ESAIM: PS 17 (2013) 698–724
DOI: 10.1051/ps/2012018

ESAIM: Probability and Statistics
www.esaim-ps.org

ADAPTIVE DENSITY ESTIMATION FOR CLUSTERING WITH GAUSSIAN MIXTURES

C. Maugis-Rabusseau[1] and B. Michel[2]

Abstract. Gaussian mixture models are widely used to study clustering problems. These model-based clustering methods require an accurate estimation of the unknown data density by Gaussian mixtures. In Maugis and Michel (2009), a penalized maximum likelihood estimator is proposed for automatically selecting the number of mixture components. In the present paper, a collection of univariate densities whose logarithm is locally β-Hölder with moment and tail conditions are considered. We show that this penalized estimator is minimax adaptive to the β regularity of such densities in the Hellinger sense.

Mathematics Subject Classification. 62G07,62G20.

Received October 17, 2011. Revised May 18, 2012.

1. Introduction

Clustering methods consists of discovering clusters among observations. Many cluster analysis methods have been proposed in statistics and learning theory, roughly fall into three categories. The first one is based on similarity or dissimilarity distances, the best-known are partitioned clustering methods as k-means and the hierarchical clustering methods (see for instance Sects. 14.3.6 and 14.3.12 in [10]). The second category consists of density level set clustering methods which consider clusters as the connected components of high density regions (see [9]). The third category is composed of model-based clustering methods which define clusters as observations having most likely the same distribution. In this last case, each subpopulation is assumed to be distributed from a parametric density, like a Gaussian one and thus the unknown data density is a mixture of these distributions (see for instance [17]). The data clustering is then deduced thanks to the maximum a posteriori (MAP) rule. The clustering problem being based on the data density estimation, it is then essential that this density be efficiently estimated.

Because of their wide range flexibility, Gaussian mixture densities are widely used to model the unknown distribution of continuous data for clustering analysis (see for instance [12, 17]). By recasting the clustering problem into a model selection problem, we have proposed in [16] a non asymptotic penalized criterion. We proved that the selected Gaussian mixture estimator fulfills an oracle inequality. The aim of this new paper is

Keywords and phrases. Rate adaptive density estimation, gaussian mixture clustering, hellinger risk, non asymptotic model selection.

[1] Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse, INSA de Toulouse, 135 avenue de Rangueil, 31077 Toulouse Cedex 4, France. cathy.maugis@insa-toulouse.fr

[2] Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie - Paris 6, 4 place Jussieu, 75252 Paris Cedex 05, France. bertrand.michel@upmc.fr

Article published by EDP Sciences

to investigate the adaptive properties of this estimator in order to justify the validity of our clustering method. More precisely, adapting a recent approximation result, we show that one version of our estimator is minimax adaptive to the regularity parameter of a particular class of Hölder spaces defined below. As far as we know, such a minimax adaptive result has never been shown for a density estimator used for model-based clustering methods.

The particular unidimensional case we study in this paper is now presented. Let us consider n independent identically distributed random variables $X_1, \dots, X_n$ with values in $\mathbb{R}$. Their common unknown density s belongs to the set $\mathcal{S}$ of all density functions with respect to the Lebesgue measure on $\mathbb{R}$. The considered unidimensional Gaussian mixtures are characterized by their number of components m and their means parameters, which are assumed to be bounded. These mixture densities are grouped into a model collection $(\mathcal{S}_m)_{m \in \mathcal{M}_n}$, subsets of $\mathcal{S}$, defined by

$$\mathcal{S}_m = \left\{ x \in \mathbb{R} \mapsto \sum_{u=1}^{m} p_u \psi_\sigma(x - \mu_u); \mu_u \in [-\bar{\mu}(m), \bar{\mu}(m)], \sigma = \lambda(m); p_u \in [0,1], \sum_{u=1}^{m} p_u = 1 \right\} \tag{1}$$

where ψ is the Gaussian kernel defined by $\psi(x) = \pi^{-\frac{1}{2}} \exp(-x^2)$ for all $x \in \mathbb{R}$ and $\psi_\sigma(\cdot) = \sigma^{-1}\psi\left(\frac{\cdot}{\sigma}\right)$ for all $\sigma > 0$. Contrary to [16], the Gaussian mixtures in a model $\mathcal{S}_m$ have a common known variance $\lambda^2(m)$ (the case of unequal variances for mixture components is discussed in Rem. 2.10). The number of free parameters, common to all the mixture densities of a given model $\mathcal{S}_m$ is called dimension and is denoted $D(m) := 2m - 1$. Considering a non asymptotic point of view (see for instance [13]), the bound $\bar{\mu}(m)$ and the variance $\lambda^2(m)$ of each model $\mathcal{S}_m$ and also the maximum number of mixture components in the collection may depend on n. The model collection is indexed by the set $\mathcal{M}_n$ which controls the number of the mixture components. For instance, $\mathcal{M}_n$ could be taken as an interval of integers of the form $[2, \dots, m_{\max}(n)]$ where $m_{\max}$ is an increasing function of n. Such mixture families are called sieves according to the terminology introduced by Grenander [7].

Over each model $\mathcal{S}_m$, a maximum likelihood estimator (MLE) $\hat{s}_m$ is obtained by minimizing the empirical contrast

$$\gamma_n(t) = -\frac{1}{n} \sum_{i=1}^{n} \ln\left\{t(X_i)\right\}.$$

The loss function associated to the likelihood contrast is the Kullback–Leibler divergence: For two densities s and t in $\mathcal{S}$, the Kullback–Leibler divergence is defined by

$$\mathrm{KL}(s,t) = \int \ln\left\{\frac{s(x)}{t(x)}\right\} s(x)\,\mathrm{d}x$$

if $s\mathrm{d}x$ is absolutely continuous with respect to $t\mathrm{d}x$ and $+\infty$ otherwise. The Hellinger distance between two densities g and h of $\mathcal{S}$ is denoted $d_H(g,h) = \frac{1}{\sqrt{2}}\left\|\sqrt{g} - \sqrt{h}\right\|_2$.

Ideally, we would like to estimate the true density s by $\hat{s}_{m^\star}$ (called oracle) where $m^\star$ minimizes over $\mathcal{M}_n$ the Kullback–Leibler risk:

$$m^\star \in \underset{m \in \mathcal{M}_n}{\operatorname{argmin}}\ \mathbb{E}_s[\mathrm{KL}(s, \hat{s}_m)].$$

Nevertheless $m^\star$ and also the associated density $\hat{s}_{m^\star}$ are unknown since they depend on the true density s. Thus we select $\hat{m}$ which minimizes over $\mathcal{M}_n$ the penalized criterion

$$\mathrm{crit}(m) = \gamma_n(\hat{s}_m) + \mathrm{pen}(m)$$

where the penalty function $\mathrm{pen} : m \in \mathcal{M}_n \mapsto \mathrm{pen}(m) \in \mathbb{R}^+$ has to be chosen such that the Kullback–Leibler risk $\mathbb{E}_s[\mathrm{KL}(s, \hat{s}_{\hat{m}})]$ of $\hat{s}_{\hat{m}}$ is close to the oracle risk $\mathbb{E}_s[\mathrm{KL}(s, \hat{s}_{m^\star})]$. The construction of such penalties is proposed in Theorem 2.2 in [16]. This result can be stated as follows for our collection of univariate Gaussian mixtures defined by (1):

Theorem 1.1. *There exists four absolute constants κ, C, c_1 and c_2 such that, if*

$$\mathrm{pen}(m) \geq \kappa \frac{D(m)}{n}\left\{1 + 2\,\mathcal{A}^2 + \ln\left(\frac{1}{1 \wedge \frac{D(m)}{n}\,\mathcal{A}^2}\right)\right\}$$

where

$$\mathcal{A} = c_2 + \sqrt{\ln\left(\frac{c_1 \bar{\mu}(m)}{\lambda(m)}\right)}, \tag{2}$$

then the model $\hat{m}$ minimizing

$$\mathrm{crit}(m) = \gamma_n(\hat{s}_m) + \mathrm{pen}(m)$$

over $\mathcal{M}_n$ exists and

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{\hat{m}})\right] \leq \mathcal{C}\left[\inf_{m \in \mathcal{M}_n}\{\mathrm{KL}(s, \mathcal{S}_m) + \mathrm{pen}(m)\} + \frac{1}{n}\right]\cdot \tag{3}$$

Note that a similar result can be found in [16] for multivariate data clustering with variable selection. The method has been successfully implemented and tested in practice in [15]. It consists of determining the estimator $\hat{s}_m$ using an Em algorithm for each model. Then the penalized criterion allows to select $\hat{s}_{\hat{m}}$. Since this penalized criterion depends on an unknown constant, this last is calibrated using a slope heuristics method as detailed in [1].

Minimax adaptive estimation has been intensively studied in nonparametric statistics, see for instance [13,18] for adaptive minimax methods based on l_0 penalization. A natural optimality criterion is the minimax risk, first introduced by [19]. Let

$$\mathcal{R}(\tilde{s}_n, \mathcal{H}_\beta) = \sup_{s \in \mathcal{H}_\beta} \mathbb{E}_s[d_H^2(s, \tilde{s}_n)]$$

be the maximal Hellinger risk of an estimator $\tilde{s}_n$ of s. The minimax Hellinger risk on a density class $\mathcal{H}_\beta$ is then defined by

$$\mathcal{R}_n(\mathcal{H}_\beta) = \inf_{\tilde{s}_n} \mathcal{R}(\tilde{s}_n, \mathcal{H}_\beta)$$

where the infimum is taken over all the possible estimators $\tilde{s}_n$ of s. An estimator is said to be minimax on $\mathcal{H}_\beta$ if its maximal risk over $\mathcal{H}_\beta$ reaches the minimax risk on this density class. Let us now consider a collection $(\mathcal{H}_\beta)_{\beta \in \mathcal{B}}$ of density classes indexed by a set $\mathcal{B}$ of regularity parameters β. An estimator is said to be minimax adaptive if it reaches the minimax risk over $\mathcal{H}_\beta$ for all β of $\mathcal{B}$, without using the knowledge of β. In order to motivate the clustering method based on Gaussian mixture estimator $\hat{s}_{\hat{m}}$ proposed in [16], we prove in this new paper that this estimator is minimax adaptive over a particular collection of Hölder density classes $(\mathcal{H}_\beta)_{\beta \in \mathcal{B}}$ defined below. Of course, adaptive density estimation in one dimension is now a classical problem and several adaptive estimators have been already proposed such as kernel estimators or thresholding wavelet estimators. Nevertheless, although these alternative methods maybe perform better than our penalized estimator $\hat{s}_{\hat{m}}$ concerning density estimation in general, these are of no interest for clustering purposes.

The link between model selection and adaptive estimation is made through approximation theory. Indeed, an adaptive estimation is possible only for functional classes $\mathcal{H}_\beta$ that can be efficiently approximated by our Gaussian mixture collection. Convolution is widely used in approximation theory and many results are known on this topic. It is well known that the convolution of a density f with scaled versions ψ_σ of the Gaussian kernel ψ converges to f (see for instance [3], Chap. 20). The so-called quasi-interpolation method consists of replacing the functions $\psi_\sigma * f$ by infinite linear combinations of scaled and translated Gaussian kernels (see for instance [3], Chap. 36). In a recent paper of Hangelbroek and Ron [8], a nonlinear approximation algorithm based on finite combinations of scaled and translated Gaussian kernels is defined to give some approximation results in L^p norm on some particular density classes. Nevertheless, all these results cannot be straightly applied to study the approximation capacities of Gaussian mixtures. Indeed, the coefficients in these linear combinations are not necessary positive and their sum is not constrained to be equal to one. Furthermore, the approximation results provided by all these methods are not given for the Kullback–Leibler divergence as required by our statistical context.

The approximation capacity of Gaussian mixtures has also been studied in non parametric Bayesian works. Lemma 3.1 in [5] gives a discretization result for Gaussian mixtures: assume that s is a location or location-scale mixture with a mixing distribution compactly supported or with sub-Gaussian tails, s can be approximated by a finite Gaussian mixture with a small number of components, the error being controlled in L_1 and L_∞ norms. In [6], these authors take advantage of this method for approximating by finite Gaussian mixtures some twice continuously differentiable functions with additional regularity conditions. More recently, Kruijer *et al.* [11] prove an approximation result for finite Gaussian mixtures for densities whose logarithm is locally Hölder. Their approximation result is given for the Kullback–Leibler divergence. This last result can be successfully adapted in our context to control the bias term in the right side term of the oracle inequality (3) on these particular density classes. Concerning approximation, the contribution of our work consists of checking that the non explicit constants of the approximation bounds given in [11] are actually uniform over a density class $\mathcal{H}_\beta$ which we define below. For easier reading, the proofs of the approximation results we need are not all given in this paper, they can be found in detail in the preprint version [14] of this work.

The paper is organized as follows: The main results are presented in Section 2. The density classes $\mathcal{H}_\beta$ are introduced in Section 2.1 and an approximation result adapted from [11] is given in Section 2.2. Next, a lower bound for the minimax risk is given in Section 2.3 and the adaptive property of our penalized Gaussian mixture estimator on the density classes $\mathcal{H}_\beta$ is addressed in Section 2.4. The approximation result, the lower bound and the adaptive result are respectively proved in Sections 4, 5 and 6. Some technical results are also given in Appendices 6 and A.2.

2. Main results

2.1. The density classes $\mathcal{H}(\beta, \mathcal{P})$

The adaptation result given below requires a slightly modified version of the approximation result by finite Gaussian mixtures proved in [11]. This approximation result concerns densities whose logarithm is locally β-Hölder and that fulfills additional tail, moments and monotonicity conditions. More precisely, let $\beta > 0$ and let $r = \lfloor \beta \rfloor$ be the largest integer less than β and $k \in \mathbb{N}$ such that $\beta \in (2k, 2k+2]$. Let also $\mathcal{P}$ be the set of parameters $\{\gamma, l^+, L, \varepsilon, C, \alpha, \xi, M\}$ where L is a polynomial function on $\mathbb{R}$ and the other parameters are positive constants. We then define the density class $\mathcal{H}(\beta, \mathcal{P})$ of all densities f satisfying the following conditions:

1. **Smoothness.** $\ln f$ is assumed to be locally β-Hölder: for all x and y such that $|y - x| \leq \gamma$,

$$\left|(\ln f)^{(r)}(x) - (\ln f)^{(r)}(y)\right| \leq r!\, L(x)|y - x|^{\beta - r}. \tag{4}$$

   Furthermore for all $j \in \{0, \ldots, r\}$,

$$|(\ln f)^{(j)}(0)| \leq l^+. \tag{5}$$

2. **Moments.** The derivative functions $(\ln f)^{(j)}$ for $j = 1, \ldots, r$ and the polynomial function L fulfill

$$\int_{\mathbb{R}} \left|(\ln f)^{(j)}(x)\right|^{\frac{2\beta+\varepsilon}{j}} f(x)\mathrm{d}x \leq C\ , \qquad \int_{\mathbb{R}} |L(x)|^{2+\frac{\varepsilon}{\beta}} f(x)\mathrm{d}x \leq C. \tag{6}$$

3. **Tail.** For all $x \in \mathbb{R}$,

$$f(x) \leq M\psi(x). \tag{7}$$

4. **Monotonicity.** f is strictly positive, f is nondecreasing on $(-\infty,-\alpha)$ and nonincreasing on (α,∞), and $f(x)\geq\xi$ for all $x\in[-\alpha,\alpha]$.

