Estimation in Inverse Problems and second-generation wavelets

Dominique Picard and Gérard Kerkyacharian

Abstract. We consider the problem of recovering a function f, when we receive a blurred (by a linear operator) and noisy version : $Y_\varepsilon = Kf + \varepsilon \dot{W}$. We will have as guides 2 famous examples of such inverse problems : the deconvolution and the Wicksell problem. The direct problem (K is the identity) isolates the denoising operation. It can't be solved unless accepting to estimate a smoothed version of f: for instance, if f has an expansion on a basis, this smoothing might correspond to stopping the expansion at some stage m. Then a crucial problem lies in finding an equilibrium for m considering the fact that for m large, the difference between f and its smoothed version is small, whereas the random effect introduces an error which is increasing with m. In the true inverse problem, in addition to denoising we have to 'inverse the operator' K, which operation not only creates the usual difficulties, but also introduces the necessity to control the additional instability due to the inversion of the random noise. Our purpose here is to emphasize the fact that in such a problem, there generally exists a basis which is fully adapted to the problem, where for instance the inversion remains very stable : this is the Singular Value Decomposition basis. On the other hand, the SVD basis might be difficult to determine and manipulate numerically, it also might not be appropriate for the accurate description of the solution with a small number of parameters. Also in many practical situations, the signal provides inhomogeneous regularity, and its local features are especially interesting to recover. In such cases, other bases (in particular localised bases such as wavelet bases) may be much more appropriate to give a good representation of the object at hand. Our approach here will be to produce estimation procedures trying to keep the advantages of localisation without loosing the stability and computability of SVD decompositions. We will especially consider two cases. In the first one (which is the case of the deconvolution example) we show that a fairly simple algorithm (WAVE-VD) using an appropriate thresholding technique performed on a standard wavelet system, enables us to estimate the object with rates which are almost optimal up to logarithm factors for any $\mathbb{L}_p$ loss function, and on the whole range of Besov spaces. In the second case (which is the case of the Wicksell example where the SVD bases lies in the range of Jacobi polynomials), we prove that quite a similar algorithm (NEED-VD) can be performed provided replacing the standard wavelet system by a second generation wavelet-type basis : the Needlets. We use here the construction (essentially following the work of Petrushev and co-authors) of a localised frame linked with a prescribed basis (here Jacobi polynomials) using a Littlewood Paley decomposition combined with a cubature formula. Section 5 describes the direct case ($K = I$). It has its own interest and will act as a guide for understanding the 'true' inverse models for a reader which is unfamiliar with nonparametric statistical estimation. It can be read first. Section 1 introduces the general inverse problem and describes the examples of deconvolution and Wicksell problem. A review of standard methods is given with a special focus on SVD methods. Section 2 describes the WAVE-VD procedure. Section 3 and 4 give a description of the needlets constructions and the performances of the NEED-VD procedure.

Mathematics Subject Classification (2000). Primary 62G07, 62G20, 62C20.

Keywords. Nonparametric estimation, denoising, inverse models, Thresholding, Meyer Wavelet, Singular value decomposition, Littlewood-Paley decomposition.

1. Inverse Models

Let $\mathbb{H}$ and $\mathbb{K}$ be two Hilbert spaces. K is a linear operator : $f \in \mathbb{H} \mapsto Kf \in \mathbb{K}$. The standard linear ill-posed inverse problem consists of recovering a good approximation f_ε of

f solution of

$$g = Kf \tag{1}$$

when only a perturbation g_ε of g is observed. In this paper, we will consider the case where this perturbation is an additive stochastic white noise. Namely we observe Y_ε defined by the following equation :

$$Y_\varepsilon = Kf + \varepsilon \dot{W},\ \mathbb{H},\ \mathbb{K}, \tag{2}$$

ε is the amplitude of the noise. It is supposed to be a small parameter which will tend to 0. Our error will be measured in terms of this small parameter.

$\dot{W}$ is a $\mathbb{K}$-white noise : i.e. for any $g,\ h$ in $\mathbb{K}$, $\xi(g) := (\dot{W}, g)_{\mathbb{K}},\ \xi(h) := (\dot{W}, h)_{\mathbb{K}}$ are forming a random gaussian vector, centered, with marginal variance $\|g\|_{\mathbb{K}},\ \|h\|_{\mathbb{K}}$, and covariance $(g,h)_{\mathbb{K}}$ (with the obvious extension when one considers k functions instead of 2).

Equation (2) means that for any g in $\mathbb{K}$, we observe : $Y_\varepsilon(g) := (Y_\varepsilon, g)_{\mathbb{K}} = (Kf, g)_{\mathbb{K}} + \varepsilon\xi(g)$ where $\xi(g) \sim N(0, \|g\|^2)$, and $Y(g),\ Y(h)$ are independent random variables for orthogonal functions g and h.

The case where K is the identity, is called the 'direct model' and summarized as a memento in section 5. The reader which is unfamiliar with nonparametric statistical estimation is invited to refer to this section, which will act as a guide, for understanding the more general inverse models. In particular, in section 5, it is recalled that the model (2) appears as an approximation of models which appear in real practical situations, for instance the case where (2) is replaced by a discretisation.

1.1. Two examples: Deconvolution and Wicksell's problems.

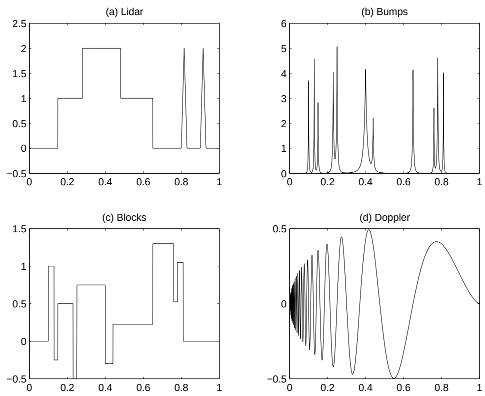
1.1.1. Deconvolution. The following problem is probably one of the famest among inverse problems in signal processing. In the deconvolution problem, we consider the following operator: let in this case $\mathbb{H} = \mathbb{K}$ be the set of square integrable periodic functions, with the standard $\mathbb{L}_2([0,1])$ norm, and consider:

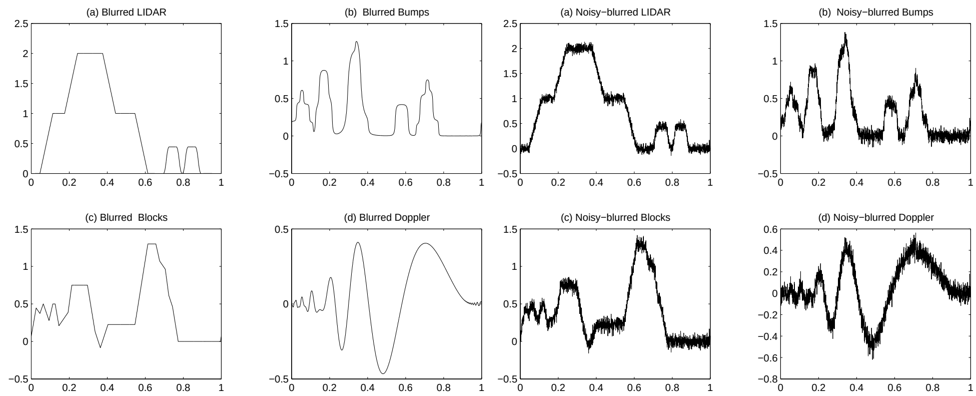
$$f \in\ \mathbb{H} \mapsto\ Kf = \int_0^1 \gamma(u-t)f(t)dt \in \mathbb{H} \tag{3}$$

where γ is a known function of $\mathbb{H}$, which is generally assumed to be a regular function (often in the sense that $\hat{\gamma}_k$, its Fourier coefficients behave like $k^{-\nu}$). A very usual example is also the box-car function.

The following figures show first four original signals to recover, which are well known test-signals of the statistical litterature. They provide typical features which are difficult to restore : bumps, blocks and Doppler effects. The second and third series of pictures show their deformation after blurring (i.e. convolution with a regular function) and addition of a noise. These figures show how the convolution regularizes the signal, making it very difficult to recover especially the high frequency features. Their statistical investigation can be found in [22].

A variant of this problem consists in observing $Y_1, \dots, Y_n$, n independent and identically distributed random variables where each Y_i may be written as the following sum $Y_i = X_i + U_i$, where X_i and U_i again are independent, the distribution of U_i is know and of density γ and we want to recover the common density of the X_i's. The direct problem is the case where $U_i = 0$, for all i, and is corresponding to a standard density estimation problem (see section 5.1) . Hence the variables U_i's are acting as perturbations of the X_i's, whose density is to be recovered.





1.1.2. Wicksell's problem. Another typical example is the following classical Wicksell's problem [42]. Suppose a population of spheres is embedded in a medium. The spheres have radii that may be assumed to be drawn independently from a density f. A random plane slice is taken through the medium and those spheres that are intersected by the plane furnish circles which radii are the points of observation $Y_1, \dots, Y_n$. The unfolding problem is then to infer the density of the sphere radii from the observed circle radii. This unfolding problem also arises in medecine, where the spheres might be tumors in an animal's liver [36], as well as in numerous contexts (biological, engineering,...) see for instance [9].

