Multivariate Intensity Estimation via Hyperbolic Wavelet Selection

Nathalie Akakpo

Multivariate Intensity Estimation via Hyperbolic Wavelet Selection

NATHALIE AKAKPO

Laboratoire de Probabilités et Modèles Aléatoires (LPMA), UMR 7599
Université Pierre et Marie Curie (UPMC), Paris

Centre de Recherches Mathématiques (CRM), UMI 3457
Université de Montréal (UdeM)

Abstract. We propose a new statistical procedure able in some way to overcome the curse of dimensionality without structural assumptions on the function to estimate. It relies on a least-squares type penalized criterion and a new collection of models built from hyperbolic biorthogonal wavelet bases. We study its properties in a unifying intensity estimation framework, where an oracle-type inequality and adaptation to mixed smoothness are shown to hold. Besides, we describe an algorithm for implementing the estimator with a quite reasonable complexity.

Keywords: Hyperbolic wavelets; Biorthogonal wavelets; Mixed smoothness; Model selection; Density; Copula; Poisson process; Lévy process.

Contents

E-mail address: nathalie.akakpo@upmc.fr.
Date: November 23, 2016.

1. Introduction

Over the last decades, many wavelet procedures have been developed in various statistical frameworks. Yet, in multivariate settings, most of them are based on isotropic wavelet bases. These indeed have the advantage of being as easily tractable as their univariate counterparts since each isotropic wavelet is a tensor product of univariate wavelets coming from the same resolution level. Notable counterexamples are [Don97], [Neu00] and [NvS97], or [ACF15] and [ACF14]. They underline the usefulness of hyperbolic wavelet bases, where coordinatewise varying resolution levels are allowed, so as to recover a wider range of functions, and in particular functions with anisotropic smoothness.

Much attention has also been paid to the so-called curse of dimensionality. A common way to overcome this problem in Statistics is to impose structural assumptions on the function to estimate. In a regression framework, beyond the well-known additive and single-index models, we may cite the work of [HM07] who propose a spline-based method in an additive model with unknown link function, or the use of ANOVA-like decompositions in [IS07] or [DIT14]. Besides, two landmark papers consider a general framework of composite functions, encompassing several classical structural assumptions: [JLT09] propose a kernel-based procedure in the white noise framework, whereas [BB14] propose a general model selection procedure with a wide scope of applications. Finally, Lepski [Lep13] (see also [Reb15b, Reb15a]) consider density estimation with adaptation to a possibly multiplicative structure of the density. In the meanwhile, in the field of Approximation Theory and Numerical Analysis, a renewed interest in function spaces with dominating mixed smoothness has been growing (see for instance [DTU16]), due to their tractability for multivariate integration for instance. Such spaces do not impose any structure, but only that the highest order derivative is a mixed derivative. Surprisingly, in the statistical literature, it seems that only the thresholding-type procedures of [Neu00] and [BPP13] deal with such spaces, either in the white noise framework or in a functional deconvolution model.

In order to fill this gap, this paper is devoted to a new statistical procedure based on wavelet selection from hyperbolic biorthogonal bases. We underline its universality by studying it in a general intensity estimation framework, encompassing many examples of interest such as density, copula density, Poisson intensity or Lévy jump intensity estimation. We first define a whole collection of linear subspaces, called models, generated by subsets of the dual hyperbolic basis, and a least-squares type criterion adapted to the norm induced by the primal hyperbolic basis. Then we describe a procedure to choose the best model from the data by using a penalized approach similar to [BBM99]. Our procedure satisfies an oracle-type inequality provided the intensity to estimate is bounded. Besides, it reaches the minimax rate up to a constant factor, or up to a logarithmic factor, over a wide range of spaces with dominating mixed smoothness, and this rate is akin to the one we would obtain in a univariate framework. Notice that, contrary to [Neu00] or [BPP13], we allow for a greater variety of such spaces (of Sobolev, Hölder or Besov type smoothness) and also for spatially nonhomogeneous smoothness. For that purpose, we prove a key result from nonlinear approximation theory, in the spirit of [BM00], that may be of interest for other types of model selection procedures (see for instance [Bir06, Bar11, BB16]). Depending on the kind of intensity to estimate, different structural assumptions might make sense, some of which have been considered in [JLT09], [BB14], [Lep13], [Reb15b, Reb15a], but not all. We explain in what respect these structural assumptions fall within the scope of estimation under dominating mixed smoothness. Yet, we emphasize that we do not need to impose any structural assumptions

on the target function. Thus in some way our method is adaptive at the same time to many structures. Besides, it can be implemented with a computational complexity linear in the sample size, up to logarithmic factors.

The plan of the paper is as follows. In Section 2, we describe the general intensity estimation framework and several examples of interest. In Section 3, we define the so-called pyramidal wavelet models and a least-squares type criterion, and provide a detailed account of estimation on a given model. Section 4 is devoted to the choice of an adequate penalty so as to perform data-driven model selection. The optimality of the resulting procedure from the minimax point of view is then discussed in Section 5, under mixed smoothness assumptions. The algorithm for implementing our wavelet procedure and an illustrative example are given in Section 6. All proofs are postponed to Section 7. Let us end with some remark about the notation. Throughout the paper, $C, C_1, \ldots$ will stand for numerical constants, and $C(\theta), C_1(\theta), \ldots$ for positive reals that only depend on some θ. Their values are allowed to change from line to line.

2. Framework and examples

2.1. **General framework.** Let $d \in \mathbb{N}, d \geq 2$, and $Q = \prod_{k=1}^{d} [a_k, b_k]$ be a given hyperrectangle in $\mathbb{R}^d$ equipped with its Borel $\sigma-$algebra $\mathcal{B}(Q)$ and the Lebesgue measure. We denote by $\mathbb{L}^2(Q)$ the space of square integrable functions on Q, equipped with its usual norm

$$\|t\| = \sqrt{\int_Q t^2(x)\mathrm{d}x} \tag{1}$$

and scalar product $\langle ., . \rangle$. In this article, we are interested in a nonnegative measure on $\mathcal{B}(Q)$ that admits a bounded density s with respect to the Lebesgue measure, and our aim is to estimate that function s over Q. Given a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, we assume that there exists some random measure M defined on $(\Omega, \mathcal{E}, \mathbb{P})$, with values in the set of Borel measures on Q such that, for all $A \in \mathcal{B}(Q)$,

$$\mathbb{E}\left[M(A)\right] = \langle \mathbb{1}_A, s \rangle. \tag{2}$$

By classical convergence theorems, this condition implies that, for all nonnegative or bounded measurable functions t,

$$\mathbb{E}\left[\int_Q t\mathrm{d}M\right] = \langle t, s \rangle. \tag{3}$$

We assume that we observe some random measure $\widehat{M}$, which is close enough to M in a sense to be made precise later. When M can be observed, we set of course $\widehat{M} = M$.

2.2. **Examples.** Our general framework encompasses several special frameworks of interest, as we shall now show.

2.2.1. *Example 1: density estimation.* Given $n \in \mathbb{N}^\star$, we observe identically distributed random variables $Y_1, \ldots, Y_n$ with common density s with respect to the Lebesgue measure on $Q = \prod_{k=1}^{d} [a_k, b_k]$. The observed empirical measure is then given by

$$\widehat{M}(A) = M(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(Y_i), \text{ for } A \in \mathcal{B}(Q),$$

and obviously satisfies (2).

2.2.2. *Example 2: copula density estimation.* Given $n \in \mathbb{N}^\star$, we observe independent and identically distributed random variables $X_1, \ldots, X_n$ with values in $\mathbb{R}^d$. For $i = 1, \ldots, n$ and $j = 1, \ldots, d$, the j-th coordinate X_{ij} of X_i has continuous distribution function F_j. We recall that, from Sklar's Theorem [Skl59] (see also [Nel06], for instance), there exists a unique distribution function C on $[0,1]^d$ with uniform marginals such that, for all $(x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$\mathbb{P}(X_{i1} \leq x_1, \ldots, X_{id} \leq x_d) = C(F_1(x_1), \ldots, F_d(x_d)).$$

This function C is called the copula of $X_{i1}, \dots, X_{id}$. We assume that it admits a density s with respect to the Lebesgue measure on $Q = [0,1]^d$. Since C is the joint distribution function of the $F_j(X_{1j}), j = 1, \dots, d$, a random measure satisfying (2) is given by

$$M(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A \left(F_1(X_{i1}), \dots, F_d(X_{id}) \right), \text{ for } A \in \mathcal{B}([0,1]^d).$$

As the marginal distributions F_j are usually unknown, we replace them by the empirical distribution functions $\hat{F}_{nj}$, where

$$\hat{F}_{nj}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_{ij} \leq t},$$

and define

$$\widehat{M}(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A \left(\hat{F}_{n1}(X_{i1}), \dots, \hat{F}_{nd}(X_{id}) \right), \text{ for } A \in \mathcal{B}([0,1]^d).$$

2.2.3. *Example 3: Poisson intensity estimation.* Let us denote by $\mathrm{Vol}_d(Q)$ the Lebesgue measure of $Q = \prod_{k=1}^{d} [a_k, b_k]$. We observe a Poisson process N on Q whose mean measure has intensity $\mathrm{Vol}_d(Q)s$. Otherwise said, for all finite family $(A_k)_{1 \leq k \leq K}$ of disjoint measurable subsets of Q, $N(A_1), \dots, N(A_K)$ are independent Poisson random variables with respective parameters $\mathrm{Vol}_d(Q) \int_{A_1} s, \dots, \mathrm{Vol}_d(Q) \int_{A_K} s$. Therefore the empirical measure

$$\widehat{M}(A) = M(A) = \frac{N(A)}{\mathrm{Vol}_d(Q)}, \text{ for } A \in \mathcal{B}(Q),$$

does satisfy (2). We do not assume s to be constant throughout Q so that the Poisson process may be nonhomogeneous.

2.2.4. *Example 4: Lévy jump intensity estimation (continuous time).* Let T be a fixed positive real, we observe on $[0,T]$ a Lévy process $\mathbf{X} = (X_t)_{t \geq 0}$ with values in $\mathbb{R}^d$. Otherwise said, $\mathbf{X}$ is a process starting at 0, with stationary and independent increments, and which is continuous in probability with càdlàg trajectories (see for instance [Ber96, Sat99, CT04]). This process may have jumps, whose sizes are ruled by the so-called jump intensity measure or Lévy measure. An important example of such process is the compound Poisson process

$$X_t = \sum_{i=1}^{N_t} \xi_i, t \geq 0,$$

where $(N_t)_{t \geq 0}$ is a univariate homogeneous Poisson process, $(\xi_i)_{i \geq i}$ are i.i.d. with values in $\mathbb{R}^d$ and distribution ρ with no mass at 0, and $(N_t)_{t \geq 0}$ and $(\xi_i)_{i \geq i}$ are independent. In this case, ρ is also the Lévy measure of $\mathbf{X}$.

Here, we assume that the Lévy measure admits a density f with respect to the Lebesgue measure on $\mathbb{R}^d \backslash \{0\}$. Given some compact hyperrectangle $Q = \prod_{k=1}^{d} [a_k, b_k] \subset \mathbb{R}^d \backslash \{0\}$, our aim is to estimate the restriction s of f to Q. For that purpose, we use the observed empirical measure

$$\widehat{M}(A) = M(A) = \frac{1}{T} \iint_{[0,T] \times A} N(\mathrm{d}t, \mathrm{d}x), \text{ for } A \in \mathcal{B}(Q).$$

A well-known property of Lévy processes states that the random measure N defined for $B \in \mathcal{B}\left([0, +\infty) \times \mathbb{R}^d \backslash \{0\}\right)$ by

$$N(B) = \sharp\{t > 0 / (t, X_t - X_{t^-}) \in B\}$$

is a Poisson process with mean measure

$$\mu(B) = \int \int_B f(x) \mathrm{d}t \mathrm{d}x,$$

so that M satisfies (2).

2.2.5. *Example 5: Lévy jump intensity estimation (discrete time).* The framework is the same as in Example 4, except that $(X_t)_{t\geq 0}$ is not observed. Given some time step $\Delta > 0$ and $n \in \mathbb{N}^\star$, we only have at our disposal the random variables

$$Y_i = X_{i\Delta} - X_{(i-1)\Delta}, i = 1, \dots n.$$

In order to estimate s on Q, we consider the random measure

$$M(A) = \frac{1}{n\Delta} \iint_{[0,n\Delta]\times A} N(\mathrm{d}t, \mathrm{d}x), \text{ for } A \in \mathcal{B}(Q),$$

which is unobserved, and replaced for estimation purpose with

$$\widehat{M}(A) = \frac{1}{n\Delta} \sum_{i=1}^{n} \mathbb{1}_A(Y_i), \text{ for } A \in \mathcal{B}(Q).$$

3. Estimation on a given pyramidal wavelet model

The first step of our estimation procedure relies on the definition of finite dimensional linear subspaces of $\mathbb{L}^2(Q)$, called models, generated by some finite families of biorthogonal wavelets. We only describe here models for $Q = [0,1]^d$. For a general hyperrectangle Q, the adequate models can be deduced by translation and scaling. We then introduce a least-squares type contrast that allows to define an estimator of s within a given wavelet model.

3.1. **Wavelets on $\mathbb{L}_2([0,1])$.** We shall first introduce a multiresolution analysis and a wavelet basis for $\mathbb{L}_2([0,1])$ satisfying the same general assumptions as in [Hoc02b] and [Hoc02a]. Concrete examples of wavelet bases satisfying those assumptions may be found in [CDV93] and [DKU99] for instance. In the sequel, we denote by κ some positive constant, that only depends on the choice of the bases. We fix the coarsest resolution level at $j_0 \in \mathbb{N}$. On the one hand, we assume that the scaling spaces

$$V_j = \mathrm{Vect}\{\phi_\lambda; \lambda \in \Delta_j\} \text{ and } V_j^\star = \mathrm{Vect}\{\phi_\lambda^\star; \lambda \in \Delta_j\}, j \geq j_0,$$

satisfy the following hypotheses:

- *S.i)* (Riesz bases) For all $j \geq j_0$, $\{\phi_\lambda; \lambda \in \Delta_j\}$ are linearly independent functions from $\mathbb{L}_2([0,1])$, so are $\{\phi_\lambda^\star; \lambda \in \Delta_j\}$, and they form Riesz bases of V_j and $V_j^\star$, *i.e.* $\left\|\sum_{\lambda\in\Delta_j} a_\lambda \phi_\lambda\right\| \sim \left(\sum_{\lambda\in\Delta_j} a_\lambda^2\right)^{1/2} \sim \left\|\sum_{\lambda\in\Delta_j} a_\lambda \phi_\lambda^\star\right\|$.
- *S.ii)* (Dimension) There exists some nonnegative integer B such that, for all $j \geq j_0$, $\dim(V_j) = \dim(V_j^\star) = \sharp\Delta_j = 2^j + B$.
- *S.iii)* (Nesting) For all $j \geq j_0$, $V_j \subset V_{j+1}$ and $V_j^\star \subset V_{j+1}^\star$.
- *S.iv)* (Density) $\overline{\cup_{j\geq j_0} V_j} = \overline{\cup_{j\geq j_0} V_j^\star} = \mathbb{L}_2([0,1])$.
- *S.v)* (Biorthogonality) Let $j \geq j_0$, for all $\lambda, \mu \in \Delta_j$, $\langle \phi_\lambda, \phi_\mu^\star \rangle = \delta_{\lambda,\mu}$.
- *S.vi)* (Localization) Let $j \geq j_0$, for all $\lambda \in \Delta_j$, $|\mathrm{Supp}(\phi_\lambda)| \sim |\mathrm{Supp}(\phi_\lambda^\star)| \sim 2^{-j}$.
- *S.vii)* (Almost disjoint supports) For all $j \geq j_0$ and all $\lambda \in \Delta_j$,

$$\max\left(\sharp\{\mu \in \Delta_j \text{ s.t. } \mathrm{Supp}(\phi_\lambda) \cap \mathrm{Supp}(\phi_\mu) \neq \varnothing\}, \sharp\{\mu \in \Delta_j \text{ s.t. } \mathrm{Supp}(\phi_\lambda^\star) \cap \mathrm{Supp}(\phi_\mu^\star) \neq \varnothing\}\right) \leq \kappa.$$

- *S.viii)* (Norms) For all $j \geq j_0$ and all $\lambda \in \Delta_j$, $\|\phi_\lambda\| = \|\phi_\lambda^\star\| = 1$ and $\max(\|\phi_\lambda\|_\infty, \|\phi_\lambda^\star\|_\infty) \leq \kappa 2^{j/2}$.
- *S.ix)* (Polynomial reproducibility) The primal scaling spaces are exact of order N, *i.e.* for all $j \geq j_0$, $\Pi_{N-1} \subset V_j$, where Π_{N-1} is the set of all polynomial functions with degree $\leq N-1$ over $[0,1]$.

On the other hand, the wavelet spaces

$$W_j = \mathrm{Vect}\{\psi_\lambda; \lambda \in \nabla_j\} \text{ and } W_j^\star = \mathrm{Vect}\{\psi_\lambda^\star; \lambda \in \nabla_j\}, j \geq j_0 + 1,$$

fulfill the following conditions:

- *W.i)* (Riesz bases) The functions $\{\psi_\lambda; \lambda \in \cup_{j\geq j_0+1} \nabla_j\}$ are linearly independent. Together with the $\{\phi_\lambda; \lambda \in \Delta_{j_0}\}$, they form a Riesz basis for $\mathbb{L}_2([0,1])$. The same holds for the $\psi^\star$ and the $\phi^\star$.

W.ii) (Orthogonality) For all $j \geq j_0$, $V_{j+1} = V_j \oplus W_{j+1}$ and $V_{j+1}^\star = V_j^\star \oplus W_{j+1}^\star$, with $V_j \perp W_{j+1}^\star$ and $V_j^\star \perp W_{j+1}$.

W.iii) (Biorthogonality) Let $j \geq j_0$, for all $\lambda, \mu \in \nabla_{j+1}$, $\langle \psi_\lambda, \psi_\mu^\star \rangle = \delta_{\lambda,\mu}$.

W.iv) (Localization) Let $j \geq j_0$, for all $\lambda \in \nabla_{j+1}$, $|\mathrm{Supp}(\psi_\lambda)| \sim |\mathrm{Supp}(\psi_\lambda^\star)| \sim 2^{-j}$.

W.v) (Almost disjoint supports) For all $j \geq j_0$ and all $\lambda \in \nabla_{j+1}$,

$$\max\left(\sharp\{\mu \in \nabla_{j+1} \text{ s.t. } \mathrm{Supp}(\psi_\lambda) \cap \mathrm{Supp}(\psi_\mu) \neq \varnothing\}, \sharp\{\mu \in \nabla_{j+1} \text{ s.t. } \mathrm{Supp}(\psi_\lambda^\star) \cap \mathrm{Supp}(\psi_\mu^\star) \neq \varnothing\}\right) \leq \kappa.$$

W.vi) (Norms) For all $j \geq j_0$ and all $\lambda \in \nabla_{j+1}$, $\|\psi_\lambda\| = \|\psi_\lambda^\star\| = 1$ and $\max(\|\psi_\lambda\|_\infty, \|\psi_\lambda^\star\|_\infty) \leq \kappa 2^{j/2}$.

W.vii) (Fast Wavelet Transform) Let $j \geq j_0$, for all $\lambda \in \nabla_{j+1}$,

$$\sharp\{\mu \in \Delta_{j+1} | \langle \psi_\lambda, \phi_\mu \rangle \neq 0\} \leq \kappa$$

and for all $\mu \in \Delta_{j+1}$

$$|\langle \psi_\lambda, \phi_\mu \rangle| \leq \kappa.$$

The same holds for the $\psi_\lambda^\star$ and the $\phi_\lambda^\star$.

Remarks:

- These properties imply that any function $f \in \mathbb{L}_2([0,1])$ may be decomposed as

$$f = \sum_{\lambda \in \Delta_{j_0}} \langle f, \phi_\lambda \rangle \phi_\lambda^\star + \sum_{j \geq j_0+1} \sum_{\lambda \in \nabla_j} \langle f, \psi_\lambda \rangle \psi_\lambda^\star. \tag{4}$$

- Properties *S.ii)* and *W.ii)* imply that $\dim(W_{j+1}) = 2^j$.
- Property *W.vii)* means in particular that, for each resolution level j, any wavelet can be represented as a linear combination of scaling functions from the same resolution level with a number of components bounded independently of the level as well as the amplitude of the coefficients.