Remark 2.1. The monotonicity assumption can be relaxed by assuming that there exist two constants $c>0$ and $\bar{\sigma}>0$ such that $\forall 0<\sigma<\bar{\sigma},\ \forall x\in\mathbb{R}$,

$$\frac{K_\sigma f(x)}{f(x)}\geq c.$$

This condition corresponds to the first point given in Lemma A.4 in Appendix 6 which is a key point to prove the approximation result. In the following, the strong monotonicity condition is assumed in the definition of the density class $\mathcal{H}(\beta,\mathcal{P})$ to simplify the proofs of the lower bound.

Remark 2.2. For easier reading, the monotonicity assumption is stated on a symmetric interval but it could be possible to consider this assumption on a general interval $[\alpha_1,\alpha_2]$ with $\alpha_1<\alpha_2$. This monotonicity assumption allows us to lower bound the convolution $f*\psi_\sigma$ by f up to a multiplicative constant according to Remark 3 in [4].

Remark 2.3. These density classes are more restrictive than those considered in [11]: Indeed the upper bounds in (6) have to be uniform over the density class $\mathcal{H}(\beta,\mathcal{P})$ and we also need the additional Condition (5). These restrictions allow us to control the Kullback–Leibler divergence between a density of $\mathcal{H}(\beta,\mathcal{P})$ and a convenient finite Gaussian mixture, uniformly over $\mathcal{H}(\beta,\mathcal{P})$. Note that Condition (7) is here assumed on $\mathbb{R}$ but it could be assumed only outside an interval as in [11].

Remark 2.4. In the sequel, $\mathcal{P}'$ is said to be "larger than" $\mathcal{P}$ if at least one of the following conditions is fulfilled:

- at least one constant among M, C or l^+ of $\mathcal{P}'$ is larger than the corresponding one of $\mathcal{P}$,
- the constant γ of $\mathcal{P}'$ is smaller than the corresponding one of $\mathcal{P}$,
- for all $x\in\mathbb{R},\ L(x)\leq L'(x)$ where L (resp. L') belongs to $\mathcal{P}$ (resp. $\mathcal{P}'$)

2.2. Approximation result

For any function f, $K_\sigma f$ denotes the convolution $f*\psi_\sigma$ and $\Delta_\sigma f$ is the error term $K_\sigma f-f$. As explained in [11], for a β-smooth density f with $\beta\leq 2$ and under reasonable regularity assumptions, it is possible to define a finite location-scale Gaussian mixture $\wp_\sigma$ such that $\mathrm{KL}(f,\wp_\sigma)=O\left(\sigma^{2\beta}\right)$. The usual approach consists of discretizing the continuous mixture $K_\sigma f$. But as $\|f-K_\sigma f\|_\infty$ remains of order σ^2 when $\beta>2$, this approach appears to be inefficient for smoother densities. An alternative strategy is proposed in Kruijer *et al.* [11], based on the following successive convolutions of f: $f_0=f$ and for all $j\geq 0$, $f_{j+1}=f-\Delta_\sigma f_j$. In their paper, the density is approximated by a discretized version of the continuous mixture $K_\sigma f_k$ where $k\in\mathbb{N}$ is such that $\beta\in(2k,2k+2]$.

In our framework, Lemma 4 in [11] cannot be directly used since the upper bound over the Kullback–Leibler divergence between f and the finite Gaussian mixture is not uniform over $\mathcal{H}(\beta,\mathcal{P})$. Thus, some additional work is necessary to obtain an uniform version of this approximation result. Another reason for revisiting the approximation results given in [11] is that these ones are stated for $\sigma\leq\bar{\sigma}$ where $\bar{\sigma}$ depends on the approximated density f. Thus we also need to check that it is possible to choose the same $\bar{\sigma}$ for all the densities of $\mathcal{H}\left(\beta,\mathcal{P}\right)$. The proof of Theorem 2.5 consists of carefully following the method proposed in [11] in order to obtain this uniform version. The main steps of the proof are given in Section 4. A complete and self-contained proof is detailed in our preprint version [14].

Theorem 2.5. *There exists a positive constant $\bar{\sigma}(\beta)<1$ such that for all $f\in\mathcal{H}(\beta,\mathcal{P})$ and for all $\sigma<\bar{\sigma}(\beta)$, there exists a finite Gaussian mixture of density $\wp_\sigma$ with less than $G_\beta\sigma^{-1}|\ln\sigma|^{\frac{3}{2}}$ support points, with the same variance σ for each component and with means belonging to $[-\mu_\sigma,\mu_\sigma]$ where*

$$\mu_\sigma\leq\tilde{G}_\beta|\ln\sigma|^{\frac{1}{2}}$$

such that

$$\mathrm{KL}(f, \wp_\sigma) = \int_{\mathbb{R}} f(x) \ln\left(\frac{f(x)}{\wp_\sigma(x)}\right) \mathrm{d}x \leq c_\beta \, \sigma^{2\beta} \tag{8}$$

where c_β is uniform on $\mathcal{H}(\beta, \mathcal{P})$ and continuous on β. The constant $\bar{\sigma}(\beta)$ only depends on $\mathcal{H}(\beta, \mathcal{P})$ and is a continuous function of β. Moreover, G_β and $\tilde{G}_\beta$ are two positive constants that only depend on $\mathcal{H}(\beta, \mathcal{P})$, and are both increasing functions of β.

The two constants G_β and $\tilde{G}_\beta$ are explicitly defined by equations (24) and (25) in the proof of Theorem 2.5 in Section 4.2.

2.3. Lower bound

In order to show that the MLE penalized estimator $\hat{s}_{\hat{m}}$ is adaptive to the smoothness parameter β, a lower bound of the minimax risk $\mathcal{R}_n(\mathcal{H}(\beta, \mathcal{P}))$ is required. For all $0 < \underline{\beta} < \bar{\beta}$, a "large enough" parameter set $\mathcal{P}(\underline{\beta}, \bar{\beta})$ is found such that for all $\beta \in [\underline{\beta}, \bar{\beta}]$, $\mathcal{H}(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta}))$ is well defined and a lower bound is given for the density classes $\mathcal{H}\left(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta})\right)$. Note that in Theorem 2.5, the constants c_β, $\bar{\sigma}(\beta)$, G_β and $\tilde{G}_\beta$ cannot be bounded uniformly for all $\beta \in \mathbb{R}^+$. Nevertheless, it can be proved that $\hat{s}_{\hat{m}}$ is minimax adaptive on a range of regularity $[\underline{\beta}, \bar{\beta}]$.

First, the parameter set $\mathcal{P}(\underline{\beta}, \bar{\beta})$ has to be defined rigorously. Its definition is rather technical since it depends on the way the lower bound is proved. The proof is based on the construction of some oscillating functions, this standard method is presented for instance in [13] (see Sect. 7.5). Let us take some infinitely differentiable function $\varphi : \mathbb{R} \to \mathbb{R}$ with compact support included into $(\frac{1}{4}, \frac{3}{4})$ such that

$$\int_{\mathbb{R}} \varphi(x) \mathrm{d}x = 0 \text{ and } \int_{\mathbb{R}} \varphi(x)^2 \mathrm{d}x = 1.$$

We set $A = \max\limits_{0 \leq k \leq r+1} \|\varphi^{(k)}\|_\infty > 1$ and let D be some positive even integer. For any positive integer $j \in \{1, \ldots, D\}$, we consider the function

$$\begin{aligned} \varphi_j : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto \tfrac{\xi D^{-\beta}}{A} \varphi\left(\tfrac{D}{\alpha}(x + \tfrac{\alpha}{2}) - (j-1)\right). \end{aligned}$$

Moreover, let $\mathcal{T}(\alpha, \xi)$ be the space of functions $\omega : \mathbb{R} \to \mathbb{R}^+$ such that w is nondecreasing on $(-\infty, -\frac{\alpha}{2})$, nonincreasing on $(\frac{\alpha}{2}, +\infty)$, $\omega(x) = 2\xi$ for all $x \in \left[-\frac{3\alpha}{4}, \frac{3\alpha}{4}\right]$, and $\omega(-\alpha) = \omega(\alpha) = \xi$.

Next, let $\tilde{\mathcal{P}} = \left\{\frac{\alpha}{4}, \ln(2\xi), \tilde{L}, \tilde{\varepsilon}, \tilde{C}, \alpha, \xi, \tilde{M}\right\}$ be a parameter set such that $\mathcal{T}(\alpha, \xi) \bigcap \mathcal{H}(\beta, \tilde{\mathcal{P}})$ is nonempty. Based on a function $\omega \in \mathcal{T}(\alpha, \xi) \bigcap \mathcal{H}(\beta, \tilde{\mathcal{P}})$ and the functions φ_j, we consider the functional space $\mathcal{J}(\beta, D) = \left\{f_\theta;\ \theta \in \{0,1\}^D\right\}$ where for all $\theta \in \{0,1\}^D$ and for all $x \in \mathbb{R}$,

$$f_\theta(x) = \omega(x) + \sum_{j=1}^{D} (2\theta_j - 1)\varphi_j(x). \tag{9}$$

Proposition 2.6. *There exists a parameter set $\mathcal{P}(\underline{\beta}, \bar{\beta})$ such that for all $D \in \mathbb{N}^*$ and for all $\beta \in [\underline{\beta}, \bar{\beta}]$,*

$$\mathcal{J}(\beta, D) \subset \mathcal{H}\left(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta})\right).$$

Remark 2.7. Note that if such a parameter set exists, Proposition 2.6 is also true for all the parameter sets larger than it (in the sense given in Rem. 2.4). A key point to prove the lower bound stated in the next theorem is that the parameter set $\mathcal{P}(\underline{\beta}, \bar{\beta})$ does not depend on D.

Theorem 2.8. *Suppose that one observes independent random variables $X_1, \dots, X_n$ with common density s with respect the Lebesgue measure on $\mathbb{R}$. For any $\beta \in [\underline{\beta}, \bar{\beta}]$ and any parameter set $\mathcal{P}(\underline{\beta}, \bar{\beta})$ given by Proposition 2.6, there exists a positive constant κ_β such that*

$$\mathcal{R}_n(\mathcal{H}(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta}))) := \inf_{\tilde{s}} \sup_{s} \mathbb{E}[d_H^2(s, \tilde{s})] \geq \kappa_\beta \; n^{-\frac{2\beta}{2\beta+1}}$$

where the supremum (resp. the infimum) is taken over all densities s in $\mathcal{H}(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta}))$ (resp. over all possible estimators $\tilde{s}$ of s).

Proposition 2.6 and Theorem 2.8 are proved in Section 5.1 and Section 5.2 respectively. After establishing Proposition 2.6, the Hellinger distance and the Kullback–Leibler divergence between two functions of $\mathcal{J}(\beta, D)$ are controlled in Lemma 5.3 and Lemma 5.4 respectively. These controls are required to combine a corollary of a Birgé's Lemma (see [2]) and the so-called Varshamov–Gilbert's Lemma. These last two results can be found in [13] (see Cor. 2.19 and Lem. 4.7), they are also reminded in Appendix A.2.

2.4. Adaptive density estimation

In a non asymptotic model selection approach, the model collection may increase with the sample size n, leading to an adaptive procedure. As it was already explained, the adaptive properties of $\hat{s}_{\hat{m}}$ are studied on a range of regularity $[\underline{\beta}, \bar{\beta}]$. Preliminary, we fix $0 < \underline{\beta} < \bar{\beta}$ and we also choose $a_{\bar{\beta}} > 1$ large enough such that

$$\frac{G_{\bar{\beta}}}{a_{\bar{\beta}}} \left(\frac{\ln a_{\bar{\beta}}}{\ln 2} + 3 \right)^{3/2} \leq 1, \tag{10}$$

where $G_{\bar{\beta}}$ is defined in Theorem 2.5. The parameters of the Gaussian mixture models $(\mathcal{S}_m)_{m \in \mathcal{M}_n}$ are now specified in order to apply the approximation results provided by Theorem 2.5:

$$\mathcal{S}_m = \left\{ x \in \mathbb{R} \mapsto \sum_{u=1}^{m} p_u \psi_\sigma(x - \mu_u); \mu_u \in [-\bar{\mu}(m), \bar{\mu}(m)], \sigma = \lambda(m), p_u \in [0,1], \sum_{u=1}^{m} p_u = 1 \right\}$$

where $\lambda(m) := a_{\bar{\beta}} m^{-1} (\ln m)^{3/2}$ and $\bar{\mu}(m) := \tilde{G}_{\bar{\beta}} |\ln \lambda(m)|^{1/2}$ ($\tilde{G}_{\bar{\beta}}$ is defined in Thm. 2.5) for all m. Since the sample size is n, it is natural to suppose that the number of mixture components m is less than n and we also assume that the mixtures have at least two components: $\mathcal{M}_n = \{2, \dots, n\}$. Note that when the sample size n increases, mixtures with small component variances and many components m are available in the model collection. This obviously improves the approximation capacity of the Gaussian mixtures.

Theorem 2.9. *Assume that $n \geq 3$ and let $\hat{s}_{\hat{m}}$ be the penalized maximum likelihood estimator minimizing the penalized criterion defined in Theorem 1.1. Then there exists a constant $c_{\underline{\beta}, \bar{\beta}}$ such that for all $\beta \in [\underline{\beta}, \bar{\beta}]$ and for all $s \in \mathcal{H}(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta}))$,*

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{\hat{m}})\right] \leq c_{\underline{\beta}, \bar{\beta}} \; (\ln n)^{\frac{5\beta}{2\beta+1}} \; n^{\frac{-2\beta}{2\beta+1}}.$$

Theorem 2.9 shows that the penalized estimator $\hat{s}_{\hat{m}}$ is minimax adaptive to the regularity β of the density classes defined in Section 2.1, up to a power of $\ln(n)$. This logarithm term is due to the penalty shape given in Theorem 1.1. It is not detected in practice as shown in [15] and we suspect that it could be removed from the penalty shape. Note that the non parametric Bayesian estimator defined in [11] has a similar rate of convergence with a greater power of the logarithm term.

Remark 2.10. Contrary to [16], the Gaussian mixtures considered in this paper have a common and known variance $\lambda(m)$. An analogous result of Theorem 2.9 for Gaussian mixtures with unknown and unequal variances can be easily stated. More precisely, let us consider

$$\tilde{\mathcal{S}}_m = \left\{ x \in \mathbb{R} \mapsto \sum_{u=1}^{m} p_u \psi_{\sigma_u}(x - \mu_u); \mu_u \in [-\bar{\mu}(m), \bar{\mu}(m)], \sigma_u \in [\lambda(m), \bar{\lambda}], p_u \in [0,1], \sum_{u=1}^{m} p_u = 1 \right\}.$$

Since $\mathcal{S}_m$ is included in $\tilde{\mathcal{S}}_m$, the approximation results for $\mathcal{S}_m$ are also valid for $\tilde{\mathcal{S}}_m$. Starting from the oracle inequality given in [16] for $\tilde{\mathcal{S}}_m$ (the constant $\mathcal{A}$ being modified, see [14]), we then deduce that the penalized estimator $\hat{s}_{\hat{m}}$, defined on the models $\tilde{\mathcal{S}}_m$, is minimax adaptive to the regularity β of the density classes defined in Section 2.1. Note that this result is actually weaker than Theorem 2.9 since the models $\tilde{\mathcal{S}}_m$ are larger than the models $\mathcal{S}_m$; one should obtain adaptation on a much larger class than uniform Hölderness.