Following [42] and [23], the Wicksell's problem corresponds to the following operator:

$$\mathbb{H} = \mathbb{L}_2([0,1], d\mu)\ d\mu(x) = (4x)^{-1}dx, \quad \mathbb{K} = \mathbb{L}_2([0,1], d\lambda)\ d\lambda(x) = 4\pi^{-1}(1-y^2)^{1/2}dy$$

$$Kf(y) = \frac{\pi}{4} y(1-y^2)^{-1/2} \int_y^1 (x^2-y^2)^{-1/2} f(x) d\mu.$$

Notice however, that, in this presentation, again in order to avoid additional technicalities, we detail this problem in the white noise framework, which is simpler than the original problem expressed above in density terms.

1.2. Singular Value Decomposition and projection methods.

Let us begin with a quick description of well known methods in inverse problems with random noise.

Under the assumption that K is compact, there exist 2 orthonormal bases (SVD bases) (e_k) of $\mathbb{H}$ and (g_k) of $\mathbb{K}$, respectively and a sequence (b_k), tending to 0, when k goes to infinity, such that

$$Ke_k = b_k g_k, \ K^* g_k = b_k e_k$$

if K^* is the adjoint operator.

For a sake of simplicity we suppose in the sequel that K and K^* are into. Otherwise we have to take care of the kernels of these operators. The bases (e_k) and (g_k) are called singular value bases, whereas the b_k's are simply called singular values.

Deconvolution In this case simple calculations prove that the SVD bases e_k and g_k both coincide with the Fourier basis. The singular values are corresponding to the Fourier coefficients of the function γ:

$$b_k = \hat{\gamma}_k. \tag{4}$$

Wicksell In this case, following [23], we have the following SVD :

$$e_k(x) = 4(k+1)^{1/2} x^2 P_k^{0,1}(2x^2 - 1)$$
$$g_k(y) = U_{2k+1}(y)$$

$P_k^{0,1}$ is the Jacobi polynomial of type $(0,1)$ with degree k U_k is the second type Chebishev polynomial with degree k. The singular values are :

$$b_k = \frac{\pi}{16}(1+k)^{-1/2}. \tag{5}$$

1.2.1. SVD Method. The Singular Value Decomposition (SVD) of K :

$$Kf = \sum_k b_k \langle f, e_k \rangle g_k$$

gives rise to approximations of the type :

$$f_\varepsilon = \sum_{k=0}^{N} b_k^{-1} \langle g_\varepsilon, g_k \rangle e_k$$

where $N = N(\varepsilon)$ has to be chosen properly. This SVD method is very attractive theoretically and can be shown to be asymptotically optimal in many situations (see Mathé and Pereversev [31], Cavalier and Tsybakov [7], Mair and Ruymgaardt [29]). It also has the big advantage of performing a quick and stable inversion of the operator. However, it suffers from different types of limitations: The SVD bases might be difficult to determine and manipulate numerically. Secondly while these bases are fully adapted to describe the operator K, they might not be appropriate for the accurate description of the solution with a small number of parameters.

Also in many practical situations, the signal provides inhomogeneous regularity, and its local features are especially interesting to recover. In such cases, other bases (in particular localised bases such as wavelet bases) may be much more appropriate to give a good representation of the object at hand.

1.2.2. Projection methods. Projection methods, which are defined as solutions of (1) restricted on finite dimensional subspaces $\mathbb{H}_N$ and $\mathbb{K}_N$ (of dimension N) also give rise to attractive approximations of f, by properly choosing the subspaces and the tuning parameter N (Dicken and Maass [10], Mathe and Pereversseh [31] together with their non linear counterparts Cavalier and Tsybakov [7], Cavalier et al [6], Tsybakov [40], Goldenschluger and Pereversev [19], Efromovich and Kolchinskii [16]. In the case where $\mathbb{H} = \mathbb{K}$, and K is a self adjoint operator, the system is particularly simple to solve since the restricted operator K_N is symmetric positive definite. This is the so-called Galerkin method. Obviously, restricting to finite subspaces has similar effects and can also be seen as a *Tykonov regularisation*, i.e. minimizing the least square functional penalised by a regularisation term.

The advantage of the Galerkin method is to allow the choice of the basis. However the Galerkin method suffers from the drawback of being unstable in many cases.

Comparing the SVD and Galerkin methods exactly states one main difficulty of the problem. The possible antagonism between the SVD basis where the inversion of the system is easy, and a 'localised' basis where the signal is sparsely represented, will be the issue we are trying to address here.

1.3. Cut-off, linear methods, thresholding. *The reader may profitably refer to subsections 5.3 and 5.4, where the linear methods and thresholding techniques are presented in details in the direct case.*

SVD as well as Galerkin methods are very sensitive to the choice of the tuning parameter $N(\varepsilon)$. This problems can be solved theoretically. However the solution depends heavily on the prior assumptions of regulatity on the solution, which have to be known in advance.

In the last ten years, many nonlinear methods have been developped especially in the direct case with the objective of automatically adapting to the unknown smoothness and local singular behavior of the solution. In the direct case, one of the most attractive methods is probably wavelet thresholding, since it allies numerical simplicity to asymptotic optimality on a large variety of functional classes such as Besov or Sobolev classes.

To adapt this approach in inverse problems, Donoho [14] introduced a wavelet-like decomposition, specifically adapted to the operator K (Wavelet-Vaguelette-Decomposition) and provided a thresholding algorithm on this decomposition. In Abramovitch and Silverman [1], this method was compared with the similar vaguelette-wavelet decomposition. Other wavelet approaches, might be mentioned such as Antoniadis and Bigot [2], Antoniadis & al [3] and especially for the deconvolution problem, Penski & Vidakovic [37], Fan & Koo [17], Kalifa & Mallat [24], Neelamani & al [34].

Later, Cohen et al [8] introduced an algorithm combining a Galerkin inversion to a thresholding algorithm.

The approach developped in the sequel is greatly influenced by these previous works. The accent we put here is over constructing (when necessary) new generation wavelet-type bases well adapted to the operator K, instead of sticking to the standard wavelet bases and reducing the range of potential operators covered by the method.

2. Wave-VD-type estimation :

We explain here the basic idea of the method, which is very simple. Let us expand f using a well suited basis (to be defined later : ' the wavelet-type' basis') :

$$f = \sum (f, \psi_\lambda)_{\mathbb{H}} \psi_\lambda$$

Using Parseval identity, we have $\beta_\lambda = (f, \psi_\lambda)_{\mathbb{H}} = \sum f_i \psi_\lambda^i$, for $f_i = (f, e_i)_{\mathbb{H}}$, $\psi_\lambda^i = (\psi_\lambda, e_i)_{\mathbb{H}}$, Let us put $Y_i = (Y_\varepsilon, g_i)_{\mathbb{K}}$, we have,

$$Y_i = (Kf, g_i)_{\mathbb{K}} + \varepsilon\xi_i = (f, K^* g_i)_{\mathbb{K}} + \varepsilon\xi_i = (\sum_j f_j e_j, K^* g_i)_{\mathbb{H}} + \varepsilon\xi_i = b_i f_i + \varepsilon\xi_i$$

where the ξ_i's are forming a sequence of iid centered gaussian variables with variance 1.

$$\hat{\beta}_\lambda = \sum_i \frac{Y_i}{b_i} \psi_\lambda^i$$

is such that $\mathbb{E}(\hat{\beta}_\lambda) = \beta_\lambda$ (i.e. its average value is β_λ). It is a plausible estimate of β_λ. Let us now put ourselves in a multiresolution setting, taking $\lambda = (j, k)$, for $j \geq 0$, k belonging to a set χ_j and consider :

$$\hat{f} = \sum_{j=-1}^{J} \sum_{k \in \chi_j} t(\hat{\beta_{jk}}) \psi_{jk}$$

where t is a thresholding operator. (*The reader which is unfamiliar with thresholding techniques may profitably refer to section 5.4.*)

$$t(\hat{\beta_{jk}}) = \hat{\beta_{jk}} I\{|\hat{\beta_{jk}}| \geq \kappa t_\varepsilon \sigma_j\} \quad t_\varepsilon = \varepsilon\sqrt{\log 1/\varepsilon}. \tag{6}$$

κ here is a tuning parameter of the method which will be properly chosen later. A main difference, here, with the direct case is the fact that the thresholding is depending on the resolution level through the constant σ_j which also will be precised later. However, our main discussion will be on how to choose the basis (ψ_{jk}). We will see that in order to perform in an optimal way, the method needs the basis (ψ_{jk}) to be 'coherent' with the SVD basis.

We will particularly focus on two situations (corresponding to the two examples discussed in the introduction). In the first type of cases, the operator has as SVD bases the Fourier basis. In this case, this 'coherence' is easily obtained with 'standard' wavelets (still, not any kind of standard wavelets as will be seen). However, more difficult problems (and typically the Wicksell's problem), require, when we need to mix these coherence conditions with the desired property of localisation of the basis, the construction of new objects: second generation-type wavelets.

