Bandlet Image Estimation with Model Selection

Charles H Dossal, Erwan Le Pennec, Stéphane Mallat

Bandlet Image Estimation with Model Selection

Ch . Dossal (Université de Bordeaux), E. Le Pennec*(Université Paris Diderot) and S. Mallat (École Polytechnique)

September 18, 2008

1 Introduction

A new estimator is introduced to reduce white noise added to images having a geometrical regularity. This estimator projects the observations on orthogonal bandlet vectors selected in a dictionary of orthonormal bases. It is proved that the resulting risk is quasi asymptotically minimax for geometrically regular images. This paper is also a tutorial on estimation with general dictionary of orthogonal bases, through model selection. It explains how to build a thresholding estimator in a adaptively chosen "best" basis and gives a simple proof of its performance with the model selection approach of Barron Birgé and Massart [1].

Section 2 gives the statistical setting of the white noise model, and describes a model of $\mathbf{C}^\alpha$ geometrically regular images. Images in this class are $\mathbf{C}^\alpha$ (Hölder regularity α) outside a set of $\mathbf{C}^\alpha$ curves in $[0,1]^2$. Korostelev and Tsybakov [12] showed that the minimax quadratic risk over this class has an asymptotic decay of the order of $\sigma^{2\alpha/(\alpha+1)}$,

*Corresponding author: Erwan LE PENNEC, LPMA / Université Paris Diderot, 175 rue du Chevaleret, 75013 PARIS (FRANCE), Tel: + 33 1 44 27 37 73, Fax : + 33 1 44 27 72 23, email: lepennec@math.jussieu.fr

for a Gaussian white noise of variance σ^2. Their estimator relies on an explicit detection of the contours and is not stable relatively to any image blurring. Later, Donoho [9] overcomes the detection issue by replacing it with an well-posed optimization problem. Nevertheless, both use a model of images with sharp edges which limits their applications since most image edges are not strict discontinuities. They are blurred because of various diffraction effects which regularize discontinuities by unknown factors.

To overcome this issue, a thresholding estimator in a best band let basis is introduced. The concept of edge is replaced by the concept of local direction which does not require a precise position. Furthermore, these directions of regularity are not estimated directly but indirectly through a bast basis search algorithm. Section 3 gives a tutorial introduction of the generic class of thresholding estimators in a best basis selected in a dictionary of orthonormal bases. Such estimators have been studied by Donoho and Johnstone [10] and fit into the framework of Birgé and Massart [2, 15] This section provides a simplified presentation and proofs of their performance. Section 4 applies these results to a dictionary of bandlet bases for geometric image estimation. The quasi minimax performance of this bandlet estimator is proved in Theorem 3

2 Image estimation

White noise model and acquisition During the digital acquisition process, a camera measures an analog image f with a filtering and sampling process, which introduces an additive white noise. In this white noise model, the process that is being observed can be written

$$\mathrm{d}X_x = f(x)\mathrm{d}x + \sigma \mathrm{d}W_x,$$

where W_x is the Wiener process and σ is a known noise level parameter.

The camera measurements project this process over a family of N functions $\mathcal{B}_0 = \{\phi_n\}_{0 \leq n < N}$ that define a Riesz basis of an approximation space V_N . These functions

are the impulse responses of the photo-sensors of the camera. The resulting noisy observations are thus

$$\mathbf{X}_N[n] = X_{\phi_n} = \langle f, \phi_n \rangle + \sigma\, W_{\phi_n} \text{ for } 0 \leq n < N.$$

Observe that W_{ϕ_n} is a gaussian field of zero mean and covariance $E[W_{\phi_n} W_{\phi_m}] = \langle \phi_n, \phi_m \rangle$. We shall write X_{ϕ_n} by $\langle X, \phi_n \rangle$.

These noisy observations specify the orthogonal projection of the observed process on V_N that we denote with a slight abuse of notation $P_{V_N} X$. To simplify explanations, in the following we suppose that $\{\phi_n\}_{0 \leq n < N}$ is an orthogonal basis, with no loss of generality.

Minimax risk and geometrically regular images We study the maximum risk of estimators for images f in a given class, depending upon σ and N. Model classes are often derived from classical regularity spaces ($\mathbf{C}^\alpha$ spaces, Besov spaces,...). This does not take into account the existence of geometrically regular structures such as edges. This paper uses a geometric image model appropriate for edges, but not for textures, where images are considered as piecewise regular functions with discontinuities along regular curves in $[0,1]^2$. This geometrical image model has been proposed by Korostelev and Tsybakov[12] in their seminal work on image estimation. It is used as a benchmark to estimate or approximate images having some form of geometric regularity (Donoho[9], Shukla *et al*[18],...). An extension of this model that incorporates a blurring kernel h has been proposed [14] to model the various diffraction effects. The resulting class of images studied in this paper is the set of $\mathbf{C}^\alpha$ geometrically regular images specified by the following definition.