As is well known, contrary to orthogonal bases, biorthogonal bases allow for both symmetric and smooth wavelets. Besides, properties of dual biorthogonal bases are usually not the same. Usually, in decomposition (4), the analysis wavelets ϕ_λ and ψ_λ are the one with most null moments, whereas the synthesis wavelets $\phi_\lambda^\star$ and $\psi_\lambda^\star$ are the one with greatest smoothness. Yet, we may sometimes need the following smoothness assumptions on the analysis wavelets (not very restrictive in practice), only to bound residual terms due to the replacement of M with $\widehat{M}$.

Assumption (L). *For all $\lambda \in \Delta_{j_0}$, for all $j \geq j_0$ and all $\mu \in \nabla_{j+1}$, ϕ_λ and ψ_μ are Lipschitz functions with Lipschitz norms satisfying $\|\phi_\lambda\|_L \leq \kappa 2^{3j_0/2}$ and $\|\psi_\mu\|_L \leq \kappa 2^{3j/2}$.*

We still refer to [CDV93] and [DKU99] for examples of wavelet bases satisfying this additional assumption.

3.2. **Hyperbolic wavelet basis on $\mathbb{L}_2([0,1]^d)$.** In the sequel, for ease of notation, we set $\mathbb{N}_{j_0} = \{j \in \mathbb{N}, j \geq j_0\}$, $\nabla_{j_0} = \Delta_{j_0}$, $W_{j_0} = V_{j_0}$ and $W_{j_0}^\star = V_{j_0}^\star$, and for $\lambda \in \nabla_{j_0}$, $\psi_\lambda = \phi_\lambda$ and $\psi_\lambda^\star = \phi_\lambda^\star$. Given a biorthogonal basis of $\mathbb{L}_2([0,1])$ chosen according to 3.1, we deduce biorthogonal wavelets of $\mathbb{L}_2([0,1]^d)$ by tensor product. More precisely, for $\boldsymbol{j} = (j_1, \ldots, j_d) \in \mathbb{N}_{j_0}^d$, we set $\boldsymbol{\nabla_j} = \nabla_{j_1} \times \ldots \times \nabla_{j_d}$ and for all $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d) \in \boldsymbol{\nabla_j}$, we define $\Psi_{\boldsymbol{\lambda}}(x_1, \ldots, x_d) = \psi_{\lambda_1}(x_1) \ldots \psi_{\lambda_d}(x_d)$ and $\Psi_{\boldsymbol{\lambda}}^\star(x_1, \ldots, x_d) = \psi_{\lambda_1}^\star(x_1) \ldots \psi_{\lambda_d}^\star(x_d)$. Contrary to most statistical works based on wavelets, we thus allow for tensor products of univariate wavelets coming from different resolution levels $j_1, \ldots, j_d$. Writing $\Lambda = \cup_{\boldsymbol{j} \in \mathbb{N}_{j_0}^d} \boldsymbol{\nabla_j}$, the families $\{\Psi_{\boldsymbol{\lambda}}; \boldsymbol{\lambda} \in \Lambda\}$ and $\{\Psi_{\boldsymbol{\lambda}}^\star; \boldsymbol{\lambda} \in \Lambda\}$ define biorthogonal bases of $\mathbb{L}_2([0,1]^d)$ called biorthogonal hyperbolic bases. Indeed,

$$\mathbb{L}_2([0,1]^d) = \overline{\cup_{j \geq j_0} V_j \otimes \ldots \otimes V_j},$$

and for all $j \geq j_0$,

$$\begin{aligned} V_j \otimes \ldots \otimes V_j &= (W_{j_0} \oplus W_{j_0+1} \oplus \ldots \oplus W_j) \otimes \ldots \otimes (W_{j_0} \oplus W_{j_0+1} \oplus \ldots \oplus W_j) \\ &= \bigoplus_{j_0 \leq k_1, \ldots, k_d \leq j} W_{k_1} \otimes \ldots \otimes W_{k_d}. \end{aligned}$$

In the same way,

$$\mathbb{L}_2([0,1]^d) = \overline{\cup_{j\geq j_0} \bigoplus_{j_0\leq k_1,\ldots,k_d\leq j} W^\star_{k_1}\otimes\ldots\otimes W^\star_{k_d}}.$$

Besides, they induce on $\mathbb{L}_2([0,1]^d)$ the norms

$$\|t\|_{\boldsymbol{\Psi}} = \sqrt{\sum_{\boldsymbol{\lambda}\in\Lambda}\langle t,\Psi_{\boldsymbol{\lambda}}\rangle^2} \text{ and } \|t\|_{\boldsymbol{\Psi}^\star} = \sqrt{\sum_{\boldsymbol{\lambda}\in\Lambda}\langle t,\Psi^\star_{\boldsymbol{\lambda}}\rangle^2}, \tag{5}$$

which are both equivalent to $\|.\|$, with equality when the wavelet basis is orthogonal. It should be noticed that the scalar product derived from $\|.\|_{\boldsymbol{\Psi}}$, for instance, is

$$\langle t,u\rangle_{\boldsymbol{\Psi}} = \sum_{\boldsymbol{\lambda}\in\Lambda}\langle t,\Psi_{\boldsymbol{\lambda}}\rangle\langle u,\Psi_{\boldsymbol{\lambda}}\rangle. \tag{6}$$

3.3. **Pyramidal models.** A wavelet basis in dimension 1 has a natural pyramidal structure when the wavelets are grouped according to their resolution level. A hyperbolic basis too, provided we define a proper notion of resolution level that takes into account anisotropy: for a wavelet $\Psi_{\boldsymbol{\lambda}}$ or $\Psi^\star_{\boldsymbol{\lambda}}$ with $\boldsymbol{\lambda}\in\boldsymbol{\nabla}_{\boldsymbol{j}}$, we define the global resolution level as $|\boldsymbol{j}| := j_1+\ldots+j_d$. Thus, the supports of all wavelets corresponding to a given global resolution level $\ell\in\mathbb{N}_{dj_0}$ have a volume of roughly $2^{-\ell}$ but exhibit very different shapes. For all $\ell\in\mathbb{N}_{dj_0}$, we define $\boldsymbol{J}_\ell = \{\boldsymbol{j}\in\mathbb{N}^d_{j_0}/|\boldsymbol{j}|=\ell\}$, and $U\boldsymbol{\nabla}(\ell) = \cup_{\boldsymbol{j}\in\boldsymbol{J}_\ell}\boldsymbol{\nabla}_{\boldsymbol{j}}$ the index set for d-variate wavelets at resolution level ℓ.

Given some maximal resolution level $L_\bullet\in\mathbb{N}_{dj_0}$, we define, for all $\ell_1\in\{dj_0+1,\ldots,L_\bullet+1\}$, the family $\mathcal{M}^{\mathcal{P}}_{\ell_1}$ of all sets m of the form

$$m = \left(\bigcup_{\ell=dj_0}^{\ell_1-1} U\boldsymbol{\nabla}(\ell)\right)\cup\left(\bigcup_{k=0}^{L_\bullet-\ell_1} m(\ell_1+k)\right),$$

where, for all $0\leq k\leq L_\bullet-\ell_1$, $m(\ell_1+k)$ may be any subset of $U\boldsymbol{\nabla}(\ell_1+k)$ with $N(\ell_1,k)$ elements. Typically, $N(\ell_1,k)$ will be chosen so as to impose some sparsity: it is expected to be smaller than the total number of wavelets at level ℓ_1+k and to decrease when the resolution level increases. An adequate choice of $N(\ell_1,k)$ will be proposed in Proposition 4. Thus, choosing a set in $\mathcal{M}^{\mathcal{P}}_{\ell_1}$ amounts to keep all hyperbolic wavelets at level at most ℓ_1-1, but only a few at deeper levels. We set $\mathcal{M}^{\mathcal{P}} = \cup_{\ell_1=dj_0+1}^{L_\bullet+1}\mathcal{M}^{\mathcal{P}}_{\ell_1}$ and define a pyramidal model as any finite dimensional subspace of the form

$$S^\star_m = \text{Vect}\{\Psi^\star_{\boldsymbol{\lambda}};\boldsymbol{\lambda}\in m\},\ \text{for } m\in\mathcal{M}^{\mathcal{P}}.$$

We denote by D_m the dimension of $S^\star_m$. Setting $m_\bullet = \bigcup_{\ell=dj_0}^{L_\bullet} U\boldsymbol{\nabla}(\ell)$, we can see that all pyramidal models are included in $S^\star_{m_\bullet}$.

3.4. **Least-squares type estimator on a pyramidal model.** Let us fix some model $m\in\mathcal{M}^{\mathcal{P}}$. If the random measure M is observed, then we can build a least-squares type estimator $\check{s}^\star_m$ for s with values in $S^\star_m$ and associated with the norm $\|.\|_{\boldsymbol{\Psi}}$ defined by (5). Indeed, setting

$$\gamma(t) = \|t\|^2_{\boldsymbol{\Psi}} - 2\sum_{\boldsymbol{\lambda}\in\Lambda}\langle t,\Psi_{\boldsymbol{\lambda}}\rangle\check{\beta}_{\boldsymbol{\lambda}},$$

where

$$\check{\beta}_{\boldsymbol{\lambda}} = \int_Q \Psi_{\boldsymbol{\lambda}}\mathrm{d}M,$$

we deduce from (3) that s minimizes over $t\in\mathbb{L}^2(Q)$

$$\|s-t\|^2_{\boldsymbol{\Psi}} - \|s\|^2_{\boldsymbol{\Psi}} = \|t\|^2_{\boldsymbol{\Psi}} - 2\sum_{\boldsymbol{\lambda}\in\Lambda}\langle t,\Psi_{\boldsymbol{\lambda}}\rangle\langle s,\Psi_{\boldsymbol{\lambda}}\rangle = \mathbb{E}[\gamma(t)],$$

so we introduce

$$\check{s}^\star_m = \underset{t\in S^\star_m}{\text{argmin}}\ \gamma(t).$$

For all sequences of reals $(\alpha_{\boldsymbol{\lambda}})_{\boldsymbol{\lambda}\in m}$,

$$\gamma\left(\sum_{\boldsymbol{\lambda}\in m}\alpha_{\boldsymbol{\lambda}}\Psi^{\star}_{\boldsymbol{\lambda}}\right)=\sum_{\boldsymbol{\lambda}\in m}\left(\alpha_{\boldsymbol{\lambda}}-\check{\beta}_{\boldsymbol{\lambda}}\right)^2-\sum_{\boldsymbol{\lambda}\in m}\check{\beta}^2_{\boldsymbol{\lambda}}, \tag{7}$$

hence

$$\check{s}^{\star}_m=\sum_{\boldsymbol{\lambda}\in m}\check{\beta}_{\boldsymbol{\lambda}}\Psi^{\star}_{\boldsymbol{\lambda}}.$$

Since we only observe the random measure $\widehat{M}$, we consider the pseudo-least-squares contrast

$$\widehat{\gamma}(t)=\|t\|^2_{\boldsymbol{\Psi}}-2\sum_{\boldsymbol{\lambda}\in\Lambda}\langle t,\Psi_{\boldsymbol{\lambda}}\rangle\widehat{\beta}_{\boldsymbol{\lambda}},$$

where

$$\widehat{\beta}_{\boldsymbol{\lambda}}=\int_Q\Psi_{\boldsymbol{\lambda}}\mathrm{d}\widehat{M},$$

and we define the best estimator of s within $S^{\star}_m$ as

$$\widehat{s}^{\star}_m=\underset{t\in S^{\star}_m}{\operatorname{argmin}}\,\widehat{\gamma}(t)=\sum_{\boldsymbol{\lambda}\in m}\widehat{\beta}_{\boldsymbol{\lambda}}\Psi^{\star}_{\boldsymbol{\lambda}}.$$

3.5. **Quadratic risk on a pyramidal model.** Let us introduce the orthogonal projection of s on $S^{\star}_m$ for the norm $\|.\|_{\boldsymbol{\Psi}}$, that is

$$s^{\star}_m=\sum_{\boldsymbol{\lambda}\in m}\beta_{\boldsymbol{\lambda}}\Psi^{\star}_{\boldsymbol{\lambda}},$$

where

$$\beta_{\boldsymbol{\lambda}}=\langle\Psi_{\boldsymbol{\lambda}},s\rangle.$$

It follows from (3) that $\check{\beta}_{\boldsymbol{\lambda}}$ is an unbiased estimator for $\beta_{\boldsymbol{\lambda}}$, so that $\check{s}^{\star}_m$ is an unbiased estimator for $s^{\star}_m$. Thanks to Pythagoras' equality, we recover for $\check{s}^{\star}_m$ the usual decomposition

$$\mathbb{E}\left[\|s-\check{s}^{\star}_m\|^2_{\boldsymbol{\Psi}}\right]=\|s-s^{\star}_m\|^2_{\boldsymbol{\Psi}}+\sum_{\boldsymbol{\lambda}\in m}\mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}}), \tag{8}$$

where the first term is a bias term or approximation error and the second term is a variance term or estimation error. When only $\widehat{M}$ is observed, combining the triangle inequality, the basic inequality (14) and (8) easily provides at least an upper-bound akin to (8), up to a residual term.

Proposition 1. *For all $\theta>0$,*

$$\mathbb{E}\left[\|s-\hat{s}^{\star}_m\|^2_{\boldsymbol{\Psi}}\right]\leq(1+\theta)\left(\|s-s^{\star}_m\|^2_{\boldsymbol{\Psi}}+\sum_{\boldsymbol{\lambda}\in m}\mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}})\right)+(1+1/\theta)\mathbb{E}\left[\|\check{s}^{\star}_m-\hat{s}^{\star}_m\|^2_{\boldsymbol{\Psi}}\right].$$

When $\widehat{M}=M$, θ can be taken equal to 0 and equality holds.

In all the examples introduced in Section 2.2, we shall verify that the quadratic risks satisfies, for all $\theta>0$,

$$\mathbb{E}\left[\|s-\hat{s}^{\star}_m\|^2_{\boldsymbol{\Psi}}\right]\leq c_1\|s-s^{\star}_m\|^2_{\boldsymbol{\Psi}}+c_2\frac{\|s\|_{\infty}D_m}{\bar{n}}+r_1(\bar{n}), \tag{9}$$

where $\bar{n}$ describes the amount of available data, and the residual term $r_1(\bar{n})$ does not weigh too much upon the estimation rate.

3.5.1. *Example 1: density estimation (continued).* In this framework, the empirical coefficients are of the form

$$\check{\beta}_{\boldsymbol{\lambda}} = \frac{1}{n}\sum_{i=1}^{n} \Psi_{\boldsymbol{\lambda}}(Y_i).$$

As the wavelets are normalized and s is bounded,

$$\mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}}) \leq \frac{1}{n}\int_Q \Psi^2_{\boldsymbol{\lambda}}(x)s(x)\mathrm{d}x \leq \frac{\|s\|_\infty}{n},$$

so

$$\mathbb{E}\left[\|s-\hat{s}^\star_m\|^2_{\boldsymbol{\Psi}}\right] \leq \|s-s^\star_m\|^2_{\boldsymbol{\Psi}} + \frac{\|s\|_\infty D_m}{n}.$$

Hence (9) is satisfied for instance with $\bar{n}=n$, $c_1=c_2=1, r_1(n)=0$.

3.5.2. *Example 2: copula density estimation (continued).* In this case,

$$\check{\beta}_{\boldsymbol{\lambda}} = \frac{1}{n}\sum_{i=1}^{n} \Psi_{\boldsymbol{\lambda}}\left(F_1(X_{i1}),\dots,F_d(X_{id})\right),$$

while

$$\widehat{\beta}_{\boldsymbol{\lambda}} = \frac{1}{n}\sum_{i=1}^{n} \Psi_{\boldsymbol{\lambda}}\left(\hat{F}_{n1}(X_{i1}),\dots,\hat{F}_{nd}(X_{id})\right).$$

As in Example 1, $\mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}}) \leq \|s\|_\infty/n$. Besides we prove in Section 7.1 the following upper-bound for the residual terms.

Proposition 2. *Under Assumption (L), for all $m \in \mathcal{M}^{\mathcal{P}}$,*

$$\mathbb{E}[\|\check{s}^\star_m - \hat{s}^\star_m\|^2] \leq C(\kappa,d)L_\bullet^{d-1}2^{4L_\bullet}\log(n)/n.$$

Hence choosing $\bar{n}=n$, $2^{4L_\bullet}=\sqrt{n}/\log(n)$, $c_1=c_2=2$, and $r_1(n)=C(\kappa,d)\log(n)^{d-1}/\sqrt{n}$ yields (9).

3.5.3. *Example 3: Poisson intensity estimation (continued).* In this case,

$$\check{\beta}_{\boldsymbol{\lambda}} = \frac{1}{\mathrm{Vol}_d(Q)}\int_Q \Psi_{\boldsymbol{\lambda}}(x)N(\mathrm{d}x).$$

From Campbell's formula,

$$\mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}}) = \frac{1}{\mathrm{Vol}_d(Q)}\int_Q \Psi^2_{\boldsymbol{\lambda}}(x)s(x)\mathrm{d}x$$

so

$$\mathbb{E}\left[\|s-\hat{s}^\star_m\|^2_{\boldsymbol{\Psi}}\right] \leq \|s-s^\star_m\|^2_{\boldsymbol{\Psi}} + \frac{\|s\|_\infty D_m}{\bar{n}},$$

with $\bar{n}=\mathrm{Vol}_d(Q)$.

3.5.4. *Example 4: Lévy jump intensity estimation with continuous time observations (continued).* In this case,

$$\check{\beta}_{\boldsymbol{\lambda}} = \frac{1}{T}\iint_{[0,T]\times Q} \Psi_{\boldsymbol{\lambda}}(x)N(\mathrm{d}t,\mathrm{d}x).$$

From Campbell's formula again,

$$\mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}}) = \frac{1}{T^2}\iint_{[0,T]\times Q} \Psi^2_{\boldsymbol{\lambda}}(x)s(x)\mathrm{d}t\mathrm{d}x$$

so

$$\mathbb{E}\left[\|s-\hat{s}^\star_m\|^2_{\boldsymbol{\Psi}}\right] \leq \|s-s^\star_m\|^2_{\boldsymbol{\Psi}} + \frac{\|s\|_\infty D_m}{\bar{n}},$$

with $\bar{n}=T$.

3.5.5. *Example 5: Lévy jump intensity estimation with discrete time observations (continued).* In this case, the empirical coefficients and their approximate counterparts are of the form

$$\check{\beta}_{\boldsymbol{\lambda}} = \frac{1}{n\Delta} \iint_{[0,n\Delta]\times Q} \Psi_{\boldsymbol{\lambda}}(x) N(\mathrm{d}t, \mathrm{d}x) \quad \text{and} \quad \widehat{\beta}_{\boldsymbol{\lambda}} = \frac{1}{n\Delta} \sum_{i=1}^{n} \Psi_{\boldsymbol{\lambda}}(X_{i\Delta} - X_{(i-1)\Delta}).$$

We deduce as previously that $\mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}}) \leq \|s\|_\infty/\bar{n}$ with $\bar{n} = n\Delta$. Besides we can bound the residual term thanks to the following proposition, proved in Section 7.2.

Proposition 3. *Under Assumption (L), for all $m \in \mathcal{M}^{\mathcal{P}}$,*

$$\mathbb{E}[\|\check{s}^\star_m - \hat{s}^\star_m\|^2] \leq 8\frac{\|s\|_\infty D_m}{n\Delta} + C(\kappa, d, f, Q) L_\bullet^{d-1} \frac{2^{4L_\bullet} n\Delta^3 + 2^{3L_\bullet}\Delta}{n\Delta}.$$

provided Δ is small enough.

Assuming $n\Delta^2$ stays bounded while $n\Delta \to \infty$ as $n \to \infty$, and choosing $2^{4L_\bullet} = n\Delta$, we deduce that (9) is satisfied under Assumption (L) with $\bar{n} = n\Delta, c_1 = 2, c_2 = 18$, and $r_1(\bar{n}) = C(\kappa, d, f, Q)\log(\bar{n})^{d-1}/\bar{n}$. Notice that these assumptions on n and Δ are classical in the so-called framework of high-frequency observations.