2.1. WAVE-VD in a wavelet scenario. It is well known (see for instance Meyer [32]) that wavelet bases provide characterisations of smoothness spaces such as Hölder spaces $Lip(s)$, Sobolev spaces W_p^s as well as Besov spaces B_{pq}^s for a range of indices s depending on the wavelet ψ. For the scale of Besov spaces which includes as particular cases $Lip(s) = B_{\infty\infty}^s$ (if $s \notin \mathbb{N}$) and $W_p^s = B_{pp}^s$ (if $p = 2$), the characterisation has the following form :

$$\text{if } f = \sum_{j\geq 0}\sum_{k\in\mathbb{Z}} \beta_{jk}\psi_{jk}$$
$$\|f\|_{B_{pq}^s} \sim \|(2^{j[s+\frac{1}{2}-\frac{1}{p}]}\|\beta_{j\cdot}\|_{l_p})_{j\geq 0}\|_{l_q}. \tag{7}$$

In this section, we take $\{\psi_{jk},\ j \geq -1,\ k \in \xi_j\}$ to be a standard wavelet basis. More presisely, we suppose as usual that ψ_{-1} stands for the scaling function, for any $j \geq -1$, ξ_j is a set included in $\mathbb{N}$ which size is of order 2^j. Moreover, we assume that the following properties are true : There exist constants c_p, C_p, d_p such that:

$$c_p 2^{j(\frac{p}{2}-1)} \leq \|\psi_{jk}\|_p^p \leq C_p 2^{j(\frac{p}{2}-1)} \tag{8}$$
$$\|\sum_{k\in\xi_j} u_k\psi_{jk}\|_p^p \leq D_p \sum_{k\in\xi_j} |u_k^p\|\psi_{jk}\|_p^p, \quad \text{for any sequence } u_k \tag{9}$$

As in section 5, we consider the loss of a decision $\hat{f}$ if the truth is f, as the $\mathbb{L}_p$ norm : $\|\hat{f} - f\|_p$ and its associated risk :

$$\mathbb{E}\|\hat{f} - f\|_p^p.$$

The following theorem is going to evaluate this risk, when the strategy is the one introduced in the previous section, and when the true function belongs to a Besov ball ($f \in B_{\pi,r}^s(M) \iff \|f\|_{B_{pq}^s} \leq M$). One nice property of this estimation procedure is that it does not need the a priori knowledge of this regularity to get a good rate of convergence.

Theorem 2.1. *If we assume that,* $1 < p < \infty$, $2\nu + 1 > 0$, *and*

$$\sigma_j^2 := \sum_i [\frac{\psi_{jk}^i}{b_i}]^2 \leq C 2^{2j\nu},\ \forall\, j \geq 0, \tag{10}$$

If we put $\kappa^2 \geq 16p$, $2^J = [t_\varepsilon]^{\frac{-2}{2\nu+1}}$, *then*
if f *belongs to* $B_{\pi,r}^s(M)$ *with* $\pi \geq 1$, $s \geq 1/\pi$, $r \geq 1$ *(with the restriction* $r \leq \pi$ *if* $s = (2\nu+1)(\frac{p}{2\pi} - \frac{1}{2})$*), we have*

$$\mathbb{E}\|\hat{f} - f\|_p^p \leq C \log(1/\varepsilon)^{p-1}[\varepsilon^2 \log(1/\varepsilon)]^{\alpha p}, \tag{11}$$

with:

$$\alpha = \frac{s}{1+2(\nu+s)}, \quad \text{if } s \geq (2\nu+1)(\frac{p}{2\pi} - \frac{1}{2})$$
$$\alpha = \frac{s-1/\pi+1/p}{1+2(\nu+s-1/\pi)}, \quad \text{if } \frac{1}{\pi} \leq s < (2\nu+1)(\frac{p}{2\pi} - \frac{1}{2}).$$

Remarks :

1. Condition (10) is essential here. As will be shown later, this condition is linking the wavelet system with the singular value decomposition of the kernel K. If we set ourself in the deconvolution case, the SVD basis is the *Fourier* basis in such a way that ψ^i_{jk} is simply the Fourier coefficients of ψ_{jk}. If we choose as wavelet basis the periodized Meyer wavelet basis (see Meyer [32] and Mallat [30]), conditions (8) and (9) are satisfied. In addition, as the Meyer wavelet has the remarkable property of being compactly supported in the Fourier domain, simple calculations prove that, for any $j \geq 0,\ k$, the number of i's such that $\psi^i_{jk} \neq 0$ is finite and of the order 2^j. Then if we assume to be in the so-called 'regular' case ($b_k \sim k^{-\nu}$, for all k), it is easy to establish that (10) is true. This condition is also true for more general cases in the deconvolution setting such as the box-car deconvolution, see [22], [28].

2. These results are minimax (see [43]) up to logarithmic factors. This means that if we consider the best estimator in its worse performance over a given Besov ball, this estimator attains a rate of convergence which is the one given in (11) up to logarithmic factors.

3. If we compare these results to the rates of convergence obtained in the direct model (see subsections 5.3 and 5.4), we see that the difference (up to logarithmic terms) essentially lies in the parameter ν which acts as a reducing factor of the rate of convergence. This parameter quantifies the extra difficulty offered by the inverse problem. It is often called coefficient of illposedness. If we recall that in the deconvolution case, the coefficients b_k's are the Fourier coefficients of the function γ, the illposedness coefficient then clearly appears to be closely related to the regularity of the blurring function.

   $\diamond$

This result has been proved in the deconvolution case in [22]. The proof of this theorem is given in Appendix I.

2.2. WAVE-VD in Jacobi scenario: NEED-VD. We have seen that the results given above are true under the condition (10) on the wavelet basis.

Let us first appreciate how the condition (10) links the 'wavelet-type' basis to the SVD basis (e_k). To see this let us put ourselves in the regular case :

$$b_i \sim i^{-\nu}.$$

(by this, we mean more precisely that there exist two positive constants c and c' such that $c' i^{-\nu} \leq b_i \leq c i^{-\nu}$.)

If (10) is true, we have :

$$C2^{2j\nu} \geq \sum_m \sum_{2^m \leq i < 2^{m+1}} [\frac{\psi^i_{jk}}{b_i}]^2$$

Hence, $\forall\ m \geq j$, :

$$\sum_{2^m \leq i < 2^{m+1}} [\psi^i_{jk}]^2 \leq c 2^{2\nu(j-m)}$$

This suggests the necessity to contruct a 'wavelet-type' basis which support, at the level j, with respect to the SVD basis (sum in i) is concentrated on the integers between 2^j and 2^{j+1} and exponentially decreasing after this band. This is exactly the case of Meyer's wavelet, when the SVD basis is the Fourier basis.

In the general case of an arbitrary linear operator giving rise to an arbitrary SVD basis (e_k), and if in addition to (10), we add a localisation condition on the basis, we do not know if such a construction can be performed. However, in some cases, even quite as far from the deconvolution as the Wicksell's problem, one can build a 'second generation wavelet-type' basis, with exactly these properties.

The following construction due to Petrushev and collaborators ([33], [39], [38]) exactly realizes the paradigm mentioned above, producing a frame (the needlet basis) in the case where the linear operator has Jacobi polynomials as SVD basis, (as well as in different other cases such as spherical harmonics, Hermite functions, Laguerre polynomials) which has the property of being localised.

3. Petrushev construction of Needlets

Frames were introduced in the 1950's by Duffin and Schaeffer [15] to represent functions via over-complete sets. Frames including tight frames arise naturally in wavelet analysis on $\mathbb{R}^d$. Tight frames which are very close to orthonormal bases are especially useful in signal and image processing.

We will see that the following construction has the advantage of being easily computable and producing well localised tight frames constructed on a specified orthonormal basis.

We recall the following definition.

Definition 3.1. Let $\mathbb{H}$ be a Hilbert space, and (e_n) a sequence in $\mathbb{H}$, (e_n) is a tight frame (with constant 1) if :

$$\forall f \in \mathbb{H}, \quad \|f\|^2 = \sum_n |\langle f, e_n \rangle|^2$$

Let now $\mathcal{Y}$ be a metric space, μ a finite measure. Let us suppose that we have the following decomposition

$$\mathbb{L}_2(\mathcal{Y}, \mu) = \bigoplus_{k=0}^{\infty} H_k$$

where the H_k's are finite dimensional spaces. For sake of simplicity, we suppose that H_0 is reduced to the constants.

Let L_k be the orthogonal projection on H_k :

$$\forall f \in \mathbb{L}_2(\mathcal{Y}, \mu), \quad L_k(f)(x) = \int_{\mathcal{Y}} f(y) L_k(x, y) d\mu(y)$$

where

$$L_k(x, y) = \sum_{i=1}^{l_k} e_i^k(x) \bar{e}_i^k(y)$$

l_k is the dimension of H_k and $(e_i^k)_{i=1,\ldots l_k}$ an orthonormal basis of H_k. Let us observe that we have the following property of the projection operators:

$$\int L_k(x, y) L_m(y, z) d\mu(z) = \delta_{k,m} L_k(x, z) \tag{12}$$

The construction, also inspired by the paper of Frazier, Jawerth and Weiss [18], is based on two fundamental steps : Littlewood-Paley decomposition and discretization, which are summarized in the two following subsections.

