_+, \delta_\mathrm{L}) = \{ L_-(1+\delta_\mathrm{L})^g | g \in \mathbb{N}, L_-(1+\delta_\mathrm{L})^g \leq L_+ \}.$$

- For any δ_D, the grid $\mathcal{G}_\mathrm{D}(p_E, \delta_\mathrm{D})$ of $SO(p_E)$ made of the elements of a δ_D-net with respect to the $\|\cdot\|_2$ operator norm (as described by Szarek [45]).

- For any δ_A, the grid $\mathcal{G}_\mathrm{A}(\lambda_-, \lambda_+, p_E, \delta_\mathrm{A})$ of $\mathcal{A}(\lambda_-, \lambda_+(1+\delta_\mathrm{A}), p_E)$:

$$\mathcal{G}_\mathrm{A}(\lambda_-, \lambda_+, p_E, \delta_\mathrm{A}) = \{ A \in \mathcal{A}(\lambda_-, \lambda_+(1+\delta_\mathrm{A}), p_E) | \forall 1 \leq i < p_E, \exists g_i \in \mathbb{N}, A_i = \lambda_-(1+\delta_\mathrm{A})^{g_i} \}.$$

Obviously, for any $\mu \in [-a, a]$, there is a $\tilde{\mu} \in \mathcal{G}_\mu(a, p_E, \delta_\mu)$ such that

$$\|\tilde{\mu} - \mu\|^2 \leq p_E \delta_\mu^2$$

while

$$|\mathcal{G}_\mu(a, p_E, \delta_\mu)| \leq \left(1 + 2\frac{a}{\delta_\mu}\right)^{p_E} \leq \max\left(2^{p_E}, \left(\frac{4a}{\delta_\mu}\right)^{p_E}\right).$$

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

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

If we further assume that $\delta_\mathrm{L} \leq \frac{1}{12}$ then $\ln(1+\delta_\mathrm{L}) \geq \frac{12}{13}\delta_\mathrm{L}$ and

$$|\mathcal{G}_\mathrm{L}(L_-, L_+, \delta_\mathrm{L})| \leq 1 + \frac{13 \ln\left(\frac{L_+}{L_-}\right)}{12\delta_\mathrm{L}}.$$

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

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

As proved by Szarek [45], it exists a universal constant c_S such that, as soon as $\delta_\mathrm{D} \leq 1$

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

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

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

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

where p_E-1 is the intrinsic dimension of $\mathcal{A}(\lambda_-,\lambda_+,p_E)$. If we further assume that $\delta_{\mathrm{A}} \leq \frac{1}{84}$ then $\ln(1+\delta_{\mathrm{A}}) \geq \frac{84}{85}\delta_{\mathrm{A}}$ and thus

$$|\mathcal{G}_{\mathrm{A}}(\lambda_-,\lambda_+,p_E,\delta_{\mathrm{A}})| \leq \left(2+\frac{85\ln\left(\frac{\lambda_+}{\lambda_-}\right)}{84\delta_{\mathrm{A}}}\right)^{p_E-1}.$$

A key to the succes of this construction is the following approximation property of this grid proved later:

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

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

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

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

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

By definition, the image of

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

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

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

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

We prove that, for γ_κ and β_κ defined in Proposition 17 and any $\kappa \geq \frac{3}{4}$, the choice

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

is such that the image of

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

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

Indeed, as proved later,

Lemma 10. *Let* $\kappa\geq\frac{3}{4}$, $\gamma_\kappa=\min\left(\frac{3(\kappa-\frac{3}{4})}{2(1+\frac{\kappa}{6})(1+\frac{1}{6})(1+\frac{1}{12})},\frac{(\kappa-\frac{1}{2})}{2(1+\frac{\kappa}{6})(1+\frac{1}{6})}\right)$ *and* $\beta_\kappa=\sqrt{\kappa^2\cosh(\frac{\kappa}{6})+\frac{1}{2}}$. *For any* $0<\delta\leq\sqrt{2}$, *any* $p_E\geq1$ *and any* $\delta_\Sigma\leq\frac{1}{9\beta_\kappa}\frac{\delta}{p_E}$,