Definition 1. *A function $f \in L^2([0,1]^2)$ is $\mathbf{C}^\alpha$ geometrically regular over $[0,1]^2$ if*

- *$f = \tilde{f}$ or $f = \tilde{f} \star h$ with $\tilde{f} \in \mathbf{C}^\alpha(\Lambda)$ for $\Lambda = [0,1]^2 - \{\mathcal{C}_\gamma\}_{1 \leq \gamma \leq G}$,*

- *the blurring kernel h is $\mathbf{C}^\alpha$, compactly supported in $[-s,s]^2$ and $\|h\|_{\mathbf{C}^\alpha} \leq s^{-(2+\alpha)}$,*
- *the edge curves $\mathcal{C}_\gamma$ are $\mathbf{C}^\alpha$ and do not intersect tangentially if $\alpha > 1$.*

Edge based estimation Korostelev and Tsybakov[12] have built an estimator that is asymptotically minimax for geometrically regular functions f, as long as their is no blurring and hence that $h = \delta$. With a detection procedure, they partition the image in regions where the image is either regular or which include a "boundary fragment" corresponding to the subpart of a single discontinuity curve. In each region, they use either an estimator tailored to this "boundary fragments" or a classical kernel estimator for the regular regions. This yields a global estimate F of the image f. If the f is $\mathbf{C}^\alpha$ outside the boundaries and if the parametrization of the curve is also $\mathbf{C}^\alpha$ then there exists a constant C such that

$$\forall \sigma \quad , \quad E\left[\|f - F\|^2\right] \leq C\sigma^{\frac{2\alpha}{\alpha+1}} \quad .$$

This rate of convergence achieves the minimax rate for uniformly $\mathbf{C}^\alpha$ functions and thus the one for $\mathbf{C}^\alpha$ geometrically regular functions that includes this class. This means that sharp edges do not alter the rate of minimax estimation. However, this estimator is not adaptive relatively to the Holder exponent α that must be known in advance. Furthermore, it uses an edge detection procedure that fails when the image is blurred or when the discontinuity jumps are not sufficiently large.

Donoho[9] and Shukla and al [18] reuse the ideas of "boundary fragment" under the name "horizon model" to construct a piecewise polynomial approximation of images. They derive efficient estimators optimized for $\alpha \in [1,2]$. These estimators use a recursive partition of the image domain in dyadic squares, each square being split in two parts by an edge curve that is a straight segment. Both optimize the recursive partition and the choice of this straight edge segment in each dyadic square by minimizing a global

function. This leads to an minimax estimator up to a logarithmic factor which is adaptive relatively to the Holder exponent as long as $\alpha \in [1, 2]$.

Korostelev and Tsybakov[12] as well as Donoho rely on the sharpness of image edges in their estimators. In both cases, the estimator is chosen amongst a family of images that are discontinuous across parametrized edges, and these estimators are therefore not appropriate when the image edges are blurred. To avoid avoid this restriction, we now consider projector estimators on adaptive subspaces.

3 Projection Estimator and Model Selection

We study projection estimators that are decomposed in two steps. First a linear projection reduces the dimensionality of the problem by projecting the signal in a finite dimensional space. This first projection is typically performed by the digital acquisition device. Then a non-linear projection estimator refines this projector by reprojecting the resulting finite dimensional observation in a space that is chosen depending upon this observation. This non-linear projection is obtained with a thresholding in a best basis selected from a dictionary of orthonormal bases. Best basis algorithms for noise removal have been introduced by Coifman and Wickerhauser [7]. As recalled by Candès[3], their risk have been studied by Donoho and Johnstone [10] and are a special case of the general framework of model selection proposed by Birgé and Massart [2]. Note that Kolaczyk and Nowak[11] have studied a similar problem in a slightly different setting. We follow here the framework of model selection. This section gives a tutorial presentation of these best basis estimators with simplified proofs on the resulting risk upper bounds.