3.1. Littlewood -Paley decomposition. Let φ be a C^∞ function supported in $|\xi| \leq 1$, such that $1 \geq \varphi(\xi) \geq 0$ and $\varphi(\xi) = 1$ if $|\xi| \leq \frac{1}{2}$. Let us define:

$$a^2(\xi) = \varphi(\xi/2) - \varphi(\xi) \geq 0$$

so that

$$\forall |\xi| \geq 1, \quad \sum_j a^2(\xi/2^j) = 1 \tag{13}$$

Let us now define the operator

$$\Lambda_j = \sum_{k \geq 0} a^2(k/2^j) L_k$$

and the associated kernel

$$\Lambda_j(x,y) = \sum_{k \geq 0} a^2(k/2^j) L_k(x,y) = \sum_{2^{j-1} < k < 2^{j+1}} a^2(k/2^j) L_k(x,y.)$$

The following proposition is true :

Proposition 3.2.

$$\forall f \in \mathbb{H}, \quad f = \lim_{J \to \infty} L_0(f) + \sum_{j=0}^{J} \Lambda_j(f) \tag{14}$$

$$\textit{For } M_j(x,y) = \sum_k a(k/2^j) L_k(x,y), \quad \Lambda_j(x,y) = \int M_j(x,z) M_j(z,y) d\mu(z) \tag{15}$$

Proof.

$$L_0(f) + \sum_{j=0}^{J} \Lambda_j(f) = L_0 + \sum_{j=0}^{J} (\sum_k a^2(k/2^j) L_k) = \sum_k \varphi(k/2^{J+1}) L_k \tag{16}$$

So:

$$\| \sum_k \varphi(k/2^{J+1}) L_k(f) - f \|^2 = \sum_{l \geq 2^{J+1}} \|L_l(f)\|^2 + \sum_{2^J \leq l < 2^{J+1}} \|L_l(f)(1 - \varphi(l/2^{J+1})\|^2$$
$$\leq \sum_{l \geq 2^J} \|L_l(f)\|^2 \longrightarrow 0, \quad \text{when} \quad J \to \infty$$

(15) is a simple consequence of (12).

□

3.2. Discretization. Let us define

$$\mathcal{K}_k = \bigoplus_{m=0}^{k} H_m,$$

and let us assume that some additional assumptions are true:

1. $$f \in \mathcal{K}_k \Longrightarrow \bar{f} \in \mathcal{K}_k$$

2. $$f \in \mathcal{K}_k, \quad g \in \mathcal{K}_l \Longrightarrow fg \in \mathcal{K}_{k+l}$$

3. Quadrature formula: for all $k \in \mathbb{N}$, there exists χ_k a finite subset of $\mathcal{Y}$ and positive real numbers $\lambda_\xi > 0$ indiced by the elements ξ of χ_k, such that

$$\forall f \in \mathcal{K}_k, \quad \int f d\mu = \sum_{\xi \in \chi_k} \lambda_\xi f(\xi)$$

Then the operator M_j defined in the subsection above is such that: $M_j(x,z) = \overline{M_j(z,x)}$ and

$$z \mapsto M_j(x,z) \in \mathcal{K}_{2^{j+1}-1}.$$

So that

$$z \mapsto M_j(x,z)M_j(z,y) \in \mathcal{K}_{2^{j+2}-2},$$

and we can write:

$$\Lambda_j(x,y) = \int M_j(x,z)M_j(z,y)d\mu(z) = \sum_{\xi \in \chi_{2^{j+2}-2}} \lambda_\xi M_j(x,\xi)M_j(\xi,y)$$

This implies:

$$\Lambda_j f(x) = \int \Lambda_j(x,y)f(y)d\mu(y) = \int \sum_{\xi \in \chi_{2^{j+2}-2}} \lambda_\xi M_j(x,\xi)M_j(\xi,y)f(y)d\mu(y)$$

$$= \sum_{\xi \in \chi_{2^{j+2}-2}} \sqrt{\lambda_\xi} M_j(x,\xi) \int \overline{\sqrt{\lambda_\xi} M_j(y,\xi)} f(y)d\mu(y)$$

This can be summarized in the following way, if we put $\chi_{2^{j+2}-2} = \mathbb{Z}_j$, $\sqrt{\lambda_\xi} M_j(x,\xi) = \psi_{j,\xi}$,

$$\Lambda_j f(x) = \sum_{\xi \in \mathbb{Z}_j} \langle f, \psi_{j,\xi} \rangle \psi_{j,\xi}(x).$$

Proposition 3.3. *The family $(\psi_{j,\xi})_{j \in \mathbb{N}, \xi \in \mathbb{Z}_j}$ is a tight frame*

Proof. As

$$f = \lim_{J \longrightarrow \infty} (L_0(f) + \sum_{j \leq J} \Lambda_j(f))$$

$$\|f\|^2 = \lim_{J \longrightarrow \infty} (\langle L_0(f), f \rangle + \sum_{j \leq J} \langle \Lambda_j(f), f \rangle)$$

but

$$\langle \Lambda_j(f), f \rangle = \sum_{\xi \in \mathbb{Z}_j} \langle f, \psi_{j,\xi} \rangle \langle \psi_{j,\xi}, f \rangle = \sum_{\xi \in \mathbb{Z}_j} |\langle f, \psi_{j,\xi} \rangle|^2$$

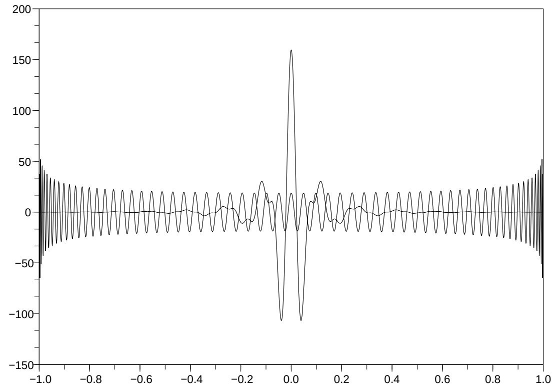
and if ψ_0 is a normalized constant $\langle L_0(f), f \rangle = |\langle f, \psi_0 \rangle|^2$ so that

$$\|f\|^2 = |\langle f, \psi_0 \rangle|^2 + \sum_{j \in \mathbb{N},\ \xi \in \mathbb{Z}_j} |\langle f, \psi_{j,\xi} \rangle|^2$$

but this is exactly the characterization of a tight frame we gave. □

3.3. Localisation properties. This construction has been performed in different frameworks by Petrushev and coauthors giving in each situation very nice localisation properties.

The following figure (thanks to Paolo Baldi) is an illustration of this phenomenum: it shows a needlet constructed as explained above using Legendre polynomials of degree 2^8. The highly oscillationg function is a Legendre polynomials of degree 2^8, whereas the localised one is a needlet centered approximately in the middel of the interval . Its localisation properties are remarquable considering the fact that both functions are polynomials of the same order.



In the case of the sphere of $\mathbb{R}^{d+1}$, where the spaces H_k are spanned by spherical harmonics, it is proved in Narcowich, Petrushev and Ward, [33] the following localisation property: For any k there exists a constant C_k such that :

$$|\psi_{j\eta}(\xi)| \leq \frac{C_k 2^{dj/2}}{[1+2^j \arccos <\eta,\xi>]^k}.$$

A similar result exists in the case of Laguerre polynomials on $\mathbb{R}_+$ [25].

In the case of Jacobi polynomials on the interval with Jacobi weight, it is proved in Petrushev and Xu [38] the following localisation property: For any k there exist constant $C,\ c$ such that :

$$|\psi_{j\eta}(\cos\theta)| \leq \frac{C 2^{j/2}}{(1+(2^j|\theta-\arccos\eta|)^k \sqrt{w_{\alpha\beta}(2^j,\cos\theta)}}.$$

where $w_{\alpha\beta}(n,x) = (1-x+n^{-2})^{\alpha+1/2}(1+x+n^{-2})^{\beta+1/2}, \quad -1 \leq x \leq 1$ if $\alpha > -1/2,\ \beta > -1/2$.

4. NEED-VD in the Jacobi case

Let us now come back to the estimation algorithm.

We consider now an operator K such that the space :

$$\mathbb{H} = \mathbb{L}_2(I, d\gamma(x))$$

$$I=[-1,1], d\gamma(x) = \omega_{\alpha,\beta}(x)dx,\ \omega_{\alpha,\beta}(x) = (1-x)^\alpha(1+x)^\beta;\ \alpha > -1/2,\ \beta > -1/2$$

For a sake of simplicity, let us suppose $\alpha \geq \beta$. (Otherwise we can exchange the parameters.)

Let P_k be the normalized Jacobi polynomials for this weight. We suppose that these polynomials appears as SVD basis of the operator K, as it is the case for the Wicksell problem, with $\beta = 0,\ \alpha = 1,\ b_k \sim k^{-1/2}$.

4.1. Needlets and condition (10). Let us define the 'needlets' as constructed above:

$$\psi_{j,\eta_k} = \sum_l \hat{a}(l/2^{j-1}) P_l(x) P_l(\eta_k) \sqrt{b_{j,\eta_k}} \tag{17}$$

The following proposition proves that such a construction always implies the condition (10) in the regular case.