3. Conclusion

In this paper, the penalized estimator $\hat{s}_{\hat{m}}$, defined in [16] for Gaussian mixture models having the same and known component variance, is shown to be adaptive to the regularity of density classes $\mathcal{H}_\beta$ whose elements are univariate densities whose logarithm is locally β-Hölder. To prove this result, the approximation result given in [11] has been adapted to control the bias term between Gaussian mixture models and the density spaces $\mathcal{H}_\beta$. A lower bound for the minimax risk on the density classes $\mathcal{H}_\beta$ has also been stated to finally prove that our estimator reaches the minimax rate.

As noted in Remark 2.10, a similar (but weaker) result can be given for the adaptation of penalized estimators defined on Gaussian mixture models with unknown and unequal component variances. In this context, the approximation method of [11] probably does not provide the most general result; one should obtain adaptation on a much larger class than uniform Hölderness. This question could be tackled in future works.

In [16], a Gaussian mixture estimator, fulfilling an oracle inequality as (3), is proposed in the context of multivariate data clustering. In a future work, it would be interesting to extend our adaptive result to this multivariate case. This requires to state an approximation result as Theorem 2.5 on multivariate density classes which have to be determined, that is obviously a technical task.

4. Proof of the approximation result

In this section, the density space $\mathcal{H}(\beta,\mathcal{P})$ is fixed. To make the proofs and the results easier to read, we use the notation c_β (resp. $\bar{\sigma}(\beta)$) for denoting constants (resp. upper bound on σ) that only depends on β and $\mathcal{P}$. We also use the notation $c_{\beta,p}$ (resp. $\bar{\sigma}(\beta,p)$) if it also depends on an other parameter p. We also denote as $l_j(.)$ the jth derivative $\frac{d^j}{\mathrm{d}x^j}\ln f(x)$ of $\ln f$ and we consider a subset A_σ defined by

$$A_\sigma := \left\{x\in\mathbb{R};\ |l_j(x)| \le \mathfrak{B}\sigma^{-j}|\ln\sigma|^{-j/2}, j=1\ldots r,\ L(x)\le \mathfrak{B}\sigma^{-\beta}|\ln\sigma|^{-\beta/2}\ \right\}$$

if $\beta>1$ and $A_\sigma := \left\{x\in\mathbb{R};\ L(x)\le\mathfrak{B}\sigma^{-\beta}|\ln\sigma|^{-\beta/2}\ \right\}$ otherwise.

4.1. Approximation by a continuous mixture

The aim of this section consists of controlling uniformly the Kullback–Leibler divergence between a density f of $\mathcal{H}(\beta,\mathcal{P})$ and a continuous Gaussian mixture. Lemmas 1 and 2, and Theorem 1 in [11] are here adapted in Lemma 4.1, Lemma 4.2 and Proposition 4.4 respectively. The goal is to obtain approximation bounds with constants which are uniform over $\mathcal{H}(\beta,\mathcal{P})$. The proofs of Lemma 4.1 and Lemma 4.2 are available in [14], they are not reported in this paper. On the other hand, an alternative and simpler proof of Proposition 4.4 is proposed in this section.

Lemma 4.1. *Let $\beta>0$ and $k\in\mathbb{N}$ such that $\beta\in(2k,2k+2]$. For all $H>0$, there exists $\bar{\sigma}(\beta,H)>0$ such that for all $\sigma<\bar{\sigma}(\beta,H)$, for all $f\in\mathcal{H}(\beta,\mathcal{P})$ and for all $x\in A_\sigma$ we have*

$$(K_\sigma f_k)(x) = f(x)\left[1+R_f(x)O_{\beta,H}(\sigma^\beta)\right] + O_{\beta,H}\left(\sigma^H\right)$$

with $R_f(x)=a_{r+1}L(x)$ if $\beta\le 1$, and

$$R_f(x) = a_{r+1}L(x) + \sum_{j=1}^{r} a_j\ |l_j(x)|^{\frac{\beta}{j}}$$

otherwise. In both cases, the a_j's are nonnegative constants that are uniform on $\mathcal{H}(\beta,\mathcal{P})$. Furthermore, $\bar{\sigma}(\beta,H)$ is a continuous function of β and H.

For a density f belonging to $\mathcal{H}(\beta,\mathcal{P})$, Lemma 4.1 shows that the convolution $K_\sigma f_k$ is close to f on a subspace of $\mathbb{R}$ where the derivative functions of $\ln f$ and L are efficiently controlled. Furthermore, the control on the difference $K_\sigma f_k - f$ is uniform over $\mathcal{H}(\beta,\mathcal{P})$, which is required to upper bound the Kullback–Leibler divergence between f and $K_\sigma f_k$. Thus $K_\sigma f_k$ seems to be a good candidate to approximate the density function f. Nevertheless, the function f_k is not a density function: Its integral over $\mathbb{R}$ is equal to 1 (see Lem. A.3) but it can take negative values. To remedy this problem, Kruijer *et al.* [11] define a density function h_k as follows: Considering the subspace

$$J_{\sigma,k} = \left\{ x\in\mathbb{R};\ f_k(x) > \frac{1}{2}f(x)\right\},$$

the following positive function is defined

$$\forall x\in\mathbb{R},\ g_k(x) = f_k(x)\mathbb{1}_{J_{\sigma,k}}(x) + \frac{1}{2}f(x)\mathbb{1}_{J^c_{\sigma,k}}(x)$$

and it is normalized to obtain a density function

$$\forall x\in\mathbb{R},\ h_k(x) = \frac{g_k(x)}{\int g_k(u)\mathrm{d}u}\cdot \tag{11}$$

Now, the result of Lemma 4.1 has to be extended for the convolution $K_\sigma h_k$. For this purpose, the integral of $K^t_\sigma f$ for all nonnegative integers $t\le k$ is controlled over A^c_σ and E^c_σ where A_σ is defined by (4) and $E_\sigma = \{x\in\mathbb{R};\ f(x)\ge\sigma^{H_1}\}$ with $H_1 > 4\beta$.

Lemma 4.2. *Let $\beta>0$ and $k\in\mathbb{N}$ such that $\beta\in(2k,2k+2]$. There exists $\bar{\sigma}(\beta,H_1)>0$ such that for all $\sigma\le\bar{\sigma}(\beta,H_1)$, for all $f\in\mathcal{H}(\beta,\mathcal{P})$ and for all nonnegative integers $t\le k$,*

$$\int_{A^c_\sigma}\left(K^t_\sigma f\right)(x)\mathrm{d}x = O_\beta\left(\sigma^{2\beta}\right) \tag{12}$$

and

$$\int_{E^c_\sigma}\left(K^t_\sigma f\right)(x)\mathrm{d}x = O_{\beta,H_1}\left(\sigma^{2\beta}\right). \tag{13}$$

Furthermore, for $\sigma\le\bar{\sigma}(\beta,H_1)$, $A_\sigma\cap E_\sigma\subset J_{\sigma,k}$ and

$$\int_{\mathbb{R}} g_k(x)\ \mathrm{d}x = 1 + O_{\beta,H_1}\left(\sigma^{2\beta}\right). \tag{14}$$

Thus, for all $H>0$, there exists $\bar{\sigma}(\beta,H_1,H)>0$ such that for all $\sigma\le\bar{\sigma}(\beta,H_1,H)$ and for all $x\in A_\sigma\cap E_\sigma$,

$$|(K_\sigma h_k)(x) - f(x)| = f(x)R_f(x)O_{\beta,H_1,H}\left(\sigma^{\beta}\right) + O_{\beta,H_1,H}\left(\sigma^{H}\right). \tag{15}$$

Furthermore, $\bar{\sigma}(\beta,H_1)$ and $\bar{\sigma}(\beta,H_1,H)$ are both continuous functions of β, H_1 and H for the last one.

Remark 4.3. The left term in (14) does not depend on H_1 whereas the right term does. Indeed, the presence of H_1 here is only technical and by choosing for instance $H_1 = 4\beta+1$, it gives that there exists a positive constant $\bar{\sigma}(\beta)$, continuous in β such that for all $\sigma\le\bar{\sigma}(\beta)$,

$$\int_{\mathbb{R}} g_k(x)\mathrm{d}x = 1 + O_\beta\left(\sigma^{2\beta}\right). \tag{16}$$

Now, a control of the Kullback–Leibler divergence between a density f of $\mathcal{H}(\beta,\mathcal{P})$ and the associated continuous mixture $K_\sigma h_k$ is established, uniformly over $\mathcal{H}(\beta,\mathcal{P})$.

Proposition 4.4. *Let $\beta > 0$ and $k \in \mathbb{N}$ such that $\beta \in (2k, 2k+2]$. There exists a positive constant $\bar{\sigma}(\beta)$ such that for all $f \in \mathcal{H}(\beta,\mathcal{P})$ and all $\sigma < \bar{\sigma}(\beta)$,*

$$KL(f, K_\sigma h_k) = \int_{\mathbb{R}} f(x) \ln\left(\frac{f(x)}{K_\sigma h_k(x)}\right) \mathrm{d}x = O_\beta\left(\sigma^{2\beta}\right)$$

where h_k is defined by (11) and where $\bar{\sigma}(\beta)$ can be chosen as a continuous function of β.

Proof. Preliminary, we remark that if p and q are two densities and S is a set, then

$$\int_S p \ln\left(\frac{p}{q}\right) \leq \int_S p\,\frac{p-q}{q} = \int_S \frac{(p-q)^2}{q} + \frac{q(p-q)}{q} = \int_S \frac{(p-q)^2}{q} + \int_{S^c} (q-p)$$

since $\int_S p = 1 - \int_{S^c} p$, $\int_S q = 1 - \int_{S^c} q$ and $\int_S (p-q) = \int_{S^c}(q-p)$. We use this inequality with the densities f and $K_\sigma h_k$, and the sets A_σ and E_σ, where E_σ is defined with $H_1 = 4\beta+1$, to obtain the following control of $\mathrm{KL}(f, K_\sigma h_k)$:

$$\begin{aligned}\int_{\mathbb{R}} f(x) \ln\left(\frac{f(x)}{K_\sigma h_k(x)}\right) \mathrm{d}x &= \int_{A_\sigma \cap E_\sigma} f(x) \ln\left(\frac{f(x)}{K_\sigma h_k(x)}\right) \mathrm{d}x + \int_{A_\sigma^c \cup E_\sigma^c} f(x) \ln\left(\frac{f(x)}{K_\sigma h_k(x)}\right) \mathrm{d}x \\ &\leq \int_{A_\sigma \cap E_\sigma} \frac{[f(x) - K_\sigma h_k(x)]^2}{K_\sigma h_k(x)} \mathrm{d}x \qquad (17) \\ &\quad + \int_{A_\sigma^c \cup E_\sigma^c} [K_\sigma h_k(x) - f(x)] \mathrm{d}x \qquad (18) \\ &\quad + \int_{A_\sigma^c \cup E_\sigma^c} f(x) \ln\left(\frac{f(x)}{K_\sigma h_k(x)}\right) \mathrm{d}x. \qquad (19)\end{aligned}$$

• **Control of (17)**: Let $H > 0$. According to Lemma 4.2 with $H_1 = 4\beta+1$, there exists $\bar{\sigma}(\beta, H) > 0$ such that for all $x \in A_\sigma \cap E_\sigma$ and for all $\sigma < \bar{\sigma}(\beta, H)$, $[K_\sigma h_k(x) - f(x)]^2 \leq \left[\Lambda_{\beta,H} f(x) R_f(x) \sigma^\beta + \Omega_{\beta,H} \sigma^H\right]^2$ where $\Lambda_{\beta,H}$ and $\Omega_{\beta,H}$ are two constants. Moreover, according to Lemma A.4, there exists $\bar{\sigma}(\beta) > 0$ such that for all $\sigma < \bar{\sigma}(\beta)$,

$$K_\sigma h_k(x) \geq \frac{D}{1 + A_\beta \sigma^{2\beta}} f(x)$$

with $D = \frac{\xi\sqrt{\pi}}{6M}$. Thus for all $\sigma < \bar{\sigma}(\beta, H) \wedge \bar{\sigma}(\beta)$,

$$\begin{aligned}\frac{[f(x) - K_\sigma h_k(x)]^2}{K_\sigma h_k(x)} \leq{}& \frac{\Lambda_{\beta,H}^2}{D}(1 + A_\beta \sigma^{2\beta})\sigma^{2\beta} R_f(x)^2 f(x) + \frac{\Omega_{\beta,H}^2}{D}(1 + A_\beta \sigma^{2\beta})\sigma^{2H} \frac{1}{f(x)} \\ &+ \frac{2\Lambda_{\beta,H}\Omega_{\beta,H}}{D}(1 + A_\beta \sigma^{2\beta})\sigma^{\beta+H} R_f(x).\end{aligned}$$

Then,

$$\begin{aligned}\int_{A_\sigma \cap E_\sigma} \frac{[f(x) - K_\sigma h_k(x)]^2}{K_\sigma h_k(x)} \mathrm{d}x \leq{}& \frac{\Lambda_{\beta,H}^2}{D}(1 + A_\beta \sigma^{2\beta})\sigma^{2\beta} \int_{A_\sigma \cap E_\sigma} R_f(x)^2 f(x) \mathrm{d}x \\ &+ \frac{\Omega_{\beta,H}^2}{D}(1 + A_\beta \sigma^{2\beta})\sigma^{2H - 2(4\beta+1)} \int_{A_\sigma \cap E_\sigma} f(x) \mathrm{d}x \\ &+ \frac{2\Lambda_{\beta,H}\Omega_{\beta,H}}{D}(1 + A_\beta \sigma^{2\beta})\sigma^{\beta+H-4\beta-1} \int_{A_\sigma \cap E_\sigma} R_f(x) f(x) \mathrm{d}x. \qquad (20)\end{aligned}$$

Thus the two integrals $\int_{A_\sigma \cap E_\sigma} R_f(x)^2 f(x) \mathrm{d}x$ and $\int_{A_\sigma \cap E_\sigma} R_f(x) f(x) \mathrm{d}x$ have to be controlled.