Remark: Proposition 3 extends [FL09] to a multivariate model with a complex structure due to the use of hyperbolic wavelets, instead of isotropic ones. Yet, the extension is not so straightforward, so we give a detailed proof in Section 7.

4. Wavelet pyramid model selection

The upper-bound (9) for the risk on one pyramidal model suggests that a good model should be large enough so that the approximation error is small, and small enough so that the estimation error is small. Without prior knowledge on the function s to estimate, choosing the best pyramidal model is thus impossible. In this section, we describe a data-driven procedure that selects the best pyramidal model from the data, without using any smoothness assumption on s. We provide theoretical results that guarantee the performance of such a procedure. We underline how these properties are linked with the structure of the collection of models.

4.1. Penalized pyramid selection.

When M is observed, we deduce from (8) that

$$\mathbb{E}\left[\|s - \check{s}^\star_m\|^2_{\boldsymbol{\Psi}}\right] - \|s\|^2_{\boldsymbol{\Psi}} = -\|s^\star_m\|^2_{\boldsymbol{\Psi}} + \sum_{\boldsymbol{\lambda}\in m} \mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}})$$

and from (7) that $\gamma(\check{s}^\star_m) = -\|\check{s}^\star_m\|^2_{\boldsymbol{\Psi}}$. Following the work of [BBM99], we introduce a penalty function $\mathrm{pen} : \mathcal{M}^{\mathcal{P}} \to \mathbb{R}^+$ and choose a best pyramidal model from the data defined as

$$\hat{m}^{\mathcal{P}} = \operatorname*{argmin}_{m\in\mathcal{M}^{\mathcal{P}}} (\hat{\gamma}(\hat{s}^\star_m) + \mathrm{pen}(m)).$$

In order to choose the pyramidal model with smallest quadratic risk, the penalty $\mathrm{pen}(m)$ is expected to behave roughly as the estimation error within model m. We provide such a penalty in the following Section. Our final estimator for s is then

$$\tilde{s}^{\mathcal{P}} = \hat{s}^\star_{\hat{m}^{\mathcal{P}}}.$$

4.2. Combinatorial complexity and choice of the penalty function.

As widely examplified in [Mas07, BGH09] for instance, the choice of an adequate penalty depends on the combinatorial complexity of the collection of models, which is measured through the index

$$\max_{dj_0+1\leq\ell_1\leq L_\bullet+1} \frac{\log\left(\sharp\mathcal{M}^{\mathcal{P}}_{\ell_1}\right)}{D(\ell_1)}, \tag{10}$$

where $D(\ell_1)$ is the common dimension of all pyramidal models in $\mathcal{M}^{\mathcal{P}}_{\ell_1}$. Ideally, this index should be upper-bounded independently of the sample size for the resulting model selection procedure to reach the optimal estimation rate. The following proposition describes the combinatorial complexity of the collection of pyramidal models.

Proposition 4. *Let $M = 2+B/2^{j_0-1}$. For all $\ell_1 \in \{dj_0+1, \dots, L_\bullet+1\}$ and all $k \in \{0, \dots, L_\bullet-\ell_1\}$, let*

$$N(\ell_1, k) = \lfloor 2\sharp U \boldsymbol{\nabla}(\ell_1+k)(k+2)^{-(d+2)}2^{-k}M^{-d} \rfloor \tag{11}$$

and $D(\ell_1)$ be the common dimension of all models in $\mathcal{M}^{\mathcal{P}}_{\ell_1}$. There exists positive reals $\kappa_1(d)$, $\kappa_2(j_0, B, d)$ and $\kappa_3(j_0, B, d)$ such that

$$\kappa_1(d)(\ell_1 - dj_0 + d - 2)^{d-1}2^{\ell_1} \le D(\ell_1) \le \kappa_2(j_0, B, d)(\ell_1 - dj_0 + d - 2)^{d-1}2^{\ell_1}$$

and

$$\log\left(\sharp\mathcal{M}^{\mathcal{P}}_{\ell_1}\right) \le \kappa_3(j_0, B, d)D(\ell_1).$$

We remind that B is defined in Section 3.1(Assumption *S.ii*)). Possible values for κ_1, κ_2 and κ_3 are given in the proof, which is postponed to Section 7.3. In the same way, we could prove a matching lower-bound for $\log\left(\sharp\mathcal{M}^{\mathcal{P}}_{\ell_1}\right)$ for large enough ℓ_1, so that the whole family $\mathcal{M}^{\mathcal{P}}$ contains of order of $L_\bullet^{d-1}2^{L_\bullet}$ models. Typically, we will choose $L_\bullet$ such that $2^{L_\bullet}$ is a power of the sample size $\bar{n}$. So while $\mathcal{M}^{\mathcal{P}}$ contains at least an exponential number of models, the number of models per dimension is moderate enough so that the combinatorial index (10) bounded.

From now on, we assume that (11) is satisfied, as well as the following hypotheses. For all subfamily $\mathcal{T}$ of $S^\star_{m_\bullet}$, let

$$\mathcal{Z}(\mathcal{T}) = \sup_{t\in\mathcal{T}}\left(\int_Q \sum_{\lambda\in m_\bullet}\langle t, \Psi_{\boldsymbol{\lambda}}\rangle\Psi_{\boldsymbol{\lambda}}\mathrm{d}M - \langle s, t\rangle_{\boldsymbol{\Psi}}\right).$$

Assumption (Conc). *There exist positive reals $\bar{n}, \kappa_1', \kappa_2', \kappa_3'$ such that, for all countable subfamily $\mathcal{T}$ of $\left\{t \in S^\star_{m_\bullet} | \|t\|_{\boldsymbol{\Psi}} = 1\right\}$ satisfying*

$$\sup_{t\in\mathcal{T}}\left\|\sum_{\lambda\in m_\bullet}\langle t, \Psi_{\boldsymbol{\lambda}}\rangle\Psi_{\boldsymbol{\lambda}}\right\|_\infty \le B(\mathcal{T})$$

for some positive constant $B(\mathcal{T})$, we have, for all $x > 0$,

$$\mathbb{P}\left(\mathcal{Z}(\mathcal{T}) \ge \kappa_1'\mathbb{E}\left[\mathcal{Z}(\mathcal{T})\right] + \sqrt{\kappa_2'\|s\|_\infty\frac{x}{\bar{n}}} + \kappa_3'B(\mathcal{T})\frac{x}{\bar{n}}\right) \le \exp(-x).$$

Assumption (Var). *There exist a nonnegative constant κ_4' and a collection of estimators $(\hat{\sigma}^2_{\boldsymbol{\lambda}})_{\boldsymbol{\lambda}\in m_\bullet}$ such that, for all $\boldsymbol{\lambda} \in m_\bullet$,*

$$\mathbb{E}\left[\hat{\sigma}^2_{\boldsymbol{\lambda}}\right] \le \kappa_4'\max(\|s\|_\infty, 1).$$

Besides there exist a nonnegative constant κ_5', a nonnegative function w such that $w(\bar{n})/\bar{n} \xrightarrow[\bar{n}\to\infty]{} 0$, and a measurable event Ω_σ on which, for all $\boldsymbol{\lambda} \in m_\bullet$,

$$\mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}}) \le \kappa_5'\frac{\max\{\hat{\sigma}^2_{\boldsymbol{\lambda}}, 1\}}{\bar{n}}$$

and such that

$$p_\sigma := \mathbb{P}(\Omega_\sigma^c) \le \frac{w(\bar{n})}{\bar{n}}.$$

Assumption (Rem). *For the same function w as in Assumption **(Var)** and some nonnegative constant κ_6',*

$$\mathbb{E}\left[\|\check{s}^\star_m - \hat{s}^\star_m\|^2_{\boldsymbol{\Psi}}\right] \le \kappa_6'\frac{\|s\|_\infty D_m}{\bar{n}} + \frac{w(\bar{n})}{\bar{n}}, \text{ for all } m \subset m_\bullet,$$

and

$$\max\left\{\frac{1}{\bar{n}(\log(\bar{n})/d)^{(d+1)/2}}\sqrt{\mathbb{E}\left[\|\check{s}^\star_{m_\bullet} - \hat{s}^\star_{m_\bullet}\|^4_{\boldsymbol{\Psi}}\right]}, \sqrt{p_\sigma\mathbb{E}\left[\|\check{s}^\star_{m_\bullet} - \hat{s}^\star_{m_\bullet}\|^4_{\boldsymbol{\Psi}}\right]}\right\} \le \frac{w(\bar{n})}{\bar{n}}.$$

Assumption **(Conc)** describes how the random measure M concentrates around the measure to estimate. Assumption **(Var)** ensures that we can estimate the variance terms $\mathbb{E}\left[\|s^\star_m - \check{s}^\star_m\|^2_{\mathbf{\Psi}}\right]$ over each $m \in \mathcal{M}^{\mathcal{P}}$. Last, Assumption **(Rem)** describes how close $\widehat{M}$ is to M.

Theorem 1. *Assume that (11), Assumptions **(Conc)**, **(Var)**,**(Rem)** are satisfied, and that* $\max(\|s\|_\infty, 1) \leq \bar{R}$. *Choose* $L_\bullet$ *such that*

$$2^{L_\bullet} \leq \frac{\bar{n}}{((\log \bar{n})/d)^{2d}}$$

and a penalty of the form

$$\mathrm{pen}(m) = \sum_{\lambda \in m} \frac{c_1 \hat{\sigma}^2_{\boldsymbol{\lambda}} + c_2 \bar{R}}{\bar{n}}, m \in \mathcal{M}^{\mathcal{P}}.$$

If c_1, c_2 *are positive and large enough, then*

$$\mathbb{E}\left[\|s - \tilde{s}^{\mathcal{P}}\|^2_{\mathbf{\Psi}}\right] \leq C_1 \min_{m \in \mathcal{M}^{\mathcal{P}}} \left(\|s - s^\star_m\|^2_{\mathbf{\Psi}} + \frac{\bar{R} D_m}{\bar{n}} \right) + C_2 \frac{\max\left\{\|s\|^2_{\mathbf{\Psi}}, \|s\|_\infty, 1\right\}}{\bar{n}} \left(1 + (\log(\bar{n})/d)^{-3(d+1)/2} + w(\bar{n})\right)$$

where C_1 *may depend on* $\kappa'_1, \kappa'_2, \kappa'_4, \kappa'_5, \kappa'_6, c_1, c_2$ *and* C_2 *may depend* $\kappa'_1, \kappa'_2, \kappa'_3, \kappa'_7, j_0, d$.

In practice, the penalty constants c_1 and c_2 are calibrated by simulation study. We may also replace $\bar{R}$ in the penalty by $\max\{\|\hat{s}^\star_{m_\bullet}\|_\infty, 1\}$, and extend Theorem 1 to a random $\bar{R}$ by using arguments similar to [AL11].

4.3. **Back to the examples.** First, two general remarks are in order. For $t \in S^\star_{m_\bullet}$, let $f_t = \sum_{\lambda \in m_\bullet} \langle t, \Psi_{\boldsymbol{\lambda}} \rangle \Psi_{\boldsymbol{\lambda}}$, then $\|f_t\| = \|t\|_{\mathbf{\Psi}}$ and by (3), for all countable subfamily $\mathcal{T}$ of $S^\star_{m_\bullet}$,

$$\mathcal{Z}(\mathcal{T}) = \sup_{t \in \mathcal{T}} \left(\int_Q f_t \mathrm{d}M - \mathbb{E}\left[\int_Q f_t \mathrm{d}M \right] \right).$$

So Assumption **(Conc)** usually proceeds from a Talagrand type concentration inequality. Besides, we have seen in Section 3.5 that in general

$$\max_{\boldsymbol{\lambda} \in m_\bullet} \mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}}) \leq \frac{\|s\|_\infty}{\bar{n}}.$$

Thus, whenever some upper-bound R_∞ for $\|s\|_\infty$ is known, Assumption **(Var)** is satisfied with $\hat{\sigma}^2_{\boldsymbol{\lambda}} = R_\infty$ for all $\boldsymbol{\lambda} \in m_\bullet$, $\Omega_\sigma = \Omega$, $w(\bar{n}) = 0, \kappa'_4 = \kappa'_5 = 1$. One may also estimate each variance term: this is what we propose in the following results, proved in Section 7.5.

Corollary 1. *In the density estimation framework (see 2.2.1), let* $\bar{R} \geq \max(\|s\|_\infty, 1)$, $2^{L_\bullet} = n((\log n)/d)^{-2d}$,

$$\hat{\sigma}^2_{\boldsymbol{\lambda}} = \frac{1}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\Psi_{\boldsymbol{\lambda}}(Y_i) - \Psi_{\boldsymbol{\lambda}}(Y_j)\right)^2 \text{ for all } \boldsymbol{\lambda} \in m_\bullet,$$

and

$$\mathrm{pen}(m) = \sum_{\lambda \in m} \frac{c_1 \hat{\sigma}^2_{\boldsymbol{\lambda}} + c_2 \bar{R}}{n}, \text{ for all } m \in \mathcal{M}^{\mathcal{P}}.$$

If c_1, c_2 *are positive and large enough, then*

$$\mathbb{E}\left[\|s - \tilde{s}^{\mathcal{P}}\|^2_{\mathbf{\Psi}}\right] \leq C_1 \min_{m \in \mathcal{M}^{\mathcal{P}}} \left(\|s - s^\star_m\|^2_{\mathbf{\Psi}} + \frac{\bar{R} D_m}{n} \right) + C_2 \frac{\max\left\{\|s\|^2_{\mathbf{\Psi}}, \bar{R}\right\}}{n}$$

where C_1 *may depend on* κ, d, c_1, c_2 *and* C_2 *may depend* κ, j_0, d.

Corollary 2. *In the copula density estimation framework (see 2.2.2), let* $\bar{R} \geq \max(\|s\|_\infty, 1)$ *and* $2^{L_\bullet} = \min\{n^{1/8}(\log n)^{-1/4}, n((\log n)/d)^{-2d}\}$. *For all* $\boldsymbol{\lambda} \in m_\bullet$, *define*

$$\hat{\sigma}^2_{\boldsymbol{\lambda}} = \frac{1}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\Psi_{\boldsymbol{\lambda}}\left(\hat{F}_{n1}(X_{i1}), \ldots, \hat{F}_{nd}(X_{id})\right) - \Psi_{\boldsymbol{\lambda}}\left(\hat{F}_{n1}(X_{j1}), \ldots, \hat{F}_{nd}(X_{jd})\right) \right)^2,$$

and for all $m \in \mathcal{M}^{\mathcal{P}}$, *let*

$$\mathrm{pen}(m) = \sum_{\lambda \in m} \frac{c_1 \hat{\sigma}^2_{\boldsymbol{\lambda}} + c_2 \bar{R}}{n}.$$

Under Assumption (L), and if c_1, c_2 *are positive and large enough, then*

$$\mathbb{E}\left[\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}}\right] \leq C_1 \min_{m \in \mathcal{M}^{\mathcal{P}}} \left(\|s - s^\star_m\|^2_{\boldsymbol{\Psi}} + \frac{\bar{R} D_m}{n} \right) + C_2 \max\left\{\|s\|^2_{\boldsymbol{\Psi}}, \bar{R}\right\} \frac{(\log n)^{d-1}}{\sqrt{n}}$$

where C_1 *may depend on* κ, d, c_1, c_2 *and* C_2 *may depend* κ, j_0, d.

Corollary 3. *In the Poisson intensity estimation framework (see 2.2.3), let* $\bar{R} \geq \max(\|s\|_\infty, 1)$, $2^{L_\bullet} = Vol_d(Q)((\log Vol_d(Q))/d)^{-2d}$,

$$\hat{\sigma}^2_{\boldsymbol{\lambda}} = \frac{1}{Vol_d(Q)} \int_Q \Psi^2_{\boldsymbol{\lambda}} \mathrm{d}N, \ \textit{for all } \boldsymbol{\lambda} \in m_\bullet,$$

and

$$\mathrm{pen}(m) = \sum_{\lambda \in m} \frac{c_1 \hat{\sigma}^2_{\boldsymbol{\lambda}} + c_2 \bar{R}}{Vol_d(Q)}, \ \textit{for all } m \in \mathcal{M}^{\mathcal{P}}.$$

If c_1, c_2 *are positive and large enough, then*

$$\mathbb{E}\left[\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}}\right] \leq C_1 \min_{m \in \mathcal{M}^{\mathcal{P}}} \left(\|s - s^\star_m\|^2_{\boldsymbol{\Psi}} + \frac{\bar{R} D_m}{Vol_d(Q)} \right) + C_2 \frac{\max\left\{\|s\|^2_{\boldsymbol{\Psi}}, \bar{R}\right\}}{Vol_d(Q)}$$

where C_1 *may depend on* κ, d, c_1, c_2 *and* C_2 *may depend* κ, j_0, d.

Corollary 4. *In the Lévy jump intensity estimation framework with continuous time observations (see 2.2.4), let* $\bar{R} \geq \max(\|s\|_\infty, 1)$, $2^{L_\bullet} = T((\log T)/d)^{-2d}$,

$$\hat{\sigma}^2_{\boldsymbol{\lambda}} = \frac{1}{T} \iint_{[0,T]\times Q} \Psi^2_{\boldsymbol{\lambda}}(x) N(\mathrm{d}t, \mathrm{d}x), \ \textit{for all } \boldsymbol{\lambda} \in m_\bullet,$$

and

$$\mathrm{pen}(m) = \sum_{\lambda \in m} \frac{c_1 \hat{\sigma}^2_{\boldsymbol{\lambda}} + c_2 \bar{R}}{T}, \ \textit{for all } m \in \mathcal{M}^{\mathcal{P}}.$$

If c_1, c_2 *are positive and large enough, then*

$$\mathbb{E}\left[\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}}\right] \leq C_1 \min_{m \in \mathcal{M}^{\mathcal{P}}} \left(\|s - s^\star_m\|^2_{\boldsymbol{\Psi}} + \frac{\bar{R} D_m}{T} \right) + C_2 \frac{\max\left\{\|s\|^2_{\boldsymbol{\Psi}}, \bar{R}\right\}}{T}$$

where C_1 *may depend on* κ, d, c_1, c_2 *and* C_2 *may depend* κ, j_0, d.

Corollary 5. *In the Lévy jump intensity estimation framework with discrete time observations (see 2.2.5), let* $\bar{R} \geq \max(\|s\|_\infty, 1)$, $2^{L_\bullet} = \min\left\{(n\Delta)^{1/4}, n\Delta(\log(n\Delta)/d)^{-2d}\right\}$,

$$\hat{\sigma}^2_{\boldsymbol{\lambda}} = \frac{1}{n\Delta} \sum_{i=1}^{n} \Psi^2_{\boldsymbol{\lambda}} \left(X_{i\Delta} - X_{(i-1)\Delta}\right), \ \textit{for all } \boldsymbol{\lambda} \in m_\bullet,$$

and

$$\mathrm{pen}(m) = \sum_{\lambda \in m} \frac{c_1 \hat{\sigma}^2_{\boldsymbol{\lambda}} + c_2 \bar{R}}{n\Delta}, \ \textit{for all } m \in \mathcal{M}^{\mathcal{P}}.$$

If Assumption (L) is satisfied, if $n\Delta^2$ *stays bounded while* $n\Delta \to \infty$ *as* $n \to \infty$, *and if* c_1, c_2 *are positive and large enough, then*

$$\mathbb{E}\left[\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}}\right] \leq C_1 \min_{m \in \mathcal{M}^{\mathcal{P}}} \left(\|s - s^\star_m\|^2_{\boldsymbol{\Psi}} + \frac{\bar{R} D_m}{T} \right) + C_2 \max\left\{\|s\|^2_{\boldsymbol{\Psi}}, \bar{R}\right\} \frac{\log^{d-1}(n\Delta)}{n\Delta}$$

where C_1 *may depend on* κ, d, c_1, c_2 *and* C_2 *may depend* κ, j_0, d, Q, f.

Corollaries 3, 4, 5 extend respectively the works of [RB03, FLH09, UK11] to a multivariate framework, with a complex family of models allowing for nonhomogeneous smoothness, and a more refined penalty.

5. Adaptivity to mixed smoothness

There remains to compare the performance of our procedure $\tilde{s}^{\mathcal{P}}$ to that of other estimators. For that purpose, we derive the estimation rate of $\tilde{s}^{\mathcal{P}}$ under smoothness assumptions that induce sparsity on the hyperbolic wavelet coefficients of s. We then compare it to the minimax rate.