Approximation space V_N and further projection The observations are given by

$$\mathbf{X}_N[n] = X_{\phi_n} = \langle f, \phi_n \rangle + \sigma W_{\phi_n} \text{ for } 0 \leq n < N$$

where $\{\phi_n\}_{0\leq n<n}$ is an orthogonal basis of V_N. The observation thus provides finite dimensional observation $P_{V_N}X = P_{V_N}f + \sigma W_{V_N}$ where W_{V_N} is a finite dimensional white noise on V_N.

The observations $P_{V_N}X$ are reprojected in a subspace $\mathfrak{M} \subset V_N$ which results in an estimator $P_{\mathfrak{M}}P_{V_N}X = P_{\mathfrak{M}}X$. The overall risk includes the errors of the linear and non-linear projections:

$$\|f - P_{\mathfrak{M}}X\|^2 = \|f - P_{V_N}f\|^2 + \|P_{V_N}f - P_{\mathfrak{M}}X\|^2.$$

The dimension N of V_N is often chosen large enough so that $\|f - P_{V_N}f\|^2 \leq \|P_{V_N}f - P_{\mathfrak{M}}X\|^2$. This means that the acquisition device resolution is set so that the approximation error due to discretization is smaller than the estimation error. Engineers may also set N so that both terms are of the same order of magnitude, to limit the cost in terms of storage and computations.

Non-linear Projector in finite dimension Since all calculations are performed in the space V_N over observed samples $X_N[n] = \langle X, \phi_n \rangle$, we concentrate on these samples which amounts to identify $\mathbf{V}_N$ to $\mathbb{C}^N$. These noisy observations can be rewritten

$$\mathbf{X}_N[n] = \mathbf{f}_N[n] + \sigma \mathbf{W}_N[n]$$

where $\mathbf{f}_N[n] = \langle f, \phi_n \rangle \in \mathbb{C}^N$ is the digital image and $\mathbf{W}_N[n] = W_{\phi_n}$ is a Gaussian white noise random vector of variance σ^2. To simplify the notation, we drop the index N of the discrete observation $\mathbf{f}_N$ in the following although the resulting $\mathbf{f}$ still depends on the space V_N and its basis $\{\phi_n\}_{0\leq n<N}$.

The orthogonal projection $P_{\mathcal{M}}\mathbf{X}$ of $\mathbf{X}$ in a subspace $\mathcal{M}$ of $\mathbb{C}^N$ defines an orthogonal

projection of X in a corresponding subspace $\mathfrak{M}$ of V_N:

$$P_{\mathfrak{M}} X = P_{\mathfrak{M}} P_{V_N} X = \sum_{n=0}^{N-1} (P_{\mathcal{M}} \mathbf{X})[n]\, \phi_n,$$

and

$$\|P_{V_N} f - P_{\mathfrak{M}} X\|^2 = \|\mathbf{f} - P_{\mathcal{M}} \mathbf{X}\|^2 = \|\mathbf{f} - P_{\mathcal{M}} \mathbf{f}\|^2 + \sigma^2 \|P_{\mathcal{M}} \mathbf{W}\|^2.$$

The first bias term is the non-linear approximation error due to the projection in a subspace $\mathcal{M}$ of $\mathbb{C}^N$, decreases with the dimensionality of this space. The second "variance" term, which gives the energy of the noise projected in $\mathcal{M}$, increases with the space dimensionality. It is thus necessary to find a trade-off between this two trends, and select a space $\mathcal{M}$ to minimize the sum of both terms.

Model Selection in a Dictionary of orthonormal bases Let $\mathcal{B} = \{\mathbf{g}_m\}_{0 \leq m < N}$ be an orthonormal basis of $\mathbb{C}^N$. From any $M \leq N$ and any sub-family $\{\mathbf{g}_{m_k}\}_{1 \leq k \leq M}$ of M vectors, one can define a projection estimator on the space $\mathcal{M}$ generated by these vectors:

$$P_{\mathcal{M}} \mathbf{X} = \sum_{k=1}^{M} \langle \mathbf{X}, \mathbf{g}_{m_k} \rangle\, \mathbf{g}_{m_k}.$$

Instead of choosing a priori the orthogonal basis $\mathcal{B}$ of $\mathbb{C}^N$ we define a dictionary $\mathcal{D}_N$ which is a family of orthonormal bases. Some bases of $\mathcal{D}_N$ may have vectors in common. This dictionary can thus also be viewed as set of $K_N \geq N$ vectors, that are regrouped to form many possible orthonormal bases. Any collection of M vectors from any orthogonal basis $\mathcal{B} \in \mathcal{D}_N$ generates a space $\mathcal{M}$ that defines a possible estimator $P_{\mathcal{M}} \mathbf{X}$ of $\mathbf{f}$. Let $\mathcal{C} = \{\mathcal{M}_\gamma\}_\Gamma$ be the family of all such projection spaces. Ideally we would like to find the space $\mathcal{M} \in \mathcal{C}$ which minimizes $\|\mathbf{f} - P_{\mathcal{M}} \mathbf{X}\|$. The space $\mathcal{M}$ can be considered as an estimation model selected from a predefined family. It is therefore a model selection problem.