The first integral can be decomposed into

$$\begin{aligned}\int_{A_\sigma\cap E_\sigma} R_f(x)^2 f(x)\mathrm{d}x &= \int_{A_\sigma\cap E_\sigma}\left[a_{r+1}L(x)+\sum_{j=1}^{r}a_j|l_j(x)|^{\frac{\beta}{j}}\right]^2 f(x)\mathrm{d}x\\ &= a_{r+1}^2\int_{A_\sigma\cap E_\sigma} L(x)^2 f(x)\mathrm{d}x+\sum_{j=1}^{r}a_j^2\int_{A_\sigma\cap E_\sigma}|l_j(x)|^{\frac{2\beta}{j}}f(x)\mathrm{d}x\\ &\quad +2\sum_{j=1}^{r}a_{r+1}a_j\int_{A_\sigma\cap E_\sigma}|l_j(x)|^{\frac{\beta}{j}}L(x)f(x)\mathrm{d}x\\ &\quad +\sum_{\substack{j,j'=1\\ j\neq j'}}^{r}a_{j'}a_j\int_{A_\sigma\cap E_\sigma}|l_j(x)|^{\frac{\beta}{j}}|l_{j'}(x)|^{\frac{\beta}{j'}}f(x)\mathrm{d}x.\end{aligned}$$

Using the Hölder inequality and Condition (6), for all $j=1,\dots,r$,

$$\int_{A_\sigma\cap E_\sigma}|l_j(x)|^{\frac{2\beta}{j}}f(x)\mathrm{d}x\le\left[\int_{\mathbb{R}}|l_j(x)|^{\frac{2\beta+\varepsilon}{j}}f(x)\mathrm{d}x\right]^{\frac{2\beta}{2\beta+\varepsilon}}\left[\int_{\mathbb{R}}f(x)\mathrm{d}x\right]^{\frac{\varepsilon}{2\beta+\varepsilon}}\le C^{\frac{2\beta}{2\beta+\varepsilon}} \tag{21}$$

and $\int_{A_\sigma\cap E_\sigma}L(x)^2f(x)\mathrm{d}x \le \left[\int_{\mathbb{R}}L(x)^{2+\frac{\varepsilon}{\beta}}f(x)\mathrm{d}x\right]^{\frac{2\beta}{2\beta+\varepsilon}}\left[\int_{\mathbb{R}}f(x)\mathrm{d}x\right]^{\frac{\varepsilon}{2\beta+\varepsilon}} \le C^{\frac{2\beta}{2\beta+\varepsilon}}$. Next, using the Cauchy–Schwarz inequality and (21), for all $j,j'\in\{1,\dots,r\}, j\neq j'$,

$$\int_{A_\sigma\cap E_\sigma}|l_j(x)|^{\frac{\beta}{j}}|l_{j'}(x)|^{\frac{\beta}{j'}}f(x)\mathrm{d}x\le\left[\int_{\mathbb{R}}|l_j(x)|^{\frac{2\beta}{j}}f(x)\mathrm{d}x\right]^{\frac{1}{2}}\left[\int_{\mathbb{R}}|l_{j'}(x)|^{\frac{2\beta}{j'}}f(x)\mathrm{d}x\right]^{\frac{1}{2}}\le C^{\frac{2\beta}{2\beta+\varepsilon}}$$

and for all $j\in\{1,\dots,r\}$,

$$\int_{A_\sigma\cap E_\sigma}|l_j(x)|^{\frac{\beta}{j}}L(x)f(x)\mathrm{d}x\le\left[\int_{\mathbb{R}}|l_j(x)|^{\frac{2\beta}{j}}f(x)\mathrm{d}x\right]^{\frac{1}{2}}\left[\int_{\mathbb{R}}L(x)^2f(x)\mathrm{d}x\right]^{\frac{1}{2}}\le C^{\frac{2\beta}{2\beta+\varepsilon}}.$$

Finally, $\int_{A_\sigma\cap E_\sigma}R_f(x)^2f(x)\mathrm{d}x\le\left(\sum_{j=1}^{r+1}a_j\right)^2C^{\frac{2\beta}{2\beta+\varepsilon}}$.

For the second integral,

$$\begin{aligned}\int_{A_\sigma\cap E_\sigma}R_f(x)f(x)\mathrm{d}x &= \int_{A_\sigma\cap E_\sigma}\left[a_{r+1}L(x)+\sum_{j=1}^{r}a_j|l_j(x)|^{\frac{\beta}{j}}\right]f(x)\mathrm{d}x\\ &\le a_{r+1}\sqrt{\int_{\mathbb{R}}L(x)^2f(x)\mathrm{d}x}\sqrt{\int_{\mathbb{R}}f(x)\mathrm{d}x}+\sum_{j=1}^{r}a_j\sqrt{\int_{\mathbb{R}}|l_j(x)|^{\frac{2\beta}{j}}f(x)\mathrm{d}x}\sqrt{\int_{\mathbb{R}}f(x)\mathrm{d}x}\\ &\le\sum_{j=1}^{r+1}a_jC^{\frac{\beta}{2\beta+\varepsilon}}.\end{aligned}$$

Finally, (20) becomes

$$\begin{aligned}\int_{A_\sigma\cap E_\sigma}\frac{[f(x)-K_\sigma h_k(x)]^2}{K_\sigma h_k(x)}\mathrm{d}x \le{}& \frac{\Lambda_{\beta,H}^2}{D}(1+A_\beta\sigma^{2\beta})\sigma^{2\beta}\left(\sum_{j=1}^{r+1}a_j\right)^2 C^{\frac{2\beta}{2\beta+\varepsilon}}\\ &+\frac{\Omega_{\beta,H}^2}{D}(1+A_\beta\sigma^{2\beta})\sigma^{2H-8\beta-2}\\ &+\frac{2\Lambda_{\beta,H}\Omega_{\beta,H}}{D}(1+A_\beta\sigma^{2\beta})\sigma^{H-3\beta-1}\left(\sum_{j=1}^{r+1}a_j\right)C^{\frac{\beta}{2\beta+\varepsilon}}.\end{aligned}$$

By taking $H=5\beta+1$, it gives that there exists $\bar{\sigma}(\beta)>0$ such that for all $\sigma<\bar{\sigma}(\beta)$,

$$\int_{A_\sigma\cap E_\sigma}\frac{[f(x)-K_\sigma h_k(x)]^2}{K_\sigma h_k(x)}\mathrm{d}x = O_\beta\left(\sigma^{2\beta}\right).$$

- **Control of (18)**: According to Lemma A.3,

$$\begin{aligned}h_k(x) &= \left(\int_{\mathbb{R}} g_k(x)\mathrm{d}x\right)^{-1}\left\{f_k(x)\mathbb{1}_{J_{\sigma,k}}(x)+\frac{1}{2}f(x)\mathbb{1}_{J^c_{\sigma,k}}(x)\right\}\\ &\le \left\{2\sum_{i=0}^{k}(-1)^i\binom{k+1}{i+1}K_\sigma^i f(x)\right\}\mathbb{1}_{J_{\sigma,k}}(x)+f(x)\mathbb{1}_{J^c_{\sigma,k}}(x)\end{aligned}$$

thus

$$K_\sigma h_k(x)\le 2\sum_{j=1}^{k+1}\binom{k+1}{j}K_\sigma^j f(x)+K_\sigma f(x).$$

According to (12) and (13) in Lemma 4.2 with $H_1=4\beta+1$, there exists $\bar{\sigma}(\beta)>0$ such that for all $\sigma<\bar{\sigma}(\beta)$,

$$\begin{aligned}\int_{A_\sigma^c\cup E_\sigma^c}[K_\sigma h_k(x)-f(x)]\mathrm{d}x &\le \int_{A_\sigma^c\cup E_\sigma^c}K_\sigma h_k(x)\mathrm{d}x+\int_{A_\sigma^c\cup E_\sigma^c}f(x)\mathrm{d}x\\ &\le 2\sum_{j=1}^{k+1}\binom{k+1}{j}\int_{A_\sigma^c\cup E_\sigma^c}K_\sigma^j f(x)\mathrm{d}x+\int_{A_\sigma^c\cup E_\sigma^c}K_\sigma f(x)\mathrm{d}x+\int_{A_\sigma^c\cup E_\sigma^c}f(x)\mathrm{d}x\\ &\le 2\sum_{j=2}^{k+1}\binom{k+1}{j}\int_{A_\sigma^c}K_\sigma^j f(x)\mathrm{d}x+[2(k+1)+1]\int_{A_\sigma^c}K_\sigma f(x)\mathrm{d}x+\int_{A_\sigma^c}K_\sigma^0 f(x)\mathrm{d}x\\ &\quad+2\sum_{j=2}^{k+1}\binom{k+1}{j}\int_{E_\sigma^c}K_\sigma^j f(x)\mathrm{d}x+[2(k+1)+1]\int_{E_\sigma^c}K_\sigma f(x)\mathrm{d}x+\int_{E_\sigma^c}K_\sigma^0 f(x)\mathrm{d}x\\ &\le 2\left[2\sum_{j=2}^{k+1}\binom{k+1}{j}+2(k+2)\right]c_\beta\sigma^{2\beta}.\end{aligned}$$

- **Control of (19)**: According to Lemma A.4, for all $\sigma < \bar{\sigma}(\beta)$, $K_\sigma h_k(x) \geq \frac{D}{1+A_\beta \sigma^{2\beta}} f(x)$ then

$$\begin{aligned}\int_{A_\sigma^c \cup E_\sigma^c} f(x) \ln\left(\frac{f(x)}{K_\sigma h_k(x)}\right) \mathrm{d}x &\leq \ln\left(\frac{1+A_\beta\sigma^{2\beta}}{D}\right) \int_{A_\sigma^c \cup E_\sigma^c} f(x)\mathrm{d}x \\ &\leq \ln\left(\frac{1+A_\beta\sigma^{2\beta}}{D}\right) \left\{\int_{A_\sigma^c} K_\sigma^0 f(x)\mathrm{d}x + \int_{E_\sigma^c} K_\sigma^0 f(x)\mathrm{d}x\right\} \\ &\leq \ln\left(\frac{1+A_\beta\sigma^{2\beta}}{D}\right) 2c_\beta\sigma^{2\beta}.\end{aligned}$$

In conclusion, there exists $\bar{\sigma}(\beta) > 0$ such that for all $\sigma < \bar{\sigma}(\beta)$, $\mathrm{KL}(f, K_\sigma h_k) = O_\beta\left(\sigma^{2\beta}\right)$.

□

4.2. Proof of Theorem 2.5

The proof of Theorem 2.5 consists of using a discretization result (Prop. A.1) to replace the continuous mixture $K_\sigma h_k$ given in Proposition 4.4 by a finite Gaussian mixture $\wp_\sigma$. To show this result, we do not exactly follow the lines of the proof of Lemma 4 in [11] and we use an alternative discretization result. It seems that some little modifications are required to make their proof correct because the argument that $f_k \leq 2^k f_0$ (middle of p. 1252) cannot be checked so easily. Furthermore, as explained before, we need an uniform version of their result. One last reason is that we need to compute precisely the constants G_β and $\tilde{G}_\beta$ to prove Theorem 2.9.

Proof. For the definition of E_σ, we choose $H_1 = 4(\beta+1)$. Let $\tilde{h}_k$ be the restriction of h_k on an interval $[-\mu_\sigma, \mu_\sigma]$, normalized in order to have a density function:

$$\tilde{h}_k : x \in \mathbb{R} \mapsto \left(\int_{[-\mu_\sigma,\mu_\sigma]} h_k(y)\mathrm{d}y\right)^{-1} h_k(x) \mathbb{1}_{[-\mu_\sigma,\mu_\sigma]}(x)$$

where μ_σ depends on σ and will be chosen below such that

$$\mu_\sigma \geq \sigma. \tag{22}$$

Let $\varepsilon \in (0, \pi^{-1/2})$. According to Proposition A.1 in Appendix A.1, there exists a discrete distribution $\tilde{F}$ on $[-\mu_\sigma, \mu_\sigma]$ with at most $54\mu_\sigma\sigma^{-1}e^2\left[-\ln\left(\sqrt{\pi}\varepsilon\right) \vee 1\right]$ support points such that

$$\|\tilde{h}_k * \psi_\sigma - \tilde{F} * \psi_\sigma\|_\infty \leq \frac{2\varepsilon}{\sigma}. \tag{23}$$

Denoting $\tilde{\wp}(x)\mathrm{d}x = \left(\int_{[-\mu_\sigma,\mu_\sigma]} h_k(x)\mathrm{d}x\right) \tilde{F} * \psi_\sigma(x)$, it gives for all $x \in \mathbb{R}$,

$$\begin{aligned}|K_\sigma h_k(x) - \tilde{\wp}(x)| &= \left(\int_{[-\mu_\sigma,\mu_\sigma]} h_k(x)\mathrm{d}x\right) \left|\frac{h_k}{\int_{[-\mu_\sigma,\mu_\sigma]} h_k(y)\mathrm{d}y} * \psi_\sigma(x) - \tilde{F} * \psi_\sigma(x)\right| \\ &\leq \left|\frac{h_k \mathbb{1}_{[-\mu_\sigma,\mu_\sigma]}}{\int_{[-\mu_\sigma,\mu_\sigma]} h_k(y)\mathrm{d}y} * \psi_\sigma(x) - \tilde{F} * \psi_\sigma(x)\right| + \left(h_k \mathbb{1}_{[-\mu_\sigma,\mu_\sigma]^c}\right) * \psi_\sigma(x).\end{aligned}$$

By applying Lemma A.6 with $p = \frac{1}{2}$, it follows that for all $\sigma \leq 1 - 2^{-1/k}$ and for all $x \in \mathbb{R}$,

$$h_k(x) \leq 4M\left(\frac{4}{\sqrt{3}}\right)^k \psi\left(\frac{x}{2}\right)$$

and thus $\left(h_k \mathbb{1}_{[-\mu_\sigma,\mu_\sigma]^c}\right) * \psi_\sigma(x) \leq 4M \left(\frac{4}{\sqrt{3}}\right)^k \psi(\frac{\mu_\sigma}{2})$. Now, we choose $\mu_\sigma := 2\sqrt{\ln\left(\frac{4M}{\sqrt{\pi}}\left(\frac{4}{\sqrt{3}}\right)^k \frac{\sigma}{\varepsilon}\right)}$ in order to obtain that $\left\|\left(h_k \mathbb{1}_{[-\mu_\sigma,\mu_\sigma]^c}\right) * \psi_\sigma\right\|_\infty \leq \frac{\varepsilon}{\sigma}$. This last inequality together with (23) yields

$$\|K_\sigma h_k - \tilde{\wp}\|_\infty \leq \frac{3\varepsilon}{\sigma}.$$

We also define the function $t := \tilde{\wp} + \sigma^{6\beta+5}\psi_\sigma$ and the finite Gaussian mixture with density

$$\wp(x) := \frac{t(x)}{\int_{\mathbb{R}} t(y)\mathrm{d}y} = \frac{\tilde{\wp}(x) + \sigma^{6\beta+5}\psi_\sigma(x)}{\int_{[-\mu_\sigma,\mu_\sigma]} h_k(y)\mathrm{d}y + \sigma^{6\beta+5}}.$$

Then we want to upper bound

$$\begin{aligned}
\mathrm{KL}(f,\wp) &= \int_{\mathbb{R}} f(x)\ln\left(\frac{f(x)}{\wp(x)}\right)\mathrm{d}x \\
&= \int_{\mathbb{R}} f(x)\ln\left(\frac{f(x)}{K_\sigma h_k(x)}\right)\mathrm{d}x + \int_{\mathbb{R}} f(x)\ln\left(\frac{K_\sigma h_k(x)}{t(x)}\right)\mathrm{d}x + \int_{\mathbb{R}} f(x)\ln\left(\frac{t(x)}{\wp(x)}\right)\mathrm{d}x \\
&= \int_{\mathbb{R}} f(x)\ln\left(\frac{f(x)}{K_\sigma h_k(x)}\right)\mathrm{d}x \\
&\quad + \int_{E_\sigma^c} f(x)\ln\left(\frac{K_\sigma h_k(x)}{t(x)}\right)\mathrm{d}x + \int_{E_\sigma} f(x)\ln\left(\frac{K_\sigma h_k(x)}{t(x)}\right)\mathrm{d}x + \int_{\mathbb{R}} f(x)\ln\left(\frac{t(x)}{\wp(x)}\right)\mathrm{d}x \\
&= \boxed{\mathrm{I1}} + \boxed{\mathrm{I2}} + \boxed{\mathrm{I3}} + \boxed{\mathrm{I4}}.
\end{aligned}$$

- **Control of $\boxed{\mathbf{I1}}$**: According to Proposition 4.4, for all $\sigma < \bar{\sigma}(\beta)$,

$$\int_{\mathbb{R}} f(x)\ln\left(\frac{f(x)}{K_\sigma h_k(x)}\right)\mathrm{d}x = O_\beta\left(\sigma^{2\beta}\right).$$

- **Control of $\boxed{\mathbf{I2}}$**: According to Lemma A.6, $K_\sigma h_k(x) \leq 4M\left(\frac{4}{\sqrt{3}}\right)^k$ for σ small enough and since $s(x) \geq \sigma^{6\beta+5}\psi_\sigma(x)$,

$$\begin{aligned}
\boxed{\mathrm{I2}} &\leq \int_{E_\sigma^c} f(x)\ln\left(\frac{4M\left(\frac{4}{\sqrt{3}}\right)^k}{\sigma^{6\beta+5}\psi_\sigma(x)}\right)\mathrm{d}x \\
&\leq \left(\int_{E_\sigma^c} f(x)\mathrm{d}x\right)\left[(6\beta+4)|\ln\sigma| + \ln\left(4M\left(\frac{4}{\sqrt{3}}\right)^k\right)\right] + \int_{E_\sigma^c} f(x)\frac{x^2}{\sigma^2}\mathrm{d}x.
\end{aligned}$$

For the second integral,

$$\begin{aligned}
\int_{E_\sigma^c} \frac{x^2}{\sigma^2} f(x)\mathrm{d}x &\leq \sigma^{\frac{H_1}{2}-2}\int_{E_\sigma^c} x^2\sqrt{f(x)}\mathrm{d}x \\
&\leq \sigma^{2\beta}\int_{\mathbb{R}} x^2\sqrt{M\psi(x)}\mathrm{d}x \\
&\leq 4\pi\sqrt{M}\sigma^{2\beta}.
\end{aligned}$$

Similarly, $\int_{E_\sigma^c} f(x)\mathrm{d}x \leq \sigma^{2\beta+2}\sqrt{2M}$ and finally

$$\boxed{\text{I2}} \leq \left\{ \ln\left(4M\left(\frac{4}{\sqrt{3}}\right)^k\right) + (6\beta+4)|\ln\sigma| \right\} \sqrt{2M}\sigma^{2\beta+2} + 4\pi\sqrt{M}\sigma^{2\beta}.$$

Thus $\boxed{\text{I2}} = O_\beta\left(\sigma^{2\beta}\right)$.