5.1. **Function spaces with dominating mixed smoothness.** For $\alpha \in \mathbb{N}^\star$ and $1 \leq p \leq \infty$, the mixed Sobolev space with smoothness α measured in the $\mathbb{L}_p-$norm is defined as

$$SW^{\alpha}_{p,(d)} = \left\{ f \in \mathbb{L}_p([0,1]^d) \middle| \|f\|_{SW^{\alpha}_{p,(d)}} := \sum_{0 \leq r_1, \dots, r_d \leq \alpha} \left\| \frac{\partial^{r_1+\dots+r_d} f}{\partial^{r_1} x_1 \dots \partial^{r_d} x_d} \right\|_p < \infty \right\},$$

while the classical Sobolev space is

$$W^{\alpha}_{p,(d)} = \left\{ f \in \mathbb{L}_p([0,1]^d) \middle| \|f\|_{W^{\alpha}_{p,(d)}} := \sum_{0 \leq r_1+\dots+r_d \leq \alpha} \left\| \frac{\partial^{r_1+\dots+r_d} f}{\partial^{r_1} x_1 \dots \partial^{r_d} x_d} \right\|_p < \infty \right\}.$$

The former contains functions whose highest order derivative is the mixed derivative $\partial^{d\alpha} f / \partial^{\alpha} x_1 \dots \partial^{\alpha} x_d$, while the latter contains all derivatives up to global order $d\alpha$. Both spaces coincide in dimension $d = 1$, and otherwise we have the obvious continuous embeddings

$$W^{d\alpha}_{p,(d)} \hookrightarrow SW^{\alpha}_{p,(d)} \hookrightarrow W^{\alpha}_{p,(d)}. \tag{12}$$

Hölder and Besov spaces with mixed dominating smoothness may be defined thanks to mixed differences. For $f : [0,1] \to \mathbb{R}, x \in [0,1]$ and $h > 0$,

$$\Delta^0_h(f,x) = f(x), \quad \Delta^1_h(f,x) = f(x+h) - f(x)$$

and more generally, for $r \in \mathbb{N}^\star$, the r-th order univariate difference operator is

$$\Delta^r_h = \Delta^1_h \circ \Delta^{r-1}_h,$$

so that

$$\Delta^r_h(f,x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x+kh). \tag{13}$$

Then for $t > 0$ the univariate modulus of continuity of order r in $\mathbb{L}_p$ is defined as

$$w_r(f,t)_p = \sup_{0<h<t} \|\Delta^r_h(f,.)\|_p.$$

For $f : [0,1]^d \to \mathbb{R}, \mathbf{x} = (x_1, \dots, x_d) \in [0,1]^d, r \in \mathbb{N}^\star$ and $h_\ell > 0$, we denote by $\Delta^r_{h_\ell,\ell}$ the univariate difference operator applied to the ℓ-th coordinate while keeping the other ones fixed, so that

$$\Delta^r_{h_\ell,\ell}(f,\mathbf{x}) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x_1, \dots, x_\ell + kh_\ell, \dots, x_d).$$

For any subset $\mathbf{e}$ of $\{1, \dots, d\}$ and $\mathbf{h} = (h_1, \dots, h_d) \in (0, +\infty)^d$, the r-th order mixed difference operator is given by

$$\Delta^{r,\mathbf{e}}_{\mathbf{h}} := \prod_{\ell \in \mathbf{e}} \Delta^r_{h_\ell,\ell}.$$

For $\mathbf{t} = (t_1, \dots, t_d) \in (0, +\infty)^d$, we set $\mathbf{t}_\mathbf{e} = (t_\ell)_{\ell \in \mathbf{e}}$, and define the mixed modulus of continuity

$$w^{\mathbf{e}}_r(f, \mathbf{t}_\mathbf{e})_p = \sup_{0<h_\ell<t_\ell, \ell \in \mathbf{e}} \|\Delta^{r,\mathbf{e}}_{\mathbf{h}}(f,.)\|_p.$$

For $\alpha > 0$ and $0 < p \leq \infty$, the mixed Hölder space $SH^{\alpha}_{p,(d)}$ is the space of all functions $f : [0,1]^d \to \mathbb{R}$ such that

$$\|f\|_{SH^{\alpha}_{p,(d)}} := \sum_{\mathbf{e} \subset \{1,\dots,d\}} \sup_{\mathbf{t}>0} \prod_{\ell \in \mathbf{e}} t_\ell^{-\alpha} w^{\mathbf{e}}_{\lfloor \alpha \rfloor + 1}(f, \mathbf{t}_\mathbf{e})_p$$

is finite, where by convention the term associated with $\mathbf{e} = \emptyset$ is $\|f\|_p$. More generally, for $\alpha > 0$ and $0 < p, q \leq \infty$, the mixed Besov space $SB^{\alpha}_{p,q,(d)}$ is the space of all functions $f : [0,1]^d \to \mathbb{R}$ such that

$$\|f\|_{SB^{\alpha}_{p,q,(d)}} := \sum_{\mathbf{e} \subset \{1,\dots,d\}} \left(\int_{(0,1)} \dots \int_{(0,1)} \left(\prod_{\ell \in \mathbf{e}} t_\ell^{-\alpha} w^{\mathbf{e}}_{\lfloor \alpha \rfloor + 1}(f, \mathbf{t}_\mathbf{e})_p \right)^q \prod_{\ell \in \mathbf{e}} \frac{\mathrm{d}t_\ell}{t_\ell} \right)^{1/q},$$

where the $\mathbb{L}_q$-norm is replaced by a sup-norm in case $q = \infty$, so that $SB^{\alpha}_{p,\infty,(d)} = SH^{\alpha}_{p,(d)}$. By comparison, the usual Besov space $B^{\alpha}_{p,q,(d)}$ may be defined as the space of all functions $f \in \mathbb{L}_p([0,1]^d)$ such that

$$\|f\|_{B^{\alpha}_{p,q,(d)}} := \begin{cases} \|f\|_p + \sum_{\ell=1}^d \left(\int_{(0,1)} \left(t_\ell^{-\alpha} w^{\{\ell\}}_{\lfloor \alpha \rfloor + 1}(f, t_\ell)_p \right)^q \frac{\mathrm{d}t_\ell}{t_\ell} \right)^{1/q} & \text{if } 0 < q < \infty \\ \|f\|_p + \sum_{\ell=1}^d \sup_{t_\ell > 0} t_\ell^{-\alpha} w^{\{\ell\}}_{\lfloor \alpha \rfloor + 1}(f, t_\ell)_p & \text{if } q = \infty \end{cases}$$

is finite. Extending (12), the recent results of [NS16a] confirm that the continuous embeddings

$$B^{d\alpha}_{p,q,(d)} \hookrightarrow SB^{\alpha}_{p,q,(d)} \hookrightarrow B^{\alpha}_{p,q,(d)},$$

hold under fairly general assumptions on $\alpha, p, q, d.$.

On the other hand, given $\alpha > 0, 0 < p < \infty, 0 < q \leq \infty$, we define

$$N_{\mathbf{\Psi},\alpha,p,q}(f) = \begin{cases} \left(\sum_{\ell \geq dj_0} 2^{q\ell(\alpha+1/2-1/p)} \sum_{\boldsymbol{j} \in \boldsymbol{J}_\ell} \left(\sum_{\boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{j}}} |\langle f, \Psi_{\boldsymbol{\lambda}} \rangle|^p \right)^{q/p} \right)^{1/q} & \text{if } 0 < q < \infty \\ \sup_{\ell \geq dj_0} 2^{\ell(\alpha+1/2-1/p)} \sup_{\boldsymbol{j} \in \boldsymbol{J}_\ell} \left(\sum_{\boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{j}}} |\langle f, \Psi_{\boldsymbol{\lambda}} \rangle|^p \right)^{1/p} & \text{if } q = \infty \end{cases}$$

and $N_{\mathbf{\Psi},\alpha,\infty,q}$ in the same way by replacing the ℓ_p-norm with a sup-norm. Then for $\alpha > 0, 0 < p, q \leq \infty, R > 0$, we denote by $\mathcal{SB}(\alpha, p, q, R)$ the set of all functions $f \in \mathbb{L}_p([0,1]^d)$ such that

$$N_{\mathbf{\Psi},\alpha,p,q}(f) \leq R.$$

Under appropriate conditions on the smoothness of $\mathbf{\Psi}^\star$, that we will assume to be satisfied in the sequel, the sets $\mathcal{SB}(\alpha, p, q, R)$ may be interpreted as balls with radius R in Besov spaces with dominating mixed smoothness $SB^{\alpha}_{p,q,(d)}$ (see for instance [ST87, Hoc02a, Hep04, DTU16]). Mixed Sobolev spaces are not easily characterized in terms of wavelet coefficients, but they satisfy the compact embeddings

$$SB^{\alpha}_{p,\min(p,2),(d)} \hookrightarrow SW^{\alpha}_{p,(d)} \hookrightarrow SB^{\alpha}_{p,\max(p,2),(d)}, \text{ for } 1 < p < \infty$$

and

$$SB^{\alpha}_{1,1,(d)} \hookrightarrow SW^{\alpha}_{1,(d)} \hookrightarrow SB^{\alpha}_{1,\infty,(d)}$$

(see [DTU16], Section 3.3). So, without loss of generality, we shall mostly turn our attention to Besov-Hölder spaces in the sequel.

5.2. **Link with structural assumptions.** The following property collects examples of composite functions with mixed dominating smoothness built from lower dimensional functions with classical Sobolev or Besov smoothness. The proof and upper-bounds for the norms of the composite functions are given in Section 7.6. An analogous property for (mixed) Sobolev smoothness instead of (mixed) Besov smoothness can be proved straightforwardly.

Proposition 5. *Let $\alpha > 0$ and $0 < p, q \leq \infty$.*

(i) If $u_1, \dots, u_d \in B^{\alpha}_{p,q,(1)}$, then $f(\mathbf{x}) = \sum_{\ell=1}^d u_\ell(x_\ell) \in SB^{\alpha}_{p,q,(d)}$.

(ii) Let $\mathfrak{P}$ be some partition of $\{1, \dots, d\}$. If, for all $I \in \mathfrak{P}, u_I \in B^{\alpha_I}_{p,q,(|I|)}$, then $f(\mathbf{x}) = \prod_{I \in \mathfrak{P}} u_I(\mathbf{x}_I) \in SB^{\bar{\alpha}}_{p,q,(d)}$ where $\bar{\alpha} = \min_{I \in \mathfrak{P}}(\alpha_I / |I|)$.

(iii) Let $\alpha \in \mathbb{N}^\star$ and $p > 1$, if $g \in W^{d\alpha}_{\infty,(1)}$ and $u_\ell \in W^{\alpha}_{p,1}$ for $\ell = 1, \dots, d$, then $f(\mathbf{x}) = g\left(\sum_{\ell=1}^d u_\ell(x_\ell) \right) \in SW^{\alpha}_{p,(d)}$.

(iv) If $f \in SB^{\alpha}_{p,q,(d)}$ with $\alpha > 1$ and $\partial^d f / \partial x_1 \dots \partial x_d \in \mathbb{L}_p([0,1]^d)$, then $\partial^d f / \partial x_1 \dots \partial x_d \in SB^{\alpha-1}_{p,q,(d)}$.

(v) If f_1 *and* $f_2 \in SB^{\alpha}_{p,p,(d)}$ *where either* $1 < p \leq \infty$ *and* $\alpha > 1/p$, *or* $p = 1$ *and* $\alpha \geq 1$, *then the product function* $\mathbf{x} \mapsto f_1(\mathbf{x})f_2(\mathbf{x}) \in SB^{\alpha}_{p,p,(d)}$.

Notice that in (i) (resp. $(ii), (iii)$), the assumptions on the component functions u_ℓ, u_I or g are not enough to ensure that $f \in B^{d\alpha}_{p,q,(d)}$ (resp. $B^{d\bar{\alpha}}_{p,q,(d)}, W^{d\alpha}_{p,(d)}$).

Remark: We believe that a generalization of (iii) to Besov or fractional Sobolev smoothness holds. Yet such a generalization would require refined arguments from Approximation Theory in the spirit of [BS11, Mou11] which are beyond the scope of that paper.

The structural assumption (ii) may be satisfied in the multivariate density estimation framework 2.2.1 whenever $Y_1 = (Y_{11}, \ldots, Y_{1d})$ can be split into independent sub-groups of coordinates, and has recently been considered in [Lep13, Reb15b, Reb15a]. Case (i) and its generalization (iii) may not be directly of use in our multivariate intensity framework, but they will allow to draw a comparison with [HM07, BB14]. Combining (iii) and (iv) is of interest for copula density estimation 2.2.2, having in mind that a wide nonparametric family of copulas are Archimedean copulas (see [Nel06], Chapter 4), which have densities of the form

$$s(x_1, \ldots, x_d) = (\phi^{-1})'(x_1) \ldots (\phi^{-1})'(x_d)\phi^{(d)}\left(\phi^{-1}(x_1) + \ldots + \phi^{-1}(x_d)\right)$$

provided the generator ϕ is smooth enough (see for instance [MN09]). Combining $(iii), (iv), (v)$ may be of interest for Lévy intensity estimation in 2.2.4 or 2.2.5. Indeed, a popular way to build multivariate Lévy intensities is based on Lévy copulas studied in [KT06] (see also [CT04], Chapter 5). The resulting Lévy intensities then have the form

$$f(x_1, \ldots, x_d) = f_1(x_1) \ldots f_d(x_d)F^{(1,\ldots,1)}(U_1(x_1) + \ldots + U_d(x_d))$$

where F is a so-called Lévy copula, $F^{(1,\ldots,1)} = \partial^d F/\partial t_1 \ldots \partial t_d$ and $U_\ell(x_\ell) = \int_{x_\ell}^{\infty} f_\ell(t)\mathrm{d}t$. Besides, a common form for F is

$$F(x) = \phi\left(\phi^{-1}(x_1) + \ldots + \phi^{-1}(x_d)\right)$$

under appropriate smoothness assumptions on ϕ. Last, let us emphasize that any linear combination (mixtures for instance) of functions in $SB^{\alpha}_{p,q,(d)}$ inherits the same smoothness. Consequently, mixed dominating smoothness may be thought as a fully nonparametric surrogate for a wide range of structural assumptions.

5.3. **Approximation qualities and minimax rate.** We provide in Section 7.7 a constructive proof for the following nonlinear approximation result, in the spirit of [BBM99].

Theorem 2. *Let* $R > 0, 0 < p < \infty, 0 < q \leq \infty, \alpha > \max(1/p - 1/2, 0)$, *and* $f \in \mathbb{L}_2([0,1]^d) \cap \mathcal{SB}(\alpha, p, q, R)$. *Under* (11), *for all* $\ell_1 \in \{dj_0+1, \ldots, L_\bullet+1\}$, *there exists some model* $m_{\ell_1}(f) \in \mathcal{M}^{\mathcal{P}}_{\ell_1}$ *and some approximation* $A(f, \ell_1) \in S^{\star}_{m_{\ell_1}(f)}$ *for* f *such that*

$$\|f - A(f,\ell_1)\|^2_{\mathbf{\Psi}}$$
$$\leq C(B, j_0, \alpha, p, d)R^2\left(L_\bullet^{2(d-1)(1/2-1/q)_+}2^{-2L_\bullet(\alpha-(1/p-1/2)_+)} + \ell_1^{2(d-1)(1/2-1/\max(p,q))}2^{-2\alpha\ell_1}\right).$$

Remark: When $p \geq 2$, the same kind of result still holds with all $N(\ell_1, k) = 0$. But Assumption (11) is really useful when $p < 2$, the so-called non-homogeneous smoothness case.

The first term in the upper-bound is a linear approximation error by the highest dimensional model $S^{\star}_{m_\bullet}$ in the collection. As $D_{m_\bullet}$ is of order $L_\bullet^{d-1}2^{L_\bullet}$, we deduce from [DTU16] (Section 4.3) that this first term is optimal over $SB^{\alpha}_{p,q}$, at least for $1 < p < \infty, 1 \leq q \leq \infty$ and $\alpha > \max(1/p - 1/2, 0)$, for instance. The second term in the upper-bound is a nonlinear approximation error of f within the model $S^{\star}_{m_{\ell_1}(f)}$, with dimension $D_{m_{\ell_1}(f)}$ of order $\ell_1^{d-1}2^{\ell_1}$. So we deduce from [DTU16] (Theorem 7.6) that this second term, which is of order $D^{-2\alpha}_{m_{\ell_1}(f)}(\log D_{m_{\ell_1}(f)})^{2(d-1)(\alpha+1/2-1/q)}$, is also optimal up to a constant factor over $SB^{\alpha}_{p,q}$, at least for $1 < p < \infty, p \leq q \leq \infty$ and $\alpha > \max(1/p - 1/2, 0)$. Notice that, under the classical Besov smoothness assumption $f \in \mathbb{L}_2([0,1]^d) \cap B^{\alpha}_{p,q,(d)}$, the best possible approximation rate for f by D-dimensional linear subspaces in the $\mathbb{L}_2$-norm would be of order $D^{-2\alpha/d}$. Thus with a mixed smoothness of order α in dimension d, we recover the same

approximation rate as with a classical smoothness of order $d\alpha$ in dimension d, up to a logarithmic factor.

Let us define, for $\alpha, p, q, R, R' > 0$,

$$\overline{\mathcal{SB}}(\alpha, p, q, R, R') = \{f \in \mathcal{SB}(\alpha, p, q, R)/\|f\|_\infty \leq R'\}.$$

In the sequel, we use the notation $a \asymp C(\theta)b$ when there exist positive reals $C_1(\theta), C_2(\theta)$ such that $C_1(\theta)b \leq a \leq C_2(\theta)b$.

Corollary 6. *Assume $L_\bullet$ is large enough, then for all $0 < p < \infty, 0 < q \leq \infty, \alpha > (1/p - 1/2)_+, R \geq \bar{n}^{-1}, R' > 0$,*

$$\sup_{s \in \overline{\mathcal{SB}}(\alpha,p,q,R,R')} \mathbb{E}_s\left[\|s - \tilde{s}^{\mathcal{P}}\|_{\mathbf{\Psi}}^2\right] \leq C(B, d, \alpha, p, R')\left(\left(\log(\bar{n}R^2)\right)^{(d-1)(\alpha+1/2-1/\max(p,q))} R\bar{n}^{-\alpha}\right)^{2/(1+2\alpha)}.$$

Proof. In order to minimize approximately the upper-bound, we choose ℓ_1 such that

$$\ell_1^{2(d-1)(1/2-1/\max(p,q))} 2^{-2\alpha\ell_1} R^2 \asymp C(\alpha, p, q, d)\ell_1^{d-1}2^{\ell_1}/\bar{n},$$

that is for instance

$$2^{\ell_1} \asymp C(\alpha, p, q, d)\left(\left(\log(\bar{n}R^2)\right)^{-2(d-1)/\max(p,q)} (\bar{n}R^2)\right)^{1/(1+2\alpha)},$$

which yields the announced upper-bound. □

Remember that a similar result holds when replacing the $\mathbf{\Psi}$-norm by the equivalent $\mathbb{L}_2$-norm. Though unusual, the upper-bound in Corollary 6 is indeed related to the minimax rate.

Proposition 6. *In the density estimation framework, assume $R^2 \geq n^{-1}, R' > 0, p > 0, 0 < q \leq \infty$, and either $\alpha > (1/p - 1/2)_+$ and $q \geq 2$ or $\alpha > (1/p - 1/2)_+ + 1/\min(p, q, 2) - 1/\min(p, 2)$, then*

$$\inf_{\hat{s}\ estimator\ of\ s} \sup_{s \in \overline{\mathcal{SB}}(\alpha,p,q,R,R')} \mathbb{E}_s\left[\|s - \hat{s}\|^2\right] \asymp C(\alpha, p, q, d)\left(\left(\log(nR^2)\right)^{(d-1)(\alpha+1/2-1/q)} Rn^{-\alpha}\right)^{2/(1+2\alpha)}.$$

Proof. One may derive from [DTU16] (Theorem 6.20), [Dun01] (proof of Theorem 1) and the link between entropy number and Kolmogorov entropy that the Kolmogorov ϵ-entropy of $\mathcal{SB}(\alpha, p, q, R)$ is

$$H_\epsilon(\alpha, p, q, R) = (R/\epsilon)^{1/\alpha}\left(\log(R/\epsilon)\right)^{(d-1)(\alpha+1/2-1/q)/\alpha}.$$

According to [YB99] (Proposition 1), in the density estimation framework, the minimax risk over $\overline{\mathcal{SB}}(\alpha, p, q, R, R')$ is of order ρ_n^2 where $\rho_n^2 = H_{\rho_n}(\alpha, p, q, R)/n$, which yields the announced rate. □

Consequently, in the density estimation framework, the penalized pyramid selection procedure is minimax over $\mathcal{SB}(\alpha, p, q, R)$ up to a constant factor if $p \leq q \leq \infty$, and only up to a logarithmic factor otherwise.

