Electronic Journal of Statistics
Vol. 11 (2017) 4065–4102
ISSN: 1935-7524
DOI: 10.1214/17-EJS1344

Posterior concentration rates for mixtures of normals in random design regression

Zacharie Naulet and Judith Rousseau

CEREMADE, Université Paris-Dauphine, France
e-mail: zacharie.naulet@dauphine.eu; rousseau@ceremade.dauphine.com

Abstract: Previous works on location and location-scale mixtures of normals have shown different upper bounds on the posterior rates of contraction, either in a density estimation context or in nonlinear regression. In both cases, the observations were assumed not too spread by considering either the true density has light tails or the regression function has compact support. It has been conjectured that in a situation where the data are diffuse, location-scale mixtures may benefit from allowing a spatially varying order of approximation. Here we test the argument on the mean regression with normal errors and random design model. Although we cannot invalidate the conjecture due to the lack of lower bound, we find slower upper bounds for location-scale mixtures, even under heavy tails assumptions on the design distribution. However, the proofs suggest to introduce *hybrid* location-scale mixtures for which faster upper bounds are derived. Finally, we show that all tails assumptions on the design distribution can be released at the price of making the prior distribution covariate dependent.

MSC 2010 subject classifications: Primary 62G20; secondary 62G08.
Keywords and phrases: Adaptive estimation, Bayesian nonparametric estimation, nonparametric regression, Hölder class, mixture prior, rate of contraction, heavy tails.

Received August 2016.

1. Introduction

Nonparametric mixtures models are highly popular in the Bayesian nonparametric literature, due to both their reknown flexibility and relative easiness of implementation, see Hjort et al. (2010) for a review. They have been used in particular for density estimation, clustering and classification and recently nonparametric mixtures models have also been proposed in nonlinear regression models, see for instance de Jonge and van Zanten (2010); Wolpert, Clyde and Tu (2011); Naulet and Barat (2015).

Letting $\mathcal{E}$ denote the set of all $d \times d$ positive definite real matrices and $\varphi_\Sigma(x) := \exp(-\frac{1}{2}x^T\Sigma^{-1}x)$ for all $x \in \mathbb{R}^d$

$$f_{M,\Sigma}(x) = \int_{\mathbb{R}^d} \det(\Sigma)^{-\frac{q}{2}} \varphi_\Sigma(x-\mu)\, dM(\mu), \tag{1}$$

while a location-scale mixture has the form

$$f_M(x) = \int_{\mathcal{E}\times\mathbb{R}^d} \det(\Sigma)^{-\frac{q}{2}} \varphi_\Sigma(x-\mu)\, dM(\Sigma,\mu). \tag{2}$$

In the context of density estimation $q = 1$ in equations (1) and (2) and M is a probability measure so that $f_{M,\Sigma}$ and f_Σ are proper density functions. In nonlinear regression q can be arbitrary and M is a signed measure.

Location and location-scale mixtures of normals are used in the Bayesian nonparametric literature to model smooth curves, typically probability densities, by putting a prior on the mixing distribution M, and on Σ for location mixtures (1). The most popular prior distributions on M are either finite with unknown number of components, as in Kruijer, Rousseau and van der Vaart (2010) and the reknown Dirichlet Process (Ferguson, 1973) or some of its extensions. In both cases M is discrete almost surely.

There is now a large literature on posterior concentration rates for nonparametric mixtures models, initiated by Ghosal and van der Vaart (2001, 2007a) and improved by Kruijer, Rousseau and van der Vaart (2010); Shen, Tokdar and Ghosal (2013); Scricciolo (2014) in the context of density estimation with location mixtures of normals. Canale and De Blasi (2017) studied location-scale mixtures of normal distributions, still in density estimation. Regarding nonlinear regression, location mixtures models have been investigated in de Jonge and van Zanten (2010) and location-scale mixtures models in Naulet and Barat (2015), both in the context of the Gaussian mean regression with fixed design.

In Kruijer, Rousseau and van der Vaart (2010) and later on in Shen, Tokdar and Ghosal (2013); Scricciolo (2014) it was proved that location mixtures of normals distributions lead to adaptive (nearly) optimal posterior concentration rates (for L_1 metrics) over collections of β-Hölder types functional classes, in the context of density estimation for independently and identically distributed random variables. Contrarywise, in Canale and De Blasi (2017), suboptimal posterior concentration rates are derived for location-scale mixtures of normals and the authors obtain rates that are at best $n^{-\beta/(2\beta+d+1)}$ up to a $\log n$ term in place of $n^{-\beta/(2\beta+d)}$. These results are obtained under strong assumptions on the tail of the true density f_0, since it is assumed that $f_0(x) \lesssim e^{-c|x|^\tau}$ when x goes to infinity, for some positive c, τ.

The same phenomenon is observed in the nonlinear regression model with normal errors and covariates lying in a compact set. While the optimal rate $n^{-2\beta/(2\beta+d)}$ (up to power of $\log n$) with respect to the empirical L^2 metric is found by de Jonge and van Zanten (2010) using location mixtures, Naulet and Barat (2015) were only able to find the slower $n^{-2\beta/(2\beta+d+1)}$ for location-scale mixtures. In both cases the design lives on $[0,1]^d$.

In density estimation, it is well known that the optimal rates with respect to L^1 metric depend heavily on the nature of the assumptions made on the tails of f_0, see for instance Juditsky et al. (2004); Reynaud-Bouret, Rivoirard and Tuleau-Malot (2011); Goldenshluger and Lepski (2014). In particular, the optimal rate is $n^{-2\beta/(2\beta+d)}$ only under some tail assumptions, and deteriorates to 1 gradually as the tails of the density become heavier. In Canale and De Blasi

(2017), the authors suggest that location-scale mixtures could perform better than location mixtures if the true density f_0 is heavy tailed, since it may benefit from approximating f_0 differently in zones of dense data than in zones of sparse data.

There is currently, however, one strong limitation in understanding the robustness to tails of mixtures of normals in density estimation. The proofs to rates of contraction involve approximating the true density f_0 by a *convex* and finite mixture in terms of Kullback-Leibler divergence. Although approximation with non convex mixtures is rather easy, the convexity constraint is painful and is dealt by imposing non classical smoothness assumptions, such as requiring that $\log f_0$ is locally β-Hölder instead of requiring f_0 β-Hölder or Besov. This seemingly innocuous fact has deep consequences. In Bochkina and Rousseau (2016), almost no tail assumption (but a moment of order strictly greater than 2 for F_0) is needed to achieve the minimax rate $n^{-\beta/(2\beta+1)}$, for estimating densities on $\mathbb{R}_+$ using mixtures of Gamma distributions. Thus, some nonexplicit tail assumptions must be hidden behind the β-Hölder assumption on $\log f_0$, which blurs the understanding of the robustness of mixtures of normals to tails.

Instead of challenging the problem of approximating a given density by a convex finite mixture with respect to KL divergence, we propose to test the intuition of Canale and De Blasi (2017) on the mean regression problem with normal errors, since the same difference in the rates between location and location-scale mixtures of normals has been observed by Naulet and Barat (2015) when measuring the contraction rates with respect to the empirical L^2 distance of the covariates. Our goal is to understand if location-scale mixtures of normals can benefit from a varying order of approximation of the true regression function f_0, where the order vary according to the density of the observations. Hence, we study the use of mixtures models in the nonparametric regression models

$$\begin{aligned} Y_i &= f(X_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} N(0, s^2), \quad i = 1, \dots, n, \\ X_1, \dots, X_n &\overset{\text{i.i.d}}{\sim} Q_0, \quad f \in L^2(Q_0), \end{aligned} \tag{3}$$

where $L^2(Q_0)$ stand for space of (equivalence classes of) functions that are square integrable with respect to Q_0 and the spreadness of the data is controlled by the design distribution and written as $\int_{\mathbb{R}^d} \|x\|^p \, dQ_0(x)$. The parameter is f with prior distribution denoted by Π. We assume that s is known and $s = 1$, which is just a matter of convenience for proofs. All the results of the paper can be translated to the case s unknown using the same methodology as Salomond (2013) or Naulet and Barat (2015).

Our aim is to study posterior concentration rates around the true regression function f_0 defined by sequences ϵ_n converging to zero with n and such that

$$\Pi\left(n^{-1} \textstyle\sum_{i=1}^n |f(x_i) - f_0(x_i)|^2 \leq \epsilon_n^2 \mid \mathbf{y}^n, \mathbf{x}^n\right) = 1 + o_p(1), \tag{4}$$

under the model f_0 for both location and location-scale mixtures of normals. By analogy to the case of density estimation of Reynaud-Bouret, Rivoirard and Tuleau-Malot (2011) and Goldenshluger and Lepski (2014) we assume that $f_0 \in L^1$ and belongs to a Hölder ball with smoothness β.

We show in Section 2, that in most cases the bounds found on ϵ_n^2 in equation (4) for location mixtures are better than the bounds for location-scale mixtures. Unless p goes to infinity, the posterior concentration rates are not as good as the minimax rate $n^{-2\beta/(2\beta+d)}$, obtained in the context of a design on $[0,1]^d$. This rate is suboptimal for light tail design points, since in this case the minimax posterior concentration rate is given by $n^{-2\beta/(2\beta+d)}$. To improve on this bound we propose a version of location-scale mixtures models, which we call the hybrid location-scale mixtures, and we show that this nonparametric mixture model leads to better bounds than location mixtures (and thus than location-scale mixtures). All these results are up to $\log n$ terms and are summarized in Table 1 which displays the value r defined by $\epsilon_n^2 = n^{-r}$.

Finally, we draw the attention of the reader to the fact that all the results in this paper are only upper bounds on the rates of contraction. In absence of corresponding lower bounds, no one should use these results to conclude definitively on the performance of each mixtures over β-Hölder balls. The computation of lower bounds on the rate of contraction for mixture priors is still an open question today. However, in the case $p > 2\beta$ for hybrid mixture and $p \to \infty$ for location mixture, the minimax rates are known to be $n^{-2\beta/(2\beta+d)}$ thus we can conclude about the optimality of these mixtures in that cases. To our knowledge, in all other cases no minimax lower bound are known.

TABLE 1

Summary of posterior rates of convergence for different types of mixtures. The rates are understood to be in the form $\epsilon_n^2 = n^{-r}$, up to powers of $\log n$ factors, where r is given below. here $\kappa > 0$ is a parameter that depends on the prior and can be made equal to 1.

	$0 < p < 2d$		$p \geq 2d$	
	$0 < p \leq 2\beta$	$p > 2\beta$	$0 < p \leq 2\beta$	$p > 2\beta$
Location	$\frac{2\beta}{2\beta + \max(\kappa, \beta + d)}$		$\frac{2\beta}{2\beta + \max(\kappa, d + 2d\beta/p)}$	
Location-scale	$\frac{2\beta}{2\beta + \max(\beta + d, 2\beta d/p) + \kappa}$	$\frac{2\beta}{2\beta + d + \kappa}$	$\frac{2\beta}{2\beta + \max(\beta + d, 2\beta d/p) + \kappa}$	$\frac{2\beta}{2\beta + d + \kappa}$
Hybrid	$\frac{2\beta}{2\beta + \max(\kappa, \min(\beta + d, 2\beta d/p))}$	$\frac{2\beta}{2\beta + \max(\kappa, d)}$	$\frac{2\beta}{2\beta + \max(\kappa, \min(\beta + d, 2\beta d/p))}$	$\frac{2\beta}{2\beta + \max(\kappa, d)}$

The main results with the description of the three types of prior models and the associated posterior concentration rates are presented in Section 2. Proofs are presented in Section 3 and some technical lemma are proved in the appendix.

1.1. Notations

In the sequel we use repeatedly the following notations.

- We call $P_f(\cdot \mid X)$ the distribution of the random variable $Y \mid X$ under the model (3), associated with the regression function f. Given $(X_1, \ldots, X_n)$, $P_f^n(\cdot \mid X_1, \ldots, X_n)$ stands for the distribution of the vector $(Y_1, \ldots, Y_n)$ of independent random variables $Y_j \sim P_f(\cdot \mid X_j)$. Also, for any random variable Z with distribution P, and any function g, $Pg(Z)$ denote the expectation of $g(Z)$.

- For any $a > 0$, we let $\mathrm{SGa}(a)$ denote the symmetric Gamma distribution with parameter a; that is $X \sim \mathrm{SGa}(a)$ has the distribution of the difference of two independent Gamma random variables with parameters $(a, 1)$.
- For any finite positive measure α on the measurable space $(X, \mathcal{X})$, let Π_α denote the symmetric Gamma process distribution with parameter α (Wolpert, Clyde and Tu, 2011; Naulet and Barat, 2015); that is, $M \sim \Pi_\alpha$ is a random signed measure on $(X, \mathcal{X})$ such that far any disjoints $B_1, \dots, B_k \in \mathcal{X}$ the random variables $M(B_1), \dots, M(B_k)$ are independent with distributions $\mathrm{SGa}(\alpha(B_i))$, $i = 1, \dots, k$.
- For all $k = (k_1, \dots, k_d) \in \mathbb{N}^d$ and all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ we write $|k| := k_1 + \cdots + k_d$, $k! := k_1! \dots k_d!$, and $x^k := x_1^{k_1} \dots x_d^{k_d}$. Moreover, for all $f : \mathbb{R}^d \to \mathbb{R}$ with continuous p-th order derivatives at $x \in \mathbb{R}^d$ we write

$$D^k f(x) := \frac{\partial^{|k|} f}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}(x), \quad |k| \leq p.$$

- For any $\beta > 0$, we let $\mathtt{C}^\beta$ denote the Hölder space of order β; that is the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\|f\|_{\mathtt{C}^\beta} := \max_{|k| \leq p} \sup_{x \in \mathbb{R}^d} |D^k(x)| + \max_{|k|=p} \sup_{x \neq y} |D^k(x) - D^k(y)|/|x - y|^{\beta - p}$ is finite, where p is the largest integer strictly smaller than β.
- We denote by $\|\cdot\|$ the standard euclidean norm on $\mathbb{R}^d$, and, for any $x, y \in \mathbb{R}^d$, xy is the standard inner product. For any $d \times d$ matrix A with real eigenvalues, we denote $\lambda_1(A) \geq \dots \lambda_d(A)$ its eigenvalues in decreasing order, $\|A\| := \sup_{x \neq 0} \|Ax\|/\|x\|$ its spectral norm, and $\|A\|_{\max} := \max_{i,j} |A_{i,j}|$, where A_{ij} are the entries of A.
- Throughout the paper C denotes a generic constant, not necessarily the same everywhere. Inequalities up to a generic constant are denoted by $\lesssim$ and $\gtrsim$.

2. Posterior convergence rates for Symmetric Gamma mixtures

In this section we present the main results of the paper. We first present the three types of priors that are studied; *i.e.* location mixtures, location-scale mixtures and hybrid location-scale mixtures and for each of these families of priors we provide the associated posterior concentration rates.

Recall that we consider observations $(Y_i, X_i)_{i=1}^n$ independent and identically distributed according to model (3) and we note $\mathbf{y}^n = (Y_1, \cdots, Y_n)$ and $\mathbf{x}^n = (X_1, \cdots, X_n)$. We denote the prior and the posterior distribution on f by $\Pi(\cdot)$ and $\Pi(\cdot \mid \mathbf{y}^n, \mathbf{x}^n)$ respectively.

2.1. Families of priors

In this section we present three variants of mixture models as defined in equation (1) or equation (2).

2.1.1. Location mixtures of normals

A symmetric Gamma process location mixture of normals prior Π is the distribution of the random function $f(x) := \int \varphi_\Sigma(x-\mu)\, dM(\mu)$ where $\Sigma \sim G_\Sigma$ and $M \sim \Pi_\alpha$, with α a finite positive measure on $\mathbb{R}^d$, and G_Σ a probability measure on $\mathcal{E}$.

We restrict our discussion to priors for which the following conditions are verified. We assume that there are positive constants $a_1, a_2, a_3,\ b_1, b_2, b_3, b_4$ and $\kappa > 0$ such that G_Σ satisfies for all $x \geq 1$ and all $t \in (0,1)$

$$G_\Sigma\left(\Sigma\ :\ \lambda_1(\Sigma) > x\right) \lesssim \exp(-a_1 x^{b_1}) \tag{5}$$

$$G_\Sigma\left(\Sigma\ :\ \lambda_d(\Sigma) < 1/x\right) \lesssim \exp(-a_2 x^{b_2}) \tag{6}$$

$$G_\Sigma\left(x^{-1} \leq \lambda_i(\Sigma^{-1}) \leq x^{-1}(1+t),\ 1 \leq i \leq d\right) \gtrsim x^{-b_3} t^{b_4} \exp(-a_3 x^{-\kappa/2}). \tag{7}$$

We let $\alpha := \overline{\alpha} G_\mu$ for a positive constant $\overline{\alpha} > 0$ and G_μ a probability distribution on $\mathbb{R}^d$. We assume that there are positive constants b_5, b_6 such that G_μ satisfies for all $x \in \mathbb{R}^d$

$$G_\mu\left(\|\mu - x\| \leq t\right) \gtrsim t^{b_5}(1 + \|x\|)^{-b_6}, \quad \forall t \in (0,1). \tag{8}$$

The heavy tail condition on G_μ is required to adapt to potential heavy tails of Q_0.

2.1.2. Location-scale mixtures of normals

A symmetric Gamma process location-scale mixture of normals prior Π is the distribution of the random function $f(x) := \int \varphi_\Sigma(x-\mu)\, dM(\Sigma,\mu)$ where $M \sim \Pi_\alpha$, with α a finite positive measure on $\mathcal{E} \times \mathbb{R}^d$. Hence in this model the prior is entirely defined by Π_α with α a measure on $\mathcal{E} \times \mathbb{R}^d$, while in Section 2.1.1 the prior isdefined by $\Pi_\alpha \times G_\Sigma$, with α a measure on $\mathbb{R}^d$. To simplify notations we keep π_α for both types of priors and the context will make clear which prior is refered to.

We restrict our discussion to priors for which $\alpha := \overline{\alpha} G_\Sigma \times G_\mu$, with $\overline{\alpha} > 0$ and G_Σ, G_μ satisfying the same assumptions as in Section 2.1.1.