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

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

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

Furthermore, let $(\mu,L,A,D)\in[-a,a]^{p_E}\times[L_-,L_+]\times\mathcal{A}(\lambda_-,\lambda_+)\times SO(p_E)$ *and define* $\Sigma=LDAD'$. *If*

$$\begin{cases}\|\mu-\tilde\mu\|^2\leq p_E\gamma_\kappa L_-\lambda_-\frac{\lambda_-}{\lambda_+}\delta_\Sigma^2\\ (1+\frac{\delta_\Sigma}{2})^{-1}\tilde L\leq L\leq\tilde L\\ \forall1\leq i\leq p_E,\quad|A_{i,i}^{-1}-\tilde A_{i,i}^{-1}|\leq\frac{1}{14}\frac{1}{\lambda_+}\delta_\Sigma\\ \forall x\in\mathbb{R}^{p_E},\quad\|Dx-\tilde Dx\|\leq\frac{1}{14}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\|x\|\end{cases}$$

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

By definition of $d^{\max}$, this implies that our choice of δ_μ, δ_{L}, δ_{D}, δ_{A} and δ_Σ is such that every K-tuple of pairs of the collections is a $\delta/9$-bracket and they cover the whole set.

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

$$\begin{aligned}&\left(\left(1+\frac{2a}{\frac{\sqrt{\gamma_\kappa L_-\lambda_-\frac{\lambda_-}{\lambda_+}}}{9\beta_\kappa}\frac{\delta}{p_E}}\right)^{p_E}\right)^{c_{\mu_\star}}\times\left(\left(1+\frac{13\ln\left(\frac{L_+}{L_-}\right)}{12\frac{1}{18\beta_\kappa}\frac{\delta}{p_E}}\right)\right)^{c_{\mathrm{L}_\star}}\\&\quad\times\left(c_S\left(\frac{1}{\frac{1}{126\beta_\kappa}\frac{\lambda_-}{\lambda_+}\frac{\delta}{p_E}}\right)^{\frac{p_E(p_E-1)}{2}}\right)^{c_{\mathrm{D}_\star}}\\&\quad\times\left(\left(2+\left(\frac{85\ln\left(\frac{\lambda_+}{\lambda_-}\right)}{84\frac{1}{126\beta_\kappa}\frac{\lambda_-}{\lambda_+}\frac{\delta}{p_E}}\right)\right)^{p_E-1}\right)^{c_{\mathrm{A}_\star}}\\&\leq\left(\left(1+\frac{18a\beta_\kappa p_E}{\sqrt{\gamma_\kappa L_-\lambda_-\frac{\lambda_-}{\lambda_+}}\delta}\right)^{p_E}\right)^{c_{\mu_\star}}\times\left(\left(1+\frac{39\beta_\kappa\ln\left(\frac{L_+}{L_-}\right)p_E}{2\delta}\right)\right)^{c_{\mathrm{L}_\star}}\\&\quad\times\left(c_S\left(\frac{126\beta_\kappa\frac{\lambda_+}{\lambda_-}p_E}{\delta}\right)^{\frac{p_E(p_E-1)}{2}}\right)^{c_{\mathrm{D}_\star}}\times\left(\left(2+\left(\frac{255\beta_\kappa\frac{\lambda_+}{\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)p_E}{2\delta}\right)\right)^{p_E-1}\right)^{c_{\mathrm{A}_\star}}\end{aligned}$$

So

$$
\begin{aligned}
&H_{[\cdot],d^{\max}}(\delta/9, \mathcal{G}_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_E^K}) \\
&\quad \leq c_{\mu_\star} p_E \left(\ln\left(1 + \frac{18\beta_\kappa a p_E}{\sqrt{\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}}\right) + \ln\frac{1}{\delta}\right) + c_{\mathrm{L}_\star}\left(\ln\left(1+\frac{39}{2}\beta_\kappa \ln\left(\frac{L_+}{L_-}\right)p_E\right) + \ln\frac{1}{\delta}\right) \\
&\quad + c_{\mathrm{D}_\star}\frac{p_E(p_E-1)}{2}\left(\frac{2\ln c_S}{p_E(p_E-1)} + \ln\left(126\beta_\kappa\frac{\lambda_+}{\lambda_-}p_E\right) + \ln\frac{1}{\delta}\right) \\
&\quad + c_{\mathrm{A}_\star}(p_E-1)\left(\ln\left(2+\frac{255}{2}\beta_\kappa\frac{\lambda_+}{\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)p_E\right) + \ln\frac{1}{\delta}\right)
\end{aligned}
$$

which concludes the proof. □

E.4 Entropy of spatial mixtures (Lemmas)

Proof of Lemma 7. This a variation around the proof of Genovese and Wasserman [24].