Let us end with some comments about these estimation rates. First, we remind that the minimax rate under the assumption $s \in B^\alpha_{p,q,(d)}$ is of order $n^{-2\alpha/d/(1+2\alpha/d)}$. Thus, under a mixed smoothness assumption of order α, we recover, up to a logarithmic factor, the same rate as with smoothness of order α in dimension 1, which can only be obtained with smoothness of order $d\alpha$ under a classical smoothness assumption in dimension d. Besides, under the multiplicative constraint (ii) of Proposition 5, we recover the same rate as [Reb15a], up to a logarithmic factor. And under the generalized additive constraint (iii) of Proposition 5, we recover the same rate as [BB14] (Section 4.3), up to a logarithmic factor. Regarding Neumann seminal work on estimation under mixed smoothness [Neu00] (see his Section 3), a first adaptive wavelet thresholding is proved to be optimal up to a logarithmic factor over $SW^r_{2,(d)} = SB^r_{2,2,(d)}$, and another, nonadaptive one, is proved to be optimal up to a constant over $SB^r_{1,\infty,(d)}$, where r is a positive integer. Our procedure thus outperforms [Neu00] by being at the same time adaptive and minimax optimal up to a constant over these two classes, and many other ones.

6. Implementing wavelet pyramid selection

We end this paper with a quick overview of practical issues related to wavelet pyramid selection. As we perform selection within a large collection of models, where typically the number of models is exponential in the sample size, we must guarantee that the estimator can still be computed in a reasonable time. Besides, we provide simulation based examples illustrating the interest of this new method.

6.1. Algorithm and computational complexity.

Theorem 1 supports the choice of an additive penalty of the form

$$\mathrm{pen}(m) = \sum_{\boldsymbol{\lambda}\in m} \hat{v}_{\boldsymbol{\lambda}}^2,$$

where detailed expressions for $\hat{v}_{\boldsymbol{\lambda}}^2$ in several statistical frameworks have been given in Section 4.3. As $\hat{\gamma}(\hat{s}_m^\star) = -\sum_{\boldsymbol{\lambda}\in m} \hat{\beta}_{\boldsymbol{\lambda}}^2$, the penalized selection procedure amounts to choose

$$\hat{m}^{\mathcal{P}} = \operatorname*{argmax}_{m\in\mathcal{M}^{\mathcal{P}}} \mathrm{crit}(m)$$

where

$$\mathrm{crit}(m) = \sum_{\boldsymbol{\lambda}\in m} (\hat{\beta}_{\boldsymbol{\lambda}}^2 - \hat{v}_{\boldsymbol{\lambda}}^2).$$

Since each $\hat{v}_{\boldsymbol{\lambda}}^2$ is roughly an (over)estimate for the variance of $\hat{\beta}_{\boldsymbol{\lambda}}^2$, our method, though different from a thresholding procedure, will mainly retain empirical wavelet coefficients $\hat{\beta}_{\boldsymbol{\lambda}}^2$ which are significantly larger than their variance.

A remarkable thing is that, due to both the structure of the collection of models and of the penalty function, the penalized estimator can be determined without computing all the preliminary estimators $(\hat{s}_m^\star)_{m\in\mathcal{M}^{\mathcal{P}}}$, which makes the computation of $\tilde{s}^{\mathcal{P}}$ feasible in practice. Indeed, we can proceed as follows.

Step 1. For each $\ell_1 \in \{dj_0+1, \dots, L_\bullet+1\}$, determine

$$\hat{m}_{\ell_1} = \operatorname*{argmax}_{m\in\mathcal{M}_{\ell_1}^{\mathcal{P}}} \sum_{\boldsymbol{\lambda}\in m} (\hat{\beta}_{\boldsymbol{\lambda}}^2 - \hat{v}_{\boldsymbol{\lambda}}^2).$$

For that purpose, it is enough, for each $k \in \{0, \dots, L_\bullet - \ell_1\}$, to

- compute and sort in decreasing order all the coefficients $(\hat{\beta}_{\boldsymbol{\lambda}}^2 - \hat{v}_{\boldsymbol{\lambda}}^2)_{\boldsymbol{\lambda}\in U\boldsymbol{\nabla}(\ell_1+k)}$;
- keep the $N(\ell_1,k)$ indices in $U\boldsymbol{\nabla}(\ell_1+k)$ that yield the $N(\ell_1,k)$ greatest such coefficients.

Step 2. Determine the integer $\hat{\ell} \in \{dj_0+1, \dots, L_\bullet+1\}$ such that

$$\hat{m}_{\hat{\ell}} = \operatorname*{argmax}_{dj_0+1\le\ell_1\le L_\bullet+1} \mathrm{crit}(\hat{m}_{\ell_1}).$$

The global computational complexity of $\tilde{s}^{\mathcal{P}}$ is thus $\mathcal{O}(\log(L_\bullet) L_\bullet^d 2^{L_\bullet})$. Typically, we will choose $L_\bullet$ at most of order $\log_2(\bar{n})$ so the resulting computational complexity will be at most of order $\mathcal{O}(\log(\log(\bar{n}))\log^d(\bar{n})\bar{n})$.

6.2. Illustrative examples.

In this section, we study two examples in dimension $d=2$ by using Haar wavelets.

First, in the density estimation framework, we consider an example where the coordinates of $Y_i = (Y_{i1}, Y_{i2})$ are independent conditionally on a K-way categorical variable Z, so that the density of Y_i may be written as

$$s(x_1,x_2) = \sum_{k=1}^{K} \pi_k s_{1,k}(x_1) s_{2,k}(x_2),$$

where $\pi = (\pi_1, \dots, \pi_K)$ is the probability vector characterizing the distribution of Z. For a compact interval I, and $a, b > 0$, let us denote by $\beta(I; a, b)$ the Beta density with parameters a, b shifted and rescaled to have support I, and by $\mathcal{U}(I)$ the uniform density on I. In our example, we take

- $K = 4$ and $\pi = (3/5, 1/10, 1/40, 11/40)$;
- $s_{1,1} = \beta([0,3/5]; 4, 4)$ and $s_{2,1} = \beta([0,2/5]; 4, 4)$;

- $s_{1,2} = \beta([2/5,1];100,100)$ and $s_{2,2} = \beta([2/5,1];20,20)$;
- $s_{1,3} = \mathcal{U}([0,1])$ and $s_{2,3} = \mathcal{U}([0,1])$;
- $s_{1,4} = \beta([3/5,1];8,4)$ and $s_{2,4} = \mathcal{U}([2/5,1])$.

The resulting mixture density s of Y_i is shown in Figure 1 (b). We choose $L_\bullet = \left[n/((\log n)/2)^2\right]$ and first compute the least-squares estimator $\hat{s}^\star_\bullet$ of s on the model $V^\star_{L_\bullet/2} \otimes \ldots \otimes V^\star_{L_\bullet/2}$, which provides the estimator $\hat{R} = \max\{\|\hat{s}^\star_\bullet\|, 1\}$ for $\bar{R}$. We then use the penalty

$$\mathrm{pen}(m) = \sum_{\boldsymbol{\lambda}\in m} \frac{1.5\hat{\sigma}^2_{\boldsymbol{\lambda}} + 0.5\hat{R}}{n}.$$

For a sample with size $n = 2000$, Figure 1 illustrates how the procedure first selects a rough model $\hat{m}_{\hat{\ell}}$ (Figure 1 (c)) and then add some details wherever needed (Figure 1 (d)). Summing up the two yields the pyramid selection estimator $\tilde{s}^{\mathcal{P}}$ (Figure 1 (e)). By way of comparison, we also represent in Figure 1 (f) a widely used estimator: the bivariate Gaussian kernel estimator, with the "known support" option, implemented in MATLAB `ksdensity` function. We observe that, contrary to the kernel density estimator, the pyramid selection estimator recovers indeed the main three modes, and in particular the sharp peak.

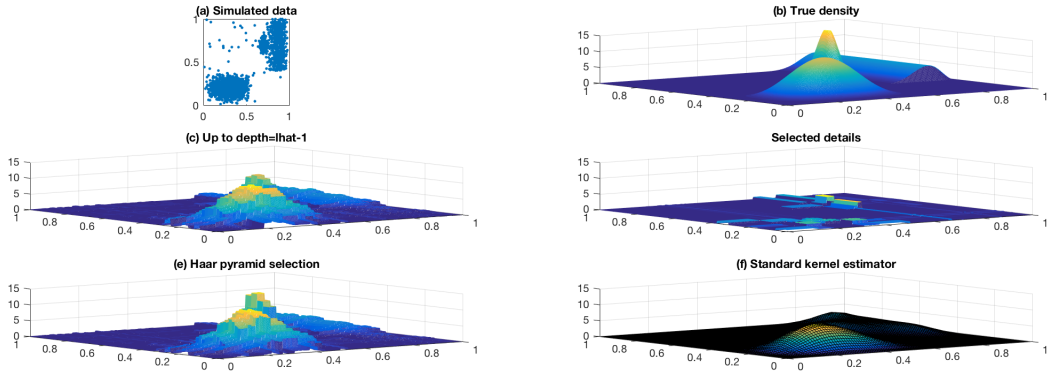

FIGURE 1. Pyramid selection and standard kernel for an example of mixture of multiplicative densities.

In the copula density estimation framework, we consider an example where the copula of $X_i = (X_{i1}, X_{i2})$ is either a Frank copula or a Clayton copula conditionally to a binary variable Z. More precisely, we consider the mixture copula

$$s(x_1, x_2) = 0.5 s_F(x_1, x_2) + 0.5 s_C(x_1, x_2)$$

where s_F is the density of a Frank copula with parameter 4 and s_C is the density of a Clayton copula with parameter 2. These two examples of Archimedean copula densities are shown in Figure 2 and the resulting mixture in Figure 3 (b). We use the same penalty as in the previous example, adapted of course to the copula density estimation framework. We illustrate in Figure 3 the pyramid selection procedure on a sample with size $n = 2000$. Though not all theoretical conditions are fully satisfied here, the pyramid selection procedure still provides a reliable estimator.

As a conclusion, those examples suggest that the Haar pyramid selection already provides a useful new estimation procedure. This is most encouraging for pyramid selection based on higher order wavelets, whose full calibration based on an extensive simulation study in each framework will be the subject of another work.

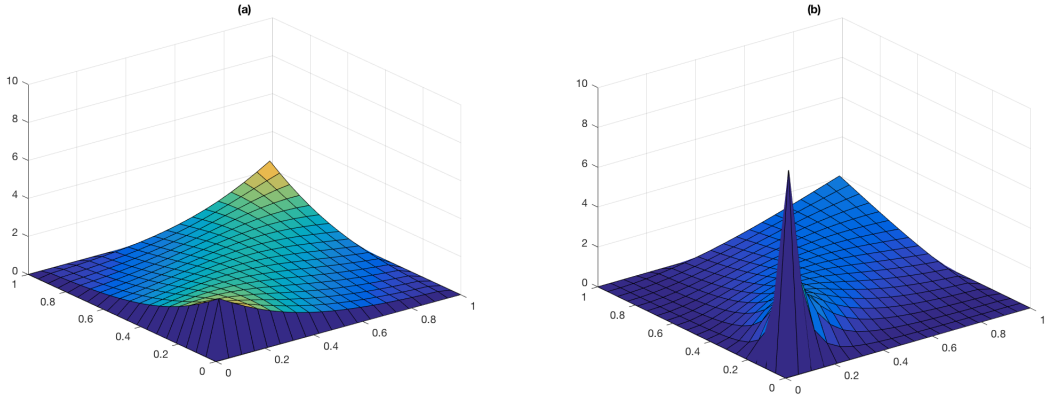

FIGURE 2. Left: Frank copula density with parameter 4; Right: Clayton copula density with parameter 2.

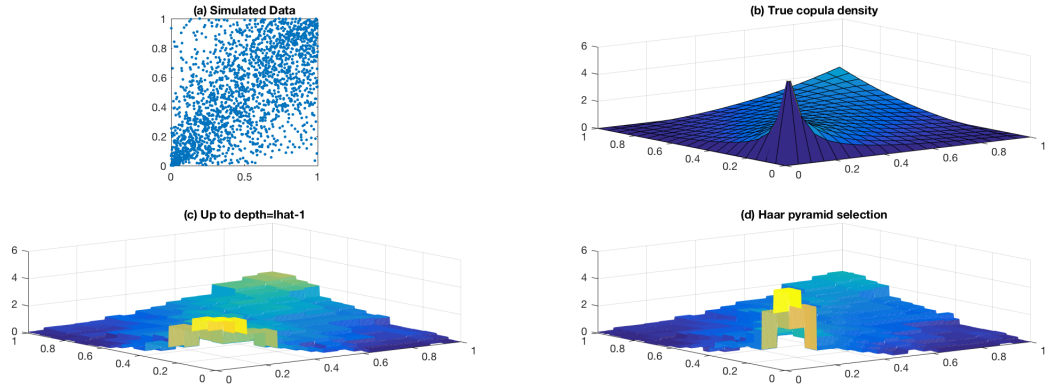

FIGURE 3. Pyramid selection for an example of mixture copula density.

7. PROOFS

We shall use repeatedly the classical inequality

$$2ab \leq \theta a^2 + \frac{1}{\theta} b^2 \tag{14}$$

for all positive θ, a, b.

7.1. **Proof of Proposition 2.** We only have to prove 2 for $m_\bullet$. Indeed, as any pyramidal model m is a subset of $m_\bullet$, a common upper-bound for the residual terms is

$$\|\check{s}^\star_m - \hat{s}^\star_m\|^2 \leq \|\check{s}^\star_{m_\bullet} - \hat{s}^\star_{m_\bullet}\|^2.$$

Under Assumption (L), and thanks to assumptions $S.iii$) and $W.v$), we have that for all $\boldsymbol{\lambda} \in m_\bullet$, $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $\boldsymbol{y} = (y_1, \ldots, y_d) \in Q$,

$$|\Psi_{\boldsymbol{\lambda}}(\boldsymbol{x}) - \Psi_{\boldsymbol{\lambda}}(\boldsymbol{y})| \leq \kappa^d 2^{3L_\bullet/2} \sum_{k=1}^{d} |x_k - y_k|.$$

According to Massart's version of Dworetzky-Kiefer-Wolfowitz inequality (see [Mas90]), for any positive z, and $1 \leq k \leq d$, there exists some event $\Omega_k(z)$ on which $\|\hat{F}_{nk} - F_k\|_\infty \leq z/\sqrt{n}$ and such that $\mathbb{P}(\Omega_k^c(z)) \leq 2\exp(-2z^2)$. Setting $\Omega(z) = \cap_{k=1}^d \Omega_k(z)$, we thus have for all $\boldsymbol{\lambda} \in m_\bullet$

$$|\check{\beta}_{\boldsymbol{\lambda}} - \widehat{\beta}_{\boldsymbol{\lambda}}| \leq \kappa^d 2^{3L_\bullet/2} d(z/\sqrt{n}) \mathbb{1}_{\Omega(z)} + \kappa^d 2^{3L_\bullet/2} d \mathbb{1}_{\Omega^c(z)},$$

hence

$$\mathbb{E}[\|\check{s}^\star_{m_\bullet} - \hat{s}^\star_{m_\bullet}\|^2] \le \kappa^{2d} 2^{3L_\bullet} d^2 (z^2/n + 2d\exp(-2z^2)) D_{m_\bullet}.$$

Finally, $D_{m_\bullet}$ is of order $L_\bullet^{d-1} 2^{L_\bullet}$ (see Proposition 4), so by choosing $2z^2 = \log(n)$,

$$\mathbb{E}[\|\check{s}^\star_{m_\bullet} - \hat{s}^\star_{m_\bullet}\|^2] \le C(\kappa, d) L_\bullet^{d-1} 2^{4L_\bullet} \log(n)/n.$$

7.2. **Proof of Proposition 3.** For all bounded measurable function g, let us denote $D_\Delta(g) = \mathbb{E}[g(X_\Delta)]/\Delta - \int_Q gs$. For all $\boldsymbol{\lambda} \in \Lambda$,

$$\mathbb{E}\left[\left(\check{\beta}_{\boldsymbol{\lambda}} - \widehat{\beta}_{\boldsymbol{\lambda}}\right)^2\right] \le 4\left(\mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}}) + \mathrm{Var}(\widehat{\beta}_{\boldsymbol{\lambda}}) + D_\Delta^2(\Psi_{\boldsymbol{\lambda}})\right) \le 8\frac{\|s\|_\infty}{n\Delta} + 4\frac{D_\Delta(\Psi_{\boldsymbol{\lambda}}^2)}{n\Delta} + 4D_\Delta^2(\Psi_{\boldsymbol{\lambda}}). \tag{15}$$

We shall bound $D_\Delta(\Psi_{\boldsymbol{\lambda}})$ by using the decomposition of a Lévy process into a big jump compound Poisson process and an independent small jump Lévy process. Let us fix $\varepsilon > 0$ small enough so that $Q = \prod_{k=1}^d [a_k, b_k] \subset \{\|x\| > \varepsilon\}$ and denote by $(\Sigma, \boldsymbol{\mu}, \nu)$ the characteristic Lévy triplet of $\mathbf{X} = (X_t)_{t\ge 0}$, where $\boldsymbol{\mu}$ stands for the drift and ν is the Lévy measure, with density f with respect to the Lebesgue measure on $\mathbb{R}^d$ (see Section 2.2). Then $\mathbf{X}$ is distributed as $\mathbf{X}^\varepsilon + \tilde{\mathbf{X}}^\varepsilon$, where $\mathbf{X}^\varepsilon$ and $\tilde{\mathbf{X}}^\varepsilon$ are independent Lévy processes with following characteristics. First, $\mathbf{X}^\varepsilon$ is a Lévy process with characteristic Lévy triplet $(\Sigma, \boldsymbol{\mu}_\varepsilon, \nu_\epsilon)$, where the drift is

$$\boldsymbol{\mu}_\varepsilon = \boldsymbol{\mu} - \int_{\varepsilon < \|x\| \le 1} x\ f(x)\mathrm{d}x.$$

and the Lévy measure is

$$\nu_\varepsilon(\mathrm{d}x) = \mathbb{1}_{\|x\|\le\varepsilon} f(x)\mathrm{d}x.$$

The process $\tilde{\mathbf{X}}^\varepsilon$ is the compound Poisson process

$$\tilde{X}_t^\varepsilon = \sum_{i=1}^{\tilde{N}_t} \xi_i,$$

where $\tilde{N}$ is a homogeneous Poisson process with intensity $\lambda_\varepsilon = \nu(\{\|x\| > \varepsilon\})$, $(\xi_i)_{i\ge 1}$ are i.i.d. with density $\lambda_\varepsilon^{-1}\mathbb{1}_{\|x\|>\varepsilon} f(x)$, and $\tilde{N}$ and $(\xi_i)_{i\ge 1}$ are independent.