2.1.3. Hybrid location-scale mixtures of normals

By hybrid location-scale mixture of normals, we mean the distribution Π of the random function $f(x) := \int \varphi_\Sigma(x-\mu)\, dM(\Sigma,\mu)$, where

$$M \sim \Pi_\alpha, \quad \alpha = \overline{\alpha} P_\Sigma \times G_\mu, \quad \overline{\alpha} > 0,$$
$$P_\Sigma \sim \Pi_\Sigma.$$

and G_μ a probability measure satisfying equation (8). Here Π_Σ is a prior distribution on the space of probability measures on $\mathcal{E}$ (endowed with Borel σ-algebra). We now formulate conditions on Π_Σ that are the random analoguous

to equations (5) and (6). For the same constants a_1, a_2, b_1, b_2 as in Section 2.1.1, we consider the existence of positive constants a_4, a_5 such that Π_Σ satisfies for $x > 0$ large enough

$$\Pi_\Sigma\left(P_\Sigma : P_\Sigma(\lambda_1(\Sigma) > x) \geq \exp(-a_1 x^{b_1}/2)\right) \lesssim \exp(-a_4 x^{b_1}), \tag{9}$$

$$\Pi_\Sigma\left(P_\Sigma : P_\Sigma(\lambda_d(\Sigma) < 1/x) \geq \exp(-a_2 x^{b_2}/2)\right) \lesssim \exp(-a_5 x^{b_2}). \tag{10}$$

For any $j, u \geq 0$, let $\mathcal{E}_{j,u} := \{\Sigma \in \mathcal{E} : \forall i : 2^{2j} \leq \lambda_i(\Sigma^{-1}) \leq 2^{2j}(1 + 2^{-u})\}$. As a replacement of equation (7), we assume that for all $\beta > 0$ there are constants a_6, b_7 and κ^* such that for any positive integer J large enough

$$\Pi_\Sigma\left(\bigcap_{j=0}^{J}\{P_\Sigma : P_\Sigma(\mathcal{E}_{j,J\beta}) \geq 2^{-J}\}\right) \gtrsim \exp(-a_6 J^{b_7} 2^{J\kappa^*}). \tag{11}$$

Equations (9) to (11) are rather restrictive and it is not clear *a priori* whether or not such distribution exists. For example, if P_Σ is chosen to be almost-surely equal to G_Σ satisfying equations (5) to (7), then equation (11) is not satisfied. However, we now show that under conditions on the base measure, Π_Σ can be chosen as a Dirichlet Process, hereafter refered to as DP.

We recall that if Π_Σ is a Dirichlet Process distribution with base measure $\alpha_\Sigma G_\Sigma(\cdot)$ on $\mathcal{E}$ (Ferguson, 1973), then $P_\Sigma \sim \Pi_\Sigma$ is a random probability measure on $\mathcal{E}$ such that for any Borel measurable partition $A_1, \ldots, A_k$ of $\mathcal{E}$, the joint distribution $P_\Sigma(A_1), \ldots, P_\Sigma(A_k)$ is the k-variate Dirichlet distribution with parameters $\alpha_\Sigma G_\Sigma(A_1), \ldots, \alpha_\Sigma G_\Sigma(A_k)$.

Proposition 1. *Let $\alpha_\Sigma > 0$, G_Σ a probability measure on $\mathcal{E}$ satisfying equations (5) to (7), and Π_Σ be a Dirichlet Process with base measure $\alpha_\Sigma G_\Sigma(\cdot)$. Then Π_Σ satisfies equations (9) to (11) with constants $a_4 = a_1/2$, $a_5 = a_2$, $b_7 = 0$, $\kappa^* = \kappa$ and a constant $a_6 > 0$ eventually depending on β.*

Proof. We first prove equation (9). It follows from the definition of the DP that $P_\Sigma(\Sigma : \lambda_1(\Sigma) > x)$ has Beta distribution with parameters $\alpha_\Sigma G_\Sigma(\Sigma : \lambda_1(\Sigma) > x)$ and $\alpha_\Sigma[1 - G_\Sigma(\Sigma : \lambda_1(\Sigma) > x)]$, then by Markov's inequality

$$\Pi_\Sigma\Big(P_\Sigma : P_\Sigma(\Sigma : \lambda_1(\Sigma) > x) \geq t\Big) \leq \frac{G_\Sigma(\Sigma : \lambda_1(\Sigma) > x)}{t}.$$

Choosing $t = \exp(-a_1 x^{b_1}/2)$ and using equations (5) to (7) leads to (9). The same steps with $G_\Sigma(\Sigma : \lambda_d(\Sigma) < 1/x)$ give the proof of equation (10). It remains to prove equation (11). For all $\beta \geq 1$ the sets $\mathcal{E}_{j,J\beta}$, $j = 0, \ldots, J$ are disjoint. Set $\mathcal{E}^c_{J\beta} := \bigcap_{j=0}^{J} \mathcal{E}^c_{j,J\beta}$, where $\mathcal{E}^c_{j,J\beta}$ are the complement of the sets $\mathcal{E}_{j,J\beta}$. If $\alpha_\Sigma G_\Sigma(\mathcal{E}^c_{J\beta}) \leq 1$ let $\mathcal{E}_{J+1,J\beta} = \mathcal{E}^c_{J\beta}$ and $N = 1$; otherwise split $\mathcal{E}^c_{J\beta}$ into $N > 1$ disjoint subsets $\mathcal{E}^c_{1,J\beta}, \ldots \mathcal{E}^c_{N,J\beta}$ such that $\exp(-2^{J\kappa}) \leq \alpha_\Sigma G_\Sigma(\mathcal{E}^c_{k,J\beta}) \leq 1$ for all $k = 1, \ldots, N$ and set $\mathcal{E}_{J+1,J\beta} = \mathcal{E}^c_{1,J\beta}$, $\mathcal{E}_{J+2,J\beta} = \mathcal{E}^c_{2,J\beta}$, $\ldots, \mathcal{E}_{J+N,J\beta} = \mathcal{E}^c_{N,J\beta}$ (since $G_\Sigma(\mathcal{E}) = 1$ this can be done with a number N independent of J). For J large enough, acting as in Ghosal, Ghosh and van der Vaart (2000, lemma 6.1), it follows

$$\Pi_\Sigma\Big(P_\Sigma : P_\Sigma(\mathcal{E}_{j,J\beta}) \geq 2^{-J} \quad \forall\, 0 \leq j \leq J\Big) \geq \frac{\Gamma(\alpha_\Sigma) 2^{-J(J+N)}}{\prod_{j=0}^{J+N} \Gamma(\alpha_\Sigma G_\Sigma(\mathcal{E}_{j,J\beta}))}.$$

Also, $\alpha_\Sigma G_\Sigma(\mathcal{E}_{j,J\beta}) \leq 1$ implies $\Gamma(\alpha_\Sigma G_\Sigma(\mathcal{E}_{j,J\beta})) \leq 1/(\alpha_\Sigma G_\Sigma(\mathcal{E}_{j,J\beta}))$, hence

$$\Pi_\Sigma\Big(P_\Sigma \ : \ P_\Sigma(\mathcal{E}_{j,J\beta}) \geq 2^{-J} \quad \forall\, 0 \leq j \leq J\Big) \\ \geq \Gamma(\alpha_\Sigma)\alpha_\Sigma^{J+N+1}2^{-J(J+N)} \prod_{j=0}^{J+N} G_\Sigma(\mathcal{E}_{j,J\beta}).$$

Since N does not depend on J, one can find a constant $C > 0$ such that

$$\Pi_\Sigma\Big(P_\Sigma : P_\Sigma(\mathcal{E}_{j,J\beta}) \geq 2^{-J} \quad \forall\, 0 \leq j \leq J\Big) \\ \geq \Gamma(\alpha_\Sigma) \exp\left\{-CJ^2 + \sum_{j=0}^{J} \log G_\Sigma(\mathcal{E}_{j,J\beta}) + \sum_{j=J+1}^{J+N} \log G_\Sigma(\mathcal{E}_{j,J\beta})\right\}.$$

By construction, the second sum in the rhs of the last equation is lower bounded by $-N2^{J\kappa}$, whereas if G_Σ satisfies equations (5) to (7), the first sum is lower bounded by $-C'2^{J\kappa}$ for a constant $C' > 0$ eventually depending on β. □

Another example that can satisfy equations (9) to (11) is to consider for P_Σ a finite mixture with unknown number of components. For instance,

$$P_\Sigma = \sum_{j=1}^{J} p_j \delta_{\Sigma_j}, \quad J \sim 1 + \mathcal{P}(\alpha_\Sigma), \quad \Sigma_j \overset{iid}{\sim} G_\Sigma.$$

This example behaves very similarly to the Dirichlet process. Note that instead of a Poisson random variable, some distribution with exponential tails like the Geometric distribution also satisfies equations (5) to (7). For the two previous examples, draws from Π_Σ are almost-surely purely atomic measures. We don't know any example of prior distribution Π_Σ such that $P_\Sigma \sim \Pi_\Sigma$ is not almost-surely purely atomic and Π_Σ satisfies equation (11). A distinctive feature of previous examples is that *a priori* the probability of having two (or more) components of the mixture sharing the same covariance matrix is positive, a fact which is not true when P_Σ is not atomic. We believe this property is the fundamental reason why the rates are improved compared to location-scale mixtures. Inspection of proofs in the present paper shows that, to improve the rates, it is sufficient that a priori, the probability of having more than $(\log M)^u$ distinct dilation matrices on any subset of M components of the mixture goes to zero fast enough for some $u > 0$ when $M \to \infty$.

Note that this is the same idea as the prior defined by equation (2.2) in Ghosal and van der Vaart (2001) in the context of density estimation for super-smooth densities with light tails. It is also worth mentioning that when Π_Σ is a Dirichlet Process, hybrid location-scale mixtures are closely related to the well-known *Hierarchical Dirichlet Processes* (Teh et al., 2006), because of the close relationship between Dirichlet Processes and (symmetric) Gamma Processes.

2.1.4. Discussion of the assumptions on G_Σ

Notice that the often used inverse Wishart distribution for G_Σ does not satisfy equation (5). However we can weaken equation (5) by using the same refinement as in Canale and De Blasi (2017); Naulet and Barat (2015) and thus obtain the same rates for the inverse-Wishart prior by using the *square-root technique* from Lijoi, Prünster and Walker (2005). The approach to rates used here is standard and involves two parts, showing the prior puts enough probability mass on certain Kullback-Leibler neighborhoods, and showing the existence of sequence of sets $\mathcal{F}_n$ capturing the essential of the prior mass and having metric entropy not growing too fast as $n \to \infty$. Equations (5) and (6) are only involved in the construction of $\mathcal{F}_n$, while equation (7) occurs in the proof the Kullback-Leibler condition, which is the essential part to understand the impact of tail conditions. The current article focus on the approximation theory needed to prove the Kullback-Leibler condition, thus we voluntary use stronger assumptions than needed to construct $\mathcal{F}_n$, to not complicate the proofs unnecessarily. However, this won't change the bounds on the rate found in this paper.

A typical example of probability distribution satisfying equations (5) to (7) is the inverse-Gaussian distribution when $d = 1$. For arbitrary d, Barndorff-Nielsen et al. (1982) propose an interesting generalization of the inverse-Gaussian, whose density is given by

$$g_\Sigma(\Sigma; \lambda, A, B) := \frac{(\det \Sigma)^{\lambda - \frac{d+1}{2}}}{H(\lambda, A, B)} \exp\left\{-\frac{1}{2}\left(\operatorname{tr}(A\Sigma) + \operatorname{tr}(B\Sigma^{-1})\right)\right\} \mathbb{1}_{\mathcal{E}}(\Sigma), \quad (12)$$

where $(\lambda, A, B) \in \mathbb{R} \times \mathcal{E} \times \mathcal{E}$ and $H(\lambda, A, B)$ is a normalizing constant that can be expressed in term of a matrix Bessel function of the second kind. Then, we have the following proposition.

Proposition 2. *The distribution G_Σ whose the density is given in equation (12) satisfies equations (5) to (7) with $\kappa = 2$.*

Proof. We first prove that G_Σ satisfies equation (6). Let $\nu_1 \geq \max(-2\lambda, d-1)$. From the definition of g_Σ, we have that

$$\begin{aligned}
&G_\Sigma(\Sigma : \lambda_d(\Sigma) < 1/x) \\
&\quad = \int_{\mathcal{E}} \mathbb{1}(\lambda_d(\Sigma) < 1/x)\, g_\Sigma(\Sigma) d\Sigma \\
&\quad \lesssim \int_{\mathcal{E}} \det(\Sigma)^{\lambda + \frac{\nu_1}{2}} e^{-\frac{1}{2}\operatorname{tr}(A\Sigma)} \det(\Sigma)^{-\frac{\nu_1+d+1}{2}} e^{-\frac{1}{2}\operatorname{tr}(B\Sigma^{-1})}\, \mathbb{1}(\lambda_d(\Sigma) < 1/x)\, d\Sigma \\
&\quad \lesssim \int_{\mathcal{E}} \det(\Sigma)^{-\frac{\nu_1+d+1}{2}} e^{-\frac{1}{2}\operatorname{tr}(B\Sigma^{-1})}\, \mathbb{1}(\lambda_d(\Sigma) < 1/x)\, d\Sigma.
\end{aligned}$$

Then, G_Σ satisfies the same bound (up to a constant) as the inverse-Wishart distribution with ν_1 degrees of freedom and scale matrix B. Thus equation (6) follows from (Shen, Tokdar and Ghosal, 2013, Lemma 1).

Similarly, with $\nu_2 \geq \max(2\lambda, d-1)$, we find that

$$G_\Sigma(\Sigma : \lambda_1(\Sigma) > x) \lesssim \int_{\mathcal{E}} \det(\Sigma)^{\frac{\nu_2-d-1}{2}} e^{-\frac{1}{2}\mathrm{tr}(A\Sigma)} \mathbb{1}(\lambda_1(\Sigma) > x)\, d\Sigma.$$

Then, G_Σ satisfies the same bound (up to constant) as the Wishart distribution with ν_2 degrees of freedom and scale matrix A^{-1}. Thus equation (5) follows from a straightforward argument using the relationship between Wishart and inverse-Wishart distribution.

It remains to prove equation (7), but this follows from (Shen, Tokdar and Ghosal, 2013, Lemma 1) using the fact that G_Σ behave like an inverse-Wishart distribution for small $\lambda_1(\Sigma)$. □

As mentioned in Shen, Tokdar and Ghosal (2013), the choice of G_Σ is crucial because the value of κ influences the bounds on the posterior rates of contraction. The smaller κ is, the better the bounds are. The example of equation (12) satisfies $\kappa = 2$. It is possible to achieve $\kappa = 1$ using a prior on diagonal matrices $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, where $\sigma_1, \ldots, \sigma_d$ are independent inverse-Gaussian random variables.

2.2. Posterior concentration rates under mixtures priors

We let $\Pi(\cdot \mid \mathbf{y}^n, \mathbf{x}^n)$ denote the posterior distribution of $f \sim \Pi$ based on n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ modelled as in Section 1. Let $(\epsilon_n)_{n\geq 1}$ be a sequence of positive numbers with $\lim_n \epsilon_n = 0$, and d_n denote the empirical L^2 distance, that is $d_n(f,g)^2 = n^{-1}\sum_{i=1}^n |f(X_i) - g(X_i)|^2$. The following theorem is proved in Section 3.

Theorem 1. *Consider the model (3), and assume that $f_0 \in L^1 \cap \mathtt{C}^\beta$ and $Q_0\|X\|^p < +\infty$. Then there exist constants $C > 0$ and $t > 0$ depending only on f_0 and Q_0 such that*

- *If the prior Π is a symmetric Gamma location mixture of normals as defined in Section 2.1.1*

$$\Pi\left(d_n(f,f_0)^2 > Cn^{-\frac{2\beta}{2\beta+\max(\kappa,\beta+d)}}(\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1)$$

when $0 < p \leq 2d$, and

$$\Pi\left(d_n(f,f_0)^2 > Cn^{-\frac{2\beta}{2\beta+\max(\kappa,d+2\beta d/p)}}(\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1)$$

when $p > 2d$.

- *If the prior Π is a symmetric Gamma location-scale mixture of normals defined in Section 2.1.2*

$$\Pi\left(d_n(f,f_0)^2 > Cn^{-\frac{2\beta}{2\beta+\min(\beta+d,2\beta d/p)+\kappa}}(\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1)$$

when $0 < p \leq 2\beta$, *and*

$$\Pi\left(d_n(f,f_0)^2 > Cn^{-\frac{2\beta}{2\beta+d+\kappa}}(\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1),$$

when $p > 2\beta$.

- *If the prior* Π *is a hybrid symmetric Gamma location-scale mixture of normals defined in Section 2.1.3*

$$\Pi\left(d_n(f,f_0)^2 > Cn^{-\frac{2\beta}{2\beta+\max(\kappa^*,\min(\beta+d,2\beta d/p))}}(\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1),$$

when $0 < p \leq 2\beta$, *and*

$$\Pi\left(d_n(f,f_0)^2 > Cn^{-\frac{2\beta}{2\beta+\max(\kappa^*,d)}}(\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1),$$

when $p > 2\beta$.

The upper bounds on the rates in the previous paragraph are no longer valid when $p = 0$. Indeed the constant $C > 0$ depends on p and might not be definite if $p = 0$; the reason is to be found in the fact that C heavily depends on the ability of the prior to draw mixture component in regions of observed data, which remains concentrated near the origin when $p > 0$. In Section 2.3, we overcome this issue by making the prior covariate dependent; this allows to derive rates under the assumption $p = 0$ (no tail assumption).

2.3. *Relaxing the tail assumption: covariate dependent prior for location mixtures*

Although the rates derived in Section 2.2 do not depend on $p > 0$ when p is small, the assumption $Q_0\|X\|^p < +\infty$ is crucial in proving the Kullback-Leibler condition. Indeed, this condition ensures that the covariates belong to a set $\mathcal{X}_n$ which is not too large, which allows us to bound from below the prior mass of Kullback-Leibler neighbourhoods of the true distribution. Surprisingly, it seems very difficult to get rid of this assumption without and covariate dependent prior, while making the prior covariates dependent allows to drop all tail conditions on Q_0. Doing so, we can adapt to the tail behaviour of Q_0, as shown in the following theorem, which is an adaptation of the general theorems of Ghosal and van der Vaart (2007b). For convenience, in the sequel we drop out the superscript n and we write $\mathbf{x}$, $\mathbf{y}$ for $\mathbf{x}^n$, $\mathbf{y}^n$, respectively. For $\epsilon > 0$ and any subset A of a metric space equipped with metric d, we let $N(\epsilon, A, d)$ denote the ϵ-covering number of A, *i.e.* $N(\epsilon, A, d)$ is the smallest number of balls of radius ϵ needed to cover A.