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

By construction,

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

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

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

Using

Lemma 11. *For any δ Hellinger brackets $[t^-(x), t^+(x)]$, if for any x $[u^-(x,y), u^+(x,y)]$ is a 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

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

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

By definition,

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

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

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

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

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

to obtain

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

and then using again Lemma 11

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

□

Proof of Lemma 11.

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

Using

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

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

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

□

Proof of Lemma 12.

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

□

Proof of Lemma 13. 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 yields

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

□

E.5 Entropy of Gaussian families (Lemma)

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

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

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

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

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

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

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

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

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

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

and thus

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

which implies

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

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

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

and $\tilde{A}_{p_E,p_E} = \frac{1}{\prod_{i=1}^{p_E-1}\tilde{A}_{i,i}}$ belongs to $\mathcal{G}_{\mathrm{A}}(\lambda_-,\lambda_+,p_E,\delta_{\mathrm{A}})$. 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

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

$$\leq \lambda_-^{-1}\delta_{\mathrm{A}}.$$

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

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

□

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

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

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

proved by Maugis and Michel [39]. This yields using eventually $\kappa \geq \frac{1}{2}$

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

Concerning the Hellinger width,

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

Using

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

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

also proved in [39], we derive

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

Combining

Lemma 16. *For any $0 < \delta \leq \sqrt{2}$ and any $p_E \geq 1$, let $\kappa \geq \frac{3}{4}$ and $\beta_\kappa = \sqrt{\kappa^2\cosh(\frac{\kappa}{6}) + \frac{1}{2}}$, if $\delta_\Sigma \leq \frac{1}{9\beta_\kappa}\frac{\delta}{p_E}$, then*

$$\delta_\Sigma \leq \frac{1}{p_E}\frac{1}{6} \leq \frac{1}{6}.$$

and

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

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

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 $c = \frac{1}{6}$ yields

$$d^2(t^-,t^+) \leq \left(\kappa^2\cosh(\frac{\kappa}{6}) + \frac{1}{2}\right)p_E^2\delta_\Sigma^2 \leq \left(\frac{\delta}{9}\right)^2$$

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

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

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

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

we can apply Lemma 14 on Gaussian density ratio to both

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

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

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

and thus

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

Now as by construction,

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

one obtains

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

It is thus sufficient to prove that

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

or equivalently

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

Now let

$$\begin{aligned}
f_1(\delta_\Sigma) &= \ln\left(\frac{1+\kappa\delta_\Sigma}{\sqrt{1+\delta_\Sigma}\sqrt{1+\frac{1}{2}\delta_\Sigma}} \right) = \ln(1+\kappa\delta_\Sigma) - \frac{1}{2}\ln(1+\delta_\Sigma) - \frac{1}{2}\ln(1+\frac{1}{2}\delta_\Sigma) \\
f_1'(\delta_\Sigma) &= \frac{\kappa}{1+\kappa\delta_\Sigma} - \frac{\frac{1}{2}}{1+\delta_\Sigma} - \frac{\frac{1}{4}}{1+\frac{1}{2}\delta_\Sigma} = \frac{\frac{3}{4}(\kappa-\frac{2}{3})\delta_\Sigma + \kappa - \frac{3}{4}}{(1+\kappa\delta_\Sigma)(1+\delta_\Sigma)(1+\frac{1}{2}\delta_\Sigma)}
\end{aligned}$$

and thus provided $\kappa > \frac{3}{4}$, as $\delta_\Sigma \leq \frac{1}{6}$

$$f_1'(\delta_\Sigma) > \frac{\kappa - \frac{3}{4}}{(1+\frac{\kappa}{6})(1+\frac{1}{6})(1+\frac{1}{12})}$$

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

$$f_1(\delta_\Sigma) > \frac{\kappa - \frac{3}{4}}{(1+\frac{\kappa}{6})(1+\frac{1}{6})(1+\frac{1}{12})}\delta_\Sigma \geq \frac{2}{3}\gamma_k\delta_\Sigma$$

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.