Conditioning by $\tilde{N}$ and using the aforementioned independence properties yields

$$\frac{\mathbb{E}[\Psi_{\boldsymbol{\lambda}}(X_\Delta)]}{\Delta} = e^{-\lambda_\varepsilon\Delta}\frac{\mathbb{E}[\Psi_{\boldsymbol{\lambda}}(X_\Delta^\varepsilon)]}{\Delta} + \lambda_\varepsilon e^{-\lambda_\varepsilon\Delta}\mathbb{E}[\Psi_{\boldsymbol{\lambda}}(X_\Delta^\varepsilon + \xi_1)] + \lambda_\varepsilon^2\Delta e^{-\lambda_\varepsilon\Delta}\sum_{j=0}^\infty \mathbb{E}\left[\Psi_{\boldsymbol{\lambda}}\left(X_\Delta^\varepsilon + \sum_{i=1}^{j+2}\xi_i\right)\right]\frac{(\lambda_\varepsilon\Delta)^j}{(j+2)!}.$$

Conditioning by ξ_1 and using independence between X_Δ^ε and ξ_1 then yields

$$\lambda_\varepsilon\mathbb{E}[\Psi_{\boldsymbol{\lambda}}(X_\Delta^\varepsilon + \xi_1)] = \int_{\|x\|>\varepsilon} \mathbb{E}[\Psi_{\boldsymbol{\lambda}}(X_\Delta^\varepsilon + x)] f(x)\mathrm{d}x.$$

Writing $\langle\Psi_{\boldsymbol{\lambda}}, s\rangle = e^{-\lambda_\varepsilon\Delta}\langle\Psi_{\boldsymbol{\lambda}}, s\rangle + (1 - e^{-\lambda_\varepsilon\Delta})\langle\Psi_{\boldsymbol{\lambda}}, s\rangle$ and using $(1 - e^{-\lambda_\varepsilon\Delta}) \le \lambda_\varepsilon\Delta$ leads to

$$|D_\Delta(\Psi_{\boldsymbol{\lambda}})| \le R_\Delta^{(1)}(\Psi_{\boldsymbol{\lambda}}) + R_\Delta^{(2)}(\Psi_{\boldsymbol{\lambda}}) + R_\Delta^{(3)}(\Psi_{\boldsymbol{\lambda}}) + R_\Delta^{(4)}(\Psi_{\boldsymbol{\lambda}}),$$

where

$$R_\Delta^{(1)}(\Psi_{\boldsymbol{\lambda}}) = e^{-\lambda_\varepsilon\Delta}\frac{\mathbb{E}[\Psi_{\boldsymbol{\lambda}}(X_\Delta^\varepsilon)]}{\Delta}, \quad R_\Delta^{(2)}(\Psi_{\boldsymbol{\lambda}}) = e^{-\lambda_\varepsilon\Delta}\int_{\|x\|>\varepsilon} |\mathbb{E}\left[\Psi_{\boldsymbol{\lambda}}(X_\Delta^\varepsilon + x) - \Psi_{\boldsymbol{\lambda}}(x)\right]| f(x)\mathrm{d}x,$$

$$R_\Delta^{(3)}(\Psi_{\boldsymbol{\lambda}}) = \lambda_\varepsilon\Delta\|\Psi_{\boldsymbol{\lambda}}\|_1\|s\|_\infty, \quad R_\Delta^{(4)}(\Psi_{\boldsymbol{\lambda}}) = \lambda_\varepsilon^2\Delta\|\Psi_{\boldsymbol{\lambda}}\|_\infty. \tag{16}$$

As $\Psi_{\boldsymbol{\lambda}}$ has compact support Q,

$$R_\Delta^{(1)}(\Psi_{\boldsymbol{\lambda}}) \le e^{-\lambda_\varepsilon\Delta}\|\Psi_{\boldsymbol{\lambda}}\|_\infty\frac{\mathbb{P}(X_\Delta^\varepsilon \in Q)}{\Delta}.$$

Let us denote by $X_{\Delta,k}^\varepsilon$ the k-th coordinate of X_Δ^ε and by d_Q the maximal distance from $[a_k, b_k]$ to 0, for $k = 1, \dots, d$, reached for instance at $k = k_0$. We deduce from the proof of Lemma 2 in [RW02] (see also [FLH09], equation (3.3)) that there exists $z_0 = z_0(\varepsilon)$ such that if $\Delta < d_Q/z_0(\varepsilon)$,

$$\mathbb{P}(X_\Delta^\varepsilon \in Q) \le \mathbb{P}(|X_{\Delta,k_0}^\varepsilon| \ge d_Q) \le \exp\left((z_0\log(z_0) + u - u\log(u))/(2\varepsilon)\right)\Delta^{d_Q/(2\varepsilon)}$$

so that

$$R_\Delta^{(1)}(\Psi_{\boldsymbol{\lambda}}) \le C(d_Q,\varepsilon)e^{-\lambda_\varepsilon \Delta}\|\Psi_{\boldsymbol{\lambda}}\|_\infty \Delta^{d_Q/(2\varepsilon)-1}. \tag{17}$$

Under Assumption (L), $\Psi_{\boldsymbol{\lambda}}$ is Lipschitz on Q, so

$$|\Psi_{\boldsymbol{\lambda}}(X_\Delta^\varepsilon + x) - \Psi_{\boldsymbol{\lambda}}(x)| \le \|\Psi_{\boldsymbol{\lambda}}\|_L \|X_\Delta^\varepsilon\|_1 \mathbb{1}_{\{X_\Delta^\varepsilon + x \in Q\}\cap\{x\in Q\}} + |\Psi_{\boldsymbol{\lambda}}(x)|\mathbb{1}_{\{X_\Delta^\varepsilon + x \notin Q\}\cap\{x\in Q\}} + \|\Psi_{\boldsymbol{\lambda}}\|_\infty \mathbb{1}_{\{X_\Delta^\varepsilon + x \in Q\}\cap\{x\notin Q\}}.$$

Besides, as Q is compact and bounded away from the origin, there exists δ_Q and $\rho_Q > 0$ such that

$$\{X_\Delta^\varepsilon + x \in Q\} \cap \{x \in Q\} \subset \{\|X_\Delta^\varepsilon\| \ge \delta_Q\}$$

$$(\{X_\Delta^\varepsilon + x \in Q\} \cap \{x \notin Q\}) \cup (\{X_\Delta^\varepsilon + x \notin Q\} \cap \{x \in Q\}) \subset \{\|X_\Delta^\varepsilon\| \ge \rho_Q\}$$

The Lévy measure of $\mathbf{X}^\varepsilon$ is compactly supported and satisfies

$$\int \|x\|^2 \nu_\varepsilon(\mathrm{d}x) = \int_{\|x\|\le\varepsilon} \|x\|^2 \nu(\mathrm{d}x)$$

which is finite since ν is a Lévy measure (see for instance [Sat99], Theorem 8.1). So we deduce from [Mil71], Theorem 2.1, that

$$\mathbb{E}\left[\|X_\Delta^\varepsilon\|^2\right] \le C(d,f)\Delta,$$

hence

$$\mathbb{E}\left[\|X_\Delta^\varepsilon\|_1 \mathbb{1}_{\|X_\Delta^\varepsilon\|\ge\delta_Q}\right] \le C(d)\delta_Q^{-1}\mathbb{E}\left[\|X_\Delta^\varepsilon\|^2\right] \le C(d,f)\delta_Q^{-1}\Delta$$

and from Markov inequality

$$\mathbb{P}(\|X_\Delta^\varepsilon\| \ge \rho_Q) \le C(d,f)\rho_Q^{-2}\Delta.$$

Finally, fixing $0 < \varepsilon < \min(d_Q/4, \inf_{x\in Q}\|x\|)$, we have for all $0 < \Delta < \min(d_Q/z_0(\epsilon), 1)$

$$R_\Delta^{(2)}(\Psi_{\boldsymbol{\lambda}}) \le C(d,f)e^{-\lambda_\varepsilon\Delta}\Delta(\delta_Q^{-1}\lambda_\varepsilon\|\Psi_{\boldsymbol{\lambda}}\|_L + \rho_Q^{-2}\|s\|_\infty\|\Psi_{\boldsymbol{\lambda}}\|_1 + \rho_Q^{-2}\lambda_\varepsilon\|\Psi_{\boldsymbol{\lambda}}\|_\infty). \tag{18}$$

For all $\boldsymbol{\lambda} \in m_\bullet$,

$$\max(\|\Psi_{\boldsymbol{\lambda}}\|_L, \|\Psi_{\boldsymbol{\lambda}}\|_\infty, \|\Psi_{\boldsymbol{\lambda}}\|_1) \le C(\kappa)2^{3L_\bullet/2}, \max(\|\Psi_{\boldsymbol{\lambda}}^2\|_L, \|\Psi_{\boldsymbol{\lambda}}^2\|_\infty, \|\Psi_{\boldsymbol{\lambda}}^2\|_1) \le C(\kappa)2^{2L_\bullet},$$

so that combining (15), (17), (18) and (16) yields

$$\mathbb{E}[\|\check{s}_m^\star - \hat{s}_m^\star\|^2] \le 8\frac{\|s\|_\infty D_m}{n\Delta} + C(\kappa,d,f,Q,\varepsilon)L_\bullet^{d-1}\frac{2^{4L_\bullet}n\Delta^3 + 2^{3L_\bullet}\Delta}{n\Delta}.$$

7.3. **Proof of Proposition 4.** Due to hypotheses $S.ii)$ and $W.ii)$, we have for all $j \ge j_0$,

$$2^{j-1} \le \sharp\nabla_j \le M2^{j-1},$$

hence, for all $\boldsymbol{j} \in \mathbb{N}_{j_0}^d$,

$$(1/2)^d 2^{|\boldsymbol{j}|} \le \sharp\boldsymbol{\nabla}_{\boldsymbol{j}} \le (M/2)^d 2^{|\boldsymbol{j}|}.$$

Let us fix $\ell \in \{dj_0, \dots, L_\bullet\}$. The number of d-uples $\boldsymbol{j} \in \mathbb{N}_{j_0}^d$ such that $|\boldsymbol{j}| = \ell$ is equal to the number of partitions of the integer $\ell - dj_0$ into d nonnegative integers, hence

$$\sharp\boldsymbol{J}_\ell = \binom{\ell - dj_0 + d - 1}{d-1} = \prod_{k=1}^{d-1}\left(1 + \frac{\ell - dj_0}{k}\right).$$

The last two displays and the classical upper-bound for binomial coefficient (see for instance [Mas07], Proposition 2.5) yield

$$c_0(d)(\ell - dj_0 + d - 1)^{d-1}2^\ell \le \sharp U\boldsymbol{\nabla}(\ell) \le c_1(M,d)(\ell - dj_0 + d - 1)^{d-1}2^\ell, \tag{19}$$

where $c_0(d) = 2^{-d}(d-1)^{-(d-1)}$ and $c_1(M,d) = (M/2)^d(e/(d-1))^{d-1}$.

Let us now fix $\ell_1 \in \{dj_0 + 1, \dots, L_\bullet + 1\}$. Any model $m \in \mathcal{M}_{\ell_1}^{\mathcal{P}}$ satisfies

$$D_m = \sum_{\ell=dj_0}^{\ell_1-1} \sharp U\boldsymbol{\nabla}(\ell) + \sum_{k=0}^{L_\bullet - \ell_1} N(\ell_1, k).$$

So we obviously have

$$D_m \ge \sharp U\boldsymbol{\nabla}(\ell_1 - 1) \ge \kappa_1(d)(\ell_1 - dj_0 + d - 2)^{d-1}2^{\ell_1},$$

with $\kappa_1(d) = c_0(d)/2 = 2^{-(d+1)}(d-1)^{-(d-1)}$. Besides, with our choice of $N(\ell_1, k)$,

$$D_m \leq c_1(M,d)(\ell_1 - dj_0 + d - 2)^{d-1} \sum_{\ell=dj_0}^{\ell_1-1} 2^\ell + 2M^{-d}c_1(M,d)s_1(d)(\ell_1 - dj_0 + d - 2)^{d-1}2^{\ell_1},$$

so that Proposition 4 holds with $\kappa_2(d, j_0, B) = c_1(M,d)(1 + 2M^{-d}c_1(M,d)s_1(d))$, where

$$s_1(d) = \sum_{k=0}^{\infty} \frac{(1 + k/(d-1))^{d-1}}{(2+k)^{d+2}}.$$

The number of subsets of Λ in $\mathcal{M}_{\ell_1}^{\mathcal{P}}$ satisfies

$$\sharp\mathcal{M}_{\ell_1}^{\mathcal{P}} = \prod_{k=0}^{L_\bullet - \ell_1} \binom{\sharp U\boldsymbol{\nabla}(\ell_1 + k)}{N(\ell_1, k)} \leq \prod_{k=0}^{L_\bullet - \ell_1} \left(\frac{e\,\sharp U\boldsymbol{\nabla}(\ell_1 + k)}{N(\ell_1, k)} \right)^{N(\ell_1,k)}.$$

For $k \in \{0, \ldots, L_\bullet - \ell_1\}$, let $f(k) = (k+2)^{d+2}2^k M^d/2$, then $N(\ell_1, k) \leq \sharp U\boldsymbol{\nabla}(\ell_1 + k)/f(k)$. As the function $x \in [0, U] \mapsto x\log(eU/x)$ is increasing, we deduce

$$\log(\sharp\mathcal{M}_{\ell_1}^{\mathcal{P}}) \leq D(\ell_1) \sum_{k=0}^{L_\bullet - \ell_1} \frac{\sharp U\boldsymbol{\nabla}(\ell_1 + k)}{\sharp U\boldsymbol{\nabla}(\ell_1 - 1)} \frac{1 + \log(f(k))}{f(k)}.$$

Setting

$$s_2 = \sum_{k=0}^{\infty} \frac{1}{(k+2)^3}, \quad s_3 = \sum_{k=0}^{\infty} \frac{\log(k+2)}{(k+2)^3}, s_4 = \sum_{k=0}^{\infty} \frac{1}{(k+2)^2},$$

one may take for instance $\kappa_3(j_0, B, d) = (\log(e/2) + d\log(M))s_2 + (d+2)s_3 + \log(2)s_4$ in Proposition 4.

7.4. **Proof of Theorem 1.**

7.4.1. *Notation and preliminary results.* Hyperbolic wavelet bases inherit from the underlying univariate wavelet bases a localization property which can be stated as follows.

Lemma 1. *Let* $\underline{D}(L_\bullet) = (e(L_\bullet - dj_0 + d - 1)/(d-1))^{d-1}\, 2^{L_\bullet/2}$, *then for all real-valued sequence* $(a_{\boldsymbol{\lambda}})_{\boldsymbol{\lambda} \in m_\bullet}$,

$$\max\left\{ \left\| \sum_{\boldsymbol{\lambda}\in m_\bullet} a_{\boldsymbol{\lambda}}\Psi_{\boldsymbol{\lambda}} \right\|_\infty, \left\| \sum_{\boldsymbol{\lambda}\in m_\bullet} a_{\boldsymbol{\lambda}}\Psi_{\boldsymbol{\lambda}}^\star \right\|_\infty \right\} \leq \kappa_7' \max_{\boldsymbol{\lambda}\in m_\bullet} |a_{\boldsymbol{\lambda}}|\underline{D}(L_\bullet),$$

where $\kappa_7' = \kappa^{2d}(2 + \sqrt{2})$ *for instance.*

Proof. For all $\mathbf{x} = (x_1, \ldots, x_d) \in [0,1]^d$, using assumptions $S.vi), S.vii), S.viii), W.iv), W.v), W.vi)$ in Section 3.1, we get

$$\begin{aligned} \left| \sum_{\boldsymbol{\lambda}\in m_\bullet} a_{\boldsymbol{\lambda}}\Psi_{\boldsymbol{\lambda}} \right| &\leq \max_{\boldsymbol{\lambda}\in m_\bullet} |a_{\boldsymbol{\lambda}}| \sum_{\ell=dj_0}^{L_\bullet} \sum_{\boldsymbol{j}\in\boldsymbol{J}_\ell} \prod_{k=1}^{d} \left(\sum_{\lambda_k \in \nabla_{j_k}} |\psi_{\lambda_k}(x_k)| \right) \\ &\leq \kappa^{2d} \max_{\boldsymbol{\lambda}\in m_\bullet} |a_{\boldsymbol{\lambda}}| \sum_{\ell=dj_0}^{L_\bullet} \sum_{\boldsymbol{j}\in\boldsymbol{J}_\ell} 2^{\ell/2}. \end{aligned}$$

We deduce from the proof of Proposition 4 the upper-bound $\sharp\boldsymbol{J}_\ell \leq (e(L_\bullet - dj_0 + d - 1)/(d-1))^{d-1}$ which allows to conclude. □

For all $t \in \mathbb{L}_2([0,1]^d)$, we define

$$\nu(t) = \sum_{\boldsymbol{\lambda}\in\Lambda} \langle t, \Psi_{\boldsymbol{\lambda}}\rangle(\check{\beta}_{\boldsymbol{\lambda}} - \langle s, \Psi_{\boldsymbol{\lambda}}\rangle), \quad \nu_R(t) = \sum_{\boldsymbol{\lambda}\in\Lambda} \langle t, \Psi_{\boldsymbol{\lambda}}\rangle(\hat{\beta}_{\boldsymbol{\lambda}} - \check{\beta}_{\boldsymbol{\lambda}}), \quad \hat{\nu}(t) = \nu(t) + \nu_R(t),$$

and for all $m \in \mathcal{M}^{\mathcal{P}}$, we set

$$\chi(m) = \sup_{t\in S_m^\star, \|t\|_\Psi = 1} \nu(t), \quad \chi_R(m) = \sup_{t\in S_m^\star, \|t\|_\Psi = 1} \nu_R(t).$$

Lemma 2. *For all $m \in \mathcal{M}^{\mathcal{P}}$, let $t^\star_m = \sum_{\boldsymbol{\lambda}\in m}(\nu(\Psi^\star_{\boldsymbol{\lambda}})/\chi(m))\Psi^\star_{\boldsymbol{\lambda}}$, then*

$$\chi(m) = \sqrt{\sum_{\boldsymbol{\lambda}\in m}\nu^2(\Psi^\star_{\boldsymbol{\lambda}})} = \|s^\star_m - \check{s}^\star_m\|_{\boldsymbol{\Psi}} = \nu(t^\star_m),$$

$$\chi_R(m) = \sqrt{\sum_{\boldsymbol{\lambda}\in m}\nu^2_R(\Psi^\star_{\boldsymbol{\lambda}})} = \|\check{s}^\star_m - \hat{s}^\star_m\|_{\boldsymbol{\Psi}}.$$

Proof. The proof follows from the linearity of ν and ν_R and Cauchy-Schwarz inequality. □

Lemma 3. *Let $\epsilon = \kappa'_2\|s\|_\infty/(\kappa'_3\kappa'_7\underline{D}(L_\bullet))$ and*

$$\Omega_T = \cap_{\boldsymbol{\lambda}\in m_\bullet}\left\{|\nu(\Psi^\star_{\boldsymbol{\lambda}})| \le \epsilon\right\}.$$

For all $x > 0$, there exists a measurable event $\Omega_m(x)$ on which

$$\chi^2(m)\mathbb{1}_{\Omega_T\cap\Omega_\sigma} \le 2\kappa'^2_1\kappa'_5\sum_{\boldsymbol{\lambda}\in m}\frac{\max\{\hat{\sigma}^2_{\boldsymbol{\lambda}},1\}}{\bar{n}} + 8\kappa'_2\|s\|_\infty\frac{x}{\bar{n}}.$$

and such that $\mathbb{P}(\Omega^c_m(x)) \le \exp(-x)$.