Theorem 2. *Let* $\Pi_{\mathbf{x}}$ *be a prior distribution that depends on the covariate vector* $\mathbf{x}$, $0 < c_2 < 1/4$ *and* $\epsilon_n \to 0$ *with* $n\epsilon_n^2 \to \infty$. *Suppose that* $\mathcal{F}_n \subseteq \mathcal{F}$ *is such that* $Q_0^n \Pi_{\mathbf{x}}(\mathcal{F}_n^c) \lesssim \exp(-\frac{1}{2}(1+2c_2)n\epsilon_n^2)$ *and* $\log N(\epsilon_n/18, \mathcal{F}_n, d_n) \leq n\epsilon_n^2/4$ *for* n *large enough. If there exists* $M_0 > 0$ *such that*

$$Q_0^n\left(\Pi_{\mathbf{x}}(f \ : \ d_n(f,f_0) \leq \epsilon_n) \leq M_0 \exp(-c_2 n\epsilon_n^2)\right) = o(1),$$

then there exists $M > 0$ *such that* $\Pi_{\mathbf{x}}(f \ : \ d_n(f,f_0) > M\epsilon_n \mid \mathbf{y}, \mathbf{x}) = o_p(1)$.

The proof of the previous theorem is to be found in Section 5. We apply Theorem 2 to symmetric Gamma process location mixture of normals in the following way. Let $\mathbb{Q}_{\mathbf{x}}^n = n^{-1}\sum_{i=1}^n \delta_{x_i}$ denote the empirical measure of the covariate vector $\mathbf{x}$. Given a probability density function g, we let $G_{\mathbf{x}}$ the probability measure which density is $z \mapsto \int g(z-x)\, d\mathbb{Q}_{\mathbf{x}}^n(x)$.

Corollary 1. *Consider the model (3) and assume that $f_0 \in L^1 \cap \mathtt{C}^\beta$. Let $\Pi_{\mathbf{x}}$ be the distribution of the random function $f(x) := \int \varphi_\Sigma(x-\mu)\, dM(\mu)$, where $\Sigma \sim G_\Sigma$ and $M \sim \Pi_\alpha$ with $\alpha = \overline{\alpha} G_{\mathbf{x}}$ for some $\overline{\alpha} > 0$. Assume that G_Σ satisfies equations (5) to (7) and g is continuous at zero with $g(0) > 0$. Then there exists $t > 0$ such that $\Pi_{\mathbf{x}}(f \ : \ d_n(f, f_0) > M\epsilon_n \mid \mathbf{y}, \mathbf{x}) = o_p(1)$ with $\epsilon_n^2 \lesssim n^{-\frac{2\beta}{2\beta+\max(\kappa,\beta+d)}} (\log n)^t$.*

The proof of Corollary 1 is given in Appendix B. Obviously, Theorem 2 can also be applied to symmetric Gamma process location-scale and hybrid mixtures following the same path as in Corollary 1, giving rates $n^{-\frac{2\beta}{3\beta+d+\kappa}}(\log n)^t$ for location-scale mixtures and $n^{-\frac{2\beta}{2\beta+\max(\beta+d,\kappa^*)}}(\log n)^t$ for hybrid mixtures.

3. Proofs

To prove Theorem 1 we follow the lines of Ghosal, Ghosh and van der Vaart (2000); Ghosal and van der Vaart (2001, 2007a). Namely we need to verify the following three conditions

- Kullback-Leibler condition : For a constant $0 < c_2 < 1/4$,

$$\Pi(\mathtt{KL}(f_0, \epsilon_n)) \geq e^{-c_2 n\epsilon_n^2}, \tag{13}$$

  where

$$\mathtt{KL}(f_0, \epsilon_n) := \left\{ f \ : \ \frac{1}{2}\int |f_0(x) - f(x)|^2\, dQ_0(x) \leq \epsilon_n^2 \right\}.$$

- Sieve condition : There exists $\mathcal{F}_n \subset \mathcal{F}$ such that

$$\Pi(\mathcal{F}_n^c) \leq e^{-\frac{1}{2}(1+2c_2)n\epsilon_n^2} \tag{14}$$

- Tests : Let $\log N(\epsilon_n/18, \mathcal{F}_n, d_n)$ be the logarithm of the covering number of $\mathcal{F}_n$ with radius $\epsilon_n/18$ in the $d_n(\cdot,\cdot)$ metric.

$$\log N(\epsilon_n/18, \mathcal{F}_n, d_n) \leq \frac{n\epsilon_n^2}{4}. \tag{15}$$

The Kullback-Leibler condition is proved by defining an approximation of f by a discrete mixture under weak tail conditions. Although the general idea is close to Kruijer, Rousseau and van der Vaart (2010) or Scricciolo (2014), the construction remains quite different to be able to handle various tail behaviours. This is detailed in the following section.

3.1. More notations

Here we define a few more notations that are used in the proof, but were not necessary to state the main theorems of the paper.

- For $1 \leq p < \infty$ we let L^p be the space of function for which the norm $\|f\|_p^p := \int |f(x)|^p \, dx$ is finite; and by L^∞ we mean the space of functions for which $\|f\|_\infty := \sup_{x \in \mathbb{R}^d} |f(x)|$ is finite. For $0 \leq p, q \leq \infty$ and functions $f \in L^p$, $g \in L^q$, we write $f * g$ the convolution of f and g, that is $f * g(x) := \int f(x-y) g(y) \, dy$ for all $x \in \mathbb{R}^d$. Moreover, we'll use repeatedly Young's inequality which state that $\|f * g\|_r \leq \|f\|_p \|g\|_q$, with $1/p + 1/q = 1/r + 1$.
- If $f \in L^1$, then we define $\widehat{f}$ as the (L^1) Fourier transform of f; that is $\widehat{f}(\xi) := \int f(x) e^{-i\xi x} \, dx$ for all $\xi \in \mathbb{R}^d$. Moreover, if $\widehat{f} \in L^1$, then the inverse Fourier transform is well-defined and $f(x) = (2\pi)^{-d} \int \widehat{f}(\xi) e^{ix\xi} \, d\xi$. Also, we denote by $\mathcal{S}$ the Schwartz space; that is the space of infinitely differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ for which $\sup_{x \in \mathbb{R}^d} |x^k D^l f(x)| < +\infty$ for all $k, l \in \mathbb{N}^d$. Then $\mathcal{S} \subset L^1$, and it is well known that the Fourier transform maps $\mathcal{S}$ onto itself, thus the Fourier transform is always invertible on $\mathcal{S}$.

3.2. Approximation theory

To describe the approximation of f_0 by a finite mixture, we first define a few notations. Let $\widehat{\ell} : \mathbb{R} \to \mathbb{R}$ be a symmetric $\mathtt{C}^\infty$ function that equals 1 on $[-1, 1]$ and 0 outside $\mathbb{R} \backslash [-2, 2]$; the existence of such function is classical. Define $\widehat{\chi} : \mathbb{R}^d \to \mathbb{R}$ such that $\widehat{\chi}(\xi) = \prod_{i=1}^d \widehat{\ell}(\xi_i)$ for all $\xi \in \mathbb{R}^d$. For any $\sigma > 0$ we use the shortened notation $\widehat{\chi}_\sigma(\xi) := \widehat{\chi}(2\sigma\xi)$, and χ_σ will stand for the inverse Fourier transform of $\widehat{\chi}_\sigma$. Define η as the function which L^1 Fourier transform satisfies $\widehat{\eta}(\xi) = \widehat{\chi}(\xi)/\widehat{\varphi}(\xi)$ for all $\xi \in [-2, 2]^d$ and $\widehat{\eta}(\xi) = 0$ elsewhere. For two positive real numbers h and σ, we define the kernel $K_{h,\sigma} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that

$$K_{h,\sigma}(x, y) := \frac{h^d}{\sigma^d} \sum_{k \in \mathbb{Z}^d} \varphi\left(\frac{x - h\sigma k}{\sigma}\right) \eta\left(\frac{y - h\sigma k}{\sigma}\right), \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d.$$

For a measurable function f we introduce the operator associated with the kernel : $K_{h,\sigma} f(x) = \int K_{h,\sigma}(x, y) f(y) \, dy$. The function $K_{h,\sigma} f$ will play the role of an *approximation* for the function f, and we will evaluate how this approximation becomes close to f given h and σ sufficiently close to zero.

More precisely, we will prove that, when choosing h appropriately, f_0 can be approximated by $K_{h,\sigma}(\chi_\sigma * f_0)$ to the order σ^β. Moreover $K_{h,\sigma}(\chi_\sigma * f_0)$ can be written as $\sum_{k \in \mathbb{Z}^d} u_k \varphi_{\sigma^2 I}(x - \mu_k)$, where $u_k = (h^d/\sigma^d) \int \eta(y/\sigma - hk) \chi_\sigma * f_0(y) dy$. Note that by symmetry of $\widehat{\ell}$, the coefficients u_k are always real valued when f_0 takes value in $\mathbb{R}$. In a second step we approximate $K_{h,\sigma}(\chi_\sigma * f_0)$ by a truncated version of it, retaining only the k's such that $|u_k|$ is large enough and $\|\mu_k\|$ not too large. In the case of location-scale and hybrid location-scale mixture we consider a modification of this approximation to better control the number of

components for which σ needs to be small. We believe that these constructions have interest in themselves. In particular they shed light on the relations between mixtures of normals and wavelet approximations.

These approximation properties are presented in the following two Lemmas which are proved in Appendix A:

Lemma 1. *There is $C > 0$ depending only on β such that for any $f_0 \in L^1 \cap \mathtt{C}^\beta$ and any $\sigma > 0$ we have $|\chi_\sigma * f_0(x) - f_0(x)| \leq C\|f\|_{\mathtt{C}^\beta}\sigma^\beta$ for all $x \in \mathbb{R}^d$.*

Lemma 2. *Let $f_\sigma := \chi_\sigma * f_0$ and $h \leq 1$. Then there are universal constants $C, c > 0$ such that $|K_{h,\sigma}f_\sigma(x) - f_\sigma(x)| \leq C\|f_0\|_1\sigma^{-d}\exp(-ch^{-2})$ for all $x \in \mathbb{R}^d$.*

We now present the approximation schemes in the context of location mixtures.

3.3. Construction of the approximation under location mixtures

Let $0 < \sigma \leq 1$ and $c_0 > 0$ be a constant. Define $h_\sigma > 0$ such that $h_\sigma\sqrt{\log\sigma^{-1}} := c_0$. Then combining the results of Lemma 1 and Lemma 2 we can conclude that if c_0 is chosen small enough, then $|K_{h_\sigma,\sigma}(\chi_\sigma * f_0)(x) - f_0(x)| \lesssim \sigma^\beta$. Now we define the coefficients u_k, $k \in \mathbb{Z}^d$ so that

$$K_{h_\sigma,\sigma}(\chi_\sigma * f_0)(x) =: \sum_{k\in\mathbb{Z}^d} u_k\,\varphi_{\sigma^2 I}(x - \mu_k), \quad \forall k \in \mathbb{Z}^d,$$

where $\mu_k := h_\sigma\sigma k$ for all $k \in \mathbb{Z}^d$. Let define

$$\Lambda_\sigma := \Big\{k \in \mathbb{Z}^d \,:\, |u_k| > \sigma^\beta, \quad \|\mu_k\| \leq \sigma^{-2\beta/p} + \sigma\sqrt{2(\beta+d)\log\sigma^{-1}}\Big\},$$

$U_\sigma := \{\Sigma \in \mathcal{E} \,:\, \sigma^{-2} \leq \lambda_i(\Sigma^{-1}) \leq \sigma^{-2}(1+\sigma^{\beta+d})\ i = 1,\dots,d\}$, and for all $k \in \Lambda_\sigma$ let define $V_k := \{\mu \in \mathbb{R}^d \,:\, \|\mu - \mu_k\| \leq \sigma^{\beta+1}\}$ and $V = \cup_{k\in\Lambda_\sigma}V_k$. We also denote

$$\mathcal{M}_\sigma := \left\{M \text{ signed measure on } \mathbb{R}^d \,:\, \begin{array}{l}\forall k \in \Lambda_\sigma : |M(V_k) - u_k| \leq \sigma^\beta, \\ |M|(V^c) \leq \sigma^\beta\end{array}\right\},$$

and for any $M \in \mathcal{M}_\sigma$ and $\Sigma \in U_\sigma$, we write $f_{M,\Sigma}(x) := \int \varphi_\Sigma(x-\mu)\,dM(\mu)$.

Proposition 3. *For $\sigma > 0$ small enough, it holds*

$$|\Lambda_\sigma| \lesssim \min(\sigma^{-(\beta+d)}, h_\sigma^{-d}\sigma^{-(2\beta/p+1)d}).$$

Proof. Because there is a separation of $h_\sigma\sigma$ between two consecutive μ_k, it is clear that $|\Lambda_\sigma| \lesssim h_\sigma^{-d}\sigma^{-(2\beta/p+1)d}$ when σ is small enough. Moreover, from Proposition 9 we have the following estimate.

$$\|f_0\|_1\sigma^{-d} \gtrsim \sum_{k\in\mathbb{Z}^d}|u_k| \geq \sum_{k\in\Lambda_\sigma}|u_k| \geq \sigma^\beta|\Lambda_\sigma|. \qquad \square$$

Proposition 4. *For all $x \in \mathbb{R}^d$, all $\sigma > 0$ small enough, all $\Sigma \in U_\sigma$ and all $M \in \mathcal{M}_\sigma$ it holds $|f_{M,\Sigma}(x) - f_0(x)| \lesssim 1$.*

Proof. We have that $|f_{M,\Sigma}(x) - f_0(x)| \leq |f_{M,\Sigma}(x)| + \|f_0\|_\infty$. But, with $\mathcal{I} \equiv \mathcal{I}(x) := \{k \in \mathbb{Z}^d : \|x - \mu_k\| \leq 2\sigma\}$, we can write

$$f_{M,\Sigma}(x) = \sum_{k \in \Lambda_\sigma \cap \mathcal{I}} \int_{V_k} \varphi_\Sigma(x-\mu)\, dM(\mu) + \sum_{k \in \Lambda_\sigma \cap \mathcal{I}^c} \int_{V_k} \varphi_{\sigma^2 I}(x-\mu)\, dM(\mu)$$
$$+ \sum_{k \in \Lambda_\sigma \cap \mathcal{I}^c} \int_{V_k} [\varphi_\Sigma(x-\mu) - \varphi_{\sigma^2 I}(x-\mu)]\, dM(\mu) + \int_{V^c} \varphi_\Sigma(x-\mu)\, dM(\mu). \tag{16}$$

Clearly the last term of this last expression is bounded above by $\|\varphi\|_\infty \sigma^\beta$. For the second term, we have for any $\mu \in V_k$ with $k \in \mathcal{I}^c$ that $\|x - \mu\| \geq \|x - \mu_k\| - \|\mu - \mu_k\| \geq \|x - \mu_k\|/2$. Then the second term of the rhs of equation (16) is bounded above by

$$\sup_{k \in \Lambda_\sigma \cap \mathcal{I}^c} |M|(V_k) \sum_{k \in \mathbb{Z}^d} \varphi_{\sigma^2 I}(\|x - h_\sigma \sigma k\|/2).$$

Proceeding as in the proof of Lemma 8, we deduce that the series in the last expression is bounded above by a constant times h_σ^{-d}, whereas $|M|(V_k) \leq |M(V_k) - u_k| + |u_k| \lesssim \sigma^\beta + h_\sigma^d \|f_0\|_\infty$ by Proposition 9. Therefore the second term of the rhs in equation (16) is bounded by a constant when σ is small enough. Regarding the first term in equation (16), it is bounded by $\|\varphi\|_\infty |\mathcal{I}| \sup_{k \in \Lambda_\sigma} |M|(V_k)$, which is in turn bounded by a constant by the same argument as previously. By Propositions 9 and 11, the remaining term is bounded by

$$\sup_{\Sigma \in U_\sigma} \|I - \sigma^2 \Sigma^{-1}\| \sum_{k \in \Lambda_\sigma \cap \mathcal{I}^c} |M|(V_k)$$
$$\lesssim \sigma^{\beta+d} \sum_{k \in \Lambda_\sigma \cap \mathcal{I}^c} |M(V_k) - u_k| + \sigma^{\beta+d} \sum_{k \in \Lambda_\sigma \cap \mathcal{I}^c} |u_k|$$
$$\lesssim \sigma^{2\beta+d} |\Lambda_\sigma| + \sigma^\beta,$$

which is in turn bounded by a multiple constant of σ^β by Proposition 3. □

Proposition 5. *For all $\sigma > 0$ small enough, all $x \in \mathbb{R}^d$ with $\|x\| \leq \sigma^{-2\beta/p}$, all $\Sigma \in U_\sigma$ and all $M \in \mathcal{M}_\sigma$ it holds $|f_{M,\Sigma}(x) - f_0(x)| \lesssim h_\sigma^{-2d} \sigma^\beta$.*

Proof. We define $A_\sigma(\beta) := \sqrt{2 \log |\Lambda_\sigma| + 2(\beta + d) \log \sigma^{-1}}$. Then for any $M \in \mathcal{M}_\sigma$, letting $\mathcal{J} \equiv \mathcal{J}(x) := \{k \in \mathbb{Z}^d : \|x - \mu_k\| \leq 2\sigma A_\sigma(\beta)\}$, we may write

$$f_{M,\Sigma}(x) - K_{h_\sigma,\sigma}(\chi_\sigma * f_0)(x) = \sum_{k \in \Lambda_\sigma \cap \mathcal{J}} \int_{V_k} [\varphi_\Sigma(x-\mu) - \varphi_{\sigma^2 I}(x-\mu_k)]\, dM(\mu)$$
$$+ \sum_{k \in \Lambda_\sigma \cap \mathcal{J}} [M(V_k) - u_k]\, \varphi_{\sigma^2 I}(x - \mu_k) + \sum_{k \in \Lambda_\sigma \cap \mathcal{J}^c} \int_{V_k} \varphi_\Sigma(x-\mu)\, dM(\mu)$$

$$- \sum_{k\in\Lambda_\sigma\cap\mathcal{J}^c} u_k\,\varphi_{\sigma^2 I}(x-\mu_k) - \sum_{k\in\Lambda_\sigma^c} u_k\,\varphi_{\sigma^2 I}(x-\mu_k) + \int_{V^c}\varphi_\Sigma(x-\mu)\,dM(\mu)$$
$$:= r_1(x) + r_2(x) + r_3(x) + r_4(x) + r_5(x) + r_6(x). \quad (17)$$