Oracle Model A projection estimator yields an estimation error

$$\|\mathbf{f} - P_{\mathcal{M}}\mathbf{X}\|^2 = \|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2 + \|P_{\mathcal{M}}\mathbf{W}\|^2.$$

The expected error of such an estimator is given by

$$E\left[\|\mathbf{f} - P_{\mathcal{M}}\mathbf{X}\|^2\right] = \|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2 + \sigma^2 M$$

and the best subspace for this criterion is the one that realizes the best trade-off between the approximation error $\|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2$ and the complexity of the models measured by $\sigma^2 M$.

This expected error can not be computed in practice since we have a single realization of X. We thus search for a subspace $\mathcal{M}$ such that the estimation error obtained by projecting $\mathbf{X}$ on this subspace is small with an overwhelming probability. An upper bound of the estimation error is obtained from an upper bound of the energy of the noise projected on $\mathcal{M}$. Each of the K_N dictionary noise coefficient $\langle \mathbf{W}, \mathbf{g}_\gamma \rangle$ is a Gaussian random variable of variance σ^2. A standard large deviation result proves that the absolute values taken by K_N such Gaussian random variables are bounded simultaneously by $T = \sigma\sqrt{2 \log K_N}$ with a probability that tends to 1 when N increases. Since the noise energy projected in $\mathcal{M}$ is the sum of M squared dictionary noise coefficients, we get $\|P_{\mathcal{M}}\mathbf{W}\|^2 \leq M\, T^2$. It results that

$$\|\mathbf{f} - P_{\mathcal{M}}\mathbf{X}\|^2 \leq \|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2 + M\, T^2. \tag{1}$$

over all subspaces $\mathcal{M}$ with a probability that tends to 1 as N increases. The estimation error is small if $\mathcal{M}$ is a space of small dimension M which yields a small approximation error $\|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|$. We denote by $\mathcal{M}_O \in \mathcal{C}$ the oracle space that minimizes the estimation

error upper bound (1)

$$\mathcal{M}_O = \arg\min_{\mathcal{M}\in\mathcal{C}}(\|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2 + M\,T^2).$$

Penalized empirical error The oracle space can not be calculated since $\mathbf{f}$ is unknown. It is thus necessary to replace it by a "best" space that will hopefully yield a close estimation error. A crude estimation of $\|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2$ is given by the empirical norm

$$\|\mathbf{X} - P_{\mathcal{M}}\mathbf{X}\|^2 = \|\mathbf{X}\|^2 - \|P_{\mathcal{M}}\mathbf{X}\|^2.$$

This may seem naive because estimating $\|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2$ with $\|\mathbf{X} - P_{\mathcal{M}}\mathbf{X}\|^2$ yields a large error

$$\|\mathbf{X} - P_{\mathcal{M}}\mathbf{X}\|^2 - \|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2 = (\|\mathbf{X}\|^2 - \|\mathbf{f}\|^2) + (\|P_{\mathcal{M}}\mathbf{f}\|^2 - \|P_{\mathcal{M}}\mathbf{X}\|^2),$$

whose expected value is $(N-M)\sigma^2$, with typically $M \ll N$. However, most of this error is in the first term on the right hand-side, which has no effect on the choice of space $\mathcal{M}$. This choice depends only upon the second term and is thus only influenced by noise projected in the space $\mathcal{M}$ of lower dimension M. The bias and the fluctuation of this term, and thus the choice of the basis, are controlled by increasing the parameter T.

We define the best empirical projection estimator $P_{\widehat{\mathcal{M}}}$ as the estimator that minimizes the resulting empirical penalized risk:

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}\in\mathcal{C}} \|\mathbf{X} - P_{\mathcal{M}}\mathbf{X}\|^2 + M\,T^2 \tag{2}$$

Thresholding in a best basis Finding the best estimator which minimizes (2) may seem computationally untractable because the number of possible spaces $\mathcal{M} \in \mathcal{C}$ is typically an exponential function of the number K_N of vectors in $\mathcal{D}_N$. We show that

this best estimator may however be found with a thresholding in a best basis.