Proof. We observe that $\chi(m) = \mathcal{Z}(\mathcal{T}_m)$ where $\mathcal{T}_m = \{t \in S^\star_m | \|t\|_{\boldsymbol{\Psi}} = 1\}$. Let us set $z = \sqrt{\kappa'_2\|s\|_\infty x/\bar{n}}$ and consider a countable and dense subset $\mathcal{T}'_m$ of $\{t \in S^\star_m | \|t\|_{\boldsymbol{\Psi}} = 1, \max_{\boldsymbol{\lambda}\in m}|\langle t, \Psi_{\boldsymbol{\lambda}}\rangle| \le \epsilon/z\}$. Thanks to the localization property in Lemma 1,

$$\sup_{t\in\mathcal{T}'_m}\left\|\sum_{\lambda\in m_\bullet}\langle t,\Psi_{\boldsymbol{\lambda}}\rangle\Psi_{\boldsymbol{\lambda}}\right\|_\infty \le \kappa'_3\frac{\sqrt{\kappa'_2\|s\|_\infty}}{\sqrt{x/\bar{n}}}.$$

So Assumption **(Conc)** ensures that there exists $\Omega_m(x)$ such that $\mathbb{P}(\Omega^c_m(x)) \le \exp(-x)$ and on which

$$\mathcal{Z}(\mathcal{T}'_m) \le \kappa'_1\mathbb{E}\left[\mathcal{Z}(\mathcal{T}'_m)\right] + 2\sqrt{\kappa'_2\|s\|_\infty\frac{x}{\bar{n}}},$$

hence

$$\mathcal{Z}^2(\mathcal{T}'_m) \le 2\kappa'^2_1\mathbb{E}^2\left[\mathcal{Z}(\mathcal{T}'_m)\right] + 8\kappa'_2\|s\|_\infty\frac{x}{\bar{n}}.$$

As $Z(\mathcal{T}'_m) \le \chi(m)$, we obtain by convexity and Lemma 2

$$\mathbb{E}^2\left[\mathcal{Z}(\mathcal{T}'_m)\right] \le \mathbb{E}\left[\chi^2(m)\right] = \sum_{\boldsymbol{\lambda}\in m}\mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}}).$$

On $\Omega_T \cap \{\chi(m) \ge z\}$, $t^\star_m$ given by Lemma 2 satisfies $\sup_{\boldsymbol{\lambda}\in m}|\langle t^\star_m, \Psi_{\boldsymbol{\lambda}}\rangle| \le \epsilon/z$, so that $\chi^2(m) = \mathcal{Z}^2(\mathcal{T}'_m)$, while on $\Omega_T\cap\{\chi(m) < z\}$, $\chi^2(m) < \kappa'_2\|s\|_\infty x/\bar{n}$. The proof then follows from Assumption **(Var)**. □

7.4.2. *Proof of Theorem 1.* Let us fix $m \in \mathcal{M}^{\mathcal{P}}$. From the definition of $\hat{m}^{\mathcal{P}}$ and of $\hat{s}^\star_m$, we get

$$\hat{\gamma}(\tilde{s}^{\mathcal{P}}) + \mathrm{pen}(\hat{m}^{\mathcal{P}}) \le \hat{\gamma}(s^\star_m) + \mathrm{pen}(m).$$

For all $t, u \in \mathbb{L}_2([0,1]^d)$,

$$\hat{\gamma}(t) - \hat{\gamma}(u) = \|t - s\|^2_{\boldsymbol{\Psi}} - \|u - s\|^2_{\boldsymbol{\Psi}} - 2\hat{\nu}(t - u),$$

so

$$\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}} \le \|s - s^\star_m\|^2_{\boldsymbol{\Psi}} + 2\hat{\nu}(\tilde{s}^{\mathcal{P}} - s^\star_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}^{\mathcal{P}}).$$

Using the triangle inequality and Inequality (14) with $\theta = 1/4$ and $\theta = 1$, we get

$$\begin{aligned}2\hat{\nu}(\tilde{s}^{\mathcal{P}} - s^\star_m) &\le 2\|\tilde{s}^{\mathcal{P}} - s^\star_m\|_{\boldsymbol{\Psi}}(\chi(m\cup\hat{m}^{\mathcal{P}}) + \chi_R(m\cup\hat{m}^{\mathcal{P}}))\\ &\le \frac{1}{2}\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}} + \frac{1}{2}\|s - s^\star_m\|^2_{\boldsymbol{\Psi}} + 8\chi^2(m\cup\hat{m}^{\mathcal{P}}) + 8\chi^2_R(m\cup\hat{m}^{\mathcal{P}}),\end{aligned}$$

hence

$$\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}} \le 3\|s - s^\star_m\|^2_{\boldsymbol{\Psi}} + 16\chi^2(m\cup\hat{m}^{\mathcal{P}}) + 2(\mathrm{pen}(m) - \mathrm{pen}(\hat{m}^{\mathcal{P}})) + 16\chi^2_R(m\cup\hat{m}^{\mathcal{P}}). \quad (20)$$

Let us fix $\zeta > 0$ and set $\omega = \kappa_3(j_0, B, d) + \log(2)$ and $\Omega_\star(\zeta) = \cap_{m' \in \mathcal{M}^{\mathcal{P}}} \Omega_{m \cup m'}(\zeta + \omega D_m)$. We deduce from Lemma 3 that on $\Omega_\star(\zeta)$

$$\chi^2(m \cup \hat{m}^{\mathcal{P}}) \mathbb{1}_{\Omega_T \cap \Omega_\sigma} \leq 2\kappa_1'^2 \kappa_5' \sum_{\boldsymbol{\lambda} \in m \cup \hat{m}^{\mathcal{P}}} \frac{\max\{\hat{\sigma}^2_{\boldsymbol{\lambda}}, 1\}}{\bar{n}} + 8\kappa_2' \|s\|_\infty \frac{\omega(D_m + D_{\hat{m}^{\mathcal{P}}})}{\bar{n}} + 8\kappa_2' \|s\|_\infty \frac{\zeta}{\bar{n}}. \tag{21}$$

Besides, given Proposition 4, our choice of ω leads to

$$\mathbb{P}(\Omega_\star^c(\zeta)) \leq e^{-\zeta} \sum_{\ell = dj_0 + 1}^{L_\bullet + 1} \exp\left(-D(\ell) \left(\omega - \frac{\log(\sharp \mathcal{M}_\ell^{\mathcal{P}})}{D(\ell)} \right) \right) \leq e^{-\zeta}.$$

Choosing for instance

$$\mathrm{pen}(m) = c_1 \sum_{\boldsymbol{\lambda} \in m} \frac{\hat{\sigma}^2_{\boldsymbol{\lambda}}}{\bar{n}} + c_2 \frac{\bar{R} D_m}{\bar{n}},$$

with $c_1 \geq 16\kappa_1'^2 \kappa_5'$ and $c_2 \geq 64\kappa_2' \omega + 8\kappa_6'$ and integrating with respect to $\zeta > 0$, we deduce from (20), (21), Assumption **(Var)** and Assumption **(Conc)** that

$$\mathbb{E}\left[\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}} \mathbb{1}_{\Omega_T \cap \Omega_\sigma} \right] \leq 3\|s - s^\star_m\|^2_{\boldsymbol{\Psi}} + C \frac{\bar{R} D_m}{\bar{n}} + 64\kappa_2' \frac{\|s\|_\infty}{\bar{n}} + 8 \frac{w(\bar{n})}{\bar{n}}, \tag{22}$$

where C may depend on $\kappa_1', \kappa_2', \kappa_4', \kappa_5', \kappa_6', c_1, c_2$.

In order to bound $\mathbb{E}\left[\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}} \mathbb{1}_{\Omega_T^c \cup \Omega_\sigma^c} \right]$, we first notice that from the triangle inequality and Lemma 2

$$\begin{aligned} \|s - \tilde{s}^{\mathcal{P}}\|_{\boldsymbol{\Psi}} &\leq \|s - s^\star_{\hat{m}^{\mathcal{P}}}\|_{\boldsymbol{\Psi}} + \|s^\star_{\hat{m}^{\mathcal{P}}} - \hat{s}^\star_{\hat{m}^{\mathcal{P}}}\|_{\boldsymbol{\Psi}} \\ &\leq \|s\|_{\boldsymbol{\Psi}} + \chi(\hat{m}) + \chi_R(\hat{m}), \end{aligned}$$

hence

$$\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}} \leq \|s\|^2_{\boldsymbol{\Psi}} + 4\chi^2(m_\bullet) + 4\chi^2_R(m_\bullet).$$

Then setting $p_T = \mathbb{P}(\Omega_T^c)$ and $p_\sigma = \mathbb{P}(\Omega_\sigma^c)$, Cauchy-Schwarz inequality entails

$$\mathbb{E}\left[\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}} \mathbb{1}_{\Omega_T^c \cup \Omega_\sigma^c} \right] \leq 2(p_T + p_\sigma)\|s\|^2_{\boldsymbol{\Psi}} + 4\sqrt{p_T + p_\sigma} \left(\sqrt{\mathbb{E}\left[\chi^4(m_\bullet)\right]} + \sqrt{\mathbb{E}\left[\chi^4_R(m_\bullet)\right]} \right).$$

Let $\boldsymbol{\lambda} \in m_\bullet$, $\|\Psi^\star_{\boldsymbol{\lambda}}\|_\infty \leq \kappa 2^{d L_\bullet / 2}$, so applying Assumption **(Conc)** with $\mathcal{T} = \{\Psi^\star_{\boldsymbol{\lambda}}\}$ and $\mathcal{T} = \{-\Psi^\star_{\boldsymbol{\lambda}}\}$, we get

$$\mathbb{P}\left(|\nu(\Psi^\star_{\boldsymbol{\lambda}})| \geq \epsilon \right) \leq 2 \exp\left(- \min\left\{ \frac{\bar{n}\epsilon^2}{4\kappa_2' \|s\|_\infty}, \frac{\bar{n}\epsilon}{2\kappa_3' \kappa^d 2^{L_\bullet/2}} \right\} \right).$$

Then setting $\iota = \left(e(L_\bullet - dj_0 + d - 1)/(d-1)\right)^{d-1}$, Proposition 4 yields

$$p_T \leq 2\iota 2^{L_\bullet} \exp\left(-C\|s\|_\infty \frac{\bar{n}}{\iota^2 2^{L_\bullet}} \right) \leq \frac{C}{\bar{n}^2 (\log(\bar{n})/d)^{d+1}},$$

where C may depend on $\kappa_2', \kappa_3', \kappa_7', j_0, d$. Besides, we deduce from Assumption **(Conc)** and Lemma 1 that, for all $x > 0$,

$$\mathbb{P}\left(\chi(m_\bullet) \geq \kappa_1' \sqrt{\frac{\|s\|_\infty D_{m_\bullet}}{\bar{n}}} + \sqrt{\kappa_2' \|s\|_\infty \frac{x}{\bar{n}}} + \kappa_3' \kappa_7' \underline{D}(L_\bullet) \frac{x}{\bar{n}} \right) \leq \exp(-x).$$

For a nonnegative random variable U, Fubini's inequality implies

$$\mathbb{E}[U^4] = \int_0^\infty 4x^{p-1} \mathbb{P}(U \geq x) \mathrm{d}x$$

so

$$\mathbb{E}[\chi^4(m_\bullet)] \leq C \max\left\{ \frac{\iota^4 2^{2L_\bullet}}{\bar{n}^4}, \frac{\iota^2 2^{2L_\bullet}}{\bar{n}^2} \right\} \leq \frac{C}{\left(\log(\bar{n})/d\right)^{2(d+1)}}$$

where C may depend on $\kappa_1', \kappa_2', \kappa_3', \kappa_7', j_0, d$. Remembering (22) , we conclude that

$$\mathbb{E}\left[\|s - \tilde{s}^{\mathcal{P}}\|^2_{\boldsymbol{\Psi}} \right] \leq 3\|s - s^\star_m\|^2_{\boldsymbol{\Psi}} + C_1 \frac{\bar{R} D_m}{\bar{n}} + C_2 \frac{\|s\|_\infty}{\bar{n}} + C_3 \max\{\|s\|^2_{\boldsymbol{\Psi}}, 1\} \left(\frac{1}{\bar{n} \left(\log(\bar{n})/d\right)^{3(d+1)/2}} + \frac{w(\bar{n})}{\bar{n}} \right),$$

where C_1 may depend on $\kappa_1', \kappa_2', \kappa_4', \kappa_5', \kappa_6', c_1, c_2, C_2$ may depend on κ_2', C_3 may depend $\kappa_1', \kappa_2', \kappa_3', \kappa_7', j_0, d$.

7.5. Proofs of Corollaries 1 to 5.

7.5.1. *Proof of Corollary 1.* Assumption **(Conc)** is a straightforward consequence of Talagrand's inequality, as stated for instance in [Mas07] (Inequality (5.50), and is satisfied, whatever $\theta > 0$, for

$$\bar{n} = n, \kappa_1' = 1 + \theta, \kappa_2' = 2, \kappa_3' = (1/3 + 1/\theta)/2. \tag{23}$$

For all $\boldsymbol{\lambda} \in m_\bullet$, $\hat{\sigma}^2_{\boldsymbol{\lambda}}$ is an unbiased estimator for $\mathrm{Var}(\Psi_{\boldsymbol{\lambda}}(Y_1))$. Besides, the existence of Ω_σ follows from Lemma 1 in [RBRTM11] with $\gamma = 2$. Thus Assumptions **(Var)** and **(Rem)** are satisfied by taking $\kappa_4' = 1, \kappa_5'$ that only depends on κ and d, $\kappa_6' = 0$, $w(n) = C(\kappa, j_0, d)/\log^{d+1}(n)$.

7.5.2. *Proof of Corollary 2.* Setting $Y_i = (F_1(X_{i1}), \dots, F_d(X_{id})), i = 1, \dots, n$, we recover the previous density estimation framework, so Assumption **(Conc)** is still satisfied with (24). Setting $\hat{Y}_i = (F_{n1}(X_{i1}), \dots, \hat{F}_{nd}(X_{id})), i = 1, \dots, n$, and

$$\check{\sigma}^2_{\boldsymbol{\lambda}} = \frac{1}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\Psi_{\boldsymbol{\lambda}}(Y_i) - \Psi_{\boldsymbol{\lambda}}(Y_j)\right)^2,$$

we observe that, for all $\boldsymbol{\lambda} \in m_\bullet$

$$\max\left\{\hat{\sigma}^2_{\boldsymbol{\lambda}} - 4\check{\sigma}^2_{\boldsymbol{\lambda}}, \check{\sigma}^2_{\boldsymbol{\lambda}} - 4\hat{\sigma}^2_{\boldsymbol{\lambda}}\right\} \le 8R_{\boldsymbol{\lambda}}(n)$$

where

$$R_{\boldsymbol{\lambda}}(n) = \frac{1}{n(n-1)} \sum_{i=2}^{n} (i-1) \left(\Psi_{\boldsymbol{\lambda}}(\hat{Y}_i) - \Psi_{\boldsymbol{\lambda}}(Y_i)\right)^2.$$

Using the same arguments as in the proof of Proposition 2, we get for all $\boldsymbol{\lambda} \in m_\bullet$ and all $m \subset m_\bullet$

$$\mathbb{E}\left[R_{\boldsymbol{\lambda}}(n)\right] \le C(\kappa, d) 2^{3L_\bullet} \log(n)/n,$$

$$R_{\boldsymbol{\lambda}}(n) \le C(\kappa, d) 2^{3L_\bullet} \log(n)/n$$

except on a set with probability smaller than $2d/n$,

$$\mathbb{E}\left[\|\check{s}^\star_m - \hat{s}^\star_m\|^2\right] \le C(\kappa, d, j_0) L_\bullet^{d-1} 2^{4L_\bullet} \log(n)/n,$$

and

$$\sqrt{\mathbb{E}\left[\|\check{s}^\star_{m_\bullet} - \hat{s}^\star_{m_\bullet}\|^4\right]} \le C(\kappa, d, j_0) L_\bullet^{d-1} 2^{4L_\bullet}/\sqrt{n}.$$

Building on the proof of Corollary 1, we conclude that Assumptions **(Var)** and **(Rem)** are satisfied with κ_4', κ_5' that only depend on κ, j_0, d, $\kappa_6' = 0$, and $w(n) = \sqrt{n}\log^{d-1}(n)$.

7.5.3. *Proof of Corollary 3.* Assumption **(Conc)** is a straightforward consequence of Talagrand's inequality for Poisson processes proved by [RB03] (Corollary 2), and is satisfied, whatever $\theta > 0$, by

$$\bar{n} = \mathrm{Vol}_d(Q), \kappa_1' = 1 + \theta, \kappa_2' = 12, \kappa_3' = (1.25 + 32/\theta). \tag{24}$$

For all $\boldsymbol{\lambda} \in m_\bullet$, $\hat{\sigma}^2_{\boldsymbol{\lambda}}$ is an unbiased estimator for $\int_Q \Psi^2_{\boldsymbol{\lambda}} s = \mathrm{Vol}_d(Q)\mathrm{Var}(\check{\beta}_{\boldsymbol{\lambda}})$. Besides, the existence of Ω_σ follows from Lemma 6.1 in [RBR10]. Thus Assumptions **(Var)** and **(Rem)** are satisfied by taking $\kappa_4' = 1, \kappa_5'$ that only depends on κ and d, $\kappa_6' = 0$, $w(\bar{n}) = C(\kappa, j_0, d)/\log^{d+1}(\bar{n})$.

7.5.4. *Proof of Corollary 4.* The proof is similar to that of Corollary 3 with $\bar{n} = T$.

7.5.5. *Proof of Corollary 5.* Regarding Assumption **(Conc)**, the proof is similar to that of Corollary 3 with $\bar{n} = n\Delta$. For all $\boldsymbol{\lambda} \in m_\bullet$, let

$$\check{\sigma}^2_{\boldsymbol{\lambda}} = \frac{1}{n\Delta} \iint_{[0,n\Delta]\times Q} \Psi^2_{\boldsymbol{\lambda}}(x) N(\mathrm{d}t, \mathrm{d}x).$$

For any bounded measurable function g on Q, let

$$R(g) = \int_Q g(\mathrm{d}\widehat{M} - \mathrm{d}M), \quad I(g) = \int_Q g\mathrm{d}M - \mathbb{E}\left[\int_Q g\mathrm{d}M\right], \quad \hat{I}(g) = \int_Q g\mathrm{d}\widehat{M} - \mathbb{E}\left[\int_Q g\mathrm{d}\widehat{M}\right],$$

then

$$R(g) = \hat{I}(g) - I(g) + D_\Delta(g)$$

where D_Δ has been defined in the proof of Proposition 3. Notice that

$$\hat{\sigma}^2_{\boldsymbol{\lambda}} - \check{\sigma}^2_{\boldsymbol{\lambda}} = R(\Psi^2_{\boldsymbol{\lambda}})$$

and

$$\|\hat{s}^\star_{m_\bullet} - \check{s}^\star_{m_\bullet}\|^2_{\boldsymbol{\Psi}} = \sum_{\boldsymbol{\lambda}\in m_\bullet} R^2(\Psi_{\boldsymbol{\lambda}}).$$

In the course of the proof of Proposition 3, we have shown that, for bounded and Lipschitz functions g on Q,

$$|D_\Delta(g)| \le C(\lambda_\varepsilon, \varepsilon, f, Q) \max\{\|g\|_1, \|g\|_\infty, \|g\|_L\}\, \Delta$$

provided Δ and ε are small enough. Besides, both $\hat{I}(g)$ and $I(g)$ satisfy Bernstein inequalities (Bernstein inequality as stated in [Mas07], Proposition 2.9, for the former, and Bernstein inequality as stated in [RB03], Proposition 7, for the latter). Combining all these arguments yields Corollary 5.

7.6. **Proof of Proposition 5.** For $\alpha > 0$, we set $r = \lfloor \alpha \rfloor + 1$.