With the same argument as in Proposition 3, we deduce that $|\mathcal{J}| \lesssim h_\sigma^{-d} A_\sigma(\beta)^d$. The same proposition implies $A_\sigma(\beta) \lesssim \sqrt{\log\sigma^{-1}}$. From the proof of Proposition 4, we have $|M|(V_k) \lesssim \sigma^\beta + h_\sigma^d \|f_0\|_\infty$ for all $k \in \Lambda_\sigma$ and all $M \in \mathcal{M}_\sigma$, it follows from Proposition 11 that $|r_1(x)| \lesssim A_\sigma(\beta)^d \sigma^{\beta+d}$. From the definition of $\mathcal{M}_\sigma$, it comes $|r_2(x)| \le \|\varphi\|_\infty |\mathcal{J}| \sigma^\beta \lesssim \|\varphi\|_\infty A_\sigma(\beta)^d h_\sigma^{-d}\sigma^\beta$. Using the definition of φ_Σ, whenever $\Sigma \in U_\sigma$ we can write that

$$\begin{aligned}\varphi_\Sigma(x) &= \exp\left\{-\frac{1}{2\sigma^2}\|x\|^2 + \frac{1}{2}x^T[\sigma^{-2}I - \Sigma^{-1}]x\right\}\\ &\le \exp\left\{-\frac{1}{2\sigma^2}\|x\|^2 + \frac{1}{2}\|\Sigma^{-1} - \sigma^2 I\|\|x\|^2\right\}\\ &\le \exp\left\{-\frac{1}{2\sigma^2}(1-\sigma^{\beta+2})\|x\|^2\right\},\end{aligned}$$

where the second line follows from Cauchy-Schwarz inequality, and the last line by the definition of U_σ. Moreover, when $k \in \Lambda_\sigma \cap \mathcal{J}^c$ and $\mu \in V_k$, it holds $\|x-\mu\| \ge \|x-\mu_k\| - \|\mu-\mu_k\| \ge \sigma A_\sigma(\beta)/\sqrt{1-\sigma^{\beta+2}}$ for σ small enough. Therefore, $|r_3(x)| \lesssim \exp(-\frac{1}{2}A_\sigma(\beta)^2)|\Lambda_\sigma| \lesssim \sigma^{\beta+d}$. With the same argument, Proposition 9 and Young's inequality we get $|r_4(x)| \lesssim \|\chi_\sigma * f_0\|_\infty \exp(-2A_\sigma(\beta)^2)|\Lambda_\sigma| \le \|\chi\|_1\|f_0\|_\infty\sigma^\beta$. Regarding r_5, we rewrite $\Lambda_\sigma^c = \Lambda_1^c \cup \Lambda_2^c$, with $\Lambda_1^c := \{k \in \mathbb{Z}^d : |u_k| \le \sigma^\beta\}$ and $\Lambda_2^c := \{k \in \mathbb{Z}^d : \|\mu_k\| > \sigma^{-2\beta/p} + \sigma\sqrt{2(\beta+d)\log\sigma^{-1}}\}$. Then,

$$\begin{aligned}|r_5(x)| &\le \sum_{k\in\Lambda_1^c} |u_k|\,\varphi_{\sigma^2 I}(x-\mu_k) + \sum_{k\in\Lambda_2^c} |u_k|\,\varphi_{\sigma^2 I}(x-\mu_k)\\ &\le \sigma^\beta \sup_{x\in\mathbb{R}^d}\sum_{k\in\mathbb{Z}^d}\varphi_{\sigma^2 I}(x-\mu_k) + \sum_{k\in\Lambda_2^c}|u_k|\,\varphi_{\sigma^2 I}(x-\mu_k). \quad (18)\end{aligned}$$

The first term of the rhs of equation (18) is bounded by a multiple constant of $h_\sigma^{-d}\sigma^\beta$, with the same argument as in the proof of Lemma 8. By definition of Λ_2^c, $\|x-\mu_k\| \ge \sigma\sqrt{2(\beta+d)\log\sigma^{-1}}$ when $k \in \Lambda_2^c$ and $\|x\| \le \sigma^{-2\beta/p}$. This implies, together with Proposition 9 and Young's inequality, that the second term of the rhs of equation (18) is bounded by a constant multiple of $\sigma^{\beta+d}\sum_{k\in\mathbb{Z}^d}|u_k| \lesssim \|\chi_\sigma * f_0\|_1\sigma^\beta \le \|\chi\|_1\|f_0\|_1\sigma^\beta$ for all $\|x\| \le \sigma^{-2\beta/p}$. Finally, we have the trivial bound $|r_6(x)| \le \|\varphi\|_\infty |M|(V^c) \le \|\varphi\|_\infty\sigma^\beta$. □

3.4. *Construction of the approximation under location-scale and hybrid location-scale mixtures*

Let $\sigma_0 := 1$ and define recursively $\sigma_{j+1} := \sigma_j/2$ for any $j \ge 0$. Let $\Delta_0 := f_0 - \chi_{\sigma_0} * f_0$, and define recursively $\Delta_{j+1} := \Delta_j - \chi_{\sigma_{j+1}} * \Delta_j$, for any $j \ge 0$.

The general idea of the construction is that $\sup_{x\in\mathbb{R}^d}|\Delta_j(x)|\lesssim\sigma_j^\beta$, as shown in Proposition 10 in appendix, and that similarly to wavelet decomposition, we approximate a function f_0 Hölder β by

$$f_1 := K_0(\chi_{\sigma_0} * f_0) + \sum_{j=1}^{J} K_j(\chi_{\sigma_j} * \Delta_{j-1}).$$

where $J \geq 1$ is a large enough integer, $h_J := c_0 J^{-1/2}$ for a small enough constant $c_0 > 0$, and $K_j := K_{h_J,\sigma_j}$. By induction, we get that $\Delta_j = \Delta_0 - \sum_{l=0}^{j-1} \chi_{\sigma_{l+1}} * \Delta_l$. It follows,

$$\begin{aligned} f_1 - f_0 &= K_0(\chi_{\sigma_0} * f_0) - f_0 + \sum_{j=1}^{J} K_j(\chi_{\sigma_j} * \Delta_{j-1}) \\ &= \Delta_J + K_0(\chi_{\sigma_0} * f_0) - \chi_{\sigma_0} * f_0 + \sum_{j=1}^{J} \left[K_j(\chi_{\sigma_j} * \Delta_{j-1}) - \chi_{\sigma_j} * \Delta_{j-1}\right]. \end{aligned}$$

Therefore, from Lemma 2, Proposition 10 and Young's inequality, the error of approximating f_0 by f_1 when c_0 is small enough is

$$\begin{aligned} &|f_1(x) - f_0(x)| \\ &\quad\leq |\Delta_J| + |K_0(\chi_{\sigma_0} * f_0) - \chi_{\sigma_0} * f_0| + \sum_{j=1}^{J} |K_j(\chi_{\sigma_j} * \Delta_{j-1}) - \chi_{\sigma_j} * \Delta_{j-1}| \\ &\quad\lesssim \|f_0\|_{\mathcal{C}^\beta}\sigma_J^\beta + \|f_0\|_1\sigma_0^{-d}e^{-ch_J^{-2}} + e^{-ch_J^{-2}}\sum_{j=1}^{J}\|\Delta_{j-1}\|_1\sigma_j^{-d} \\ &\quad\lesssim \|f_0\|_{\mathcal{C}^\beta}\sigma_J^\beta + \|f_0\|_1 e^{-ch_J^{-2}} + \|f_0\|_1 e^{-ch_J^{-2}}\sum_{j=1}^{J} 2^j \\ &\quad\lesssim \|f_0\|_{\mathcal{C}^\beta}\sigma_J^\beta + \|f_0\|_1(1+2^J)e^{-ch_J^{-2}} \lesssim \sigma_J^\beta. \end{aligned}$$

The reason for considering different scale parameters in the construction, is to deal with fat tail, the heuristic being that in the tails we do not require as precise an approximation as in the center. In particular small values of j will be used to estimate the function far off in the tails. To formalize this, we define $A_j := \{x \in \mathbb{R}^d : \|x\| \leq 2^{2\beta(J-j)/p}\}$, for all $j = 0, \dots J$. We also define $I_J := \{x \in \mathbb{R}^d : \|x\| \leq 1\}$, and for all $j = 0, \dots, J-1$ we set $I_j := A_j \backslash A_{j+1}$. Notice that by definition of K_j, we can write,

$$\begin{aligned} K_0(\chi_{\sigma_0} * f_0)(x) &=: \sum_{k\in\mathbb{Z}^d} u_{0k}\,\varphi_{\sigma_0^2 I}(x - h_J\sigma_0 k) \\ K_j(\chi_{\sigma_j} * \Delta_{j-1})(x) &=: \sum_{k\in\mathbb{Z}^d} u_{jk}\,\varphi_{\sigma_j^2 I}(x - h_J\sigma_j k), \quad \forall j \geq 1. \end{aligned}$$

To ease notation, we define $\mu_{jk} := h_J \sigma_j k$ for all $j \geq 0$ and all $k \in \mathbb{Z}^d$. In the sequel we shall need the following subset of indexes,

$$\Lambda_J := \left\{ (j,k) \in \{0,\dots,J\} \times \mathbb{Z}^d \, : \, \begin{array}{l} |u_{jk}| > \sigma_J^\beta, \\ \|\mu_{jk}\| \leq 2^{\frac{2\beta}{p}(J-j)} + \sigma_j \sqrt{2(\beta+d)\log \sigma_J^{-1}} \end{array} \right\}.$$

We prove below that we can approximate f_1 by a finite mixture corresponding to retaining only the components associated to indices in Λ_J and that we can bound the cardinality of Λ_J by $O(J \log J \sigma_J^{-2\beta/p})$

To any $(j,k) \in \Lambda_J$ we associate $U_{jk} := \{\Sigma \in \mathcal{E} \, : \, \sigma_j^{-2} \leq \lambda_i(\Sigma^{-1}) \leq \sigma_j^{-2}(1+\sigma_J^{\beta+d}),\ i = 1,\dots d\}$, $V_{jk} := \{\mu \in \mathbb{R}^d \, : \, \|\mu - \mu_{jk}\| \leq \sigma_j \sigma_J^\beta\}$ and $W_{jk} := U_{jk} \times V_{jk}$. We denote by $\mathcal{M}$ the set of signed measures M on $\mathcal{E} \times \mathbb{R}^d$ such that $|M(W_{jk}) - u_{jk}| \leq \sigma_J^\beta$ for all $(j,k) \in \Lambda_J$, and $|M|(W^c) \leq \sigma_J^\beta$, where W^c is the relative complement of the union of all W_{jk} for $(j,k) \in \Lambda_J$. For any $M \in \mathcal{M}$, we write

$$f_M(x) := \int \varphi_\Sigma(x-\mu)\, dM(\Sigma,\mu).$$

In Proposition 6 we control the cardinality of Λ_J while in Proposition 8 we control the error between f_M and f_1 on the decreasing sequence of balls A_j. Proposition 7 provides a crude uniform upper bound on f_M and f_0.

Proposition 6. *There is a constant $C > 0$ depending only on f_0 and Q_0 such that $|\Lambda_J| \leq C \min[\sigma_J^{-(\beta+d)}, J^{d/2} \sigma_J^{-2d\beta/p}]$ if $p \leq 2\beta$, and $|\Lambda_J| \leq C J^{d/2} \sigma_J^{-d}$ if $p > 2\beta$.*

Proof. First notice that because of Propositions 9 and 10, we always have the bound

$$\|f_0\|_1 \sigma_J^{-d} \gtrsim \|f_0\|_1 \sum_{j=0}^J \sigma_j^{-d} \geq \sum_{j=0}^J \sum_{k \in \mathbb{Z}^d} |u_{jk}| \geq \sum_{(j,k) \in \Lambda_J} |u_{jk}| \geq \sigma_J^\beta |\Lambda_J|. \tag{19}$$

We define $B := \sqrt{2(\beta+d)\log 2}$, so that $\sqrt{2(\beta+d)\log \sigma_J^{-1}} = B\sqrt{J}$. Now consider those indexes j with $2^{2\beta(J-j)/p} \leq \sigma_j B \sqrt{J}$. An elementary computation shows that there are at most a multiple constant of such indexes for J large enough. Therefore, recalling that there is a separation of $h_J \sigma_j$ between two consecutive μ_{jk},

$$\begin{aligned} |\Lambda_J| &\lesssim \sum_{j=0}^J \left(\frac{2^{2\beta(J-j)/p}}{h_J \sigma_j} \right)^d + \sup_j \left(\frac{2\sigma_j B \sqrt{J}}{h_J \sigma_j} \right)^d \\ &\lesssim h_J^{-d} \sigma_J^{-2d\beta/p} \sum_{j=0}^J 2^{-jd(\frac{2\beta}{p} - 1)} + J^d. \end{aligned} \tag{20}$$

Because $h_J \sqrt{J} \lesssim 1$ by definition, and if $p \leq 2\beta$, the result follows from the last equation and equation (19). If $p > 2\beta$, the reasoning is the same as in the first

part, but we can rewrite in this situation the equation (20) as

$$|\Lambda_J| \lesssim h_J^{-d}\sigma_J^{-d}\sum_{j=0}^{J} 2^{d(j-J)(1-\frac{2\beta}{p})} + J^d.$$

Since $p > 2\beta$, the conclusion is immediate. □

Proposition 7. *For all $x \in \mathbb{R}^d$, all $J > 0$ large enough and all $M \in \mathcal{M}$, it holds $|f_M(x) - f_0(x)| \lesssim J$.*

Proof. Let $\mathcal{I} \equiv \mathcal{I}(x) := \{(j,k) \in \{0,\dots,J\} \times \mathbb{Z}^d : \|x - \mu_{jk}\| \leq \sigma_j\}$. Then the proof is almost identical to Proposition 4. It suffices to notice that for J large enough:

- $|M|(W_{jk}) \leq |M(W_{jk}) - u_{jk}| + |u_{jk}|$ is always bounded above by a multiple constant of h_J^{-d}, because of the definition of $\mathcal{M}$, of Propositions 9 and 10.
- $\|x - \mu\| > (1 - \sigma_J^\beta)\|x - \mu_{jk}\|$ whenever $(\Sigma, \mu) \in W_{jk}$ and $(j,k) \in \Lambda_J \cap \mathcal{I}^c$.
- $|\mathcal{I}| \lesssim Jh_J^{-1}$. □

Proposition 8. *If $f_0 \in \mathtt{C}^\beta$, for all $J > 0$ large enough, all $0 \leq j \leq J$, all $x \in A_j$ and all $M \in \mathcal{M}$, it holds $|f_M(x) - f_0(x)| \lesssim J^{1+d}\sigma_j^\beta$.*

The proof of Proposition 8 is given in Appendix C.

4. Proof of Theorem 1

As mentioned earlier, the proof of Theorem 1 boils down to verifying conditions (13), (14) and (15) for the three types of priors.

4.1. Location mixtures

4.1.1. Kullback-Leibler condition for location mixtures

In this Section we verify condition (13) in the case of a location mixture prior, using the results of Section 3.3. We use the notations Λ_σ, U_σ, $\mathcal{M}_\sigma$ and $f_{M,\Sigma}$ defined in Section 3.3.

By Chebychev inequality, we have

$$Q_0\left(\|X\| > \sigma^{2\beta/p}\right) \leq \sigma^{2\beta}\int_{\mathbb{R}^d} \|x\|^p dQ_0(x).$$

Then by bringing together results from Propositions 4 and 5, we can find a constant $C > 0$ such that for all $M \in \mathcal{M}_\sigma$ and all $\Sigma \in U_\sigma$

$$\int |f_{M,\Sigma}(x) - f_0(x)|^2\, dQ_0(x) \leq \sup_{\|x\| \leq \sigma^{-2\beta/p}} |f_{M,\Sigma}(x) - f_0(x)|^2 + \sup_{\|x\| > \sigma^{-2\beta/p}} |f_{M,\Sigma}(x) - f_0(x)|^2\, Q_0\left(\|X\| > \sigma^{2\beta/p}\right) \leq C\sigma^{2\beta}(\log \sigma^{-1})^{2d}.$$

By equation (7), we have $G_\Sigma(U_\sigma) \gtrsim \sigma^{-2b_3}\sigma^{b_4(\beta+d)}\exp(-a_3\sigma^{-\kappa})$. Moreover, there is a separation of at least $h_\sigma\sigma$ between two consecutive μ_k and $h_\sigma\sigma \ll \sigma$, thus all the V_k with $k \in \Lambda_\sigma$ are disjoint. By assumptions on G_μ (see equation (8)), $\alpha_k := \overline{\alpha}G_\mu(V_k) \gtrsim \sigma^{b_5(\beta+1)}(1+\|\mu_k\|)^{-b_6}$ for all $k \in \Lambda_\sigma$. We also define $\alpha^c := \alpha(V^c)$. For σ small enough, there is a constant $C' > 0$ not depending on σ such that $\alpha^c > C'$. Moreover, since α has finite variation we can assume without loss of generality that $C' \leq \alpha^c \leq 1$, otherwise we split V^c into disjoint parts, each of them having α-measure smaller than one. With $\epsilon_n^2 := C\sigma^{2\beta}(\log\sigma^{-1})^{2d}$, using that $\Gamma(\alpha) \leq 2\alpha^{\alpha-1}$ for $\alpha \leq 1$, it follows from Proposition 12 the lower bound

$$
\begin{aligned}
\Pi(\mathtt{KL}(f_0,\epsilon_n)) &\geq G_\Sigma(U_\sigma)\Pi_\alpha(\mathcal{M}_\sigma) \\
&\gtrsim \sigma^{-2b_3+b_4(\beta+d)}e^{-a_3\sigma^{-\kappa}}\frac{\sigma^\beta}{3e\Gamma(\alpha^c)}\prod_{k\in\Lambda_\sigma}\left(\frac{\sigma^\beta e^{-2|u_k|}}{3e\Gamma(\alpha_k)}\right) \\
&\gtrsim \exp\left\{-K|\Lambda_\sigma|\log\sigma^{-1} - a_3\sigma^{-\kappa} - 2\sum_{k\in\Lambda_\sigma}|u_k| - \sum_{k\in\Lambda_\sigma}\log\frac{1}{\alpha_k}\right\} \\
&\gtrsim \exp\left\{-K|\Lambda_\sigma|\log\sigma^{-1} - K\sigma^{-\max(\kappa,d)} - \sum_{k\in\Lambda_\sigma}\log\frac{1}{\alpha_k}\right\},
\end{aligned}
$$

for a generic constant $K > 0$. From the definition of α_k, it holds

$$
\sum_{k\in\Lambda_\sigma}\log\frac{1}{\alpha_k} \lesssim |\Lambda_\sigma|\log\sigma^{-1} + \sum_{k\in\Lambda_\sigma}\log\left(1+\|\mu_k\|\right),
$$

when σ is small enough. Also,

$$
\begin{aligned}
&\sum_{k\in\Lambda_\sigma}\log\left(1+\|\mu_k\|\right) \\
&\quad= \sum_{k\in\Lambda_\sigma}\log\left(1+\|\mu_k\|\right)\mathbb{1}\{\|\mu_k\| \leq 1\} + \sum_{k\in\Lambda_\sigma}\log\left(1+\|\mu_k\|\right)\mathbb{1}\{\|\mu_k\| > 1\} \\
&\quad\leq |\{k\in\Lambda_\sigma \,:\, \|\mu_k\| \leq 1\}| + |\Lambda_\sigma|\log 2 + \sum_{k\in\Lambda_\sigma}\log\|\mu_k\| \\
&\quad\lesssim h_\sigma^{-d}\sigma^{-d} + |\Lambda_\sigma|\frac{2\beta}{p}\log\sigma^{-1} \lesssim |\Lambda_\sigma|\log\sigma^{-1} + h_\sigma^{-d}\sigma^{-d}.
\end{aligned}
$$

Because $|\Lambda_\sigma| > \sigma^{-d}$ for σ small enough (see Proposition 3), it follows from all of the above the existence of a constant $K' > 0$, depending only on f, φ and Π, such that

$$
\begin{aligned}
\Pi(\mathtt{KL}(f_0,\epsilon_n)) &\geq \exp\left\{-K'\max(\sigma^{-\max(\kappa,d)}, |\Lambda_\sigma|\log\sigma^{-1})\right\} \\
&\geq \exp\left\{-K'\max(\sigma^{-\kappa}, |\Lambda_\sigma|\log\sigma^{-1})\right\}.
\end{aligned}
$$

Then for appropriate constants $C''', t > 0$, as a consequence of Proposition 3, we can have $\Pi(\mathtt{KL}(f_0, \epsilon_n)) \geq e^{-c_2 n\epsilon_n^2}$ if

$$\epsilon_n^2 = \begin{cases} C''' n^{-\frac{2\beta}{2\beta+\max(\kappa,\beta+d)}} (\log n)^t & 0 < p \leq 2d, \\ C''' n^{-\frac{2\beta}{2\beta+\max(\kappa,d+2d\beta/p)}} (\log n)^t & p > 2d. \end{cases}$$

4.1.2. Sieve construction for location mixtures

We construct the following sequence of subsets of $\mathcal{F}$, also called *a sieve*. With the notation $f_{M,\Sigma}(x) := \int \varphi_\Sigma(x-\mu)\, dM(\mu)$,

$$\mathcal{F}_n(H,\epsilon) := \left\{ f = f_{M,\Sigma} \,:\, \begin{array}{l} M = \sum_{i=1}^\infty u_i \delta_{\mu_i},\ \forall j:\ n^{-1/b_2} \leq \lambda_j(\Sigma) \leq n^{1/b_1} \\ \sum_{i=1}^\infty |u_i| \leq n, \quad \sum_{i=1}^\infty |u_i| \mathbb{1}\{|u_i| \leq n^{-1}\} \leq \epsilon \\ |\{i\,:\, |u_i| > n^{-1}\}| \leq H n\epsilon^2 / \log n \end{array} \right\}.$$

The next two lemmas show that $\mathcal{F}_n(H,\epsilon)$ defined as above satisfies all the condition stated in equations (14) and (15) if H and δ are chosen small enough.