Suppose that we impose that $\mathcal{M}$ are generated by M vectors from a particular basis $\mathcal{B} \in \mathcal{D}_N$. The following lemma proves that among all such spaces, the best projection estimator is obtained with a thresholding at T.

Lemma 1. *Among all spaces $\mathcal{M}$ that are families of any number $M \leq N$ of vectors in an orthonormal basis $\mathcal{B} = \{\mathbf{g}_m\}_{0 \leq m < N}$, the best estimator which minimizes $\|\mathbf{X} - P_{\mathcal{M}}\mathbf{X}\|^2 + M\,T^2$ is the thresholding estimator*

$$P_{\mathcal{M}_{\mathcal{B},T}}\mathbf{X} = \sum_{|\langle \mathbf{X}, \mathbf{g}_m \rangle| > T} \langle \mathbf{X}, \mathbf{g}_m \rangle\, \mathbf{g}_m. \tag{3}$$

Proof. Let $\mathcal{M} = \mathrm{Span}\{\mathbf{g}_m\}_{m \in I}$ with $I \subset [0, N)$, as $\mathcal{B}$ is an orthonormal basis,

$$\|\mathbf{X} - P_{\mathcal{M}}\mathbf{X}\|^2 + M\,T^2 = \sum_{m \notin I} |\langle \mathbf{X}, \mathbf{g}_m \rangle|^2 + \sum_{m \in I} T^2$$

which is minimal if $I = \{m, |\langle \mathbf{X}, \mathbf{g}_m \rangle|^2 > T^2\}$. □

The thresholding estimator (3) projects X in the space $\mathcal{M}_{\mathcal{B},T}$ generated by the M vectors $\{\mathbf{g}_m\}_{|\langle \mathbf{X}, \mathbf{g}_m \rangle| > T}$ of $\mathcal{B}$ which produce coefficients above threshold. This lemma implies that best projection estimators are necessarily thresholding estimators in some basis. Minimizing $\|\mathbf{X} - P_{\mathcal{M}}\mathbf{X}\|^2 + M\,T^2$ over $\mathcal{M} \in \mathcal{C}$ is thus equivalent to find the best basis $\widehat{\mathcal{B}}$ which minimizes the thresholding penalized empirical risk:

$$\widehat{\mathcal{B}} = \arg \min_{\mathcal{B} \in \mathcal{D}_N} \|\mathbf{X} - P_{\mathcal{M}_{\mathcal{B},T}}\mathbf{X}\|^2 + M\,T^2.$$

The best space which minimizes the empirical penalized risk in (2) is derived from a thresholding in the best basis $\widehat{\mathcal{M}} = \mathcal{M}_{\widehat{\mathcal{B}},T}$.

The following theorem, first proved by Barron, Birgé and Massart[1], proves that the thresholding estimation error in the best basis is bounded by the estimation error by

projecting in the oracle space $\mathcal{M}_O$, up to a factor 4.

Theorem 1. *Let $\mathcal{C} = \{\mathcal{M}_\gamma\}_\Gamma$ be the family of projection spaces generated by vectors in the orthogonal bases of a dictionary $\mathcal{D}_N$. Let K_N be the number of different vectors in $\mathcal{D}_N$. Let $\sigma > 0$ and $T = \lambda\sqrt{\log(K_N)}\,\sigma$ with Let $\lambda \geq \sqrt{32 + \frac{8}{\log(K_N)}}$. For any $\mathbf{f} \in \mathbb{C}^N$, the thresholding estimator $P_{\mathcal{M}_{\widehat{\mathcal{B}},T}}\mathbf{X}$ in the best basis*

$$\widehat{\mathcal{B}} = \arg\min_{\mathcal{B}\in\mathcal{D}_N} \|\mathbf{X} - P_{\mathcal{M}_{\mathcal{B},T}}\mathbf{X}\|^2 + M\,T^2$$

satisfies

$$E\left[\|\mathbf{f} - P_{\mathcal{M}_{\widehat{\mathcal{B}},T}}\mathbf{X}\|^2\right] \leq 4\left(\min_{\mathcal{M}\in\mathcal{C}} \|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2 + M\,T^2\right) + \frac{64}{K_N}\sigma^2.$$

The appendix gives a simple proof of Theorem 1, inspired by Birgé and Massart[2], which requires only a concentration lemma for the norm of the noise in all the subspaces spanned by the K_N generators of $\mathcal{D}_N$. Birgé and Massart [2] obtain a better lower bound condition for λ (roughly $\lambda > \sqrt{2}$) and a multiplicative factor smaller than 4, with a more complex proof using Talagrand's inequalities.