(i). From (13), it is easy to see that $\Delta^r_{h_\ell,\ell}(f, \mathbf{x}) = \Delta^r_{h_\ell}(u_\ell, x_\ell)$. Thus $w^{\{\ell\}}_r(f, t_\ell)_p = w_r(u_\ell, t_\ell)_p$ and $w^{\mathbf{e}}_r(f, \mathbf{t}_\mathbf{e})_p = 0$ as soon as $\mathbf{e} \subset \{1, \dots, d\}$ contains at least two elements. Therefore,

$$\|f\|_{SB^\alpha_{p,q,(d)}} \le C(p) \sum_{\ell=1}^d \|u_\ell\|_{B^\alpha_{p,q,(1)}}.$$

(ii). For the sake of readability, we shall detail only two special cases. Let us first deal with the case $f(\mathbf{x}) = \prod_{\ell=1}^d u_\ell(x_\ell)$ where each $u_\ell \in B^\alpha_{p,q,(1)}$. From (13),

$$\Delta^{r,\mathbf{e}}_{\mathbf{h}}(f, \mathbf{x}) = \prod_{\ell\in\mathbf{e}} \Delta^r_{h_\ell}(u_\ell, x_\ell) \prod_{\ell\notin\mathbf{e}} u_\ell(x_\ell),$$

so

$$\|f\|_{SB^\alpha_{p,q,(d)}} \le 2^d \prod_{\ell=1}^d \|u_\ell\|_{B^\alpha_{p,q,(1)}}.$$

Let us now assume that $d = 3$ and that $f(\mathbf{x}) = u_1(x_1)u_{2,3}(x_2, x_3)$ where $u_1 \in B^{\alpha_1}_{p,q,(1)}$ and $u_{2,3} \in B^{\alpha_2}_{p,q,(2)}$. We set $r_\ell = \lfloor \alpha_\ell \rfloor + 1$ for $\ell = 1, 2$, and $\bar{r} = \lfloor \bar{\alpha} \rfloor + 1$, where $\bar{\alpha} = \min(\alpha_1, \alpha_2/2)$. For $0 < t_1, t_2, t_3 < 1$, we easily have

$$\|f\|_p = \|u_1\|_p \|u_{2,3}\|_p$$
$$t_1^{-\bar{\alpha}} w^{\{1\}}_{\bar{r}}(f, t_1)_p \le t_1^{-\alpha_1} w_{r_1}(u_1, t_1)_p \|u_{2,3}\|_p$$
$$t_\ell^{-\bar{\alpha}} w^{\{\ell\}}_{\bar{r}}(f, t_\ell)_p \le \|u_1\|_p t_\ell^{-\alpha_\ell} w^{\{\ell\}}_{r_\ell}(u_{2,3}, t_\ell)_p, \text{ for } \ell = 2, 3$$
$$t_1^{-\bar{\alpha}} t_\ell^{-\bar{\alpha}} w^{\{1,\ell\}}_{\bar{r}}(f, t_1, t_\ell)_p \le t_1^{-\alpha_1} w_{r_1}(u_1, t_1)_p t_\ell^{-\alpha_\ell} w^{\{\ell\}}_{r_\ell}(u_{2,3}, t_\ell)_p, \text{ for } \ell = 2, 3.$$

Besides, we deduce from (13) that

$$\|\Delta^{\bar{r}}_h(g, .)\|_p \le C(\bar{r}, p)\|g\|_p,$$

and as operators $\Delta^{\bar{r}}_{h_\ell,\ell}$ commute, we have

$$t_2^{-\bar{\alpha}} t_3^{-\bar{\alpha}} w^{\{2,3\}}_{\bar{r}}(f, t_2, t_3)_p \le C(p, \bar{r}) \|u_1\|_p t_2^{-\bar{\alpha}} t_3^{-\bar{\alpha}} \min\left\{w^{\{2\}}_{\bar{r}}(u_{2,3}, t_2)_p, w^{\{3\}}_{\bar{r}}(u_{2,3}, t_3)_p\right\}.$$

The inequality of arithmetic and geometric means entails that $2t_2^{-\bar\alpha}t_3^{-\bar\alpha} \le t_2^{-2\bar\alpha} + t_3^{-2\bar\alpha}$, so

$$t_2^{-\bar\alpha}t_3^{-\bar\alpha} w_{\bar r}^{\{2,3\}}(f,t_2,t_3)_p \le C(p,\bar r)\|u_1\|_p \left(t_2^{-\alpha_2} w_{r_2}^{\{2\}}(u_{2,3},t_2)_p + t_3^{-\alpha_3} w_{r_2}^{\{3\}}(u_{2,3},t_3)_p\right).$$

In the same way,

$$t_1^{-\bar\alpha}t_2^{-\bar\alpha}t_3^{-\bar\alpha} w_{\bar r}^{\{1,2,3\}}(f,t_1,t_2,t_3)_p \le C(p,\bar r)t_1^{-\alpha_1} w_{r_1}(u_1,t_1)_p \left(t_2^{-\alpha_2} w_{r_2}^{\{2\}}(u_{2,3},t_2)_p + t_3^{-\alpha_3} w_{r_2}^{\{3\}}(u_{2,3},t_3)_p\right).$$

Consequently,

$$\|f\|_{SB^{\bar\alpha}_{p,q,(d)}} \le C(p,\bar r)\|u_1\|_{B^{\alpha_1}_{p,q,(1)}}\|u_{2,3}\|_{B^{\alpha_2}_{p,q,(2)}}.$$

(iii). The proof follows from the chain rule for higher order derivatives of a composite function. Notice that for all $1 \le \ell \le d$ and $1 \le r \le \alpha - 1$, $u_\ell^{(r)} \in W^{\alpha-r}_{p,(1)}$, with $\alpha - r > 1/p$, so $u_\ell^{(r)}$ is bounded.

(iv). The proof follows from a d-variate extension of Theorem 4.1, Inequality (10) in [PST13] (see also [DL93] Chapter 6, Theorem 3.1).

(v). See Theorem 3.10 in [NS16b].

7.7. **Proof of Theorem 2.** We recall that for any finite sequence $(a_i)_{i\in I}$, and $0 < p_1, p_2 < \infty$,

$$\left(\sum_{i\in I}|a_i|^{p_2}\right)^{1/p_2} \le |I|^{(1/p_2 - 1/p_1)_+}\left(\sum_{i\in I}|a_i|^{p_1}\right)^{1/p_1}.$$

Besides, we have proved in the course of the proof of Proposition 4 that

$$\boldsymbol{J}_\ell \le c_1(M,d)(\ell - dj_0 + d - 1)^{d-1}.$$

In the hyperbolic basis, f admits a unique decomposition of the form

$$f = \sum_{\ell=dj_0}^{\infty}\sum_{\boldsymbol{\lambda}\in U\boldsymbol{\nabla}(\ell)}\langle f,\Psi_{\boldsymbol{\lambda}}\rangle\Psi^\star_{\boldsymbol{\lambda}}.$$

Defining

$$f_\bullet = \sum_{\ell=dj_0}^{L_\bullet}\sum_{\boldsymbol{\lambda}\in U\boldsymbol{\nabla}(\ell)}\langle f,\Psi_{\boldsymbol{\lambda}}\rangle\Psi^\star_{\boldsymbol{\lambda}},$$

we have for finite $q > 0$, using the aforementioned reminders,

$$\begin{aligned}
\|f - f_\bullet\|^2_{\boldsymbol{\Psi}} &= \sum_{\ell=L_\bullet+1}^{\infty}\sum_{\boldsymbol{j}\in\boldsymbol{J}_\ell}\sum_{\boldsymbol{\lambda}\in\boldsymbol{\nabla}_{\boldsymbol{j}}}\langle f,\Psi_{\boldsymbol{\lambda}}\rangle^2 \\
&\le \sum_{\ell=L_\bullet+1}^{\infty}\sum_{\boldsymbol{j}\in\boldsymbol{J}_\ell}(\sharp\boldsymbol{\nabla}_{\boldsymbol{j}})^{2(1/2-1/p)_+}\left(\sum_{\boldsymbol{\lambda}\in\boldsymbol{\nabla}_{\boldsymbol{j}}}|\langle f,\Psi_{\boldsymbol{\lambda}}\rangle|^p\right)^{2/p} \\
&\le C(B,d,p)\sum_{\ell=L_\bullet+1}^{\infty}2^{2\ell(1/2-1/p)_+}\sum_{\boldsymbol{j}\in\boldsymbol{J}_\ell}\left(\sum_{\boldsymbol{\lambda}\in\boldsymbol{\nabla}_{\boldsymbol{j}}}|\langle f,\Psi_{\boldsymbol{\lambda}}\rangle|^p\right)^{2/p} \\
&\le C(B,d,p)\sum_{\ell=L_\bullet+1}^{\infty}2^{2\ell(1/2-1/p)_+}\sharp\boldsymbol{J}_\ell^{2(1/2-1/q)_+}\left(\sum_{\boldsymbol{j}\in\boldsymbol{J}_\ell}\left(\sum_{\boldsymbol{\lambda}\in\boldsymbol{\nabla}_{\boldsymbol{j}}}|\langle f,\Psi_{\boldsymbol{\lambda}}\rangle|^p\right)^{q/p}\right)^{2/q} \\
&\le C(B,d,p)\sum_{\ell=L_\bullet+1}^{\infty}2^{2\ell(1/2-1/p)_+}(\ell - dj_0 + d - 1)^{2(d-1)(1/2-1/q)_+}R^2 2^{-2\ell(\alpha+1/2-1/p)} \\
&\le C(B,d,p)R^2\sum_{\ell=L_\bullet+1}^{\infty}(\ell - dj_0 + d - 1)^{2(d-1)(1/2-1/q)_+}2^{-2\ell(\alpha-(1/p-1/2)_+)} \\
&\le C(B,\alpha,p,d)R^2 L_\bullet^{2(d-1)(1/2-1/q)_+}2^{-2L_\bullet(\alpha-(1/p-1/2)_+)}.
\end{aligned}$$

The case $q = \infty$ can be treated in the same way.

Let us fix $k \in \{0, \dots, L_\bullet - \ell_1\}$ and define $\bar{m}(\ell_1 + k, f)$ as the subset of $U\boldsymbol{\nabla}(\ell_1 + k)$ such that $\{|\langle f, \Psi_{\boldsymbol{\lambda}}\rangle|; \boldsymbol{\lambda} \in \bar{m}(\ell_1 + k, f)\}$ are the $N(\ell_1, k)$ largest elements among $\{|\langle f, \Psi_{\boldsymbol{\lambda}}\rangle|; \boldsymbol{\lambda} \in U\boldsymbol{\nabla}(\ell_1 + k)\}$. We then consider the approximation for f given by

$$A(\ell_1, f) = \sum_{\ell=dj_0}^{\ell_1-1} \sum_{\boldsymbol{\lambda}\in U\boldsymbol{\nabla}(\ell)} \langle f, \Psi_{\boldsymbol{\lambda}}\rangle \Psi^\star_{\boldsymbol{\lambda}} + \sum_{k=0}^{L_\bullet-\ell_1} \sum_{\boldsymbol{\lambda}\in \bar{m}(\ell_1+k,f)} \langle f, \Psi_{\boldsymbol{\lambda}}\rangle^2$$

and the set

$$m_{\ell_1}(f) = \left(\bigcup_{\ell=dj_0}^{\ell_1-1} U\boldsymbol{\nabla}(\ell) \right) \cup \left(\bigcup_{k=0}^{L_\bullet-\ell_1} \bar{m}(\ell_1 + k, f) \right).$$

Let us first assume that $0 < p \le 2$. Using Lemma 4.16 in [Mas07] and (7.7), we get

$$\begin{aligned}
\|f_\bullet - A(\ell_1, f))\|^2_{\boldsymbol{\Psi}} &= \sum_{k=0}^{L_\bullet-\ell_1} \sum_{\boldsymbol{\lambda}\in U\boldsymbol{\nabla}(\ell_1+k)\setminus \bar{m}(\ell_1+k,f)} \langle f, \Psi_{\boldsymbol{\lambda}}\rangle^2 \\
&\le \sum_{k=0}^{L_\bullet-\ell_1} \left(\sum_{\boldsymbol{\lambda}\in U\boldsymbol{\nabla}(\ell_1+k)} |\langle f, \Psi_{\boldsymbol{\lambda}}\rangle|^p \right)^{2/p} / (N(\ell_1, k) + 1)^{2(1/p-1/2)} \\
&\le \sum_{k=0}^{L_\bullet-\ell_1} \sharp \boldsymbol{J}^{2(1/p-1/q)+}_{\ell_1+k} \left(\sum_{\boldsymbol{j}\in \boldsymbol{J}_{\ell_1+k}} \left(\sum_{\boldsymbol{\lambda}\in \boldsymbol{\nabla}_{\boldsymbol{j}}} |\langle f, \Psi_{\boldsymbol{\lambda}}\rangle|^p \right)^{q/p} \right)^{2/q} / (N(\ell_1, k) + 1)^{2(1/p-1/2)}.
\end{aligned}$$

Besides, it follows from (11) that

$$N(\ell_1, k) + 1 \ge 2M^{-d}2^{-d}(d-1)^{-(d-1)}(\ell_1 + k - dj_0 + d - 1)^{d-1}2^{\ell_1}(k+2)^{-(d+2)}.$$

Therefore

$$\|f_\bullet - A(\ell_1, f))\|^2_{\boldsymbol{\Psi}} \le C(\alpha, p, d)R^2(\ell_1 - dj_0 + d - 1)^{2(d-1)(1/2-1/\max(p,q))}2^{-2\alpha\ell_1}.$$

In case $p \ge 2$, the same kind of upper-bound follows from

$$\|f_\bullet - A(\ell_1, f))\|^2_{\boldsymbol{\Psi}} \le \sum_{k=0}^{L_\bullet-\ell_1} \sharp U\boldsymbol{\nabla}(\ell_1 + k)^{2(1/2-1/p)} \left(\sum_{\boldsymbol{\lambda}\in U\boldsymbol{\nabla}(\ell_1+k)} |\langle f, \Psi_{\boldsymbol{\lambda}}\rangle|^p \right)^{2/p}.$$

Last,

$$\|f - A(\ell_1, f))\|^2_{\boldsymbol{\Psi}} = \|f - f_\bullet\|^2_{\boldsymbol{\Psi}} + \|f_\bullet - A(\ell_1, f))\|^2_{\boldsymbol{\Psi}}$$

which completes the proof.

References

[ACF14] F. Autin, G. Claeskens, and J.-M. Freyermuth. Hyperbolic wavelet thresholding methods and the curse of dimensionality through the maxiset approach. *Applied and Computational Harmonic Analysis*, 36(2):239 – 255, 2014.

[ACF15] Florent Autin, Gerda Claeskens, and Jean-Marc Freyermuth. Asymptotic performance of projection estimators in standard and hyperbolic wavelet bases. *Electron. J. Statist.*, 9(2):1852–1883, 2015.

[AL11] Nathalie Akakpo and Claire Lacour. Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electronic journal of statistics*, 5:1618–1653, 2011.

[Bar11] Yannick Baraud. Estimator selection with respect to hellinger-type risks. *Probability Theory and Related Fields*, 151(1):353–401, 2011.

[BB14] Yannick Baraud and Lucien Birgé. Estimating composite functions by model selection. *Ann. Inst. H. Poincaré Probab. Statist.*, 50(1):285–314, 02 2014.

[BB16] Yannick Baraud and Lucien Birgé. Rho-estimators revisited: general theory and applications. Working paper or preprint, June 2016.

[BBM99] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[Ber96] Jean Bertoin. *Lévy processes*, volume 121 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1996.

[BGH09] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009.

[Bir06] Lucien Birgé. Model selection via testing : an alternative to (penalized) maximum likelihood estimators. *Annales de l'I.H.P. Probabilités et statistiques*, 42(3):273–325, 2006.

[BM00] L. Birgé and P. Massart. An adaptive compression algorithm in besov spaces. *Constructive Approximation*, 16(1):1–36, 2000.

[BPP13] Rida Benhaddou, Marianna Pensky, and Dominique Picard. Anisotropic de-noising in functional deconvolution model with dimension-free convergence rates. *Electron. J. Statist.*, 7:1686–1715, 2013.

[BS11] Gérard Bourdaud and Winfried Sickel. Composition operators on function spaces with fractional order of smoothness. *RIMS Kokyuroku Bessatsu B*, 26:93–132, 2011.

[CDV93] Albert Cohen, Ingrid Daubechies, and Pierre Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1):54–81, 1993.

[CT04] Rama Cont and Peter Tankov. *Financial modelling with jump processes*. Chapman & Hall/CRC Financial Mathematics Series. Chapman & Hall/CRC, Boca Raton, FL, 2004.

[DIT14] Arnak Dalalyan, Yuri Ingster, and Alexandre B. Tsybakov. Statistical inference in compound functional models. *Probability Theory and Related Fields*, 158(3):513–532, 2014.

[DKU99] Wolfgang Dahmen, Angela Kunoth, and Karsten Urban. Biorthogonal spline wavelets on the interval—stability and moment conditions. *Appl. Comput. Harmon. Anal.*, 6(2):132–196, 1999.

[DL93] Ronald A. DeVore and George G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993.

[Don97] D. L. Donoho. CART and best-ortho-basis: a connection. *Ann. Statist.*, 25(5):1870–1911, 1997.

[DTU16] Dinh Dung, Vladimir N. Temlyakov, and Tino Ullrich. Hyperbolic cross approximation. *arXiv preprint arXiv:1601.03978v1*, 2016.

[Dun01] Dinh Dung. Non-linear approximations using sets of finite cardinality or finite pseudo-dimension. *Journal of Complexity*, 17(2):467 – 492, 2001.

[FL09] José E Figueroa-López. Nonparametric estimation for lévy models based on discrete-sampling. *Lecture notes-monograph series*, pages 117–146, 2009.

[FLH09] José E. Figueroa-López and Christian Houdré. Small-time expansions for the transition distributions of Lévy processes. *Stochastic Process. Appl.*, 119(11):3862–3889, 2009.

[Hep04] Wang Heping. Representation and approximation of multivariate functions with mixed smoothness by hyperbolic wavelets. *J. Math. Anal. Appl.*, 291(2):698–715, 2004.

[HM07] Joel L. Horowitz and Enno Mammen. Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Ann. Statist.*, 35(6):2589–2619, 12 2007.

[Hoc02a] Reinhard Hochmuth. N-term approximation in anisotropic function spaces. *Math. Nachr.*, 244:131–149, 2002.

[Hoc02b] Reinhard Hochmuth. Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2):179–208, 2002.

[IS07] Yu. Ingster and I. Suslina. Estimation and detection of high-variable functions from Sloan–Woźniakowski space. *Mathematical Methods of Statistics*, 16(4):318–353, 2007.

[JLT09] Anatoli B. Juditsky, Oleg V. Lepski, and Alexandre B. Tsybakov. Nonparametric estimation of composite functions. *Ann. Statist.*, 37(3):1360–1404, 06 2009.

[KT06] Jan Kallsen and Peter Tankov. Characterization of dependence of multidimensional Lévy processes using Lévy copulas. *Journal of Multivariate Analysis*, 97(7):1551 – 1572, 2006.

[Lep13] Oleg Lepski. Multivariate density estimation under sup-norm loss: Oracle approach, adaptation and independence structure. *Ann. Statist.*, 41(2):1005–1034, 04 2013.

[Mas90] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990.

[Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[Mil71] P. W. Millar. Path behavior of processes with stationary independent increments. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 17:53–73, 1971.

[MN09] Alexander J. McNeil and Johanna Nešlehová. Multivariate Archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions. *Ann. Statist.*, 37(5B):3059–3097, 10 2009.

[Mou11] Madani Moussai. The composition in multidimensional Triebel–Lizorkin spaces. *Mathematische Nachrichten*, 284(2-3):317–331, 2011.

[Nel06] Roger B. Nelsen. *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition, 2006.

[Neu00] Michael H. Neumann. Multivariate wavelet thresholding in anisotropic function spaces. *Statist. Sinica*, 10(2):399–431, 2000.

[NS16a] Van Kien Nguyen and Winfried Sickel. Isotropic and dominating mixed Besov spaces: a comparison. *arXiv preprint arXiv:1601.04000*, 2016.

[NS16b] Van Kien Nguyen and Winfried Sickel. Pointwise multipliers for Sobolev and Besov spaces of dominating mixed smoothness. *arXiv preprint arXiv:1608.03414*, 2016.

[NvS97] Michael H. Neumann and Rainer von Sachs. Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.*, 25(1):38–76, 1997.

[PST13] MK Potapov, BV Simonov, and S Yu Tikhonov. Mixed moduli of smoothness in $\mathbb{L}_p, 1 < p < \infty$: a survey. *Surveys in Approximation Theory*, 8(18), 2013.

[RB03] Patricia Reynaud-Bouret. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields*, 126(1):103–153, 2003.

[RBR10] Patricia Reynaud-Bouret and Vincent Rivoirard. Near optimal thresholding estimation of a Poisson intensity on the real line. *Electron. J. Stat.*, 4:172–238, 2010.

[RBRTM11] Patricia Reynaud-Bouret, Vincent Rivoirard, and Christine Tuleau-Malot. Adaptive density estimation: a curse of support? *J. Statist. Plann. Inference*, 141(1):115–139, 2011.

[Reb15a] Gilles Rebelles. $\mathbb{L}_p$- adaptive estimation of an anisotropic density under independence hypothesis. *Electron. J. Statist.*, 9(1):106–134, 2015.

[Reb15b] Gilles Rebelles. Pointwise adaptive estimation of a multivariate density under independence hypothesis. *Bernoulli*, 21(4):1984–2023, 11 2015.

[RW02] Ludger Rüschendorf and Jeannette H. C. Woerner. Expansion of transition distributions of Lévy processes in small time. *Bernoulli*, 8(1):81–96, 2002.

[Sat99] Ken-iti Sato. *Lévy processes and infinitely divisible distributions*, volume 68 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999. Translated from the 1990 Japanese original, Revised by the author.

[Skl59] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

[ST87] Hans-Jürgen Schmeisser and Hans Triebel. *Topics in Fourier analysis and function spaces*. A Wiley-Interscience Publication. John Wiley & Sons, Ltd., Chichester, 1987.

[UK11] Florian AJ Ueltzhöfer and Claudia Klüppelberg. An oracle inequality for penalised projection estimation of lévy densities from high-frequency observations. *Journal of Nonparametric Statistics*, 23(4):967–989, 2011.

[YB99] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.