Lemma 3. *Let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^{d\times n}$ be arbitrary and d_n be the empirical L^2-distance associated with $\mathbf{x}$. Then for any $n^{-1/2} < \epsilon_n \leq 1$, $0 < H \leq 1$ and n sufficiently large there is a constant $C > 0$ not depending on n such that $\log N(\epsilon_n, \mathcal{F}_n(H,\epsilon_n), d_n) \leq CHn\epsilon_n^2$.*

Proof. We write $\mathcal{F}_n \equiv \mathcal{F}_n(H,\epsilon_n)$ to ease notations. The proof is based on arguments from Shen, Tokdar and Ghosal (2013), it uses the fact that the covering number $N(\epsilon_n, \mathcal{F}_n, d_n)$ is the minimal cardinality of an ϵ_n-net over $(\mathcal{F}_n, d_n)$. We recall that $(\mathcal{F}_n, d_n)$ has ϵ_n-net $\mathcal{F}_{n,\epsilon}$, if for any $f \in \mathcal{F}_n$ we have $m \in \mathcal{F}_{n,\epsilon}$ such that $d_n(f,m) < \epsilon_n$. Let $S_n := \bigcup_{i=1}^n \{x \in \mathbb{R}^d \,:\, \|x - x_i\|^2 \leq 4n^{1/b_1} \log n\}$, $R_n := \{\mu \in \mathbb{R}^d \,:\, \mu = kn^{-(3/2+b_2^{-1})} d^{-1/2},\ k \in \mathbb{Z}^d,\ \mu \in S_n\}$, $\mathcal{O}_n$ be a $n^{-(1+2b_1^{-1}+2b_2^{-1})}\epsilon_n$-net of the group of $d \times d$ orthogonal matrices equipped with spectral norm $\|\cdot\|$, and,

$$Q_n := \left\{ \Sigma \in \mathcal{E} \,:\, \begin{array}{l} \Sigma = P^T D P,\ P \in \mathcal{O}_n,\ D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \\ \forall j:\ \lambda_j = n^{-1/b_2}(1 + n^{-(1+b_1^{-1}+b_2^{-1})}\epsilon_n)^k,\ k \in \mathbb{N} \\ \forall j:\ n^{-1/b_2} \leq \lambda_j \leq n^{1/b_1}. \end{array} \right\}.$$

Then we define the following finite subset of $\mathcal{F}_n(H,\epsilon)$.

$$\mathcal{F}_{n,\epsilon} := \left\{ f = f_{M,\Sigma} \,:\, \begin{array}{l} M = \sum_{i\in\mathcal{I}} u_i \delta_{\mu_i},\ |\mathcal{I}| \leq Hn\epsilon_n^2/\log n, \\ \forall i \in \mathcal{I}:\ \mu_i \in R_n,\ \Sigma \in Q_n,\ \sum_{i\in\mathcal{I}} |u_i| \leq n, \\ \forall i \in \mathcal{I}:\ u_i = kn^{-3/2} H^{-1},\ k \in \mathbb{Z} \end{array} \right\}.$$

We claim that there is a constant $\delta > 0$ such that $\mathcal{F}_{n,\epsilon}$ is a $\delta\epsilon_n$-net over $(\mathcal{F}_n, d_n)$. Indeed, let $f \in \mathcal{F}_n$ be arbitrary, so that $f = \sum_{i=1}^\infty u_i \varphi_\Sigma(\cdot - \mu_i)$. We define $\mathcal{J} := \mathbb{N} \cup \{\infty\}$, $\mathcal{K} := \{i \,:\, |u_i| > n^{-1}\}$, and $\mathcal{L} := \{i \,:\, \mu_i \in S_n\}$. Now choose $\mathcal{I} = \mathcal{J} \cap \mathcal{K} \cap \mathcal{L}$, and notice that $|\mathcal{I}| \leq |\mathcal{K}| \leq Hn\epsilon_n^2/\log n$. Hence we can pick a $m \in \mathcal{F}_{n,\epsilon}$ with $m(x) = \sum_{i\in\mathcal{I}} u_i' \varphi_{\Sigma'}(x - \mu_i')$. Moreover, for any $j = 1, \ldots, n$

$$|f(x_j) - m(x_j)| \leq \sum_{\mathcal{J}\cap\mathcal{K}^c} |u_i|\varphi_\Sigma(x_j-\mu_i) + \sum_{\mathcal{J}\cap\mathcal{K}\cap\mathcal{L}^c} |u_i|\varphi_\Sigma(x_j-\mu_i) + \sum_{i\in\mathcal{I}} |u_i||\varphi_\Sigma(x_j-\mu_i) - \varphi_{\Sigma'}(x_j-\mu_i')| + \sum_{i\in\mathcal{I}} |u_i-u_i'|\,\varphi_{\Sigma'}(x_j-\mu_i').$$

The first term in the rhs of the last equation is bounded above by ϵ_n. Regarding the second term, for any $i \in \mathcal{L}^c$ we have $(x_j-\mu_i)^T\Sigma^{-1}(x_j-\mu_i) > 4\log n$ for all $j = 1,\dots,n$. Then the second term is bounded by $|\mathcal{K}|n\exp(-2\log n) \leq H\epsilon_n^2/\log n \leq \epsilon_n$ for n large enough. Clearly, we can always choose $m \in \mathcal{F}_{n,\epsilon}$ with $|u_i-u_i'| \leq n^{-3/2}H^{-1}$ and $\|\mu_i-\mu_i'\| \leq n^{-(3/2-b_2^{-1})}$ for all $i \in \mathcal{I}$. Furthermore, we claim that Σ' can be chosen so that $\|I-\Sigma'\Sigma^{-1}\| \leq n^{-(1+b_1^{-1}+b_2^{-1})}\epsilon_n$. If so, Proposition 11 implies

$$\begin{aligned}
&|f(x_j)-m(x_j)| \\
&\quad\leq 2\epsilon_n + \sum_{i\in\mathcal{I}}|u_i-u_i'| + \sum_{i\in\mathcal{I}}|u_i||\varphi_\Sigma(x_j-\mu_i)-\varphi_{\Sigma'}(x_j-\mu_i')| \\
&\quad\lesssim \epsilon_n + \sum_{i\in\mathcal{I}}|u_i-u_i'| + \sum_{i\in I}|u_i|\frac{\lambda_1(\Sigma')}{\lambda_d(\Sigma')}\|I-\Sigma'\Sigma^{-1}\| + \sum_{i\in\mathcal{I}}|u_i|\frac{\|\mu_i-\mu_i'\|}{\lambda_d(\Sigma')} \\
&\quad\lesssim \epsilon_n,
\end{aligned}$$

for all $j = 1,\dots,n$. Therefore $d_n(f,m) \lesssim \epsilon_n$, and the claim is proved. A straightforward computation shows that we can find constants $0 < c_0, c_1 < \infty$ such that $|R_n| \leq n^{c_0}$ and $|Q_n| \leq n^{c_1}$, then

$$\log N(\delta\epsilon_n, \mathcal{F}_n, d_n) \leq |\mathcal{I}|\log\left(\frac{n}{n^{-3/2}}\times n^{c_0}\right) + \log\left(n^{c_1}\right) \lesssim Hn\epsilon_n^2,$$

when n is large enough. In view of the previous computations, it is clear that δ can be chosen to be $\delta = 1$.

It remains to prove that for any $\Sigma \in \mathcal{E}$ with $n^{-1/b_2} \leq \lambda_j(\Sigma) \leq n^{-1/b_1}$ we can find $\Sigma' \in Q_n$ such that $\|I-\Sigma'\Sigma^{-1}\| \leq n^{-(1+b_1^{-1}+b_2^{-1})}\epsilon_n$. Let $\Sigma = P^TDP$ be the spectral decomposition of Σ. There is $\Sigma' = P'^TD'P'$ in Q_n with $\|P-P'\| \leq n^{-(1+2b_1^{-1}+2b_2^{-1})}\epsilon_n$ and $1 \leq \lambda_j(\Sigma')/\lambda_j(\Sigma) \leq 1 + n^{-(1+b_1^{-1}+b_2^{-2})}\epsilon_n$. Writing $\widetilde{\Sigma} := P'^TDP'$, we get the bound

$$\|I-\Sigma'\Sigma^{-1}\|(1+\|I-\widetilde{\Sigma}\Sigma^{-1}\|).$$

The first term of the rhs is bounded by $n^{-(1+b_1^{-1}+b_2^{-1})}\epsilon_n$ because $I-\Sigma'\widetilde{\Sigma}^{-1} = P'^T(I-D'D^{-1})P'$. Regarding the second term, we have $I-\widetilde{\Sigma}\Sigma^{-1} = P'^T(B-DBD^{-1})P$ for $B := P'P^T - I$. Then $\|I-\widetilde{\Sigma}\Sigma^{-1}\| \leq \|B-DBD^{-1}\| \leq d\|B\|_{\max}\frac{\lambda_1(\Sigma)}{\lambda_d(\Sigma)}$, where

$$\|B\|_{\max} \leq \|B\| \leq \|P-P'\| \leq n^{-(1+2b_1^{-1}+2b_2^{-1})}\epsilon_n. \qquad \square$$

Lemma 4. *Assume that there is $n_0 \in \mathbb{N}$, and $0 < \gamma_1 \leq \gamma_2 < 1$ such that $n^{-\gamma_2/2} \leq \epsilon_n \leq n^{-\gamma_1/2}$ for all $n \geq n_0$. Then $\Pi(\mathcal{F}_n(H,\epsilon_n)^c) \lesssim \exp(-\frac{H}{4}(1-\gamma_2)n\epsilon_n^2)$ for all $n \geq n_0$.*

Proof. We use the fact that $M \sim \Pi_\alpha$ is almost surely purely-atomic (Kingman, 1992)). Then from the definition of $\mathcal{F}_n$ it follows

$$\begin{aligned}\Pi(\mathcal{F}_n^c) \leq\ & dG_\Sigma(\Sigma : \lambda_d(\Sigma) < n^{-1/b_2}) + dG_\Sigma(\Sigma : \lambda_1(\Sigma) > n^{1/b_1}) \\ & + \Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| > n\Big) + \Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i|\, \mathbb{1}\{|u_i| \leq n^{-1}\} > \epsilon_n\Big) \\ & + \Pi_\alpha\Big(|\{i : |u_i| > n^{-1}\}| > Hn\epsilon_n^2/\log n\Big).\end{aligned}$$

We bound each of the terms as follows. By assumption the first two terms are bounded by $d(e^{-a_1 n} + e^{-a_2 n})$. Notice that $\sum_{i=1}^\infty |u_i| = |M|$, where $|M|$ denote the total variation of the measure M. Since by definition we have $M \stackrel{d}{=} M_1 - M_2$, with M_1, M_2 independent Gamma random measures with same base measure $\alpha(\cdot)$, it follows that $|Q|$ has the distribution of a Gamma random variable with shape parameter $2\overline{\alpha}$. Then by Markov's inequality,

$$\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| > n\Big) = \Pi_\alpha\left(e^{\frac{1}{2}|M|} > e^{\frac{1}{2}n}\right) \leq 2^{2\overline{\alpha}} e^{-\frac{1}{2}n}.$$

Also, by the superposition theorem (Kingman, 1992, section 2), for any $M \sim \Pi_\alpha$ we have $M \stackrel{d}{=} M_3 + M_4$, where M_3 and M_4 are independent random measures with total variation $|M_3|$ and $|M_4|$ having Laplace transforms (for all $t \in \mathbb{R}$ for which the integrals in the expressions converge)

$$\begin{aligned}Ee^{t|M_3|} &:= \exp\left\{2\overline{\alpha}\int_{1/n}^\infty (e^{tx} - 1)x^{-1}e^{-x}\,dx\right\}, \\ Ee^{t|M_4|} &:= \exp\left\{2\overline{\alpha}\int_0^{1/n} (e^{tx} - 1)x^{-1}e^{-x}\,dx\right\}.\end{aligned}$$

M_3 and M_4 are almost-surely purely atomic, M_3 has only jumps greater than $1/n$ (almost surely) which number is distributed according to a Poisson distribution with intensity $2\overline{\alpha}E_1(n^{-1})$, where E_1 denotes the exponential integral E_1 function: $E_1(x) = \int_x^\infty \frac{e^{-t}}{t}dt$. Likewise, M_4 has only jumps smaller or equal to $1/n$ (almost-surely) which number is almost-surely infinite. Recalling that $E_1(x) = \gamma + \log(1/x) + o(1)$ for x small, it holds $2\overline{\alpha}\gamma \leq 2\overline{\alpha}E_1(1/n) \leq 6\overline{\alpha}\log n \leq x_n$ for n sufficiently large, with $x_n := Hn\epsilon_n^2/\log n$. Thus using Chernoff's bound on Poisson distribution, we get

$$\begin{aligned}\Pi_\alpha\Big(|\{i : |u_i| > n^{-1}\}| > Hn\epsilon_n^2/\log n\Big) &\leq e^{-2\overline{\alpha}E_1(1/n)}\frac{(e2\overline{\alpha}E_1(1/n))^{x_n}}{x_n^{x_n}} \\ &\leq \exp\left\{-\frac{1}{2}x_n\log x_n\right\}.\end{aligned}$$

But, $\log x_n = \log n + \log H - 2\log\epsilon_n^{-1} - \log\log n \geq (1-\gamma_2)\log n + \log H - \log\log n \geq \frac{1}{2}(1-\gamma_2)\log n$ for large n. Therefore, as $n \to \infty$

$$\Pi_\alpha\Big(|\{i : |u_i| > n^{-1}\}| > Hn\epsilon_n^2/\log n\Big) \leq \exp\left\{-\frac{H}{4}(1-\gamma_2)n\epsilon_n^2\right\}.$$

Finally, we use again Markov's inequality to get

$$\begin{aligned}\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i|\,\mathbb{1}\{|u_i|\le n^{-1}\} > \epsilon_n\Big) &= \Pi_\alpha\Big(e^{n\epsilon_n|M_4|} > e^{n\epsilon_n^2}\Big)\\ &\le e^{-n\epsilon_n^2}\exp\left\{2\overline{\alpha}\int_0^{1/n}(e^{n\epsilon_n x}-1)x^{-1}e^{-x}\,dx\right\}.\end{aligned}$$

But for $x\in(0,1/n)$, we have $e^{n\epsilon_n x}-1\le n(e^{n\epsilon_n\delta_n}-1)x$, thus the integral in the previous expression is bounded by $2\overline{\alpha}(e^{\epsilon_n}-1)$, which is in turn bounded by $2\overline{\alpha}(e-1)$ because $\epsilon_n\le 1$ if $n\ge n_0$. □

4.2. Location-scale mixtures

4.2.1. Kullback-Leibler condition

We use the notations $\mathcal{M}$, Λ_J, I_j, ... defined in the Section 3.4. By Chebychev inequality, we have $Q_0(\|X\|>\zeta_j)\le\zeta_j^{-p}Q_0\|X\|^p$. Therefore, bringing together results from Propositions 7 and 8,

$$\begin{aligned}\int |f_M(x)-f_0(x)|^2\,dQ_0(x) &= \sum_{j=0}^J\int_{I_j}|f_M(x)-f_0(x)|^2\,dQ_0(x)+\int_{A_0^c}|f_M(x)-f(x)|^2\,dQ_0(x)\\ &\lesssim J^3\sum_{j=0}^J\sigma_j^{2\beta}Q_0(I_j)+J^3Q_0(A_0^c).\end{aligned}$$

Then we can find a constant $C>0$ such that $\int|f_M(x)-f_0(x)|^2\,dQ_0(x)\le CJ^4\sigma_J^{2\beta}$ for all $M\in\mathcal{M}$ and J large enough.