It results that any bound on $\min_{\mathcal{M}\in\mathcal{C}} \|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2 + M\,T^2$, gives a bound on the risk of the best basis estimator in $\mathbb{C}^N$. This translates into an estimator $F = P_{\mathfrak{M}_{\widehat{\mathcal{B}},T}}X$ of f which satisfies

$$E\left[\|f - F\|^2\right] \leq 4\left(\min_{\mathfrak{M}\in\mathfrak{C}} \|f - P_{\mathfrak{M}}f\|^2 + M\,T^2\right) + \frac{64}{K_N}\sigma^2.$$

where $\mathfrak{M}$ and $\mathfrak{C}$ corresponds to $\mathcal{M}$ and $\mathcal{C}$ through the mapping between V_N and $\mathbb{C}^N$.

To obtain a computational estimator, the minimization

$$\widehat{\mathcal{B}} = \arg\min_{\mathcal{B}\in\mathcal{D}_N} \|\mathbf{X} - P_{\mathcal{M}_{\mathcal{B},T}}\mathbf{X}\|^2 + M\,T^2 \quad ,$$

should be performed with a number of operations typically proportional to the number

K_N of vectors in the dictionary. This requires to construct appropriate dictionaries of orthogonal bases. Examples of such dictionaries have been proposed by Coifman and Wickerhauser [7] with wavelet packets or by Coifman and Meyer [6] with local cosine bases for signals having localized time-frequency structures. Next section reviews the dictionary of bandlet orthogonal bases that is adapted to the estimation of geometrically regular images.

4 Best basis image estimation and bandlets

Thresholding in a single basis Theorem 1 clearly applies when the dictionary $\mathcal{D}_N$ is reduced to a single basis $\mathcal{B}$. The corresponding estimator is the classical thresholding estimator which quadratic risk satisfies thus

$$E\left[\|\mathbf{f} - P_{\mathcal{M}_{\mathcal{B},T}}\mathbf{X}\|^2\right] \leq 4\left(\min_{\mathcal{M}\in\mathcal{C}} \|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2 + M\,T^2\right) + \frac{64}{N}\sigma^2$$

It remains only to choose the basis and the space $\mathbf{V}_N$ with respect to σ.

Wavelet bases provide a first family of estimators used commonly in image processing. Such a two dimensional wavelet basis is constructed from two real functions, a one dimensional wavelet ψ and a corresponding one dimensional scaling function ϕ, which are both dilated and translated:

$$\psi_{j,k}(x) = \frac{1}{2^{j/2}}\psi\left(\frac{x - 2^j k}{2^j}\right) \text{ and } \phi_{j,k}(x) = \frac{1}{2^{j/2}}\phi\left(\frac{x - 2^j k}{2^j}\right) \quad .$$

Note that the index j goes to $-\infty$ when the wavelet scale 2^j decreases. For a suitable choice of ψ and ϕ, the family $\{\psi_{j,k}(x)\}_{j,k}$ is an orthogonal basis of $L^2([0,1])$ and the

following family constructed by tensorization

$$\left\{ \begin{array}{l} \psi^V_{j,k}(x) = \psi^V_{j,k}(x_1,x_2) = \phi_{j,k_1}(x_1)\,\psi_{j,k_2}(x_2), \\ \psi^H_{j,k}(x) = \psi^H_{j,k}(x_1,x_2) = \psi_{j,k_1}(x_1)\,\phi_{j,k_2}(x_2), \\ \psi^D_{j,k}(x) = \psi^D_{j,k}(x_1,x_2) = \psi_{j,k_1}(x_1)\,\psi_{j,k_2}(x_2) \end{array} \right\}_{(j,k_1,k_2)}$$

is an orthonormal basis of the square $[0,1]^2$. Furthermore, the spaces $V_j = \mathrm{Span}\{\phi^o_{j,k}\}_{o,k_1,k_2}$, called approximation spaces of scale 2^j, admits $\{\psi^o_{l,k}\}_{o,l\geq j,k_1,k_2}$ as an orthogonal basis.

The approximation space $\mathbf{V}_N$ of the previous section coincides with the classical wavelet approximation space V_j when $N = 2^{-j/2}$. Note that, through the identification between $\mathbf{V}_N$ and $\mathbb{C}^N$, the orthogonal wavelet basis of $\mathbf{V}_n = V_j$, $\{\psi^o_{l,k}\}_{o,l>j,k_1,k_2}$, becomes an orthogonal discrete wavelet basis of $\mathbb{C}^N$, $\mathcal{B} = \{\Psi^o_{j',k}\}_{o,l>j,k_1,k_2}$.