- **Control of** $\boxed{\textbf{I3}}$: On the one hand,

$$\begin{aligned} |K_\sigma h_k(x) - t(x)| &\leq |K_\sigma h_k(x) - \tilde{\wp}(x)| + |\tilde{\wp}(x) - t(x)| \\ &\leq 3\varepsilon\sigma^{-1} + \sigma^{6\beta+5}\psi_\sigma(x) \\ &\leq 3\varepsilon\sigma^{-1} + \sigma^{6\beta+4}\pi^{-1/2}. \end{aligned}$$

On the other hand, according to Lemma A.4, for all $x \in \mathbb{R}$ and for all $\sigma < \bar{\sigma}(\beta)$, $K_\sigma h_k(x) \geq \frac{\xi\sqrt{\pi}}{6M(1+A_\beta\sigma^{2\beta})}f(x)$. Since $x \in E_\sigma$ then $K_\sigma h_k(x) \geq \frac{\xi\sqrt{\pi}}{6M(1+A_\beta\sigma^{2\beta})}\sigma^{4(\beta+1)}$. Thus, $t(x) \geq \tilde{\wp}(x) \geq K_\sigma h_k(x) - 3\varepsilon\sigma^{-1} \geq \frac{\sigma^{4(\beta+1)}\sqrt{\pi}}{6M(1+A_\beta\sigma^{2\beta})} - 3\varepsilon\sigma^{-1}$. Finally,

$$\begin{aligned} \boxed{\text{I3}} &\leq \int_{E_\sigma} f(x)\frac{K_\sigma h_k(x) - t(x)}{t(x)}\ \mathrm{d}x \\ &\leq \frac{3\varepsilon\sigma^{-1} + \sigma^{6\beta+4}\pi^{-1/2}}{\frac{\sigma^{4(\beta+1)}}{2(1+A_\beta\sigma^{2\beta})} - 3\varepsilon\sigma^{-1}} \int_{E_\sigma} f(x)\ \mathrm{d}x \\ &\leq \frac{3\varepsilon\sigma^{-1} + \sigma^{6\beta+4}\pi^{-1/2}}{\frac{\sigma^{4(\beta+1)}}{2(1+A_\beta\sigma^{2\beta})} - 3\varepsilon\sigma^{-1}}. \end{aligned}$$

Let $\delta' := 1 + \frac{\beta}{2(\beta+1)}$ and we set $\varepsilon := \sigma^{\delta' 4(\beta+1)+1}$. It yields

$$\boxed{\text{I3}} \leq \frac{(\pi^{-1/2}+3)\sigma^{6\beta+4}}{\frac{\sigma^{4(\beta+1)}}{2(1+A_\beta\sigma^{2\beta})} - 3\sigma^{6\beta+4}} = O_\beta\left(\sigma^{2\beta}\right).$$

- **Control of** $\boxed{\textbf{I4}}$: Note that $\frac{t(x)}{\wp(x)} = \int_{[-\mu_\sigma,\mu_\sigma]} h_k(y)\mathrm{d}y + \sigma^{6\beta+5} \leq 1 + \sigma^{6\beta+5}$ and thus

$$\begin{aligned} \boxed{\text{I4}} &\leq \int_{\mathbb{R}} f(x)\ln\left(1+\sigma^{6\beta+5}\right)\ \mathrm{d}x \\ &\leq \sigma^{6\beta+5} \leq \sigma^{2\beta}. \end{aligned}$$

Finally, we obtain that $\mathrm{KL}(f,\wp) = O_\beta\left(\sigma^{2\beta}\right)$. Moreover, according to the choice of ε, we have that

$$\begin{aligned} \mu_\sigma &= 2\sqrt{\ln\left(\frac{4M}{\sqrt{\pi}}\left(\frac{4}{\sqrt{3}}\right)^k\frac{\sigma}{\varepsilon}\right)} \\ &= 2\sqrt{\ln\left(\frac{4M}{\sqrt{\pi}}\left(\frac{4}{\sqrt{3}}\right)^k\sigma^{-(6\beta+4)}\right)} \\ &= \tilde{G}_\beta|\ln\sigma|^{\frac{1}{2}} \end{aligned}$$

where

$$\tilde{G}_\beta = 2\sqrt{\ln\left(\frac{4M}{\sqrt{\pi}}\right) + k\ln\left(\frac{4}{\sqrt{3}}\right) + (6\beta+4)}. \tag{24}$$

Thus there exists $\bar{\sigma}(\beta)$ continuous in β such that (22) is fulfilled for $\sigma < \bar{\sigma}(\beta)$. Furthermore, the mixture $\wp$ has k_σ components such that

$$\begin{aligned} k_\sigma &\leq 54\mu_\sigma \sigma^{-1} e^2 \left[1 \vee \ln\left(\frac{1}{\sqrt{\pi}\varepsilon}\right)\right] + 1 \\ &\leq \tilde{G}_\beta |\ln\sigma|^{\frac{1}{2}} 54\sigma^{-1} e^2 \left[1 \vee \ln\left(\frac{1}{\sqrt{\pi}\sigma^{6\beta+5}}\right)\right] + 1 \\ &= G_\beta \sigma^{-1} |\ln\sigma|^{\frac{3}{2}}. \end{aligned} \tag{25}$$

□

5. Proof of the lower bound

5.1. Proof of Proposition 2.6

Note that for every j, φ_j is supported by

$$J_j := \left[-\frac{\alpha}{2} + \frac{\alpha}{D}(j-1) + \frac{\alpha}{4D}, -\frac{\alpha}{2} + \frac{\alpha}{D}j - \frac{\alpha}{4D}\right] \subsetneq I_j = \left[-\frac{\alpha}{2} + \frac{\alpha}{D}(j-1), -\frac{\alpha}{2} + \frac{\alpha}{D}j\right]$$

and thus the supports of the $\varphi_j, 1 \leq j \leq D$ are disjoint. We also note that for all $x \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]^c$, $f_\theta(x) = \omega(x)$ and for all $x \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]$, there exists an unique $j \in \{1, \ldots, D\}$ such that $f_\theta(x) = 2\xi + (2\theta_j - 1)\varphi_j(x)$ where $\varphi_j(x) = 0$ if $x \in I_j \backslash J_j$. The proof of Proposition 2.6 is decomposed into two lemmas.

Lemma 5.1. *Density function and monotonicity conditions.*
For all $D \in \mathbb{N}^*$ and all $\theta \in \{0,1\}^D$, the function f_θ defined by (9) is a positive density function such that for all $x \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]$, $f_\theta(x) \in [\xi, 3\xi]$. This function fulfills also the following monotonicity conditions:

1. *$\forall x \in [-\alpha, \alpha]$, $f_\theta(x) \geq \xi$ and $\forall x \in [-\alpha, \alpha]^c$, $f_\theta(x) \leq \xi$.*
2. *f_θ is nondecreasing on $(-\infty, -\alpha)$ and nonincreasing on (α, ∞).*
3. *$\forall x \in \mathbb{R}$, $f_\theta(x) \leq M\psi(x)$ with $M = \tilde{M} \vee 3\sqrt{\pi}\xi \exp(\alpha^2/4)$.*

Proof. For all $x \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]^c$, $f_\theta(x) = \omega(x) > 0$ since ω is positive. Moreover, for all $x \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]$, $\exists! j \in \{1, \ldots, D\}$ such that $x \in I_j$. Then,

$$f_\theta(x) = \omega(x) + (2\theta_j - 1)\varphi_j(x) = 2\xi + (2\theta_j - 1)\varphi_j(x).$$

Thus

$$\begin{aligned} |f_\theta(x) - 2\xi| &= |(2\theta_j - 1)|\ |\varphi_j(x)| \\ &= \left|\frac{\xi D^{-\beta}}{A}\ \varphi\left(\frac{D}{\alpha}\left(x + \frac{\alpha}{2}\right) - (j-1)\right)\right| \\ &\leq \xi D^{-\beta} \\ &\leq \xi \end{aligned}$$

since $D^{-\beta} \leq 1$. Thus for all $x \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]$, $f_\theta(x) \in [\xi, 3\xi]$. Finally, f_θ is a positive function on $\mathbb{R}$. Moreover,

$$\begin{aligned} \int_{\mathbb{R}} f_\theta(x)\mathrm{d}x &= \int_{\mathbb{R}} \omega(x)\mathrm{d}x + \sum_{j=1}^{D} (2\theta_j - 1) \int_{I_j} \varphi_j(x)\mathrm{d}x \\ &= \int_{\mathbb{R}} \omega(x)\mathrm{d}x + \sum_{j=1}^{D} (2\theta_j - 1) \frac{\xi D^{-\beta}}{A} \frac{\alpha}{D} \int_{\mathbb{R}} \varphi(y)\mathrm{d}y \\ &= 1 \end{aligned}$$

because $\int_{\mathbb{R}} \omega(x)\mathrm{d}x = 1$ and $\int_{\mathbb{R}} \varphi(y)\mathrm{d}y = 0$. Thus, f_θ is a density function.

Since $f_\theta(x) = \omega(x)$ and ω is nondecreasing on $(-\infty, -\alpha)$, the function f_θ is a nondecreasing function on $(-\infty, -\alpha)$. Moreover,

$$\forall x < -\alpha, f_\theta(x) \leq f_\theta(-\alpha) = \omega(-\alpha) = \xi.$$

In the same way, the function f_θ is nonincreasing on (α, ∞) and

$$\forall x > \alpha, f_\theta(x) \leq f_\theta(\alpha) = \omega(\alpha) = \xi.$$

For all $x \in [-\alpha, \alpha]$,

- if $x \in [-\alpha, -\frac{\alpha}{2})$, $f_\theta(x) = \omega(x) \geq \omega(-\alpha) = \xi$ because ω is nondecreasing and $\omega(-\alpha) = \xi$;
- if $x \in (\frac{\alpha}{2}, \alpha]$, $f_\theta(x) = \omega(x) \geq \omega(\alpha) = \xi$ because ω is nonincreasing and $\omega(\alpha) = \xi$;
- if $x \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]$, $f_\theta(x) \in [\xi, 3\xi]$ thus $f_\theta(x) \geq \xi$.

For the last point, we have that for all $x \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]^c$, $f_\theta(x) = \omega(x) \leq \tilde{M}\psi(x)$. Moreover, for all $x \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]$, $f_\theta(x) \leq 3\xi \leq 3\xi\sqrt{\pi}\exp(\alpha^2/4)\psi(x)$. Finally, for all $x \in \mathbb{R}$, $f_\theta(x) \leq M(\xi, \alpha, \tilde{M})\psi(x)$ with $M(\xi, \alpha, \tilde{M}) := \tilde{M} \vee 3\sqrt{\pi}\xi\exp(\alpha^2/4)$. □

Lemma 5.2. *Let $\beta \in [\underline{\beta}, \bar{\beta}]$. For all $\theta \in \{0,1\}^D$, the function $\ln f_\theta$ is locally β-Hölder: for all x, y such that $|x - y| \leq \frac{\alpha}{4}$,*

$$|(\ln f_\theta)^{(r)}(x) - (\ln f_\theta)^{(r)}(y)| \leq L(\underline{\beta}, \bar{\beta}, \tilde{L}, \alpha) r! |x - y|^{\beta - r}$$

where $L(\underline{\beta}, \bar{\beta}, \tilde{L}, \alpha)$ does not depend on D. Moreover, there exists a constant $C(\underline{\beta}, \bar{\beta}, \tilde{C}, \alpha)$, which can be taken identical for every D, such that for any integer $j = 1, \ldots, r$ and for all $D \in \mathbb{N}^*$,

$$\int_{\mathbb{R}} |(\ln f_\theta)^{(j)}(x)|^{\frac{2\beta+\tilde{\varepsilon}}{j}} f_\theta(x)\mathrm{d}x \leq C(\underline{\beta}, \bar{\beta}, \tilde{C}, \alpha),$$

and

$$\int_{\mathbb{R}} |L(\underline{\beta}, \bar{\beta}, \tilde{L}, \alpha)|^{\frac{2\beta+\tilde{\varepsilon}}{\beta}} f_\theta(x)\mathrm{d}x \leq C(\underline{\beta}, \bar{\beta}, \tilde{C}, \alpha).$$

If D is a positive even integer, for any integer $j = 0, \ldots, r$, $|(\ln f_\theta)^{(j)}(0)| \leq \ln(2\xi)$.