By equation (7), we have $G_\Sigma(U_{jk})\gtrsim\sigma_j^{-2b_3}\sigma_J^{b_4(\beta+d)}\exp(-a_3\sigma_j^{-\kappa})$ for all $j=0,\dots J$. Moreover, there is a separation of $h_J\sigma_j$ between two consecutive μ_{jk} and $h_J\sigma_j\ll\sigma_j$, thus all the W_{jk} with $(j,k)\in\Lambda_J$ are disjoint. By equation (8), we have $\alpha_{jk}:=\overline{\alpha}G_\Sigma(U_{jk})G_\mu(V_{jk})\gtrsim\sigma_j^{b_5-2b_3}\sigma_J^{b_5\beta+b_4(\beta+d)}\exp(-a_3\sigma_j^{-\kappa})(1+\|\mu_{jk}\|)^{-b_6}$ for all $(j,k)\in\Lambda_J$. We also define $\alpha^c:=\alpha(W^c)$. For J large enough, there is a constant $C'>0$ not depending on J such that $\alpha^c>C'$. Moreover, since α has finite variation we can assume without loss of generality that $C'\le\alpha^c\le 1$, otherwise we split W^c into disjoint parts, each of them having α-measure smaller than one. With $\epsilon_n^2:=CJ^4\sigma_J^{2\beta}$, using that $\Gamma(\alpha)\le 2\alpha^{\alpha-1}$ for $\alpha\le 1$ and $\mathcal{M}\subset\mathtt{KL}(f_0,\epsilon_n)$, it follows the lower bound

$$\Pi(\mathtt{KL}(f_0,\epsilon_n))\ge\frac{\sigma_J^\beta}{3e\Gamma(\alpha^c)}\prod_{(j,k)\in\Lambda_J}\left(\frac{\sigma_J^\beta e^{-2|u_{jk}|}}{3e\Gamma(\alpha_{jk})}\right)$$

$$\geq \frac{\sigma_J^\beta}{3e\Gamma(\alpha^c)} \prod_{(j,k)\in\Lambda_J} \exp\left\{-2|u_{jk}| - \beta\log\sigma_J^{-1} + \log\frac{1}{6e} + (\alpha_{jk}-1)\log\alpha_{jk}\right\}$$
$$\geq \exp\left\{-KJ|\Lambda_J| - 2\textstyle\sum_{(j,k)\in\Lambda_J}|u_{jk}| - \sum_{(j,k)\in\Lambda_J}\log\alpha_{jk}^{-1}\right\}, \tag{21}$$

for a constant $K > 0$ depending only on C and β. We now evaluate the sums involved in the rhs of equation (21). As before, be have that $\sum_{(j,k)\in\Lambda_J}|u_{jk}| \leq 4\|f_0\|_1\sigma_J^{-d}$ (see for instance the proof of Proposition 8). Act as in Section 4.1.1 to find that

$$\sum_{(j,k)\in\Lambda_J}\log\alpha_{jk}^{-1} \lesssim J|\Lambda_J| + J^{1+d/2}\sigma_J^{-d} + |\Lambda_J|\sigma_J^{-\kappa}.$$

The term proportional to $|\Lambda_J|\sigma_J^{-\kappa}$ is entirely responsible for the bad rates in location-scale mixtures, and the aim of the hybridation of next section is to get rid of it. For a constant $K' > 0$,

$$\Pi(\mathtt{KL}(f_0,\epsilon_n)) \geq \exp\left\{-K'\max(|\Lambda_J|\sigma_J^{-\kappa}, J^{1+d/2}\sigma_J^{-d})\right\}.$$

Then for appropriate constants $C', t > 0$ we can have $\Pi(\mathtt{KL}(f_0,\epsilon_n)) \geq e^{-c_2 n\epsilon_n^2}$ if

$$\epsilon_n^2 = \begin{cases} C' n^{-\frac{2\beta}{2\beta+\min(\beta+d,2\beta d/p)+\kappa}}(\log n)^t, & p \leq 2\beta, \\ C' n^{-\frac{2\beta}{2\beta+d+\kappa}}(\log n)^t, & p > 2\beta. \end{cases}$$

4.2.2. Sieve construction

Using the notation $f_M(x) := \int \varphi_\Sigma(x-\mu)\,dM(\Sigma,\mu)$, we construct the following sieve.

$$\mathcal{F}_n(H,\epsilon) := \left\{ f = f_M : \begin{array}{l} M = \sum_{i=1}^\infty u_i\delta_{\Sigma_i,\mu_i},\ \sum_{i=1}^\infty |u_i| \leq n,\ \sum_{i=1}^\infty |u_i|\,\mathbb{1}\{|u_i| \leq n^{-1}\} \leq \epsilon, \\ \sum_{i=1}^\infty |u_i|\,\mathbb{1}\{\lambda_d(\Sigma_i) < n^{-1/b_2}\} \leq \epsilon,\ \sum_{i=1}^\infty |u_i|\,\mathbb{1}\{\lambda_1(\Sigma_i) > n^{1/b_1}\} \leq \epsilon \\ |\{i : |u_i| > n^{-1},\ \forall j : n^{-1/b_2} \leq \lambda_j(\Sigma_i) \leq n^{1/b_1}\}| \leq Hn\epsilon^2/\log n, \end{array} \right\}. \tag{22}$$

Lemma 5. *Let $\mathbf{x} = (x_1,\ldots,x_n) \in \mathbb{R}^n$ be arbitrary and d_n be the empirical L^2-distance associated with x. Then for any $n^{-1/2} < \epsilon_n \leq 1$, $0 < H \leq 1$ and n sufficiently large there is a constant $C > 0$ not depending on n such that $\log N(\epsilon_n, \mathcal{F}_n(H,\epsilon_n), d_n) \leq CHn\epsilon_n^2$.*

Proof. The proof is almost identical to Lemma 3. □

Lemma 6. *Assume that there is $n_0 \in \mathbb{N}$, and $0 < \gamma_1 \leq \gamma_2 < 1$ such that $n^{-\gamma_2/2} \leq \epsilon_n \leq n^{-\gamma_1/2}$ for all $n \geq n_0$. Then $\Pi(\mathcal{F}_n(H,\epsilon_n)^c) \lesssim \exp(-\frac{H}{4}(1-\gamma_2)n\epsilon_n^2)$ for all $n \geq n_0$.*

Proof. We first write the estimate

$$\Pi(\mathcal{F}_n^c) \leq \Pi_\alpha\Big(\sum_{i=1}^\infty |u_i| > n\Big) + \Pi_\alpha\Big(\sum_{i=1}^\infty |u_i|\,\mathbb{1}\{\lambda_d(\Sigma_i) < n^{-1/b_2}\} > \epsilon_n\Big)$$
$$+ \Pi_\alpha\Big(\sum_{i=1}^\infty |u_i|\,\mathbb{1}\{|u_i| \leq n^{-1}\} > \epsilon_n\Big) + \Pi_\alpha\Big(\sum_{i=1}^\infty |u_i|\,\mathbb{1}\{\lambda_1(\Sigma_i) > n^{1/b_1}\} > \epsilon_n\Big)$$
$$+ \Pi_\alpha\Big(|\{i : |u_i| > n^{-1},\ \forall j : n^{-1/b_2} < \lambda_j(\Sigma_i) \leq n^{1/b_1}\}| > Hn\epsilon_n^2/\log n\Big).$$

The first, third and last terms in the rhs above obeys the same bounds as in the proof of Lemma 4, using the same arguments. The two remaining terms are bounded using the same trick. For instance, note that the random variable $U := \sum_{i=1}^\infty |u_i|\,\mathbb{1}\{\lambda_1(\Sigma_i) > n^{1/b_1}\}$ has Gamma distribution with parameters $(2\alpha(A_n), 1)$, where $A_n := \{(\Sigma, \mu) : \lambda_1(\Sigma) > n^{1/b_1}\}$. For n large, by assumptions on G_Σ, it holds $\alpha(A_n) \ll \epsilon_n$. Then by Chebychev inequality, for n large enough

$$\begin{aligned}\Pi_\alpha\Big(\sum_{i=1}^\infty |u_i|\,\mathbb{1}\{\lambda_1(\Sigma_i) > n^{1/b_1}\} > \epsilon_n\Big) &\leq \Pr(U - EU > \epsilon_n - EU)\\ &\leq \Pr(U - EU > \epsilon_n/2)\\ &\leq 16\epsilon_n^{-2}\alpha(A_n)^2. \qquad (23)\end{aligned}$$

The conclusion follows from the assumptions on G_Σ which imply $\alpha(A_n) = \overline{\alpha} G_\Sigma(\Sigma : \lambda_1(\Sigma) > n^{1/b_1}) \lesssim \exp(-a_1 n)$. The other term is left to the reader. □

4.3. Hybrid location-scale mixtures

Obviously, given the definition of a hybrid mixture (see Section 4.3), most of the proof is redundant with the location-scale case, and in the sequel we deal only with the parts that differ.

4.3.1. Kullback-leibler condition

Let $\mathcal{M} \equiv \mathcal{M}(\beta, J, f, \Lambda_J)$ be the set of signed measures constructed in Section 3.4. For any integer $J > 0$ let Ω_J be the event

$$\Omega_J := \big\{P_\Sigma : P_\Sigma(\mathcal{E}_{j,J\beta}) \geq 2^{-J} \quad \forall\, 0 \leq j \leq J\big\}.$$

Then with arguments and constant $C > 0$ from Section 4.2.1, letting $\epsilon_n^2 := CJ^4\sigma_J^{2\beta}$, we have

$$\Pi(\mathtt{KL}(f_0, \epsilon_n)) \geq \Pi(\mathcal{M}) \geq \Pi(\mathcal{M} \mid \Omega_J)\Pi_\Sigma(\Omega_J).$$

But by equation (11) there are constants a_6, b_7, κ^* eventually depending on β such that $\Pi_\Sigma(\Omega_J) \gtrsim \exp(-a_6 J^{b_7} 2^{J\kappa^*})$ and on Ω_J it holds

$$\alpha(W_{jk}) = \overline{\alpha} P_\Sigma(U_j) G_\mu(V_{jk}) \geq \overline{\alpha} 2^{-J} G_\mu(V_{jk}),$$

for all $(j,k) \in \Lambda_J$. Then act as in equation (21) to find a constant $K > 0$ such that (recalling that $\sigma_J = 2^{-J}$)

$$\Pi(\mathtt{KL}(f_0, \epsilon_n)) \gtrsim \exp\left\{-K \max(J^{b_7} \sigma_J^{-\kappa^*}, J^{1+d/2} \sigma_J^{-d}) - KJ|\Lambda_J|\right\}.$$

Because of Proposition 6 we can have $\Pi(\mathtt{KL}(f_0, \epsilon_n)) \geq e^{-c_2 n \epsilon_n^2}$ if for appropriate constants $C', t > 0$

$$\epsilon_n^2 = \begin{cases} C' n^{-\frac{2\beta}{2\beta + \max(\kappa^*, \min(\beta+d, 2\beta d/p))}} & p \leq 2\beta, \\ C' n^{-\frac{2\beta}{2\beta+\max(\kappa^*,d)}} (\log n)^t & p > 2\beta. \end{cases}$$

4.3.2. Sieve construction

We use the same sieve $\mathcal{F}_n(H, \epsilon)$ as in equation (22). The definition of $\mathcal{F}_n(j, \epsilon)$ is independent of Π thus the conclusion of Lemma 3 holds for hybrid location-scale mixtures. It remains to show that $\Pi(\mathcal{F}_n(H, \epsilon)^c) \leq \exp(-2c_2 n \epsilon_n^2)$, which is the object of the next lemma.

Lemma 7. *Assume that there is $n_0 \in \mathbb{N}$, and $0 < \gamma_1 \leq \gamma_2 < 1$ such that $n^{-\gamma_2/2} \leq \epsilon_n \leq n^{-\gamma_1/2}$ for all $n \geq n_0$. Then there is a constant a constant $\gamma_2 < \gamma < 1$ such that $\Pi(\mathcal{F}_n(H, \epsilon_n)^c) \lesssim \exp(-\frac{H}{4}(1-\gamma) n \epsilon_n^2)$ for all $n \geq n_0$.*

Proof. We proceed as in the proof of Lemma 6. Following the same steps, we deduce that it is sufficient to prove that

$$\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| \, \mathbb{1}\{\lambda_1(\Sigma_i) > n^{1/b_1}\} > \epsilon_n\Big) \lesssim e^{-2c_2 n},$$
$$\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| \, \mathbb{1}\{\lambda_d(\Sigma_i) < n^{-1/b_2}\} > \epsilon_n\Big) \lesssim e^{-2c_2 n}.$$

Since the proofs are almost identical for the two previous conditions, we only prove the first and leave the second to the reader. Notice that by equation (23) we have

$$\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| \, \mathbb{1}\{\lambda_1(\Sigma_i) > n^{1/b_1}\} > \epsilon_n \;\Big|\; P_\Sigma\Big) \leq 16 \overline{\alpha} \epsilon_n^{-2} P_\Sigma(\Sigma : \lambda_1(\Sigma) > n^{1/b_1})^2.$$

Letting $\Omega := \{P_\Sigma : P_\Sigma(\lambda_1(\Sigma) > n^{1/b_1}) < \exp(-a_1 n/2)\}$, with a slight abuse of notation, it follows from equation (9)

$$\begin{aligned} &\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| \, \mathbb{1}\{\lambda_1(\Sigma_i) > n^{1/b_1}\} > \epsilon_n\Big) \\ &\qquad \leq \Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| \, \mathbb{1}\{\lambda_1(\Sigma_i) > n^{1/b_1}\} > \epsilon_n \;\Big|\; \Omega\Big) + \Pi_\Sigma(\Omega^c) \\ &\qquad \lesssim \epsilon_n^{-2} \exp(-a_1 n) + \exp(-a_4 n). \end{aligned}$$

□

5. Proof of Theorem 2

The proof follows the same lines as Ghosal and van der Vaart (2007b) with additional cares. The first step consists on rewriting expectation of the posterior distribution as follows. Let $(\phi_n(\cdot \mid \cdot))_{n\geq 0}$ be a sequence of test functions such that for n large enough

$$Q_0^n\left[P_0^n[\phi_n(\mathbf{y} \mid \mathbf{x}) \mid \mathbf{x}]\right] \lesssim N(\epsilon/18, \mathcal{F}_n, d_n) \exp\left(-\frac{n\epsilon_n^2}{2}\right),$$
$$\sup_{\{f: d_n(f,f_0)\geq 17\epsilon_n/18\}\cap \mathcal{F}_n} Q_0^n\left[P_f^n[1-\phi_n(\mathbf{y} \mid \mathbf{x})] \mid \mathbf{x}\right] \lesssim \exp\left(-\frac{n\epsilon_n^2}{2}\right).$$

The existence of such test functions is standard and follows for instance from Birgé (2006, proposition 4), or Ghosal and van der Vaart (2007b, section 7.7). From here, we bound the posterior distribution in a standard fashion,

$$Q_0^n\left[P_0^n[\Pi_{\mathbf{x}}(\{f : d_n(f,f_0) > \epsilon_n\} \mid \mathbf{y}, \mathbf{x}) \mid \mathbf{x}]\right] \leq Q_0^n[P_0^n[\Pi_{\mathbf{x}}(\mathcal{F}_n^c \mid \mathbf{y}, \mathbf{x}) \mid \mathbf{x}]] + Q_0^n[P_0^n[\Pi(\{f : d_n(f,f_0) > \epsilon_n\} \cap \mathcal{F}_n \mid \mathbf{y}, \mathbf{x}) \mid \mathbf{x}]].$$

So that,

$$\begin{aligned} Q_0^n\left[P_0^n[\Pi_{\mathbf{x}}(\{f : d_n(f,f_0) > \epsilon_n\} \mid \mathbf{y}, \mathbf{x}) \mid \mathbf{x}]\right] &\leq Q_0^n[P_0^n[\Pi_{\mathbf{x}}(\mathcal{F}_n^c \mid \mathbf{y}, \mathbf{x}) \mid \mathbf{x}]] \\ &+ Q_0^n\left[P_0^n\left[\phi_n(\mathbf{y} \mid \mathbf{x})\Pi_{\mathbf{x}}(\{f : d_n(f,f_0) > \epsilon_n\} \cap \mathcal{F}_n \mid \mathbf{y}, \mathbf{x}) \mid \mathbf{x}\right]\right] \\ &+ Q_0^n\left[P_0^n\left[\left(1-\phi_n(\mathbf{y} \mid \mathbf{x})\right)\Pi_{\mathbf{x}}(\{f : d_n(f,f_0) > \epsilon_n\} \cap \mathcal{F}_n \mid \mathbf{y}, \mathbf{x}) \mid \mathbf{x}\right]\right]. \end{aligned} \tag{24}$$

Now, to any $\mathbf{x} \in \mathbb{R}^{d\times n}$, we associate the events

$$E_n(\mathbf{x}) := \left\{ y \in \mathbb{R}^n : \int_{\mathcal{F}} \prod_{i=1}^n \frac{p_f(x_i,y_i)}{p_{f_0}(x_i,y_i)}\, d\Pi_{\mathbf{x}}(f) \geq \exp\left(-(1+4c_2)\frac{n\epsilon_n^2}{4}\right)\right\}. \tag{25}$$

and we define

$$\tilde{E}_n = \left\{\mathbf{x} : \Pi_{\mathbf{x}}\left(\{f : d_n(f,f_0) \leq \epsilon_n\} \geq \delta_0 \exp(-c_2 n\epsilon_n^2)\right)\right\}.$$

By assumption $Q_0^n(\tilde{E}_n^c) = o(1)$. Consider the first term of the rhs of equation (24). We can rewrite,

$$\begin{aligned} &Q_0^n[P_0^n[\Pi_{\mathbf{x}}(\mathcal{F}_n^c \mid \mathbf{y}, \mathbf{x}) \mid \mathbf{x}]] \\ &\leq \frac{e^{\frac{1}{4}(4c_2+1)n\epsilon_n^2}}{\delta_0} \int_{\mathbb{R}^{d\times n}} \mathbb{1}_{\tilde{E}_n} \int_{E_n(\mathbf{x})} \int_{\mathcal{F}_n^c} \prod_{i=1}^n \frac{p_f(x_i,y_i)}{p_{f_0}(x_i,y_i)}\, d\Pi_{\mathbf{x}}(f) dP_0^n(\mathbf{y} \mid \mathbf{x}) dQ_0^n(\mathbf{x}) \\ &\quad + \int_{\mathbb{R}^{d\times n}} \mathbb{1}_{\tilde{E}_n} \int_{E_n(\mathbf{x})^c} dP_0^n(\mathbf{y} \mid \mathbf{x}) dQ_0^n(\mathbf{x}) + Q_0^n(\tilde{E}_n^c) \end{aligned}$$

$$
\begin{aligned}
&= \frac{e^{\frac{1}{4}(4c_2+1)n\epsilon_n^2}}{\delta_0} \int_{\mathbb{R}^{d\times n}} \int_{\mathcal{F}_n^c} \int_{E_n(\mathbf{x})} dP^n(\mathbf{y} \mid \mathbf{x}) d\Pi_{\mathbf{x}}(f) dQ_0^n(\mathbf{x}) \\
&\quad + \int_{\mathbb{R}^{d\times n}} \mathbb{1}_{\tilde{E}_n} \int_{E_n(\mathbf{x})^c} dP_0^n(\mathbf{y} \mid \mathbf{x}) dQ_0^n(\mathbf{x}) + o(1) \\
&\leq \frac{e^{\frac{1}{4}(4c_2+1)n\epsilon_n^2}}{\delta_0} \int_{\mathbb{R}^{d\times n}} \Pi_{\mathbf{x}}(\mathcal{F}_n^c)\, dQ_0^n(\mathbf{x}) \\
&\quad + \int_{\mathbb{R}^{d\times n}} \mathbb{1}_{\tilde{E}_n} \int_{E_n(\mathbf{x})^c} dP_0^n(\mathbf{y} \mid \mathbf{x}) dQ_0^n(\mathbf{x}) + o(1),
\end{aligned}
$$

where the third line follows from Fubini's theorem. The same reasoning applies to the other terms of equation (24), using the test functions introduced above and $0 < c_2 < 1/4$. Hence the theorem is proved if we show that

$$
\int_{\mathbb{R}^{d\times n}} \mathbb{1}_{\tilde{E}_n} \int_{E_n(\mathbf{x})^c} dP_0^n(\mathbf{y} \mid \mathbf{x}) dQ_0^n(\mathbf{x}) = o(1).
$$

Moreover Ghosal and van der Vaart (2007b, Lemma 10) implies that on $\tilde{E}_n$

$$
P_0^n \left(\int_{\mathcal{F}} \prod_{i=1}^n \frac{p_f(x_i, Y_i)}{p_{f_0}(x_i, Y_i)}\, d\Pi_{\mathbf{x}}(f) < \delta_0 \exp\left(-\frac{1}{4}(1+4c_2)\epsilon_n^2 \right) \;\middle|\; \mathbf{x} \right) \leq \frac{1}{n\epsilon_n^2},
$$

which terminates the proof.