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

Now as

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

and thus

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

one deduces

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

All we need to prove is thus

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

or equivalently

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

Let

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

and thus provided $\kappa > \frac{3}{4}$, as $\delta_\Sigma \leq \frac{1}{6}$

$$f_2'(\delta_\Sigma) > \frac{\kappa - \frac{1}{2}}{(1+\frac{\kappa}{6})(1+\frac{1}{6})}$$

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

$$f_2(\delta_\Sigma) > \frac{\kappa - \frac{1}{2}}{(1+\frac{\kappa}{6})(1+\frac{1}{6})}\delta_\Sigma \geq 2\gamma_\kappa(1+\frac{1}{6})\delta_\Sigma \geq 2\gamma_\kappa(1+\delta_\Sigma)\delta_\Sigma$$

which implies

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

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

Proof of Lemma 16. A straightforward computation yields

$$\delta_\Sigma \leq \frac{1}{9\beta_\kappa}\frac{\delta}{p_E} \leq \frac{1}{p_E}\frac{\sqrt{2}}{9\sqrt{\left(\frac{3}{4}\right)^2 + \frac{1}{2}}} \leq \frac{1}{p_E}\frac{1}{6} \leq \frac{1}{6}.$$

□

Proof of Lemma 17.

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

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

$$f'(x) = \frac{d}{2}\sinh(x)\cosh(x)^{-d/2-1}$$

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

and, as $\cosh(x) \geq 1$,

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

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

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

We deduce thus that

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

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

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

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 18. We deduce this result from a slightly more general:

Lemma 19. *Let $\delta_\Sigma > 0$.*

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

If

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

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

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

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

Indeed Lemma 16 ensures that $\delta_\Sigma \leq \frac{1}{6}$. Hence, if we let $\delta_{\mathrm{L}} = \frac{1}{2}\delta_\Sigma$ and $\delta_{\mathrm{A}} = \delta_{\mathrm{D}} = \frac{1}{14}\frac{\lambda_-}{\lambda_+}\delta_\Sigma$, bounds of the previous Lemma become $\forall x \in \mathbb{R}^{p_E}$,

$$\begin{aligned}
x'\left((1+\delta_\Sigma)\tilde{\Sigma}^{-1} - \Sigma^{-1}\right)x &\geq \tilde{L}^{-1}\left((\delta_\Sigma - \delta_{\mathrm{L}})\lambda_+^{-1} - (1+\delta_\Sigma)\lambda_-^{-1}\left(2\delta_{\mathrm{D}} + \delta_{\mathrm{A}}\right)\right)\|x\|^2 \\
&\geq \tilde{L}^{-1}\left(\left(\delta_\Sigma - \frac{1}{2}\delta_\Sigma\right)\lambda_+^{-1} - (1+\delta_\Sigma)\lambda_-^{-1}3\frac{1}{14}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\right)\|x\|^2 \\
&\geq \frac{1}{4}\tilde{L}^{-1}\frac{1}{\lambda_+}\delta_\Sigma\|x\|^2
\end{aligned}$$

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

$$\begin{aligned}
x'\left(\Sigma^{-1} - (1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1}\right)x &\geq \frac{\tilde{L}^{-1}}{1+\delta_\Sigma}\left(\delta_\Sigma\lambda_+^{-1} - \lambda_-^{-1}\left(2\delta_{\mathrm{D}} + 1\delta_{\mathrm{A}}\right)\right)\|x\|^2 \\
&\geq \frac{\tilde{L}^{-1}}{1+\delta_\Sigma}\left(\delta_\Sigma\lambda_+^{-1} - \lambda_-^{-1}3\frac{1}{14}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\right)\|x\|^2 \\
&\geq \frac{3}{4(1+\delta_\Sigma)}\tilde{L}^{-1}\frac{1}{\lambda_+}\delta_\Sigma\|x\|^2.
\end{aligned}$$

□

Proof of Lemma 19. By definition,

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

Similarly,

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

$$+ (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\sum_{i=1}^{p_E}\tilde{A}_{i,i}^{-1}|D_i'x|^2 - (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\sum_{i=1}^{p_E}\tilde{A}_{i,i}^{-1}|D_{j_D,i}'x|^2$$

Now

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

Furthermore,

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

We then notice that

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

while

$$\begin{aligned}
L^{-1}\sum_{i=1}^{p_E}A_{i,i}^{-1}|D_i'x|^2 - (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\sum_{i=1}^{p_E}A_{i,i}^{-1}|D_i'x|^2 &= \left(L^{-1} - (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\right)\sum_{i=1}^{p_E}A_{i,i}^{-1}|D_i'x|^2 \\
&\geq \left(1 - (1+\delta_\Sigma)^{-1}\right)\tilde{L}^{-1}\lambda_+^{-1}\|x\|^2 \\
&\geq \frac{\delta_\Sigma}{1+\delta_\Sigma}\lambda_+^{-1}\|x\|^2
\end{aligned}$$