Proof. Let $j \in \{1, \ldots, D\}$ and $1 \leq t \leq r + 1$. We start by upper bounding $\sup_{x \in I_j} |(\ln f_\theta)^{(t)}(x)|$. According to Lemma B.3, for all $x \in I_j$,

$$(\ln f_\theta)^{(t)}(x) = f_\theta(x)^{-2^{t-1}} \sum_{(\eta_0, \ldots, \eta_t) \in \Xi_t} \rho(\eta_0, \ldots, \eta_t) \prod_{u=0}^{t} \left(f_\theta^{(u)}(x)\right)^{\eta_u}$$

with

$$\Xi_t = \left\{(\eta_0, \ldots, \eta_t) \in \mathbb{N}^{t+1}; \sum_{u=0}^{t} u\eta_u = t, \ \sum_{u=0}^{t} \eta_u = 2^{t-1}\right\}.$$

For all $u \in \{1, \ldots, t\}$,

$$|f_\theta^{(u)}| \leq \frac{\xi D^{-\beta}}{A}\left(\frac{D}{\alpha}\right)^u \|\varphi^{(u)}\|_\infty \leq \frac{\xi D^{u-\beta}}{\alpha^u}.$$

Then, for all $(\eta_0, \ldots, \eta_t) \in \Xi_t$,

$$\begin{aligned}\left|\prod_{u=0}^{t} (f_\theta^{(u)})^{\eta_u}\right| &\leq D^{\sum_{u=1}^{t} u\eta_u - \beta\sum_{u=1}^{t}\eta_u} \xi^{\sum_{u=1}^{t}\eta_u} \alpha^{-\sum_{u=1}^{t} u\eta_u} \times |f_\theta|^{\eta_0} \\ &\leq \xi^{2^{t-1}-\eta_0} D^{t-\beta(2^{t-1}-\eta_0)} \alpha^{-t} \times |f_\theta(x)|^{\eta_0}\end{aligned}$$

since $\sum_{u=1}^{t} u\eta_u = t$ and $\sum_{u=1}^{t} \eta_u = 2^{t-1} - \eta_0$. Since $f_\theta(x) \in [\xi, 3\xi]$ and $2^{t-1} - \eta_0 \geq 1$,

$$\begin{aligned}
|(\ln f_\theta)^{(t)}(x)| &\leq \sum_{(\eta_0,\dots,\eta_t)\in\Xi_t} |\rho(\eta_0,\dots,\eta_t)| \left| \prod_{u=1}^{t} \left(f_\theta^{(u)}(x)\right)^{\eta_u} \right| |f_\theta(x)|^{\eta_0 - 2^{t-1}} \\
&\leq \sum_{(\eta_0,\dots,\eta_t)\in\Xi_t} |\rho(\eta_0,\dots,\eta_t)| \xi^{2^{t-1}-\eta_0} D^{t-\beta(2^{t-1}-\eta_0)} \alpha^{-t} \xi^{\eta_0 - 2^{t-1}} \\
&\leq \sum_{(\eta_0,\dots,\eta_t)\in\Xi_t} |\rho(\eta_0,\dots,\eta_t)| D^{t-\beta(2^{t-1}-\eta_0)} \alpha^{-t}.
\end{aligned}$$

Denoting $\mathcal{B}(t) := \mathrm{card}(\Xi_t)$ and $B(t) := \max_{(\eta_0,\dots,\eta_t)\in\Xi_t} |\rho(\eta_0,\dots,\eta_t)|$, it leads to

$$\sup_{x\in I_j} |(\ln f_\theta)^{(t)}(x)| \leq \mathcal{B}(t) B(t) D^{t-\beta} \alpha^{-t}. \tag{26}$$

We now use this preliminary result to prove that $\ln f_\theta$ is locally β-Holder. Let $(x, y) \in \mathbb{R}^2$ such that $|x-y| \leq \frac{\alpha}{4}$.

- If $x, y \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]^c$,

  $$\begin{aligned}
  |(\ln f_\theta)^{(r)}(x) - (\ln f_\theta)^{(r)}(y)| &= |(\ln\omega)^{(r)}(x) - (\ln\omega)^{(r)}(y)| \\
  &\leq \tilde{L} r! |x-y|^{\beta - r},
  \end{aligned}$$

  since $\ln\omega$ is locally β-Holder with $\gamma_\omega = \frac{\alpha}{4}$ and a constant $\tilde{L}$.
- If $y \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]^c$ and $x \in I_j$:
  - If $|x-y| < \frac{\alpha}{4D}$ then $x \in I_j \backslash J_j$. Thus, $\ln f_\theta(x) = \ln\omega(x)$ and

    $$\begin{aligned}
    |(\ln f_\theta)^{(r)}(x) - (\ln f_\theta)^{(r)}(y)| &= |(\ln\omega)^{(r)}(x) - (\ln\omega)^{(r)}(y)| \\
    &\leq \tilde{L} r! |x-y|^{\beta - r}.
    \end{aligned}$$

  - If $\frac{\alpha}{4D} \leq |x-y| < \frac{\alpha}{4}$, $\ln\omega(y) = \ln(2\xi)$ since $x \in [-3\alpha/4, -\alpha/2] \cup [\alpha/2, 3\alpha/4]$ thus if $r \geq 1$,

    $$\begin{aligned}
    |(\ln f_\theta)^{(r)}(x) - (\ln f_\theta)^{(r)}(y)| &\leq \|(\ln f_\theta)^{(r)}\|_{\infty,[-\alpha/2,\alpha/2]} + \|(\ln\omega)^{(r)}\|_{\infty,[\alpha/2,3\alpha/4]} \\
    &\leq \mathcal{B}(r) B(r) D^{r-\beta} \alpha^{-r} \left(\frac{4D}{\alpha}\right)^{\beta - r} |x-y|^{\beta - r} + 0 \\
    &\leq \frac{\mathcal{B}(r) B(r)}{r!} 4^{\beta - r} \alpha^{-\beta}\ r!\ |x-y|^{\beta - r}
    \end{aligned}$$

    and if $r = 0$,

    $$\begin{aligned}
    |(\ln f_\theta)(x) - (\ln f_\theta)(y)| &\leq |\ln(2\xi) - \ln(2\xi + (2\theta_j - 1)\varphi_j(y))| \\
    &\leq \left| -\ln\left(1 + (2\xi)^{-1}(2\theta_j - 1)\varphi_j(y)\right) \right| \\
    &\leq \left| (2\xi)^{-1}(2\theta_j - 1)\varphi_j(y) \right| \\
    &\leq (2\xi)^{-1} \xi D^{-\beta} (4D)^{\beta} \alpha^{-\beta} |x-y|^{\beta} \\
    &\leq 4^{\beta} \alpha^{-\beta} |x-y|^{\beta} = \frac{\mathcal{B}(1) B(1)}{0!} 4^{\beta} \alpha^{-\beta} 0!\ |x-y|^{\beta - r}.
    \end{aligned}$$

- For all $x, y \in [-\alpha/2, \alpha/2]$, $\exists!(j, j') \in \{1, \dots, D\}^2$ such that $x \in I_j$ and $y \in I_{j'}$.
  - If $|x-y| \leq \frac{\alpha}{4D}$,
    - if $j' \neq j$, $x \in I_j \backslash J_j$ and $y \in I_{j'} \backslash J_{j'}$, thus

      $$|(\ln f_\theta)^{(r)}(x) - (\ln f_\theta)^{(r)}(y)| = 0.$$

- if $j' = j$,

$$\begin{aligned}|(\ln f_\theta)^{(r)}(x) - (\ln f_\theta)^{(r)}(y)| &\leq |x-y|^{\beta-r}|x-y|^{r+1-\beta} \; \|\ln f_\theta^{(r+1)}\|_{\infty,[-\alpha/2,\alpha/2]}\\ &\leq \alpha^{-\beta+r+1}(4D)^{\beta-r-1}\frac{\mathcal{B}(r+1)B(r+1)}{r!}\frac{D^{r+1-\beta}}{\alpha^{r+1}}r!|x-y|^{\beta-r}\\ &\leq \frac{\mathcal{B}(r+1)B(r+1)}{r!}4^{\beta-r-1}\alpha^{-\beta}r!|x-y|^{\beta-r}\end{aligned}$$

– If $\frac{\alpha}{4D} < |x-y| < \frac{\alpha}{4}$: if $r=0$,

$$\begin{aligned}|(\ln f_\theta)(x) - (\ln f_\theta)(y)| &= \left|\ln\left(\frac{1+(2\xi)^{-1}(2\theta_j-1)\varphi_j(x)}{1+(2\xi)^{-1}(2\theta_j-1)\varphi_j(y)}\right)\right|\\ &\leq \left|\frac{(2\xi)^{-1}(2\theta_j-1)[\varphi_j(x)-\varphi_j(y)]}{1+(2\xi)^{-1}(2\theta_j-1)\varphi_j(y)}\right|\\ &\leq \frac{2\|\varphi_j\|_\infty}{\xi}\\ &\leq 2D^{-\beta}(4D)^\beta\alpha^{-\beta}|x-y|^\beta\\ &\leq 24^\beta\alpha^{-\beta}|x-y|^\beta = 24^\beta\alpha^{-\beta}\frac{\mathcal{B}(1)B(1)}{0!}0!|x-y|^\beta\end{aligned}$$

and if $r \geq 1$

$$\begin{aligned}|(\ln f_\theta)^{(r)}(x) - (\ln f_\theta)^{(r)}(y)| &\leq 2\|(\ln f_\theta)^{(r)}\|_{\infty,[-\alpha/2,\alpha/2]}\\ &\leq 2\mathcal{B}(r)B(r)\frac{D^{r-\beta}}{\alpha^r}\left(\frac{4D}{\alpha}\right)^{\beta-r}|x-y|^{\beta-r}\\ &\leq 2\frac{\mathcal{B}(r)B(r)}{r!}4^\beta\alpha^{-\beta}r!|x-y|^{\beta-r}.\end{aligned}$$

Finally, for all $\beta \in [\underline{\beta}, \bar{\beta}]$, for all $(x,y)\in\mathbb{R}^2$ such that $|x-y| < \frac{\alpha}{4}$,

$$|(\ln f_\theta)^{(r)}(x) - (\ln f_\theta)^{(r)}(y)| \leq L(\underline{\beta},\bar{\beta},\alpha)r!|x-y|^{\beta-r}$$

with

$$L(\underline{\beta},\bar{\beta},\tilde{L},\alpha) := \tilde{L} \vee \max_{\beta\in[\underline{\beta},\bar{\beta}]}\left(2\frac{\mathcal{B}(\lceil\beta\rceil)B(\lceil\beta\rceil)}{\lfloor\beta\rfloor!}\left(\frac{4}{\alpha}\right)^\beta\right).$$

According to (26), for any integer $j\in\{1,\dots,r\}$, $\|(\ln f_\theta)^{(j)}\|_{\infty,[-\alpha/2,\alpha/2]} \leq B(j)\mathcal{B}(j)\alpha^{-j}$ thus it yields

$$\begin{aligned}\int_{\mathbb{R}}|(\ln f_\theta)^{(j)}(x)|^{\frac{2\beta+\tilde{\varepsilon}}{j}}f_\theta(x)\mathrm{d}x &\leq \int_{[-\alpha/2,\alpha/2]^c}|(\ln\omega)^{(j)}(x)|^{\frac{2\beta+\tilde{\varepsilon}}{j}}\omega(x)\mathrm{d}x + \left[B(j)\mathcal{B}(j)\alpha^{-j}\right]^{\frac{2\beta+\tilde{\varepsilon}}{j}}\int_{[-\alpha/2,\alpha/2]}f_\theta(x)\mathrm{d}x\\ &\leq \tilde{C} + \left[B(j)\mathcal{B}(j)\alpha^{-j}\right]^{\frac{2\beta+\tilde{\varepsilon}}{j}}.\end{aligned}$$

Thus there exists a constant $C(\underline{\beta},\bar{\beta},\tilde{C},\tilde{\varepsilon},\alpha)$ such that for any integer $j\in\{1,\dots,r\}$,

$$\int_{\mathbb{R}}|(\ln f_\theta)^{(j)}(x)|^{\frac{2\beta+\varepsilon}{j}}f_\theta(x)\mathrm{d}x \leq \tilde{C} + \max_{1\leq j\leq r+1}[B(j)\mathcal{B}(j)]^{\frac{2\beta+\tilde{\varepsilon}}{j}} \leq C(\underline{\beta},\bar{\beta},\tilde{C},\tilde{\varepsilon},\alpha)$$

and

$$\int_{\mathbb{R}}|L(\underline{\beta},\bar{\beta},\tilde{L},\alpha)|^{2+\frac{\tilde{\varepsilon}}{\beta}}f_\theta(x)\mathrm{d}x = |L(\underline{\beta},\bar{\beta},\tilde{L},\alpha)|^{2+\frac{\tilde{\varepsilon}}{\beta}} \leq C(\underline{\beta},\bar{\beta},\tilde{C},\tilde{\varepsilon},\alpha).$$

The last point assumes that D is even, thus $0\in I_{D/2}\backslash J_{D/2}$. Then, $\ln f_\theta$ is equal to $\ln(2\xi)$ in a neighborhood of 0 and for all $j\in\{1,\dots,r\}$, $|(\ln f_\theta)^{(j)}(0)| = 0$. □

Lemmas 5.1 and 5.2 show that for any positive even integer D and for all $\beta \in [\underline{\beta}, \bar{\beta}]$, $\mathcal{J}(\beta, D) \subset \mathcal{H}\left(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta})\right)$ where

$$\mathcal{P}(\underline{\beta}, \bar{\beta}) = \left\{\frac{\alpha}{4}, \ln(2\xi), L(\underline{\beta}, \bar{\beta}, \tilde{L}, \alpha), \tilde{\varepsilon}, C(\underline{\beta}, \bar{\beta}, \tilde{C}, \tilde{\varepsilon}, \alpha), \alpha, \xi, M(\xi, \alpha, \tilde{M})\right\}.$$

5.2. Proof of Theorem 2.8

Lemma 5.3. *Let $\theta, \theta' \in \{0,1\}^D$. The Hellinger distance between two functions f_θ and $f_{\theta'}$ of $\mathcal{J}(\beta, D)$ fulfills*

1. $d_H^2(f_\theta, f_{\theta'}) \leq \frac{\xi\alpha}{8A^2} D^{-2\beta}$,
2. $\forall \theta \neq \theta'$, $d_H^2(f_\theta, f_{\theta'}) \geq \xi\alpha(2A)^{-2}\delta(\theta, \theta')D^{-(2\beta+1)}$ *where* $\delta(\theta, \theta') = \sum_{j=1}^{D} \mathbb{1}_{\theta_j \neq \theta'_j}$ *is the Hamming distance between θ and θ'.*

Proof.
The Hellinger distance between f_θ and $f_{\theta'}$ can be decomposed as follows:

$$\begin{aligned} d_H^2(f_\theta, f_{\theta'}) &= \frac{1}{2}\int_{[-\alpha/2,\alpha/2]} \left[\sqrt{f_\theta(x)} - \sqrt{f_{\theta'}(x)}\right]^2 \mathrm{d}x + \frac{1}{2}\int_{[-\alpha/2,\alpha/2]^c} \left[\sqrt{\omega(x)} - \sqrt{\omega(x)}\right]^2 \mathrm{d}x \\ &= \frac{1}{2}\sum_{j=1}^{D} \int_{I_j} \left[\sqrt{2\xi + (2\theta_j - 1)\varphi_j(x)} - \sqrt{2\xi + (2\theta'_j - 1)\varphi_j(x)}\right]^2 \mathrm{d}x. \end{aligned}$$

Since the quantity between the brackets is equal to zero if $\theta_j = \theta'_j$, it gives

$$\begin{aligned} d_H^2(f_\theta, f_{\theta'}) &= \frac{1}{2}\sum_{j=1}^{D} \int_{I_j} \left[\sqrt{2\xi + \varphi_j(x)} - \sqrt{2\xi - \varphi_j(x)}\right]^2 \mathrm{d}x \; \mathbb{1}_{\theta_j \neq \theta'_j} \\ &= \frac{1}{2}\sum_{j=1}^{D} \int_{I_j} \left[4\xi - 2\sqrt{(2\xi)^2 - \varphi_j(x)^2}\right] \mathrm{d}x \; \mathbb{1}_{\theta_j \neq \theta'_j}. \end{aligned}$$

Note that $\left(\frac{\varphi_j(x)}{2\xi}\right)^2 \leq 1$ for all $x \in I_j$ and $\|\varphi_j\|_\infty = \frac{\xi D^{-\beta}}{A}\|\varphi\|_\infty \leq \xi$. Then,

$$\sqrt{(2\xi)^2 - \varphi_j(x)^2} = 2\xi\sqrt{1 - \left(\frac{\varphi_j(x)}{2\xi}\right)^2} \geq \frac{1}{4}\left[1 - \left(\frac{\varphi_j(x)}{2\xi}\right)^2\right]$$

since $\sqrt{1-y} \geq 1 - y$ for all $y \in [0,1]$. Thus,

$$\begin{aligned} \int_{I_j} \left[4\xi - 2\sqrt{(2\xi)^2 - \varphi_j(x)^2}\right] \mathrm{d}x &\leq \int_{I_j} \left[4\xi - 4\xi + \frac{\varphi_j^2(x)}{4\xi}\right] \mathrm{d}x \\ &\leq (4\xi)^{-1} \int_{I_j} \left[\left(\frac{\xi D^{-\beta}}{A}\right)^2 \varphi^2\left(\frac{D}{\alpha}(x+1) - (j-1)\right)\right] \mathrm{d}x \\ &\leq (4\xi)^{-1}\left(\frac{\xi D^{-\beta}}{A}\right)^2 \frac{\alpha}{D} \end{aligned}$$

since $\int_{\mathbb{R}} \varphi^2(y)\mathrm{d}y = 1$. Finally,

$$\begin{aligned} d_H^2(f_\theta, f_{\theta'}) &\leq (4\xi)^{-1}\left(\frac{\xi D^{-\beta}}{A}\right)^2 \frac{\alpha}{D}\frac{1}{2}\delta(\theta, \theta') \\ &\leq \frac{\xi\alpha}{8A^2} D^{-2\beta} \end{aligned}$$

since $\delta(\theta, \theta') \leq D$.