Appendix A: Proofs of Lemma 1 and 2 and some technical results on the kernels

A.1. Proof of Lemma 1

Clearly, $\|\chi_\sigma * f\|_1 \leq \|\chi_\sigma\|_1 \|f\|_1$ by Young's inequality, so that $\chi_\sigma * f \in L^1$ and $\widehat{(\chi_\sigma * f)}(\xi) = \widehat{\chi}_\sigma(\xi)\widehat{f}(\xi)$, showing that the support of the Fourier transform of $\chi_\sigma * f$ is included in $[-1/\sigma, 1/\sigma]^d$. Moreover, using again Young's inequality we get that $\|\chi_\sigma * f\|_\infty \leq \|\chi_\sigma\|_1 \|f\|_\infty$, thus $\chi_\sigma * f \in L^\infty$.

By construction of $\widehat{\chi}$, it follows by integrating by parts that for any $q \in \mathbb{N}^d$ we have $(-i)^{|q|} u^q \chi(u) = (2\pi)^{-d} \int D^q \widehat{\chi}(\xi) e^{i\xi u} d\xi$. Clearly $\widehat{\chi}$ is Schwartz, hence by Fourier inversion we have that

$$
(-i)^{|q|} \int u^q \chi(u) e^{-i\xi u} du = D^q \widehat{\chi}(\xi), \quad \forall \xi \in \mathbb{R}^d.
$$

But, by construction $\widehat{\chi}(0) = 1$, and for any $q \in \mathbb{N}^d$ with $|q| \geq 1$ we have $D^q \widehat{\chi}(0) = 0$. It follows that $\int \chi(u) du = 1$, and $\int u^q \chi(u) du = 0$ for any $|q| \geq 1$. Whence, letting m be the largest integer smaller than β, and using Taylor's formula with exact remainder term

$$
\begin{aligned}
&\chi_\sigma * f(x) - f(x) \\
&= \int \chi_\sigma(y)\,[f(x-y) - f(x)]\,dy = \int \chi(y)\,[f(x-\sigma y) - f(x)]\,dy \\
&= \sum_{1\le|k|\le m-1} \frac{(-\sigma)^{|k|}}{k!} \int u^k \chi(u)\,du \\
&\quad + \sum_{|k|=m} \frac{m(-\sigma)^m}{k!} \int y^k \chi(y) \int_0^1 (1-u)^{m-1} D^k f(x-\sigma u y)\,du dy \\
&= \sum_{|k|=m} \frac{m(-\sigma)^m}{k!} \int y^k \chi(y) \int_0^1 (1-u)^{m-1} \left[D^k f(x-\sigma u y) - D^k f(x)\right] du dy.
\end{aligned}
$$

Therefore, because $f \in \mathtt{C}^\beta$,

$$
\begin{aligned}
&|\chi_\sigma * f(x) - f(x)| \\
&\quad \le \sigma^m \sum_{|k|=m} \frac{m}{k!} \int |y^k \chi(y)| \int_0^1 (1-u)^{m-1} |D^k f(x-\sigma u y) - D^k f(x)|\,du dy \\
&\quad \le \|f\|_{\mathtt{C}^\beta} \sigma^\beta \sum_{|k|=m} \frac{m}{k!} \int |y^k \chi(y)|\,dy \int_0^1 (1-u)^{m-1}\,du.
\end{aligned}
$$

A.2. Proof of Lemma 2

We mostly follow the proof of Hangelbroek and Ron (2010, proposition 1). Writing,

$$
\begin{aligned}
K_{h,\sigma} f_\sigma(x) &= \int \frac{h^d}{\sigma^d} \sum_{k\in\mathbb{Z}^d} \varphi\left(\frac{x-h\sigma k}{\sigma}\right) \eta\left(\frac{y-h\sigma k}{\sigma}\right) f_\sigma(y)\,dy \\
&= \frac{h^d}{\sigma^d} \sum_{k\in\mathbb{Z}^d} \varphi\left(\frac{x-h\sigma k}{\sigma}\right) \int \eta\left(\frac{y-h\sigma k}{\sigma}\right) f_\sigma(y)\,dy \\
&= \frac{h^d}{(2\pi)^d} \sum_{k\in\mathbb{Z}^d} \varphi\left(\frac{x-h\sigma k}{\sigma}\right) \int \widehat{\eta}(\sigma\xi) \widehat{f}_\sigma(\xi) e^{ih\sigma\xi k}\,d\xi \\
&= \int \widehat{\eta}(\sigma\xi) \widehat{f}_\sigma(\xi) \frac{h^d}{(2\pi)^d} \sum_{k\in\mathbb{Z}^d} \varphi\left(\frac{x-h\sigma k}{\sigma}\right) e^{ih\sigma\xi k}\,d\xi.
\end{aligned}
$$

Then we can invoke the *Poisson summation formula* (Härdle et al., 1998, theorem 4.1), which is obviously valid for φ, and

$$
\sum_{k\in\mathbb{Z}^d} \varphi\left(\frac{x-h\sigma k}{\sigma}\right) e^{ih\sigma\xi k} = \frac{1}{h^d} \sum_{m\in\mathbb{Z}^d} \widehat{\varphi}\left(\sigma\xi + \frac{2\pi m}{h}\right) e^{i(\sigma\xi + \frac{2\pi m}{h})x/\sigma}.
$$

Therefore, recalling that $\widehat{f_\sigma}$ is supported on $[-1/\sigma, 1/\sigma]^d$ and $\widehat{\chi}$ equals 1 on $[-1,1]^d$,

$$\begin{aligned} K_{h,\sigma} f_\sigma(x) &= \frac{1}{(2\pi)^d} \int \widehat{\chi}(\sigma\xi) \widehat{f}_\sigma(\xi) \sum_{m \in \mathbb{Z}^d} \frac{\widehat{\varphi}(\sigma\xi + 2\pi m/h)}{\widehat{\varphi}(\sigma\xi)} e^{i(\sigma\xi + \frac{2\pi m}{h})x/\sigma} \, d\xi \\ &= f_\sigma(x) + \frac{1}{(2\pi)^d} \sum_{m \in \mathbb{Z}^d \setminus \{0\}} \int \widehat{f}_\sigma(\xi) \frac{\widehat{\varphi}(\sigma\xi + 2\pi m/h)}{\widehat{\varphi}(\sigma\xi)} e^{i(\sigma\xi + \frac{2\pi m}{h})x/\sigma} \, d\xi. \end{aligned}$$

It follows that,

$$|K_{h,\sigma} f_\sigma(x) - f_\sigma(x)| \leq \frac{1}{(2\pi)^d} \|\widehat{f}_\sigma\|_1 \sup_{\xi \in [-1,1]^d} \sum_{m \in \mathbb{Z}^d \setminus \{0\}} \left| \frac{\widehat{\varphi}(\xi + 2\pi m/h)}{\widehat{\varphi}(\xi)} \right|.$$

Now, $\|\widehat{f_\sigma}\|_1 \leq 2\sigma^{-d} \|\widehat{f_\sigma}\|_\infty \leq 2\sigma^{-d} \|f_\sigma\|_1 \leq 2\sigma^{-d} \|\chi_\sigma\|_1 \|f\|_1$, which is finite by assumption. Recalling that by assumption $\widehat{\varphi}$ is Gaussian, it follows for all $\xi \in [-1,1]^d$ and all $h \leq 1$,

$$\begin{aligned} \sum_{m \in \mathbb{Z}^d \setminus \{0\}} \left| \frac{\widehat{\varphi}(\xi + 2\pi m/h)}{\widehat{\varphi}(\xi)} \right| &\leq \sum_{m \in \mathbb{Z}^d \setminus \{0\}} \exp\left\{ -\frac{1}{2} \|\xi + 2\pi m/h\|^2 + \frac{1}{2} \|\xi\|^2 \right\} \\ &\leq \sum_{m \in \mathbb{Z}^d \setminus \{0\}} \exp\left\{ -\frac{2\pi^2}{h^2} \left(1 - \frac{h}{2\pi \|m\|} \right) \|m\|^2 \right\}. \end{aligned}$$

This concludes the proof of the lemma.

A.3. Some other technical results on $K_{h,\sigma}$

Lemma 8. *There is a universal constant $C > 0$ such that for all $x \in \mathbb{R}^d$, all $0 < h \leq 1$ and all $\sigma > 0$, $\sum_{k \in \mathbb{Z}^d} |\eta((x - h\sigma k)/\sigma)| \leq C h^{-d}$. Moreover, $\eta \in \mathcal{S}$.*

Proof. We first prove that $\eta \in \mathcal{S}$. Obviously $\widehat{\varphi} \in \mathcal{S}$, and therefore so is $\widehat{\eta}$. Since the Fourier transform and the inverse Fourier transform are continuous mapping of $\mathcal{S}$ onto itself, it is immediate that $\eta \in \mathcal{S}$.

We finish the proof by remarking that $x \mapsto \sum_{k \in \mathbb{Z}^d} |\eta((x - h\sigma k)/\sigma)|$ is periodic with period $h\sigma$, hence it suffices to check that it is bounded for $x \in [0, h\sigma]^d$. If $x \in [0, h\sigma]^d$, then $\|x - h\sigma k\| \geq h\sigma \|k\|/2$ for any $\|k\| \geq 2$, so that

$$\sum_{k \in \mathbb{Z}^d} |\eta((x - h\sigma k)/\sigma)| \leq 3 \sup_{u \in \mathbb{R}^d} |\eta(u)| + \sum_{\|k\| \geq 2} |\eta((x - h\sigma k)/\sigma)|.$$

Because $\eta \in \mathcal{S}$, we can find a constant $B > 0$ such that

$$\sup_{x \in \mathbb{R}^d} (1 + \|x\|)^{d+1} |\eta(x)| \leq B.$$

Therefore,

$$\begin{aligned} \sum_{k \in \mathbb{Z}^d} |\eta((x - h\sigma k)/\sigma)| &\leq 3\|\eta\|_\infty + B \sum_{\|k\| \geq 2} (1 + h\|k\|/2)^{-2} \\ &\leq 3\|\eta\|_\infty + B(4h^{-1})^d, \end{aligned}$$

which concludes the proof of the first assertion with $C := 3\|\eta\|_\infty + 4^d B$, because of the assumption $h \leq 1$. □

The following Lemma gives some control on the coefficients of f on η.

Proposition 9. *Let $0 < h \leq 1$ and $a_k(f) := (h/\sigma)^d \int \eta((y - h\sigma k)/\sigma) f(y)\, dy$ for all $k \in \mathbb{Z}^d$. Then there are universal constants $C, C' > 0$, depending only on φ, such that $\sum_{k\in\mathbb{Z}^d} |a_k(f)| \leq C\|f\|_1 \sigma^{-d}$, and $|a_k(f)| \leq C'\|f\|_\infty h^d$ for all $k \in \mathbb{Z}^d$.*

Proof. For the first assertion of the proposition, we write,

$$\begin{aligned}\sum_{k\in\mathbb{Z}^d} |a_k(f)| &\leq \frac{h^d}{\sigma^d} \sum_{k\in\mathbb{Z}^d} \int |f(y)||\eta((y - h\sigma k)/\sigma)|\, dy \\ &\leq \sigma^{-d}\|f\|_1 \sup_{y\in\mathbb{R}^d} h^d \sum_{k\in\mathbb{Z}} |\eta((y - h\sigma k)/\sigma)|,\end{aligned}$$

and the conclusion follows from Lemma 8. The proof of the second assertion is simpler. Indeed,

$$|a_k(f)| \leq \frac{h^d}{\sigma^d} \int |f(y)||\eta((y - h\sigma k)/\sigma)|\, dy \leq h^d \|f\|_\infty \int |\eta(u)|\, du,$$

where the last integral is bounded because $\eta \in \mathcal{S}$ by Lemma 8. □

Appendix B: Proof of Corollary 1

Take $\mathcal{F}_n$ as in Section 4.1.2. Then by Lemma 3 we have $\log N(\epsilon_n/18, \mathcal{F}_n, d_n) \leq n\epsilon_n^2/4$ for H chosen small enough. As for $Q_0^n \Pi_{\mathbf{x}}(\mathcal{F}_n^c)$, it is immediate from the proof of Lemma 4 that $Q_0^n \Pi_{\mathbf{x}}(\mathcal{F}_n^c) \lesssim \exp(-\frac{1}{2}(1 + 2c_2)n\epsilon_n^2)$ for some $c_2 > 0$ when ϵ_n^2 is as in the corollary. Hence to apply Theorem 2 it remains to prove that

$$Q_0^n \left(\Pi_{\mathbf{x}}(f : d_n(f, f_0) \leq \epsilon_n) \leq \delta_0 \exp(-c_2 n\epsilon_n^2)\right) = o(1).$$

To do so, let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^{d\times n}$ arbitrary, $\sigma > 0$ and $h_\sigma \sqrt{\log \sigma^{-1}} := c_0$ for a constant $c_0 > 0$ small enough. Recall that from Lemma 1 and 2 we have $\|K_{h_\sigma,\sigma}(\chi_\sigma * f_0) - f_0\|_\infty \lesssim \sigma^\beta$, where $K_{h_\sigma,\sigma}(\chi_\sigma * f_0)(z) := \sum_{k\in\mathbb{Z}^d} u_k\, \varphi_{\sigma^2 I}(z - h_\sigma\sigma k)$. Define $S_n(\mathbf{x}) := \bigcup_{i=1}^n \{z \in \mathbb{R}^d : \|z - x_i\| \leq \sigma\sqrt{2(\beta + d)\log \sigma^{-1}}\}$ and

$$\Lambda(\mathbf{x}) := \{k \in \mathbb{Z}^d : |u_k| > \sigma^\beta, \quad h_\sigma\sigma k \in S_n(\mathbf{x})\}.$$

Also define $U_\sigma := \{\sigma' : \sigma \leq \sigma' \leq \sigma(1 + \sigma^\beta)\}$, and for all $k \in \Lambda(\mathbf{x})$ define $V_k := \{\mu : \|\mu - h_\sigma\sigma k\| \leq \sigma^{\beta+d}\}$. We denote by $\mathcal{M}_\sigma$ the set of signed measures M on $\mathbb{R}^d$ such that $|M(V_k) - u_k| \leq \sigma^\beta$ for all $k \in \Lambda(\mathbf{x})$ and $|M|(V^c) \leq \sigma^\beta$, where V^c is the relative complement of the union of all V_k for $k \in \Lambda(\mathbf{x})$. For any $M \in \mathcal{M}_\sigma$, we write $f_{M,\sigma}(z) := \int \varphi_{\sigma^2 I}(z - \mu)\, dM(\mu)$. Act as in Proposition 5 to find that $d_n(f_{M,\sigma}, f_0) \leq C h_\sigma^{-2} \sigma^\beta$ for any $M \in \mathcal{M}_\sigma$, with a constant $C > 0$

not depending on $\mathbf{x}$, M and σ. By construction of $S_n(\mathbf{x})$, for all $k \in \Lambda(\mathbf{x})$ there is at least one x_i such that $\|h_\sigma \sigma k - x_i\| \leq \sigma\sqrt{2(\beta+d)\log\sigma^{-1}}$. Then for any $k \in \Lambda(\mathbf{x})$, because $g(0) > 0$ and g is continuous at zero,

$$\overline{\alpha} G_{\mathbf{x}}(V_k) \geq n^{-1} \int_{\|z - h_\sigma \sigma k\| \leq \sigma} g(z - x_i)\, dz \geq C n^{-1} \sigma^d.$$

The constants $C > 0$ in the previous equation does not depend on $\mathbf{x}$ nor σ nor n. Remarking that $|\Lambda(\mathbf{x})| \lesssim \sigma^{-(\beta+d)}$ independently of $\mathbf{x}$ (see Proposition 3) and letting $\epsilon_n = C' h_\sigma^{-2} \sigma^\beta$ we can mimic the steps of Section 4.1.1 to find that

$$\begin{aligned}\Pi_{\mathbf{x}}(f \,:\, d_n(f, f_0) \leq \epsilon) &\gtrsim \exp\left\{-C''|\Lambda(x)|\log\sigma^{-1} - C''|\Lambda(x)|\log n\right\} \\ &\gtrsim \exp(-c_2 n \epsilon_n^2),\end{aligned}$$

for a constant $C'' > 0$ not depending on $\mathbf{x}$ and ϵ_n^2 defined in the corollary. Hence

$$\inf_{\mathbf{x}} \Pi_{\mathbf{x}}(f \,:\, d_n(f, f_0) \leq \epsilon_n) \geq \delta_0 \exp(-c_2 n \epsilon_n^2)$$

Appendix C: Some technical results on the construction of the approximation in the case of location-scale mixtures