A classical approximation result ensures that for any function f $\mathbf{C}^\alpha$, as soon as the wavelet has more than $\lfloor\alpha\rfloor$ vanishing moments, there is a constant C such that, for any T, $\min_{\mathcal{M}\in\mathcal{C}} \|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2 + M\,T^2 \leq C(T^2)^{\frac{\alpha}{\alpha+1}}$, and, for any N, $\|P_{V_N} f - f\|^2 \leq CN^{-\alpha}$. For $N = 2^{-j/2}$ with $\sigma^2 = [2^j, 2^{j+1}]$, Theorem 1 thus implies

$$E[\|f - F\|^2] \leq C(|\log(\sigma)|\sigma^2)^{\frac{\alpha}{\alpha+1}}\ .$$

This is up to the logarithmic term the best possible rate for $\mathbf{C}^\alpha$ functions. Unfortunately, wavelets bases do not provides such an optimal representation for the $\mathbf{C}^\alpha$ geometrically regular functions specified by Definition 1. Wavelets fail to capture the geometrical regularity of edges: near them, the wavelets coefficients remains large. One can verify that the rate of convergence in a wavelet basis decays like $(|\log(\sigma)|\sigma^2)^{1/2}$, which is far from the minimax rate.

A remarkably efficient representation was introduced by Candès and Donoho[4]. Their curvelets are not isotropic like wavelets but are more elongated along a preferential direction and have two vanishing moments along this direction. They are dilated

and translated like wavelets but they are also rotated. The resulting family of curvelets $\mathcal{C} = \{c_n\}_n$ is not a basis but a tight frame, which means that for any $f \in L^2([0,1]^2)$

$$\sum_{c_n \in \mathcal{C}} |\langle f, c_n \rangle|^2 = A\|f\|^2 \quad \text{with } A > 1.$$

Although this is not an orthonormal basis, the results of Section 3 can be extended to this setting. Projecting the data on the first $N = \sigma^{-1/2}$ first coefficients and thresholding the remaining coefficients with a threshold $\lambda\sqrt{\log N}\sigma$ yields an estimator F that satisfies

$$E\left[\|f - F\|^2\right] \leq C(|\log \sigma|\sigma^2)^{\frac{\alpha}{\alpha+1}}$$

with a constant C that depends only on f. This is the optimal decay rate for the risk up to the logarithmic factor for $\alpha \in [1,2]$. No such fixed representation is known to achieve a similar result for α larger than 2.

Dictionary of orthogonal bandlet bases To cope with a geometric regularity of order $\alpha > 2$, one needs basis elements which are more anisotropic than the curvelets, are more adapted to the geometry of edges and have more vanishing moments in the direction of regularity. Bandlet bases[13, 14, 17] are orthogonal bases whose elements have such properties. Their construction is based on the observation that even, if the wavelet coefficients are large in the neighborhood of an edge, these wavelets coefficients are regular along the direction of the edge as illustrated by Fig 1.

To capture this geometric regularity, wavelets coefficients are locally recombined along the direction of regularity to produce more small coefficients. A local orthogonal transform, inspired by the work of Alpert, combines locally the wavelets along the direction of regularity, represented by arrows in the right most image of Fig 1), to produce a new orthogonal basis, a basis of bandlets that are elongated along the direction of regularity and have the corresponding vanishing moments. The construction of a bandlet

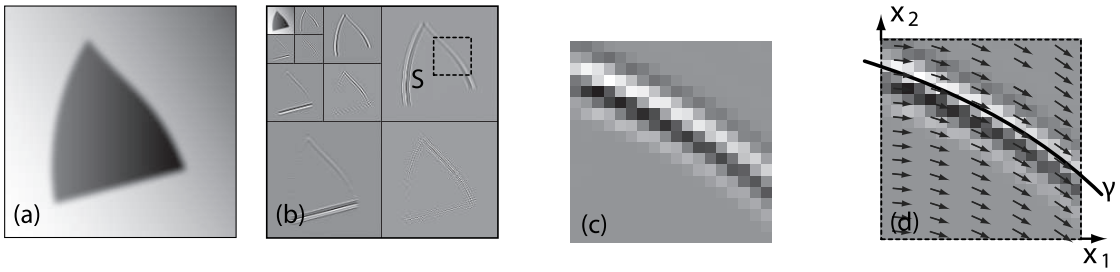

Figure 1: a) a geometrically regular image, b) the associated wavelet coefficients, c) a close-up of wavelet coefficients in a detail space W_j^o that shows their remaining regularity, d) the geometrical flow adapted to this square of coefficients, here it is vertically constant and parametrized by a polynomial curve γ