Proposition 4.1. *As soon as ψ_{j,η_k} is a frame, if $b_i \sim i^{-\nu}$, then*

$$\sigma_j^2 := \sum_i [\frac{\psi^i_{jk}}{b_i}]^2 \leq C 2^{2j\nu}$$

Proof. : As soon as the family ψ_{j,η_k} is a frame (not necessarily tight), as the elements of a frame are bounded, and the set $\{i, \psi^i_{jk} \neq 0\}$ is included into the set $\{C_1 2^j, \dots, C_2 2^j\}$, we have,

$$\sum_i [\frac{\psi^i_{jk}}{b_i}]^2 \leq C 2^{j\nu} \|\psi_{j,\eta_k}\|^2 \leq C' 2^{j\nu}$$

□

4.2. Convergence results in the Jacobi case. The following theorem is the analogous of theorem 2.1, in this case. As can be seen, the results there are at the same time more difficult to obtain (the following theorem do not cover the same range as the previous one), and richer since they furnish new rates of convergence.

Theorem 4.2. *Let us suppose that we are in the Jacobi case as stated above ($\alpha \geq \beta > -\frac{1}{2}$): We put*

$$t_\varepsilon = \varepsilon \sqrt{\log 1/\varepsilon}$$
$$2^J = t_\varepsilon^{-\frac{2}{1+2\nu}},$$

choose $\kappa \geq 16p[1 + (\frac{\alpha}{2} - \frac{\alpha+1}{p})_+]$, and suppose that we are in the regular case i.e.

$$b_i \sim i^{-\nu}, \quad \nu > -\frac{1}{2}.$$

Then if $f = \sum_j \sum_k \beta_{j,\eta_k} \psi_{j,\eta_k}$ is such that,

$$(\sum |\beta_{j,\eta_k}|^p \|\psi_{j,\eta_k}\|_p^p)^{1/p} \leq \rho_j 2^{-js}, \ \rho_\cdot \in l_r$$

then

$$\mathbb{E}\|\hat{f} - f\|_p^p \leq C[\log(1/\varepsilon)]^{p-1}[\varepsilon\sqrt{\log(1/\varepsilon)}]^{\mu p},$$

with :

1. *If $p < 2 + \frac{1}{\alpha+1/2}$*

$$\mu = \frac{s}{s + \nu + \frac{1}{2}}$$

2. *If $p > 2 + \frac{1}{\alpha+1/2}$*

$$\mu = \frac{s}{s + \nu + \alpha + 1 - \frac{2(1+\alpha)}{p}}$$

This theorem is proved in Kerkyacharian et al. [27]. Simulation results on these methods are given there, showing that their performances are far above the usual SVD methods in several cases. It is interesting to notice that the rate of convergence which are obtained here agree with the minimax rates evaluated in Johnstone and Silverman [23] where the case $p = 2$ is considered. But the second case ($p > 2 + \frac{1}{\alpha+1/2}$) show a rate of convergence which is new in the litterature. In [27], where the whole range of Besov bodies is considered, more atypical rates are given.

5. Direct Models ($K = I$): a memento

5.1. The density model. The famest nonparametric model consists in observing n i.i.d. random variables having a common density f on the interval $[0, 1]$, and trying to give an estimation of f.

A standard route to perform this estimation consists in expanding the density f in an orthonormal basis $\{e_k,\ k \in \mathbb{N}\}$ of an Hilbert space $\mathbb{H}$ -assuming implicitely that f belongs to $\mathbb{H}$.

$$f = \sum_{l \in \mathbb{N}} \theta_l e_l$$

If $\mathbb{H}$ happens to be the space $\mathbb{L}_2 = \{g :\ [0, 1] \mapsto \mathbb{R},\ \|g\|_2^2 := \int_0^1 g^2 < \infty\}$, we observe that

$$\theta_l = \int_0^1 e_l(x) f(x) dx = \mathbb{E} e_l(X_i).$$

Replacing the expectation by the empirical one leads to a standard estimate for θ_l:

$$\hat{\theta}_l = \frac{1}{n} \sum_{i=1}^{n} e_l(X_i).$$

At this step, the simplest choice of estimate for f is obviously :

$$\hat{f}_m = \sum_{i=1}^{m} \hat{\theta}_l e_l \tag{18}$$

5.2. From the density to the white noise model. Before analysing the properties of the estimator defined above, let us observe that the previous approach (representing f by its coefficients $\{\theta_k,\ k \geq 0\}$, leads to summarize the information in the following sequence model :

$$\{\hat{\theta}_k,\ k \geq 0\}. \tag{19}$$

We can write $\hat{\theta}_k =: \theta_k + u_k$, with

$$u_k = \frac{1}{n} \sum_{i=1}^{n} [e_k(X_i) - \theta_k],$$

Central limit theorem is a relatively convincing argument that the model (19), may be approximated by the following one

$$\{\hat{\theta}_k = \theta_k + \frac{\eta_k}{\sqrt{n}},\ k \geq 0\}. \tag{20}$$

where the $\eta_k' s$ are forming a sequence of i.i.d. gaussian, centered variables with fixed variance σ^2 say. Such an approximation requires more delicate calculations than these quick arguments and is rigourously proved in Nussbaum [35], see also Brown and Low [5].

This model is the sequence space model associated to the following global observation, so called white noise model (with $\varepsilon = n^{-1/2}$):

$$dY_t = f(t)dt + \varepsilon dW_t, \quad t \in [0,1]$$

where for any $\varphi \in \mathbb{L}^2([0,1], dt)$, $\int_{[0,1]} \varphi dY_t = \int_{[0,1]} f\varphi dt + \varepsilon \int_{[0,1]} \varphi dW_t$ is observable.

(20) formally consists in considering, all the observables obtained for $\varphi = e_k$, for all k in $\mathbb{N}$. Among nonparametric situations, the white noise model considered above is one of the simplest, at least technically. Mostly for this reason, this model has been given a central place in statistics, particularly by the Russian school, following Ibraguimov and Has'minskii (see for instance their book [20]). However it arises as an appropriate large sample limit to more general nonparametric models, such as regression with random design, or non independent spectrum estimation, diffusion models -see for instance [21], [4]...-

5.3. The linear estimation: how to choose the tuning parameter m ?

. In (18), the choice of m is crucial.

To better understand the situation, let us have a look to the risk of the strategy $\hat{f}_m$. If we consider that, when deciding $\hat{f}_m$, when f is the truth, we have a loss of order $\|\hat{f}_m - f\|_2^2$, then, our risk will be the following mathematical expectation :

$$\mathbb{E}\|\hat{f}_m - f\|_2^2.$$

Of course this way of measuring our risk is arguable since there is no particular reason for the $\mathbb{L}_2$ norm to reflect well the features we want to recover in the signal. For instance, an $\mathbb{L}_\infty$-norm could be prefered because it is easier to visualize. In general, several $\mathbb{L}_p$ norms are considered (as it is the case in sections 2.1 and 4.2). Here we restrict to the $\mathbb{L}_2$ case for a sake of simplicity.

To avoid technical difficulties, we set ourselves in the case of a white noise model, considering that we observe the sequence defined in (20). Hence,

$$\mathbb{E}(\hat{\theta}_l - \theta_l)^2 = \frac{1}{n}\int_0^1 e_l(x)^2 dx = \frac{1}{n} := \varepsilon^2.$$

We are now able to obtain :

$$\begin{aligned}\mathbb{E}\|\hat{f}_m - f\|_2^2 &= \sum_{l\le m} (\hat{\theta}_l - \theta_l)^2 + \sum_{l>m} \theta_l^2 \\ &\le m\varepsilon^2 + \sum_{l>m} \theta_l^2\end{aligned}$$

If we now assume that :

$$\forall k \in \mathbb{N}_*, \ \sum_{l>k} \theta_l^2 \le M k^{-2s} \tag{21}$$

for some $s > 0$ which is here an index of regularity directly connected to the size of the compact in l_2 in which the function f is supposed to lie, we get

$$\mathbb{E}\|\hat{f}_m - f\|_2^2 \le m\varepsilon^2 + Mm^{-2s}$$

We observe that the RHS is the sum of two factors: one (called the stochatic term) is increasing in m and reflects the fact that because of the noise, the more coefficients we have to estimate, the larger the global error will be. The second one (called the bias term or approximation term) does not depend on the noise and is decreasing in m. The RHS is optimised by choosing $m = m_*(s) =: c(s, M)\varepsilon^{\frac{-2}{1+2s}}$. Then

$$\mathbb{E}\|\hat{f}_{m_*(s)} - f\|_2 \le c'(s, M)\varepsilon^{\frac{-4s}{1+2s}}.$$

Let us observe that the more f is supposed to be regular (in the sense the larger s is), the less coefficients we need to estimate : a very irregular function (s close to 0) requires almost as much as $\varepsilon^{-2} = n$ coefficients, which corresponds to estimate as many coefficients as the number of available observations -in the density model for instance-. The rate obtained in (5.3) can be proved to be optimal in the following sense (minimax) : if we consider the best estimator in its worse performance over the class of functions verifying (21), this estimator attains a rate of convergence which is (up to a constant) the one given in (5.3). See Tsybakov [41] for a detailed review of the minimax point of view.

5.4. The thresholding estimation. Let us now suppose that the constant s, which plays an essential role in the construction of the previous estimator is not known. This is realistic, since it is extremely rare to know in advance that the function we are seeking has a specified regularity. Also, the previous approach takes very seriously into account the order in which the basis is taken. Let us now present a very elegant way of addressing at the same time both of these issues. The thresholding techniques which have been known for long by engineers in electronic and telecommunications, was introduced in statistics in Donoho and Johnstone [11] and later in a series of papers on wavelet thresholoding [12], [13]. It allies numerical simplicity to asymptotic optimality.