For the lower bound, we have

$$\sqrt{(2\xi)^2-\varphi_j(x)^2}=2\xi\sqrt{1-\left(\frac{\varphi_j(x)}{2\xi}\right)^2}\leq 2\xi\left[1-\frac{1}{2}\left(\frac{\varphi_j(x)}{2\xi}\right)^2\right]$$

since $\sqrt{1-y}\leq 1-\frac{1}{2}y$ for all $y\in[0,1]$. Thus,

$$\begin{aligned}\int_{I_j}\left[4\xi-2\sqrt{(2\xi)^2-\varphi_j(x)^2}\right]\mathrm{d}x &\geq \int_{I_j}\left[4\xi-4\xi+\frac{\varphi_j^2(x)}{2\xi}\right]\mathrm{d}x\\ &\geq (2\xi)^{-1}\left(\frac{\xi D^{-\beta}}{A}\right)^2\frac{\alpha}{D}\int_{\mathbb{R}}\varphi^2(y)\mathrm{d}y\\ &\geq (2\xi)^{-1}\left(\frac{\xi D^{-\beta}}{A}\right)^2\frac{\alpha}{D}\end{aligned}$$

and finally

$$\begin{aligned}d_H^2(f_\theta,f_{\theta'}) &\geq (2\xi)^{-1}\left(\xi\frac{D^{-\beta}}{A}\right)^2\frac{\alpha}{D}\frac{1}{2}\sum_{j=1}^{D}\mathbb{1}_{\theta_j\neq\theta'_j}\\ &\geq \xi\alpha(2A)^{-2}D^{-(2\beta+1)}\delta(\theta,\theta').\end{aligned}$$

□

Lemma 5.4. *Let $\theta,\theta'\in\{0,1\}^D$. The Kullback–Leibler divergence between two functions f_θ and $f_{\theta'}$ of $\mathcal{J}(\beta,D)$ fulfills*

$$\mathrm{KL}(f_\theta,f_{\theta'})\leq\frac{5\xi\alpha}{4A^2}D^{-2\beta}.$$

Proof. According to equation (9) and since $f_\theta(x)=f_{\theta'}(x)=\omega(x)$ for all $x\in[-\alpha/2,\alpha/2]^c$, the Kullback–Leibler divergence between f_θ and $f_{\theta'}$ is given by

$$\begin{aligned}\mathrm{KL}(f_\theta,f_{\theta'}) &= \int_{\mathbb{R}}f_\theta(x)\ln\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right)\mathrm{d}x\\ &= \int_{[-\alpha/2,\alpha/2]}f_\theta(x)\ln\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right)\mathrm{d}x+\int_{[-\alpha/2,\alpha/2]^c}\omega(x)\ln\left(\frac{\omega(x)}{\omega(x)}\right)\mathrm{d}x\\ &= \int_{[-\alpha/2,\alpha/2]}f_\theta(x)\ln\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right)\mathrm{d}x.\end{aligned}$$

Then for all $x\in[-\alpha/2,\alpha/2]$ and for all $\theta\in\{0,1\}^D$, $f_\theta(x)\in[\xi,3\xi]$ according to Lemma 5.1 thus $\left\|\frac{f_\theta}{f_{\theta'}}\right\|_{\infty,[-1,1]}\leq 3$. According to Lemma 7.23 in [13],

$$\mathrm{KL}(f_\theta,f_{\theta'})\leq 2\left[2+\ln\left(\left\|\frac{f_\theta}{f_{\theta'}}\right\|_\infty\right)\right]d_H^2\left(f_\theta,f_{\theta'}\right).$$

Lemma 5.1 gives that for all $x\in[-\alpha/2,\alpha/2]$, $f_\theta(x)\in[\xi,3\xi]$ and furthermore, $f_\theta=f_{\theta'}$ on $[-\alpha/2,\alpha/2]^c$. Thus,

$$\begin{aligned}\mathrm{KL}(f_\theta,f_{\theta'}) &\leq 2\left[2+\ln\left(\sup_{[-\alpha/2,\alpha/2]}\left|\frac{f_\theta(x)}{f_{\theta'}(x)}\right|\right)\right]d_H^2\left(f_\theta,f_{\theta'}\right)\\ &\leq 10\ d_H^2\left(f_\theta,f_{\theta'}\right)\\ &\leq \frac{5\xi\alpha}{4A^2}D^{-2\beta}\end{aligned}$$

according to Lemma 5.3. □

Proof of Theorem 2.8. The proof consists of applying Corollary B.2 given in Appendix A.2 with the space $\mathcal{J}(\beta, D)$, the Hellinger distance d_H, $p = 2$ and the finite subset $\mathcal{C} = \{f_\theta,\ \theta \in \Theta\}$ where Θ is the subset of $\{0,1\}^D$ provided by Lemma B.1. Then, it has to be checked that

$$n \max_{\theta,\theta' \in \Theta} \mathrm{KL}(f_\theta, f_{\theta'}) \le \kappa \ln |\Theta|.$$

According to Lemma B.1, $\ln|\Theta| > \frac{D}{8}$ and $\kappa \ge \frac{1}{2}$. Moreover, $\mathrm{KL}(f_\theta, f_{\theta'}) \le \frac{5\xi\alpha}{4A^2} D^{-2\beta}$ and thus D is chosen such that

$$n\frac{5\xi\alpha}{4A^2} D^{-2\beta} \le \frac{D}{16} \Leftrightarrow \frac{20\xi\alpha n}{A^2} \le D^{2\beta+1}.$$

Since $3\xi\alpha \le 1$, $20\xi\alpha n A^{-2} \le \frac{20}{3} n \le 7n$ and we finally choose $D = \min\{2k; k \in \mathbb{N}^*, (2k)^{2\beta+1} \ge 7n\}$. It follows that for any estimator $\tilde{s}$,

$$\begin{aligned}
\sup_{\theta\in\Theta} \mathbb{E}_s[d_H^2(f_\theta, \tilde{s})] &\ge 2^{-2}(1-\kappa)\left[\min_{\theta,\theta'\in\Theta,\theta\neq\theta'} d_H(f_\theta, f_{\theta'})\right]^2 \\
&\ge 2^{-2}(1-\kappa)\xi\alpha(2A)^{-2}D^{-(2\beta+1)} \min_{\theta,\theta'\in\Theta,\theta\neq\theta'} \delta(\theta,\theta') \\
&\ge 2^{-2}(1-\kappa)\xi\alpha(2A)^{-2}D^{-(2\beta+1)}\frac{D}{4} \\
&\ge \frac{(1-\kappa)\xi\alpha}{A^2}\, 2^{-6-2\beta}\, (7n)^{-\frac{2\beta}{2\beta+1}}
\end{aligned}$$

according to Lemma B.1. □

6. Proof of Theorem 2.9

Under the hypotheses of Section 2.4, let $\mathcal{P}(\underline{\beta}, \bar{\beta})$ be the parameter set given in Proposition 2.6. In order to prove Theorem 2.9, we start with the following lemma that makes the connection between the models $\mathcal{S}_m$ and the approximation result given in Theorem 2.5.

Lemma 6.1. *There exists a positive constant $c_{\underline{\beta},\bar{\beta}}$ such that for all $\beta \in [\underline{\beta}, \bar{\beta}]$ and for all $s \in \mathcal{H}\left(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta})\right)$,*

$$\mathrm{KL}(s, \mathcal{S}_m) \le c_{\underline{\beta},\bar{\beta}}\ \lambda(m)^{2\beta}.$$

Proof. According to Theorem 2.5, the level $\bar{\sigma}(\beta)$ under which the approximation (8) is valid is a continuous function of β. Thus we can define the positive constant $\bar{\sigma}(\underline{\beta}, \bar{\beta}) := \inf_{\beta\in[\underline{\beta},\bar{\beta}]} \bar{\sigma}(\beta)$. Next, let

$$m_0(\underline{\beta}, \bar{\beta}) := \inf\left\{m \ge 2;\ \lambda(m) < \bar{\sigma}(\underline{\beta}, \bar{\beta})\right\}$$

and consider $m \ge m_0(\underline{\beta}, \bar{\beta})$. Then Theorem 2.5 can be applied for $\sigma = \lambda(m)$: for all $\beta \in [\underline{\beta}, \bar{\beta}]$ and for all $s \in \mathcal{H}\left(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta})\right)$, there exists a mixture $\wp$ with less than $G_\beta \lambda(m)^{-1} |\ln \lambda(m)|^{\frac{3}{2}}$ components, with means belonging to $[-\bar{\mu}(m), \bar{\mu}(m)]$ and with the same variance $\lambda^2(m)$ for each component such that

$$\mathrm{KL}(s, \wp) \le c_\beta\ \lambda(m)^{2\beta}.$$

Since G_β is a non decreasing function of β, the number of components is less than

$$\begin{aligned}
G_{\bar{\beta}}\lambda(m)^{-1}|\ln\lambda(m)|^{\frac{3}{2}} &\le G_{\bar{\beta}}\left[\frac{a_{\bar{\beta}}}{m}(\ln m)^{\frac{3}{2}}\right]^{-1}\left|\ln\left\{\frac{a_{\bar{\beta}}}{m}(\ln m)^{\frac{3}{2}}\right\}\right|^{\frac{3}{2}} \\
&\le m\frac{G_{\bar{\beta}}}{a_{\bar{\beta}}}\left[\frac{\ln a_{\bar{\beta}}}{\ln m} + 1 + \frac{3}{2}\frac{|\ln\ln m|}{\ln m}\right]^{\frac{3}{2}} \\
&\le m
\end{aligned}$$

according to the definition of $\lambda(m)$ and Condition (10). This shows that $\wp \in \mathcal{S}_m$ and thus $\mathrm{KL}(s, \mathcal{S}_m) \leq c_\beta \, \lambda(m)^{2\beta}$ for all $m \geq m_0(\underline{\beta}, \bar{\beta})$. Since c_β is continuous in β, there exists $c_{\underline{\beta},\bar{\beta}} > 0$ such that for all $\beta \in [\underline{\beta}, \bar{\beta}]$, for all $s \in \mathcal{H}\left(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta})\right)$, and for all $m \geq m_0(\underline{\beta}, \bar{\beta})$,

$$\mathrm{KL}(s, \mathcal{S}_m) \leq c_{\underline{\beta},\bar{\beta}} \lambda(m)^{2\beta}. \tag{27}$$

It remains to show the same result for $m \leq m_0(\underline{\beta}, \bar{\beta})$: let t_m be a mixture of $\mathcal{S}_m$, for all $\beta \in [\underline{\beta}, \bar{\beta}]$ and for all $s \in \mathcal{H}\left(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta})\right)$,

$$\begin{aligned} \mathrm{KL}(s, \mathcal{S}_m) &\leq \mathrm{KL}(s, t_m) \\ &\leq \int M\psi(x) \ln\left(\frac{M\psi(x)}{t_m(x)}\right) < +\infty. \end{aligned}$$

Then it can be easily shown that (27) is valid for all $m \geq 1$ by changing the constant $c_{\underline{\beta},\bar{\beta}}$. □

Proof of Theorem 2.9. In order to upper bound the right-hand side of the oracle inequality (3), we first control the constant $\mathcal{A}$ defined by (2) that depends on the parameters of the Gaussian mixture model $\mathcal{S}_m$:

$$\mathcal{A}^2 \leq 4\left\{c_2^2 + \ln\left(\frac{c_1 \bar{\mu}(m)}{\lambda(m)}\right)\right\}.$$

For the last term, we note that

$$\begin{aligned} \ln\left(\frac{c_1 \bar{\mu}(m)}{\lambda(m)}\right) &= \frac{1}{2} \ln\left((c_1 \tilde{G}_{\bar{\beta}})^2 \frac{|\ln \lambda(m)|}{\lambda^2(m)}\right) \\ &\leq c_{\bar{\beta}} \ln(m) \end{aligned}$$

since $\lambda(m) := \frac{a_{\bar{\beta}}}{m} (\ln m)^{\frac{3}{2}}$. Thus $\mathcal{A}^2$ is upper bounded by $c_{\bar{\beta}} \ln(m)$. For the observation of a sample with size n, the model collection is indexed by $\mathcal{M}_n = \{2, \ldots, n\}$ and then $m \leq n$. Thus for all $m \in \mathcal{M}_n$,

$$\begin{aligned} \mathrm{pen}(m) &= \kappa \frac{2m-1}{n} \left\{1 + 2\,\mathcal{A}^2 + \ln\left(\frac{1}{1 \wedge \frac{D(m)}{n} \mathcal{A}^2}\right)\right\} \\ &\leq c_{\bar{\beta}} \frac{m}{n} [\ln n + \ln m] \\ &\leq 2c_{\bar{\beta}} \frac{m}{n} \ln(n). \end{aligned}$$

According to Lemma 6.1 and the definition of $\lambda(m)$, the oracle inequality is upper bounded by

$$\begin{aligned} \mathbb{E}\left[d_H^2(s, \hat{s}_{\hat{m}})\right] &\leq \mathcal{C} \inf_{m \in \mathcal{M}_n} \left[\mathrm{KL}(s, \mathcal{S}_m) + \mathrm{pen}(m) + \frac{1}{n}\right] \\ &\leq c_{\underline{\beta},\bar{\beta}} \inf_{m \in \mathcal{M}_n} \left[\frac{(\ln m)^{3\beta}}{m^{2\beta}} + m \frac{\ln n}{n}\right]. \end{aligned}$$

Let $m_n := \inf\left\{m \geq 2\,; m \in \mathbb{N}\,; \,; \frac{(\ln m)^{3\beta}}{m^{2\beta}} \leq m \frac{\ln n}{n}\right\}$. Note that if $m_n = 2$, then $\mathbb{E}\left[d_H^2(s, \hat{s}_{\hat{m}})\right] \leq 4 c_{\underline{\beta},\bar{\beta}} \frac{\ln n}{n}$ and this case is completed. Assuming now that $m_n > 2$, we want to check that $m_n \leq n$. According to the definition of m_n,

$$\frac{(m_n - 1)^{2\beta+1}}{[\ln(m_n - 1)]^{3\beta}} < \frac{n}{\ln n}$$

thus

$$\left[\frac{(m_n-1)}{\ln(m_n-1)}\right]^{\square} < \frac{n}{\ln n}$$

where $\square = 3\beta$ if $\beta > 1$ and $\square = 2\beta + 1$ otherwise. Next, since $\frac{(m_n-1)}{\ln(m_n-1)} > 1$,

$$\frac{(m_n-1)}{\ln(m_n-1)} < \frac{n}{\ln n}$$

in all cases. Assuming that $n \geq 3$, it follows that $m_n \leq n$. Since $m_n \in \mathcal{M}_n$,

$$\begin{aligned}
\mathbb{E}\left[d_H^2(s, \hat{s}_{\hat{m}})\right] &\leq 2c_{\underline{\beta},\bar{\beta}} \frac{[\ln(m_n-1)]^{3\beta}}{(m_n-1)^{2\beta}} \\
&\leq 2c_{\underline{\beta},\bar{\beta}} 2^{2\beta} \frac{[\ln m_n]^{3\beta}}{m_n^{2\beta}} \\
&\leq 2^{2\beta+1} c_{\underline{\beta},\bar{\beta}} [\ln m_n]^{3\beta} \left[\frac{\ln n}{n} (\ln m)^{-3\beta}\right]^{\frac{2\beta}{2\beta+1}} \\
&\leq \tilde{c}_{\underline{\beta},\bar{\beta}}\, n^{-\frac{2\beta}{2\beta+1}} (\ln n)^{\frac{5\beta}{2\beta+1}}\,. \qquad \square
\end{aligned}$$

Appendix A. Appendices for the approximation result

A.1. Measure discretization

The following result is adapted from Lemma 2 in [6]. It allows us to approximate a general Gaussian mixture by a finite Gaussian mixture with a limited number of components.