Proposition 10. *Let $f_0 \in \mathcal{C}_\beta$. For any $j \geq 0$, we have $|\Delta_j(x)| \leq C\|f_0\|_{\mathcal{C}^\beta}\sigma_j^\beta$, with the same constant $C > 0$ as in Lemma 1. Moreover, $\|\Delta_j\|_1 \leq 2\|f_0\|_1$ for all $j \geq 0$.*

Proof. Notice that $\|\Delta_{j+1}\|_1 \leq \|\Delta_j\|_1 + \|\chi_{\sigma_{j+1}} * \Delta_j\| \leq (1 + \|\chi\|_1)\|\Delta_j\|_1$, by Young's inequality. Since $f_0 \in L^1$, this implies $\Delta_j \in L^1$ for all $j \geq 0$. Since $\widehat{\Delta}_{j+1}(\xi) = \widehat{\Delta}_j(\xi) - \widehat{\chi}_{\sigma_{j+1}}(\xi)\widehat{\Delta}_j(\xi)$, we get $\widehat{\Delta}_j(\xi) = \widehat{f}_0(\xi)\prod_{l=1}^j (1 - \chi_{\sigma_l}(\xi))$, by induction. Because $\sigma_{j+1} = \sigma_j/2$, and by construction of χ_{σ_l} we have $\widehat{\chi}_{\sigma_m}(\xi)\widehat{\chi}_{\sigma_l}(\xi) = \widehat{\chi}_{\sigma_m}(\xi)$ for any $m > l$, hence the last equation can be rewritten as $\widehat{\Delta}_j(\xi) = \widehat{f}_0(\xi)(1 - \widehat{\chi}_{\sigma_j}(\xi))$. Then we deduce that $\Delta_j = f_0 - \chi_{\sigma_j} * f_0$. By Lemma 1, this implies that $|\Delta_j(x)| \leq C\|f_0\|_{\mathcal{C}^\beta}\sigma_j^\beta$. From the same estimate, it is clear that $\|\Delta_j\|_1 \leq \|f_0\|_1 + \|\chi_{\sigma_j} * f_0\|_1 \leq 2\|f_0\|_1$. □

C.1. Proof of Proposition 8

Let define $A(\beta, J) := (2\log|\Lambda_J| + 2\beta\log\sigma_J^{-1})^{1/2}$ and $\mathcal{J} \equiv \mathcal{J}(x) := \{(j,k) \in \{0,\ldots,J\} \times \mathbb{Z}^d \,:\, \|x - \mu_{jk}\| \leq 4A(\beta,J)\sigma_j\}$. For any $M \in \mathcal{M}$ we can write

$$\begin{aligned}f_M(x) - f_0(x) = &\sum_{(j,k)\in\Lambda_J\cap\mathcal{J}} \int_{W_{jk}} \left[\varphi_\Sigma(x-\mu) - \varphi_{\sigma_j^2 I}(x - \mu_{jk})\right] dM(\Sigma,\mu) \\ + &\sum_{(j,k)\in\Lambda_J\cap\mathcal{J}} [M(W_{jk}) - u_{jk}]\,\varphi_{\sigma_j^2 I}(x-\mu_{jk}) + \sum_{(j,k)\in\Lambda_J\cap\mathcal{J}^c} \int_{W_{jk}} \varphi_\Sigma(x-\mu)\, dM(\Sigma,\mu) \\ &- \sum_{(j,k)\in\Lambda_J\cap\mathcal{J}^c} u_{jk}\,\varphi_{\sigma_j^2 I}(x-\mu_{jk}) - \sum_{(j,k)\notin\Lambda_J} u_{jk}\,\varphi_{\sigma_j^2 I}(x-\mu_{jk})\end{aligned}$$

$$+\int_{W^c}\varphi_\Sigma(x-\mu)\,dM(\sigma,\mu)$$
$$:= r_1(x)+r_2(x)+r_3(x)+r_4(x)+r_5(x)+r_6(x).$$

The proof follows similar steps as the proof of Proposition 5. From the definition of $A(\beta,J)$ and Proposition 6, we deduce that $A(\beta,J)\lesssim\sqrt{J}$ for J large enough. Also, there is a separation of $h_J\sigma_j$ between two consecutive μ_{jk}. Then there are no more than $16A(\beta,J)\sigma_j/(h_J\sigma_j)=16A(\beta,J)h_J^{-1}$ distinct values of μ_{jk} in an interval of length $8A(\beta,J)\sigma_j$. Thus the bound $|\Lambda_J\cap\mathcal{J}|\lesssim(J+1)A(\beta,J)^d h_J^{-d}\lesssim J^{1+d}$ holds. It follows from Proposition 11 that $|r_1(x)|\lesssim|\Lambda_J\cap\mathcal{J}|\sigma_J^{\beta+d}\lesssim J^{1+d}\sigma_J^{\beta+d}$. Obviously, $|r_2(x)|\le\|\varphi\|_\infty|\Lambda_J\cap\mathcal{J}|\sigma_J^\beta\lesssim J^{1+d}\sigma_J^\beta$. Acting as in the proof of Proposition 5, we get for any $\Sigma\in U_{jk}$ that

$$\varphi_\Sigma(x)\le\exp\left\{-\frac{1}{2\sigma_j^2}(1-\sigma_j^2\sigma_J^\beta)\|x\|^2\right\}.$$

Whenever $(j,k)\in\Lambda_J\cap\mathcal{J}^c$ and $(\Sigma,\mu)\in W_{jk}$ we have $\|\mu-\mu_{jk}\|\le\sigma_j A(\beta,J)$ for J large enough. Then $\|x-\mu\|\ge 3A(\beta,J)\sigma_j\ge\sigma_j A(\beta,J)/(1-\sigma_j^2\sigma_J^2)^{1/2}$ for J large. Therefore, $|r_3(x)|\lesssim\exp(-\frac{1}{2}A(\beta,J)^2)|\Lambda_J|\le\sigma_J^\beta$. With the same reasoning we get $|r_4(x)|\lesssim\|f_0\|_\infty\sigma_J^\beta$. Regarding r_6, we have the obvious bound $|r_6(x)|\le\|\varphi\|_\infty\sigma_J^\beta$. The r_5 term is more subtle and constitutes the remainder of the proof. Let $\Lambda^c:=\{(j,k)\in\{0,\dots,J\}\times\mathbb{Z}^d\,:\,|u_{jk}|\le\sigma_J^\beta\}$ and $\mathcal{K}_j:=\{k\in\mathbb{Z}^d\,:\,\|\mu_{jk}\|>2^{2\beta(J-j)/p}+\sigma_j\sqrt{2(\beta+d)\log\sigma_J^{-1}}\}$. Recall that $A_j:=\{x\in\mathbb{R}^d\,:\,\|x\|\le 2^{2\beta(J-j)/p}\}$. Assuming that $x\in A_q$ for some $0\le q\le J$, we can bound $r_5(x)$ as follows,

$$|r_5(x)|\le\sum_{(j,k)\in\Lambda^c}|u_{jk}|\,\varphi_{\sigma_j^2 I}(x-\mu_{jk})$$
$$+\sum_{j\le q}\sum_{k\in\mathcal{K}_j}|u_{jk}|\,\varphi_{\sigma_j^2 I}(x-\mu_{jk})+\sum_{j>q}\sum_{k\in\mathcal{K}_j}|u_{jk}|\,\varphi_{\sigma_j^2 I}(x-\mu_{jk}),\qquad(26)$$

where the third term of the rhs does not exist if $q=J$. The first term of the rhs of equation (26) is bounded by $\sigma_J^\beta\sup_{x\in\mathbb{R}^d}\sum_{j=0}^J\sum_{k\in\mathbb{Z}^d}\varphi_{\sigma_j^2 I}(x-\mu_{jk})$, which is in turn bounded by a constant multiple of $J^{1+d/2}\sigma_J^\beta$ (see for instance the proof of Lemma 8). Because of Propositions 9 and 10, when $x\in A_q$ we always have

$$\sum_{j\le q}\sum_{k\in\mathcal{K}_j}|u_{jk}|\,\varphi_{\sigma_j^2 I}(x-\mu_{jk})\le\sup_{\substack{j\le q\\ k\in\mathcal{K}_j}}\varphi_{\sigma_j^2 I}(x-\mu_{jk})\sum_{j\le J}\sum_{k\in\mathbb{Z}^d}|u_{jk}|$$
$$\le\sigma_J^{\beta+d}\sum_{j\le J}2\|f_0\|_1\sigma_j^{-d}\lesssim\sigma_J^\beta.$$

Regarding the second term of the rhs of equation (26), we introduce the sets of indexes $\mathcal{L}_j\equiv\mathcal{L}_j(x):=\{k\in\mathcal{K}_j\,:\,\|x-\mu_{jk}\|\le\sigma_j\sqrt{2(\beta+d)\log\sigma_J^{-1}}\}$. Then, we can split again the sum as

$$\sum_{j>q}\sum_{k\in\mathcal{K}_j} u_{jk}\,\varphi_{\sigma_j^2 I}(x-\mu_{jk}) = \sum_{j>q}\sum_{k\notin\mathcal{L}_j} u_{jk}\,\varphi_{\sigma_j^2 I}(x-\mu_{jk}) + \sum_{j>q}\sum_{k\in\mathcal{L}_j} u_{jk}\,\varphi_{\sigma_j^2 I}(x-\mu_{jk}).$$

With exactly the same reasoning as before, we get that the first sum of the rhs of the last expression is bounded above by a multiple constant of σ_J^β. Concerning the second term, for any $j \geq 1$ we get from Propositions 9 and 10, together with the definition of u_{jk}, that $|u_{jk}| \lesssim \|f\|_{\mathcal{C}^\beta}\sigma_j^\beta$. Since there is $h_J\sigma_j$ separation between two consecutive μ_{jk}, we deduce that $|\mathcal{L}_j| \lesssim h_J^{-d}[2(\beta+1)\log\sigma_J^{-1}]^{d/2}$. Therefore, for J large enough and $x \in A_q$ with $0 \leq q \leq J$,

$$|r_5(x)| \lesssim \sigma_J^\beta + h_J^{-d}[2(\beta+1)\log\sigma_J^{-1}]^{d/2}\sum_{j>q}\sigma_j^\beta \lesssim J^d\sigma_q^\beta.$$

The conclusion of the proposition follows by combining all the preceding points.

Appendix D: Elementary results

Proposition 11. *Let $\varphi_\Sigma(x) = \exp(-\frac{1}{2}x^T\Sigma^{-1}x)$ for $x \in \mathbb{R}^d$ and $\xi \in \mathcal{E}$. Then, for all $\mu_1, \mu_2 \in \mathbb{R}^d$, and all $\Sigma_1, \Sigma_2 \in \mathcal{E}$*

$$\sup_{x\in\mathbb{R}^d}|\varphi_{\Sigma_1}(x-\mu_1) - \varphi_{\Sigma_2}(x-\mu_2)| \lesssim \frac{\|\mu_1-\mu_2\|}{\lambda_d(\Sigma_2)} + \frac{\lambda_1(\Sigma_2)}{\lambda_d(\Sigma_2)}\|I - \Sigma_2\Sigma_1^{-1}\|.$$

Proof. Using the triangle inequality, we write

$$\begin{aligned}|\varphi_{\Sigma_1}(x-\mu_1) - \varphi_{\Sigma_2}(x-\mu_2)| \\ \leq |\varphi_{\Sigma_2}(x-\mu_1) - \varphi_{\Sigma_2}(x-\mu_2)| + |\varphi_{\Sigma_1}(x-\mu_2) - \varphi_{\Sigma_2}(x-\mu_2)|\,. \end{aligned} \quad (27)$$

We start with a bound on the second term of rhs of equation (27). For any $x \in \mathbb{R}^d$, assume first that $x^T\Sigma_1^{-1}x > x^T\Sigma_2^{-1}x$, then by a Taylor expansion of $\exp(-u)$ around $u = x^T\Sigma_1^{-1}x$,

$$\varphi_{\Sigma_1}(x) = \varphi_{\Sigma_2}(x) - \varphi_{\Sigma_2}(x)\int_0^{x^T(\Sigma_1^{-1}-\Sigma_2^{-1})x}\exp(-t)[x^T(\Sigma_1^{-1}-\Sigma_2^{-1})x - t]\,dt.$$

Hence,

$$|\varphi_{\Sigma_1}(x) - \varphi_{\Sigma_2}(x)| \leq \varphi_{\Sigma_2}(x)|x^T(\Sigma_1^{-1}-\Sigma_2^{-1})x| \leq \|\Sigma_1^{-1}-\Sigma_2^{-1}\|\|x\|^2\varphi_{\Sigma_2}(x).$$

Note that Σ_2^{-1} is positive-definite and symmetric, hence $\Sigma_2^{-1} = Q^TDQ$ for some diagonal matrix D and an orthogonal matrix Q. It follows $\|x\|^2\varphi_{\Sigma_2}(x) = \|Q^TQx\|^2\exp(-(Qx)^TD(Qx))$ and thus,

$$\begin{aligned}\sup_{x\in\mathbb{R}^d}\|x\|^2\varphi_{\Sigma_2}(x) &= \sup_{x\in\mathbb{R}^d}\|x\|^2\exp(-x^TDx) \\ &\leq d\sup_{x\in\mathbb{R}}x^2\exp\left(-\frac{\lambda_d(\Sigma_2^{-1})}{2}x^2\right) \leq \frac{2de^{-1}}{\lambda_d(\Sigma_2^{-1})}.\end{aligned}$$

The conclusion follows because $\|\Sigma_1^{-1} - \Sigma_2^{-1}\| \leq \|\Sigma_2^{-1}\|\|I - \Sigma_2\Sigma_1^{-1}\| \leq \lambda_1(\Sigma_2^{-1})\|I - \Sigma_2\Sigma_1^{-1}\|$. The case $x^T\Sigma_1^{-1}x < x^T\Sigma_2^{-1}x$ is handled similarly, while $x^T\Sigma_1^{-1}x = x^T\Sigma_2^{-1}x$ is trivial. The first term of equation (27) is bounded similarly. □

Proposition 12. *Let $X \sim \mathrm{SGa}(\alpha, 1)$, with $0 < \alpha \leq 1$. Then for any $x \in \mathbb{R}$ and any $0 < \delta \leq 1/2$ we have $\Pr\{|X - x| \leq \delta\} \geq \frac{\delta e^{-2|x|}}{3e\Gamma(\alpha)}$.*

Proof. Assume for instance that $x \geq 0$. Recalling that X is distributed as the difference of two independent $\mathrm{Ga}(\alpha, 1)$ distributed random variables, it follows

$$\Pr\{|X - x| \leq \delta\} \geq \frac{1}{\Gamma(\alpha)} \int_0^\infty y^{\alpha-1}e^{-y} \frac{1}{\Gamma(\alpha)} \int_{x+y}^{x+y+\delta} z^{\alpha-1}e^{-z}\, dz\, dy.$$

Because $\alpha \leq 1$, the mapping $z \mapsto z^{\alpha-1}e^{-z}$ is monotonically decreasing on $\mathbb{R}^+$, then the last integral in the rhs of the previous equation is lower bounded by $\delta(x + y + \delta)^{\alpha-1}e^{-(x+y+\delta)} \geq \delta e^{-2(x+y+\delta)}$. Then

$$\Pr\{|X - x| \leq \delta\} \geq \frac{\delta e^{-2(x+\delta)}}{\Gamma(\alpha)^2} \int_0^\infty y^{\alpha-1}e^{-3y}\, dy \geq \frac{3^{-\alpha}e^{-2(x+\delta)}}{\Gamma(\alpha)}\delta \geq \frac{\delta e^{-2|x|}}{3e\Gamma(\alpha)}.$$

The proof when $x < 0$ is obvious. □

Acknowledgments

Part of this work has been supported by the BNPSI ANR project no ANR-13-BS-03-0006-01, the ANR-11-BS01-0010 grant "Calibration", and the "Chaire Havas" at University Paris Dauphine. The first author also acknowledge the support of the French Commissariat à l'Energie Atomique (CEA).

References

Barndorff-Nielsen, O., Blaesild, P., Jensen, J. L. and Jorgensen, B. (1982). Exponential Transformation Models. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **379** 41–65. MR0643215

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'IHP Probabilités et statistiques* **42** 273–325. MR2219712

Bochkina, N. and Rousseau, J. (2016). Adaptive density estimation based on a mixture of Gammas. *ArXiv e-prints*. MR3629019

Canale, A. and De Blasi, P. (2017). Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli* **23** 379–404. MR3556776

de Jonge, R. and van Zanten, J. H. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *The Annals of Statistics* **38** 3300–3320. MR2766853

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 209–230. MR0350949

GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* **28** 500–531. MR1790007

GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics* 1233–1263. MR1873329

GHOSAL, S. and VAN DER VAART, A. W. (2007a). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics* **35** 697–723. MR2336864

GHOSAL, S. and VAN DER VAART, A. W. (2007b). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics* **35** 192–223. MR2332274

GOLDENSHLUGER, A. and LEPSKI, O. (2014). On adaptive minimax density estimation on R^d. *Probability Theory and Related Fields* **159** 479–543. MR3230001

HANGELBROEK, T. and RON, A. (2010). Nonlinear approximation using Gaussian kernels. *Journal of Functional Analysis* **259** 203–219. MR2610384

HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (1998). Wavelets. In *Wavelets, Approximation, and Statistical Applications* 1–16. Springer. MR1618204

HJORT, N. L., HOLMES, C., MÜLLER, P. and WALKER, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, UK. MR2722988

JUDITSKY, A., LAMBERT-LACROIX, S. et al. (2004). On minimax density estimation on $\mathbb{R}$. *Bernoulli* **10** 187–220. MR2046772

KINGMAN, J. F. C. (1992). *Poisson processes* **3**. Oxford university press. MR1207584

KRUIJER, W., ROUSSEAU, J. and VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. MR2735885

LIJOI, A., PRÜNSTER, I. and WALKER, S. G. (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *Journal of the American Statistical Association* **100** 1292–1296. MR2236442

NAULET, Z. and BARAT, E. (2015). Some aspects of symmetric Gamma process mixtures. *arXiv preprint arXiv:1504.00476*.

REYNAUD-BOURET, P., RIVOIRARD, V. and TULEAU-MALOT, C. (2011). Adaptive density estimation: a curse of support? *Journal of Statistical Planning and Inference* **141** 115–139. MR2719482

SALOMOND, J.-B. (2013). Bayesian testing for embedded hypotheses with application to shape constrains. *arXiv preprint arXiv:1303.6466*.

SCRICCIOLO, C. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Analysis* **9** 475–520. MR3217004

SHEN, W., TOKDAR, S. T. and GHOSAL, S. (2013). Adaptive Bayesian multi-

variate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* **101** 1566–1581. MR2279480

Wolpert, R. L., Clyde, M. A. and Tu, C. (2011). Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels. *The Annals of Statistics* 1916–1962. MR2893857