We deduce thus that

$$\begin{aligned}
x'\left((1+\delta_\Sigma)\tilde{\Sigma}^{-1} - \Sigma^{-1}\right)x &\geq (\delta_\Sigma - \delta_{\mathrm{L}})\tilde{L}^{-1}\lambda_+^{-1}\|x\|^2 - (1+\delta_\Sigma)\tilde{L}^{-1}\lambda_-^{-1}\left(2\delta_{\mathrm{D}} + 2\delta_{\mathrm{A}}\right)\|x\|^2 \\
&\geq \tilde{L}^{-1}\left((\delta_\Sigma - \delta_{\mathrm{L}})\lambda_+^{-1} - (1+\delta_\Sigma)\lambda_-^{-1}\left(2\delta_{\mathrm{D}} + \delta_{\mathrm{A}}\right)\right)\|x\|^2
\end{aligned}$$

and

$$\begin{aligned}
x'\left(\Sigma^{-1} - (1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1}\right)x &\geq \frac{\delta_\Sigma}{1+\delta_\Sigma}\lambda_+^{-1}\|x\|^2 - (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\lambda_-^{-1}\left(2\delta_{\mathrm{D}} + \delta_{\mathrm{A}}\right)\|x\|^2 \\
&\geq \frac{\tilde{L}^{-1}}{1+\delta_\Sigma}\left(\delta_\Sigma\lambda_+^{-1} - \lambda_-^{-1}\left(2\delta_{\mathrm{D}} + \delta_{\mathrm{A}}\right)\right)\|x\|^2
\end{aligned}$$

□

References

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

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

[3] A. Antoniadis, J. Bigot, and R. von Sachs. A multiscale approach for statistical characterization of functional images. *J. Comput. Graph. Statist.*, 18(1):216–237, 2008.

[4] A. Barron, C. Huang, J. Li, and X. Luo. *MDL Principle, Penalized Likelihood, and Statistical Risk*, chapter in Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday. Tampere University Press, 2008.

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

[6] L. Bertrand, M.-A. Languille, S. Cohen, L. Robinet, C. Gervais, S. Leroy, D. Bernard, E. Le Pennec, W. Josse, J. Doucet, and S. Schöder. European research platform IPANEMA at the SOLEIL synchrotron for ancient and historical materials. *Journal of Synchrotron Radiation*, 18(5), September 2011.

[7] Ch. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. Model-based cluster and discriminant analysis with the MIXMOD software. *Comput. Statist. Data Anal.*, 51(2):587–600, 2006.

[8] L. Birgé. Model selection for gaussian regression with random design. *Bernoulli*, 10(6): 1039–1051, 2004.

[9] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[10] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73, 2007.

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

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

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

[14] O. Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007.

[15] S. Cohen and E. Le Pennec. Conditional density estimation by penalized likelihood model selection and applications. ArXiv 1103.2021, 2011.

[16] S. Cohen and E. Le Pennec. Unsupervised segmentation of hyperspectral images with spatialized gaussian mixture model and model selection. ArXiv, 2012.

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

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

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

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

[21] J. Fan and T. Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.

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

[23] J. Fan, C. Zhang, and J. Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *Ann. Stat.*, 29(1):153–193, 2001.

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

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

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

[27] P. Hall, J. Racine, and Q. Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.

[28] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence*, 1999.

[29] 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.

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

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

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

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

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

[35] M. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

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

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

[38] 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.

[39] C. Maugis and B. Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: P & S*, 2010. To appear.

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

[41] C. Maugis and B. Michel. Data-driven penalty calibration: a case study for Gaussian mixture model selection. *ESAIM: P & S*, 2011. To appear.

[42] 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.

[43] L. Si and R. Jin. Adjusting mixture weights of gaussian mixture model via regularized probabilistic latent semantic analysis. In *Advances in Knowledge Discovery and Data Mining*, pages 218–252, 2005.

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

[45] S. Szarek. Metric entropy of homogeneous spaces. *Quantum Probability (Gdansk 1997)*, pages 395–410, 1998.

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

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

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

[49] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1): 1–25, 1992.

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

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

[52] D. Young and D. Hunter. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, 54(10):2253–2266, 2010.