It starts from a different kind of observation : let us introduce the following estimate :

$$\widetilde{f} = \sum_{k=0}^{B} \hat{\theta}_k \mathbb{I}\{|\hat{\theta}_k| \geq \kappa t_\varepsilon\} e_k \tag{22}$$

Here the point of view is the following. We choose B very large (i.e. almost corresponding to $s = 0$)

$$B = \varepsilon^{-2} \log 1/\varepsilon.$$

But instead of keeping all the coefficients which numbers are between 0 and B, we decide to kill those which are not above the threshold t_ε. The intuitive justification of this choice is the following. Assuming that f has some kind of regularity conditions like (21) (unknown, but real...), essentially means that the coefficients θ_k of f are small except a small number of them. Hence, in the reconstruction of f, only these ones will be significant. t_ε is chosen, in such a way that the noise $\hat{\theta}_k - \theta_k$ due to the randomness of the observation might be neglected:

$$t_\varepsilon = \varepsilon[\log 1/\varepsilon]^{-1/2}.$$

Now, let us assume another type of condition on f, namely : there exists a positive constant $0 < q < 2$, such that,

$$\forall\, k \in \mathbb{N}_*,\ \sup_{\lambda > 0} \lambda^q \#\{k,\ |\theta_k| \geq \lambda\} \leq M \tag{23}$$

$$\begin{aligned}
\mathbb{E}\|\widetilde{f} - f\|_2^2 &= \sum_{l \leq B} (\hat{\theta}_l \mathbb{I}\{|\hat{\theta}_l| \geq \kappa t_\varepsilon\} - \theta_l)^2 + \sum_{l > B} \theta_l^2 \\
&\leq \sum_{l} (\hat{\theta}_l - \theta_l)^2 \mathbb{I}\{|\theta_l| \geq \kappa t_\varepsilon/2\} + \sum_{l} \theta_l^2 \mathbb{I}\{|\theta_l| \leq 2\kappa t_\varepsilon\} \\
&\quad + \sum_{l \leq B} [(\hat{\theta}_l - \theta_l)^2 + \theta_l^2] \mathbb{I}\{|\hat{\theta}_l - \theta_l| \geq \kappa t_\varepsilon/2\} + \sum_{l > B} \theta_l^2
\end{aligned}$$

Now, using the following probabilist bounds,

$$\mathbb{E}(\hat{\theta}_l - \theta_l)^2 = \varepsilon^2,\ \mathbb{P}(|\hat{\theta}_l - \theta_l| \geq \lambda) \leq 2 \exp -\lambda^2 \varepsilon^2,\ \forall \lambda > 0$$

and the fact that condition (23) implies :

$$\sum_{l} \theta_l^2 \mathbb{I}\{|\theta_l| \leq 2\kappa t_\varepsilon\} \leq C t_\varepsilon^{2-q}$$

we get,

$$\mathbb{E}\|\widetilde{f}-f\|_2^2 \le M\varepsilon^2 t_\varepsilon^{-q} + C' t_\varepsilon^{2-q} + \varepsilon^{\kappa^2/8} B + \sum_{l>B} \theta_l^2$$

It remains now to choose $\kappa^2 \ge 32$, to get,

$$\mathbb{E}\|\widetilde{f}-f\|_2^2 \le C' t_\varepsilon^{2-q} + \sum_{l>B} \theta_l^2$$

and if we assume in addition to (23),

$$\forall k \in \mathbb{N}_*,\ \sum_{l>k} \theta_l^2 \le M k^{-\frac{2-q}{2}} \tag{24}$$

we get,

$$\mathbb{E}\|\widetilde{f}-f\|_2^2 \le C" t_\varepsilon^{2-q}$$

Note that the interesting point in this construction is that the regularity conditions which are assumed on the function f are *not known* from the statistician, since they do not enter into the construction of the procedure. This property is called adaptation.

Now, to compare with the previous section, let us take $q = \frac{2}{1+2s}$, it is not difficult to prove that as soon as f verifies (21), it automatically verifies (23) and (24). Hence $\widetilde{f}$ and $\hat{f}_{m*(s)}$ have the same rate of convergence up to a logarithmic term. If we neglect this logarithmic loss, we substantially gain here the fact that we need not know the apriori regularity conditions on the aim function. It can also be proved that in fact conditions (23) and (24) are defining a set which is substantially larger than the set defined by condition (21): for instance its entropy is strictly larger (see [26]).

6. Appendix I : Proof of Theorem 2.1

In this proof, C will denote an absolute constant which may change from one line to the other.

We can always suppose $p \ge \pi$. Indeed, if $\pi \ge p$ it is very simple to see that $B^s_{\pi,r}(M)$ is included into $B^s_{p,r}(M)$: as $2^{j[s+\frac{1}{2}-\frac{1}{p}]}\|\beta_{j\cdot}\|_{l_p} \le 2^{j[s+\frac{1}{2}-\frac{1}{\pi}]}\|\beta_{j\cdot}\|_{l_\pi}$ (since the cardinal of ξ_j is of order 2^j).

First we have the following decomposition :

$$\begin{aligned}\mathbb{E}\|\hat{f}-f\|_p^p &\le 2^{p-1}\{\mathbb{E}\|\sum_{j=-1}^{J}\sum_{k\in\chi_j}(t(\hat{\beta_{jk}})-\beta_{jk})\psi_{jk}\|_p^p + \|\sum_{j>J}\sum_{k\in\chi_j}\beta_{jk}\psi_{jk}\|_p^p\}\\ &=: I + II\end{aligned}$$

The term II is easy to analyse, as follows: Since f belongs to $B^s_{\pi,r}(M)$, using standard embedding results (which in this case simply follows from direct comparisons between l_q norms) we have that f also belong to $B^{s-(\frac{1}{\pi}-\frac{1}{p})_+}_{p,r}(M')$, for some constant M'. Hence

$$\|\sum_{j>J}\sum_{k\in\chi_j}\beta_{jk}\psi_{jk}\|_p \le C 2^{-J[s-(\frac{1}{\pi}-\frac{1}{p})_+]}.$$

Then we only need to verify that $\frac{s-(\frac{1}{\pi}-\frac{1}{p})_+}{1+2\nu}$ is always larger that α, which not difficult.

Bounding the term I is more involved. Using the triangular inequality together with Hölder inequality, and property (9) for the second line, we get

$$I \leq 2^{p-1} J^{p-1} \sum_{j=-1}^{J} \mathbb{E} \| \sum_{k \in \chi_j} (t(\hat{\beta_{jk}}) - \beta_{jk}) \psi_{jk} \|_p^p$$
$$\leq 2^{p-1} J^{p-1} D_p \sum_{j=-1}^{J} \sum_{k \in \chi_j} \mathbb{E} | t(\hat{\beta_{jk}}) - \beta_{jk} |^p \| \psi_{jk} \|_p^p$$

Now, we separate four cases :

$$\begin{aligned}
\sum_{j=-1}^{J} \sum_{k \in \chi_j} \mathbb{E} | t(\hat{\beta_{jk}}) - \beta_{jk} |^p \| \psi_{jk} \|_p^p &= \sum_{j=-1}^{J} \sum_{k \in \chi_j} \mathbb{E} | t(\hat{\beta_{jk}}) - \beta_{jk} |^p \| \psi_{jk} \|_p^p \Big\{ I\{|\hat{\beta_{jk}}| \geq \kappa t_\varepsilon \sigma_j\} \\
&\quad + I\{|\hat{\beta_{jk}}| < \kappa t_\varepsilon \sigma_j\} \Big\} \\
&\leq \sum_{j=-1}^{J} \sum_{k \in \chi_j} \Big[\mathbb{E} | \hat{\beta_{jk}} - \beta_{jk} |^p \| \psi_{jk} \|_p^p I\{|\hat{\beta_{jk}}| \geq \kappa t_\varepsilon \sigma_j\} \\
&\quad \Big\{ I\{|\beta_{jk}| \geq \tfrac{\kappa}{2} t_\varepsilon \sigma_j\} + I\{|\beta_{jk}| < \tfrac{\kappa}{2} t_\varepsilon \sigma_j\} \Big\} \\
&\quad + |\beta_{jk}|^p \| \psi_{jk} \|_p^p I\{|\hat{\beta_{jk}}| \leq \kappa t_\varepsilon \sigma_j\} \Big\{ I\{|\beta_{jk}| \geq 2\kappa t_\varepsilon \sigma_j\} \\
&\quad + I\{|\beta_{jk}| < 2\kappa t_\varepsilon \sigma_j\} \Big\} \Big] \\
&\leq \; : Bb + Bs + Sb + Ss
\end{aligned}$$