Proposition A.1. *Let F be a probability measure on $[-a, a]$ and $\sigma > 0$ such that $\sigma < a$. Let $\varepsilon \in (0, \pi^{-\frac{1}{2}})$. Then there exists a discrete distribution F' on $[-a, a]$ with at most $54a\sigma^{-1}e^2 \left[1 \vee \ln\left(\frac{1}{\sqrt{\pi}\varepsilon}\right)\right]$ support points such that*

$$\|F * \psi_\sigma - F' * \psi_\sigma\|_\infty \leq \frac{2\varepsilon}{\sigma}.$$

Proof. The interval $[-a, a]$ can be partitioned into $k = \lfloor \frac{2a}{\sigma} \rfloor$ disjoint consecutive subintervals $I_1, \dots, I_k$ of length σ and a final subinterval I_{k+1} of length $l \leq \sigma$: $I_i = [a_i, a_i + \sigma[,\ i = 1, \dots, k$ and $I_{k+1} = [a_{k+1}, a_{k+1} + l]$. We decompose F on this partition $F = \sum_{i=1}^{k+1} F(I_i) F_i$ where each F_i is a probability measure concentrated on I_i. Then, $F * \psi_\sigma(x) = \sum_{i=1}^{k+1} F(I_i)(F_i * \psi_\sigma)(x)$. Let Z_i be a random variable distributed according to F_i, and let G_i be the law of $W_i = (Z_i - a_i)/\sigma$. Thus G_i is a probability measure on $[0, 1]$ for $i = 1, \dots, k$ and on $[0, l/\sigma] \subset [0, 1]$ for $i = k + 1$. Lemma A.2 is applied for each measure G_i and with $D = \ln\left(\frac{1}{\sqrt{\pi}\varepsilon}\right)^{-\frac{1}{2}}$. We obtain discrete distributions G'_i such that $\|G_i * \psi - G'_i * \psi\|_\infty \leq 2\varepsilon$. Let F'_i be the law of $a_i + \sigma W'_i$ if W'_i has law G'_i and set $F' = \sum_{i=1}^{k+1} F(I_i) F'_i$. We have

$$F_i * \psi_\sigma(x) = \mathbb{E}\left[\psi_\sigma(x - Z_i)\right] = \mathbb{E}\left[\frac{1}{\sigma}\psi\left(\frac{x - Z_i}{\sigma}\right)\right] = \mathbb{E}\left[\frac{1}{\sigma}\psi\left(\frac{x - a_i}{\sigma} - W_i\right)\right] = \frac{1}{\sigma} G_i * \psi\left(\frac{x - a_i}{\sigma}\right)$$

and $F'_i * \psi_\sigma(x) = \frac{1}{\sigma} G'_i * \psi\left(\frac{x - a_i}{\sigma}\right)$. Thus

$$|F_i * \psi_\sigma(x) - F'_i * \psi_\sigma(x)| = \frac{1}{\sigma}\left|G_i * \psi\left(\frac{x - a_i}{\sigma}\right) - G'_i * \psi\left(\frac{x - a_i}{\sigma}\right)\right| \leq \frac{1}{\sigma}\|G_i * \psi - G'_i * \psi\|_\infty \leq \frac{2\varepsilon}{\sigma}.$$

Then

$$|F * \psi_\sigma(x) - F' * \psi_\sigma(x)| = \left|\sum_{i=1}^{k+1} F(I_i)\left[F_i * \psi_\sigma(x) - F'_i * \psi_\sigma(x)\right]\right|$$
$$\leq \frac{2\varepsilon}{\sigma}\sum_{i=1}^{k+1} F(I_i).$$

Thus $\|F * \psi_\sigma - F' * \psi_\sigma\|_\infty \leq \frac{2\varepsilon}{\sigma}$ and the number of support points of the discrete distribution F' is upper bounded by

$$\sum_{i=1}^{k+1} 18\left[1 \vee \ln\left(\frac{1}{\sqrt{\pi}\varepsilon}\right)^{-1/2}\right]^2 e^2 \ln\left(\frac{1}{\sqrt{\pi}\varepsilon}\right) = (k+1)18\left[1 \vee \ln\left(\frac{1}{\sqrt{\pi}\varepsilon}\right)^{-1}\right] e^2 \ln\left(\frac{1}{\sqrt{\pi}\varepsilon}\right)$$
$$\leq 54a\sigma^{-1}e^2\left[1 \vee \ln\left(\frac{1}{\sqrt{\pi}\varepsilon}\right)\right]. \qquad \square$$

The following lemma is an adaptation of Lemma 3.1 in [5], the complete proof can be found in [14]. For this lemma, one introduces the inverse function of $\psi_\sigma(.)$ defined by $\psi_\sigma^{-1}(y) = \sigma\sqrt{-\ln(\sqrt{\pi}y)}$ on $(0, \pi^{-\frac{1}{2}}]$.

Lemma A.2. *Let F be a probability measure on $[0,B]$. Let $\varepsilon \in (0, \pi^{-\frac{1}{2}})$ and let D be a positive constant such that $B \leq D\psi^{-1}(\varepsilon)$. Then there exists a discrete distribution F' on $[0,B]$ with at most $18(1 \vee D)^2 e^2 \ln\left(\frac{1}{\sqrt{\pi}\varepsilon}\right)$ support points such that*

$$\|F * \psi - F' * \psi\|_\infty \leq 2\varepsilon.$$

A.2. Technical results for f, f_k, g_k, h_k and their convolutions

The proofs of the following technical lemmas are given in [14].

Lemma A.3. *Let $f_0 = f$ and $\forall k \in \mathbb{N}^*$, $f_{k+1} = f - \Delta_\sigma f_k$ with $\Delta_\sigma f_k = K_\sigma f_k - f_k$.*

1. *For all $x \in \mathbb{R}$, $f_k(x) = \sum_{i=0}^{k} \binom{k+1}{i+1}(-1)^i K_\sigma^i f(x)$.*
2. *For all $k \in \mathbb{N}$, $\int_\mathbb{R} f_k(x)\mathrm{d}x = 1$.*
3. *For all $i \in \mathbb{N}$ and for all $x \in \mathbb{R}$, $K_\sigma^i f(x) \leq \frac{M}{\sqrt{\pi}}$ and thus $|f_k(x)| \leq (2^{k+1}-1)\frac{M}{\sqrt{\pi}}$.*

Lemma A.4. *Let $\beta > 0$ and $k \in \mathbb{N}$ such that $\beta \in (2k, 2k+2]$. Let f be a density function belonging to $\mathcal{H}(\beta, \mathcal{P})$ where $\mathcal{P} = \{\gamma, l^+, L, \varepsilon, C, \alpha, \xi, M\}$.*

1. *Let $\bar{\sigma} > 0$ such that if Y is distributed from a centered Gaussian density with variance $\bar{\sigma}^2$, then $P(0 < Y < 2\alpha) = \frac{1}{3}$. For all $\sigma < \bar{\sigma}$,*
$$K_\sigma f(x) \geq \frac{\xi\sqrt{\pi}}{3M} f(x).$$
2. *There exists $\bar{\sigma}(\beta) > 0$ and $A_\beta > 0$ such that for all $\sigma < \bar{\sigma}(\beta)$,*
$$K_\sigma h_k(x) \geq \frac{\xi\sqrt{\pi}}{6M(1 + A_\beta\sigma^{2\beta})} f(x).$$

Furthermore, $\bar{\sigma}(\beta)$ can be chosen as a continuous function of β.

Remark A.5. The first result of Lemma A.4 is based on the monotonicity assumption on f. It comes from Remark 3 of [4]. In the second result, the constants $\bar{\sigma}(\beta)$ and A_β are due to the result (14) in Lemma 4.2.

Lemma A.6. *Let $p \in (0,1)$. For all $x \in \mathbb{R}$, we have that*

- *for all $i \in \mathbb{N}$ and for all $\sigma < 1 - p^{1/i}$, $K_\sigma^i f(x) \le M \left(\frac{2}{\sqrt{3}}\right)^i \psi(px)$.*
- *for all $\sigma < 1 - p^{1/k}$,*

$$\max\left(f_k(x), g_k(x), \frac{1}{2}h_k(x)\right) \le 2M\left(\frac{4}{\sqrt{3}}\right)^k \psi(px).$$

Appendix B. Appendices for the lower bound result

The two following results are crucial for establishing the lower bound: The first one is the so-called Varshamov–Gilbert's lemma and the second one is a corollary of a lemma given in [2]. They correspond to Lemma 4.7 and Corollary 2.19 in [13] respectively.

Lemma B.1. *Let $\{0,1\}^D$ be equipped with Hamming distance δ. Given $\alpha \in (0,1)$, there exists some subset Θ of $\{0,1\}^D$ with the following properties*

$$\begin{cases} \delta(\theta,\theta') > \frac{(1-\alpha)D}{2} \text{ for every } (\theta,\theta') \in \Theta, \theta \neq \theta' \\ \ln|\Theta| > \frac{\rho D}{2} \end{cases}$$

where $\rho = (1+\alpha)\ln(1+\alpha) + (1-\alpha)\ln(1-\alpha)$. In particular $\rho > \frac{1}{4}$ when $\alpha = \frac{1}{2}$.

Corollary B.2. *Let (S,d) be some pseudo-metric space, $\{\mathbb{P}_s, s \in S\}$ be some statistical model. Let κ denote an absolute constant (given in Cor. 2.18 of [13]). Then for any estimator $\tilde{s}$ and any finite subset $\mathcal{C}$ of S such that $\max\limits_{s,t\in\mathcal{C}} \mathrm{KL}(\mathbb{P}_s,\mathbb{P}_t) \le \kappa \ln|\mathcal{C}|$, the following lower bound holds for every $p > 1$*

$$\sup_{s\in\mathcal{C}} \mathbb{E}_s[d^p(s,\tilde{s})] \ge 2^{-p}(1-\kappa)\left[\min_{s,t\in\mathcal{C}, s\neq t} d(s,t)\right]^p.$$

The following lemma, used to prove Proposition 2.6, gives an expression of the derivatives of the logarithm of a function. The proof can be found in [14].

Lemma B.3. *Let $i \in \mathbb{N}^*$ and let t be a strictly positive function, $t \in \mathcal{C}^i$. Then*

$$(\ln t)^{(i)}(x) = \frac{P_i(x)}{t(x)^{2^{i-1}}}$$

where

$$P_i(x) = \sum_{(\eta_0,\ldots,\eta_i)\in\Xi_i} \rho(\eta_0,\ldots,\eta_i) \prod_{j=0}^{i} \left[t^{(j)}(x)\right]^{\eta_j}$$

with

$$\Xi_i = \left\{(\eta_0,\ldots,\eta_i) \in \mathbb{N}^{i+1}; \sum_{j=0}^{i} \eta_j = 2^{i-1},\ \sum_{j=0}^{i} j\eta_j = i\right\}$$

and $\rho(\eta_0,\ldots,\eta_i)$'s are the polynomial coefficients.

Acknowledgements. The authors thank two anonymous referees for relevant remarks and constructive comments on a previous version of the paper. The authors are grateful to Willem Kruijer and Judith Rousseau for helpful discussions.

References

[1] J.-P. Baudry, C. Maugis and B. Michel, Slope heuristics: overview and implementation. *Stat. Comput.* **22** (2011) 455–470.

[2] L. Birgé, A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory.* **51** (2005) 1611–1615.

[3] W. Cheney and W. Light, A course in approximation theory, Graduate Studies in Mathematics, vol. 101 of *Amer. Math. Soc.* Providence, RI (2009).

[4] S. Ghosal, J.K. Ghosh and R.V. Ramamoorthi, Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Stat.* **27** (1999) 143–158.

[5] S. Ghosal and A. van der Vaart, Entropy and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Stat.* **29** (2001) 1233–1263,.

[6] S. Ghosal and A. van der Vaart, Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Stat.* **35** (2007) 697–723.

[7] U. Grenander, *Abstract inference.* John Wiley and Sons Inc., New York (1981).

[8] T. Hangelbroek and A. Ron, Nonlinear approximation using Gaussian kernels. *J. Functional Anal.* **259** (2010) 203–219.

[9] J.A. Hartigan, Clustering algorithms, *Probab. Math. Stat.* John Wiley and Sons, New York-London-Sydney (1975).

[10] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning, Data mining, inference, and prediction. *Statistics.* Springer, New York, 2nd edition (2009).

[11] W. Kruijer, J. Rousseau and A van der Vaart, Adaptive Bayesian Density Estimation with Location-Scale Mixtures. *Electron. J. Statist.* **4** (2010) 1225–1257.

[12] B. Lindsay, *Mixtures Models: Theory, Geometry and Applications.* IMS, Hayward, CA (1995).

[13] P. Massart, Concentration Inequalities and Model Selection. École d'été de Probabilités de Saint-Flour, 2003. *Lect. Notes Math.* Springer (2007).

[14] C. Maugis and B. Michel, Adaptive density estimation for clustering with Gaussian mixtures (2011). `arXiv:1103.4253v2`.

[15] C. Maugis and B. Michel, Data-driven penalty calibration: a case study for Gaussian mixture model selection. *ESAIM: PS* **15** (2011) 320–339.

[16] C. Maugis and B. Michel, A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: PS* **15** (2011) 41–68.

[17] G. McLachlan and D. Peel, *Finite Mixture Models.* Wiley (2000).

[18] A.B. Tsybakov, Introduction to nonparametric estimation. *Statistics.* Springer, New York (2009).

[19] J. Wolfowitz, Minimax estimation of the mean of a normal distribution with known variance. *Ann. Math. Stat.* **21** (1950) 218–230.