If we notice that $\hat{\beta_{jk}} - \beta_{jk} = \sum_i \frac{Y_i - b_i f_i}{b_i} \psi_{jk}^i = \varepsilon \sum_i \xi_i \frac{\psi_{jk}^i}{b_i}$ is a gaussian random variable centered, and with variance $\varepsilon^2 \sum_i [\frac{\psi_{jk}^i}{b_i}]^2$, we have using standard properties of the gaussian distribution, for any $q \geq 1$, if we recall that we set $\sigma_j^2 =: \sum_i [\frac{\psi_{jk}^i}{b_i}]^2 \leq C 2^{2j\nu}$, and denote by s_q the qth absolute moment of the gaussian distribution when centered and with variance 1:

$$\mathbb{E} | \hat{\beta_{jk}} - \beta_{jk} |^q \leq s_q \sigma_j^q \varepsilon^q$$
$$\mathbb{P}\{ |\hat{\beta_{jk}} - \beta_{jk}| \geq \frac{\kappa}{2} t_\varepsilon \sigma_j \} \leq 2 \varepsilon^{\kappa^2/8}$$

Hence,

$$Bb \leq \sum_{j=-1}^{J} \sum_{k \in \chi_j} s_p \sigma_j^p \varepsilon^p \| \psi_{jk} \|_p^p I\{|\beta_{jk}| \geq \frac{\kappa}{2} t_\varepsilon \sigma_j\}$$
$$Ss \leq \sum_{j=-1}^{J} \sum_{k \in \chi_j} |\beta_{jk}|^p \| \psi_{jk} \|_p^p I\{|\beta_{jk}| < 2\kappa t_\varepsilon \sigma_j\}$$

And,

$$\begin{aligned}
Bs &\leq \sum_{j=-1}^{J} \sum_{k \in \chi_j} [\mathbb{E} | \hat{\beta_{jk}} - \beta_{jk} |^{2p}]^{1/2} [\mathbb{P}\{ |\hat{\beta_{jk}} - \beta_{jk}| \geq \frac{\kappa}{2} t_\varepsilon \sigma_j \}]^{1/2} \| \psi_{jk} \|_p^p I\{|\beta_{jk}| < \frac{\kappa}{2} t_\varepsilon \sigma_j\} \\
&\leq \sum_{j=-1}^{J} \sum_{k \in \chi_j} s_{2p}^{1/2} \sigma_j^p \varepsilon^p 2^{1/2} \varepsilon^{\kappa^2/16} \| \psi_{jk} \|_p^p I\{|\beta_{jk}| < \frac{\kappa}{2} t_\varepsilon \sigma_j\} \\
&\leq C \sum_{j=-1}^{J} 2^{jp(\nu + \frac{1}{2})} \varepsilon^p \varepsilon^{\kappa^2/16} \leq C \varepsilon^{\kappa^2/16}
\end{aligned}$$

Now, if we remark that the β_{jk} are necessarily all bounded by some constant (depending on M) since f belongs to $B^s_{\pi,r}(M)$, and using (8),

$$\begin{aligned}
Sb &\le \sum_{j=-1}^{J} \sum_{k\in\chi_j} |\beta_{jk}|^p \|\psi_{jk}\|_p^p \mathbb{P}\{|\hat{\beta_{jk}} - \beta_{jk}| \ge 2\kappa t_\varepsilon \sigma_j\} I\{|\beta_{jk}| \ge 2\kappa t_\varepsilon \sigma_j\} \\
&\le \sum_{j=-1}^{J} \sum_{k\in\chi_j} |\beta_{jk}|^p \|\psi_{jk}\|_p^p 2\varepsilon^{\kappa^2/8} I\{|\beta_{jk}| \ge 2\kappa t_\varepsilon \sigma_j\} \\
&\le C \sum_{j=-1}^{J} 2^{j\frac{p}{2}} \varepsilon^{\kappa^2/8} \le C\varepsilon^{\frac{\kappa^2}{8} - \frac{p}{2(2\nu+1)}}
\end{aligned}$$

It is easy to check that in any cases if $\kappa^2 \ge 16p$ the terms Bs and Sb are smaller than the rates announced in the theorem.

We have using (8) and condition (10) for any $z \ge 0$:

$$\begin{aligned}
Bb &\le C\varepsilon^p \sum_{j=-1}^{J} 2^{j(\nu p + \frac{p}{2} - 1)} \sum_{k\in\chi_j} I\{|\beta_{jk}| \ge \frac{\kappa}{2} t_\varepsilon \sigma_j\} \\
&\le C\varepsilon^p \sum_{j=-1}^{J} 2^{j(\nu p + \frac{p}{2} - 1)} \sum_{k\in\chi_j} |\beta_{jk}|^z [t_\varepsilon \sigma_j]^{-z} \\
&\le C{t_\varepsilon}^{p-z} \sum_{j=-1}^{J} 2^{j[\nu(p-z) + \frac{p}{2} - 1]} \sum_{k\in\chi_j} |\beta_{jk}|^z
\end{aligned}$$

Also, for any $p \ge z \ge 0$

$$\begin{aligned}
Ss &\le C \sum_{j=-1}^{J} 2^{j(\frac{p}{2} - 1)} \sum_{k\in\chi_j} |\beta_{jk}|^z \sigma_j^{p-z} [t_\varepsilon]^{p-z} \\
&\le C[t_\varepsilon]^{p-z} \sum_{j=-1}^{J} 2^{j(\nu(p-z) + \frac{p}{2} - 1)} \sum_{k\in\chi_j} |\beta_{jk}|^z
\end{aligned}$$

So in both cases we have the same bound to investigate. We will write this bound on the following form (forgetting the constant) :

$$I + II = {t_\varepsilon}^{p-z_1} [\sum_{j=-1}^{j_0} 2^{j[\nu(p-z_1) + \frac{p}{2} - 1]} \sum_{k\in\chi_j} |\beta_{jk}|^{z_1}] + {t_\varepsilon}^{p-z_2} [\sum_{j=j_0+1}^{J} 2^{j[\nu(p-z_2) + \frac{p}{2} - 1]} \sum_{k\in\chi_j} |\beta_{jk}|^{z_2}]$$

The constants z_i and j_0 will be chosen depending on the cases.

Let us first consider the case where $s \ge (\nu + \frac{1}{2})(\frac{p}{\pi} - 1)$, put

$$q = \frac{p(2\nu+1)}{2(s+\nu)+1}$$

and observe that on the considered domain, $q \le \pi$ and $p > q$. In the sequel it will be used that we have automatically $s = (\nu + \frac{1}{2})(\frac{p}{q} - 1)$. Now, taking $z_2 = \pi$, we get :

$$II \le {t_\varepsilon}^{p-\pi} [\sum_{j=j_0+1}^{J} 2^{j[\nu(p-\pi) + \frac{p}{2} - 1]} \sum_{k\in\chi_j} |\beta_{jk}|^{\pi}]$$

Now, as

$$\frac{p}{2q}-\frac{1}{\pi}+\nu(\frac{p}{q}-1)=s+\frac{1}{2}-\frac{1}{\pi}$$

and

$$\sum_{k\in\chi_j}|\beta_{jk}|^{\pi}=2^{-j(s+\frac{1}{2}-\frac{1}{\pi})}\tau_j$$

with $(\tau_j)_j \in l_r$ (this is a consequence of the fact that $f\in B^s_{\pi,r}(M)$ and (6)). Hence, we can write :

$$\begin{aligned}II &\le {t_\varepsilon}^{p-\pi}\sum_{j=j_0+1}2^{jp(1-\frac{\pi}{q})(\nu+\frac{1}{2})}\tau_j^{\pi}\\ &\le C{t_\varepsilon}^{p-\pi}2^{j_0p(1-\frac{\pi}{q})(\nu+\frac{1}{2})}\end{aligned}$$

The last inequality is true for any $r\ge 1$ if $\pi>q$ and for $r\le\pi$ if $\pi=q$. Notice that $\pi=q$ is equivalent to $s=(\nu+\frac{1}{2})(\frac{p}{\pi}-1)$. Now if we choose j_0 such that $2^{j_0\frac{p}{q}(\nu+\frac{1}{2})}\sim {t_\varepsilon}^{-1}$ we get the bound

$${t_\varepsilon}^{p-q}$$

which exactly gives the rate announced in the theorem for this case.

As for the first part of the sum (before j_0), we have, taking now $z_1=\widetilde{q}$, with $\widetilde{q}\le\pi$, so that $[\frac{1}{2^j}\sum_{k\in\chi_j}|\beta_{jk}|^{\widetilde{q}}]^{\frac{1}{\widetilde{q}}}\le[\frac{1}{2^j}\sum_{k\in\chi_j}|\beta_{jk}|^{\pi}]^{\frac{1}{\pi}}$, and using again (6),

$$\begin{aligned}I &\le {t_\varepsilon}^{p-\widetilde{q}}[\sum_{-1}^{j_0}2^{j[\nu(p-\widetilde{q})+\frac{p}{2}-1]}\sum_{k\in\chi_j}|\beta_{jk}|^{\widetilde{q}}]\\ &\le {t_\varepsilon}^{p-\widetilde{q}}[\sum_{-1}^{j_0}2^{j[\nu(p-\widetilde{q})+\frac{p}{2}-\frac{\widetilde{q}}{\pi}]}[\sum_{k\in\chi_j}|\beta_{jk}|^{\pi}]^{\frac{\widetilde{q}}{\pi}}]\\ &\le {t_\varepsilon}^{p-\widetilde{q}}\sum_{-1}^{j_0}2^{j[(\nu+\frac{1}{2})p(1-\frac{\widetilde{q}}{q})]}\tau_j^{\widetilde{q}}\\ &\le C{t_\varepsilon}^{p-\widetilde{q}}2^{j_0[(\nu+\frac{1}{2})p(1-\frac{\widetilde{q}}{q})]}\\ &\le C{t_\varepsilon}^{p-q}\end{aligned}$$

The last two lines are valid if $\widetilde{q}$ is chosen strictly smaller than q (this is possible since $\pi\ge q$).
Let us now consider the case where $s<(\nu+\frac{1}{2})(\frac{p}{q}-1)$, and choose now

$$q=\frac{p}{2(s+\nu-\frac{1}{\pi})+1},$$

in such a way that we easily verify that $p-q=2\frac{s-1/\pi+1/p}{1+2(\nu+s-1/\pi)}$, $q-\pi=\frac{(p-\pi)(1+2\nu)}{2(s+\nu-\frac{1}{\pi})+1}>0$, because s is supposed to be larger that $\frac{1}{\pi}$. Furthermore we also have $s+\frac{1}{2}-\frac{1}{\pi}=\frac{p}{2q}-\frac{1}{q}+\nu(\frac{p}{q}-1)$.

Hence taking $z_1=\pi$ and using again the fact that f belongs to $B^s_{\pi,r}(M)$,

$$\begin{aligned}I &\le {t_\varepsilon}^{p-\pi}[\sum_{-1}^{j_0}2^{j[\nu(p-\pi)+\frac{p}{2}-1]}\sum_{k\in\chi_j}|\beta_{jk}|^{\pi}]\\ &\le {t_\varepsilon}^{p-\pi}\sum_{-1}^{j_0}2^{j[(\nu+\frac{1}{2}-\frac{1}{p})\frac{p}{q}(q-\pi)]}\tau_j^{\pi}\\ &\le C{t_\varepsilon}^{p-\pi}2^{j_0[(\nu+\frac{1}{2}-\frac{1}{p})\frac{p}{q}(q-\pi)]}\end{aligned}$$

This is true since $\nu + \frac{1}{2} - \frac{1}{p}$ is also strictly positive because of our constraints. If we now take $2^{j_0 \frac{p}{q}(\nu+\frac{1}{2}-\frac{1}{p})} \sim {t_\varepsilon}^{-1}$ we get the bound

$${t_\varepsilon}^{p-q}$$

which is the rate announced in the theorem for this case.

Again, for II, we have, taking now $z_2 = \widetilde{q} > q \ (> \pi)$

$$\begin{aligned} II &\leq {t_\varepsilon}^{p-\widetilde{q}} [\sum_{j=j_0+1}^{J} 2^{j[\nu(p-\widetilde{q})+\frac{p}{2}-1]} \sum_{k \in \chi_j} |\beta_{jk}|^{\widetilde{q}}] \\ &\leq C {t_\varepsilon}^{p-\widetilde{q}} \sum_{j=j_0+1} 2^{j[(\nu+\frac{1}{2}-\frac{1}{p})\frac{p}{q}(q-\widetilde{q})]} z_j^{\frac{\widetilde{q}}{\pi}} \\ &\leq C {t_\varepsilon}^{p-\widetilde{q}} 2^{j_0[(\nu+\frac{1}{2}-\frac{1}{p})\frac{p}{q}(q-\widetilde{q})]} \\ &\leq C {t_\varepsilon}^{p-q} \end{aligned}$$

References

[1] F. Abramovich and B. W. Silverman. Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, 85(1):115–129, 1998.

[2] A. Antoniadis and J. Bigot. Poisson inverse models. 2004. Preprint Grenoble.

[3] A. Antoniadis, J. Fan, and I. Gijbels. A wavelet method for unfolding sphere size distributions. 29:265–290, 2001.

[4] Lawrence D. Brown, T. Tony Cai, Mark G. Low, and Cun-Hui Zhang. Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.*, 30(3):688–707, 2002.

[5] Lawrence D. Brown and Mark G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398, 1996.

[6] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874, 2002.

[7] Laurent Cavalier and Alexandre Tsybakov. Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*, 123(3):323–354, 2002.

[8] Albert Cohen, Marc Hoffmann, and Markus Reiß. Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.*, 42(4):1479–1501 (electronic), 2004.

[9] L.M. Cruz-Orive. Distribution-free estimation of sphere size distributions from slabs showing overprojections and truncations, with a review of previous methods. *J. Microscopy*, 131:265–290, 1983.

[10] V. Dicken and P. Maass. Wavelet-Galerkin methods for ill-posed problems. *J. Inverse Ill-Posed Probl.*, 4(3):203–221, 1996.

[11] D. L. Donoho and I. M. Johnstone. Minimax risk over ℓ_p-balls for ℓ_q-error. *Probability Theory and Related Fields*, 99:277–303, 1994.

[12] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, 57:301–369, 1995. With Discussion.

[13] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Annals of Statistics*, 24:508–539, 1996.

[14] David L. Donoho. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.*, 2(2):101–126, 1995.

[15] R. J. Duffin and A. C. Schaeffer. A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc.*, 72:341–366, 1952.

[16] Sam Efromovich and Vladimir Koltchinskii. On inverse problems with unknown operators. *IEEE Trans. Inform. Theory*, 47(7):2876–2894, 2001.

[17] J. Fan and J.K. Koo. Wavelet deconvolution. *IEEE Transactions on Information Theory*, 48(3):734–747, 2002.

[18] M. Frazier, B. Jawerth, and G. Weiss. Littlewood paley theory and the study of functions spaces. *CMBS*, 79, 1991. AMS.

[19] Alexander Goldenshluger and Sergei V. Pereverzev. On adaptive inverse estimation of linear functionals in Hilbert scales. *Bernoulli*, 9(5):783–807, 2003.

[20] I. A. Ibragimov and R. Z. Hasminskii. *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York, 1981.

[21] Michael Jähnisch and Michael Nussbaum. Asymptotic equivalence for a model of independent non identically distributed observations. *Statist. Decisions*, 21(3):197–218, 2003.

[22] Iain M. Johnstone, Gérard Kerkyacharian, Dominique Picard, and Marc Raimondo. Wavelet deconvolution in a periodic setting. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(3):547–573, 2004.

[23] Iain M. Johnstone and Bernard W. Silverman. Discretization effects in statistical inverse problems. *J. Complexity*, 7(1):1–34, 1991.

[24] Jérôme Kalifa and Stéphane Mallat. Thresholding estimators for linear inverse problems and deconvolutions. *Ann. Statist.*, 31(1):58–109, 2003.

[25] G. Kerkyacharian, P. Petrushev, D. Picard, and Y. Xu. Localized polynomials and frames induced by laguerre functions. 2005.

[26] G. Kerkyacharian and D. Picard. Thresholding algorithms and well-concentrated bases. *Test*, 9(2), 2000.

[27] G. Kerkyacharian, D. Picard, P. Petrushev, and T. Willer. Needvd: second generation wavelets for estimation in inverse problems. 2006. LPMA 2006.

[28] G. Kerkyacharian, D. Picard, and M. Raimondo. Adaptive boxcar deconvolution on full lebesgue measure sets. 2005. Preprint LPMA.

[29] Bernard A. Mair and Frits H. Ruymgaart. Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.*, 56(5):1424–1444, 1996.

[30] Stéphane Mallat. *A wavelet tour of signal processing*. Academic Press Inc., San Diego, CA, 1998.

[31] Peter Mathé and Sergei V. Pereverzev. Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(3):789–803, 2003.

[32] Yves Meyer. *Ondelettes et opérateurs. I.* Actualités Mathématiques. [Current Mathematical Topics]. Hermann, Paris, 1990.

[33] F. J. Narcowich, P. Petrushev, and J.M. Ward. Wavelet-based deconvolution for ill-conditioned systems. 2005. IMI 2005:04.

[34] R. Neelamani, H. Choi, and R. Baranuik. Wavelet-based deconvolution for ill-conditioned systems. 2000. http://www-dsp.rice.edu/publications/pub/neelsh98icassp.pdf.

[35] Michael Nussbaum. Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.*, 24(6):2399–2430, 1996.

[36] D. Nyshka, G. Wahba, S. Goldfarb, and T. Pugh. Cross validated spline methods for the estimation of three-dimensional tumor size distributions from observations on two-dimensional cross sections. *J. Amer. Statist. Assoc.*, 79:832–846, 1984.

[37] M. Pensky and B. Vidakovic. Adaptive wavelet estimator for nonparametric density deconvolution. *Annals of Statistics*, 27:2033–2053, 1999.

[38] P. Petrushev and Y. Xu. Localized polynomials frames on the interval with jacobi weights. 2005. IMI 2005.

[39] P. Petrushev and Y. Xu. Localized polynomials kernels and frames (needlets) on the ball. 2005. IMI 2005.

[40] Alexandre Tsybakov. On the best rate of adaptive estimation in some inverse problems. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(9):835–840, 2000.

[41] Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004.

[42] S. D. Wicksell. The corpuscle problem: a mathematical study of a biometric problem. 17:84–99, 1925.

[43] T. Willer. Deconvolution in white noise with a random blurring effect. 2005. LPMA 2005.

LPMA, Universités Paris VII et Paris X, CNRS, 175 rue du Chevaleret, 75013 Paris, France