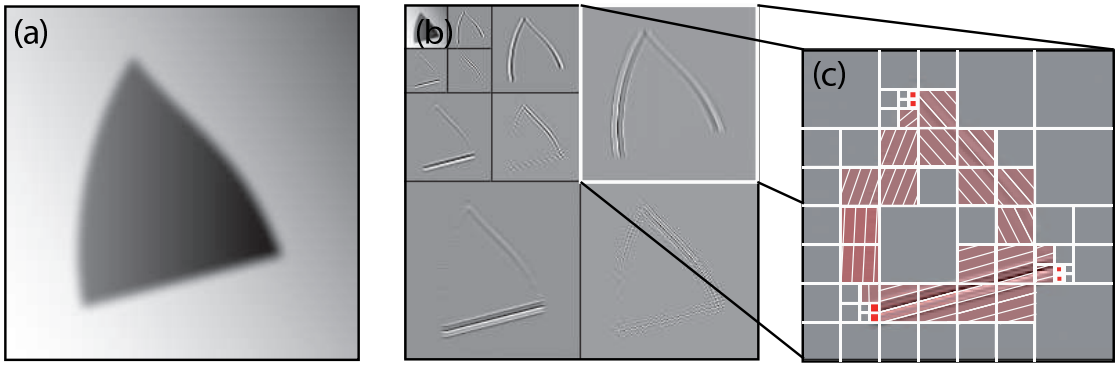

Figure 2: a) a geometrically regular image b) the corresponding wavelet coefficients c) the quadtree associated to the segmentation of a detail space W_j^o. In each square where the image is not uniformly regular, the flow is shown.

basis of a wavelet multiresolution space $V_j = \text{Span}\{\phi_{j,k_1,k_2}\}_{k_1,k_2}$ starts by decomposing this space into detail wavelet spaces

$$V_j = \bigoplus_{o,l>j} W_l^o \quad \text{with} \quad W_l^o = \text{Span}\{\psi_{l,k_1,k_2}^o\}_{k_1,k_2} \ .$$

For any level l and orientation o, the detail space W_l^o is a space of dimension $(2^{-l})^2$ in which the coefficients will be recombined along a geometric flow, a vector fields meant to follow the geometric direction of regularity. As illustrated by Fig. 2, this flow is

structured by a partition into dyadic squares in which the flow, if it exists, is vertically or horizontally constant, so that it can be easily parametrized by its tangent. For each geometry choice, a specific orthogonalization process[17] yields an orthogonal basis of bandlets that have vanishing moments in the direction of the geometric flow. This geometry should obviously be adapted to each image: the partition and the flow direction should match the image structures. This choice of geometry can be seen as an ill posed problem of estimation of the edges or of the direction of regularity. To avoid this issue, the problem is recasted as a best basis search in a dictionary. The geometry chosen will be the one of the best basis.

The dictionary $\mathcal{D}_{(2^{-j})^2}$ of orthogonal bandlets bases of V_j should be finite for both approximation and estimation: this requires a discretization of the geometry. As proved in [17], this is not an issue: the flow does not have to follow exactly the direction of regularity but only up to a known precision. More precisely, it is sufficient to parametrized the flow in any dyadic square by the tangent of a polynomial of degree p (the number of vanishing moments of the wavelets) chosen amongst a family of size $O(2^{-j(p+1)})$.

Any basis of the dictionary $\mathcal{D}_{(2^{-j})^2}$ is specified by a set of dyadic squares partitions for each details spaces W_l^o, $l > j$, and, for each square of the partition, a flow parametrized by a direction and one of these $O(2^{-j(p+1)})$ polynomials. The number of bases in the dictionary $\mathcal{D}_{(2^{-j})^2}$ grows exponentially with 2^{-j}, but the total number of different bandlets $K_{(2^{-j})^2}$ grows only polynomially like $O(2^{-j(2p+5)})$. Indeed the bandlets in a given dyadic square with a given geometry are reused in numerous bases. The total number of bandlets in the dictionary is thus bounded by the sum over all $O(2^{-2j})$ dyadic squares and all $O(2^{-j(p+1)})$ choices for the flow of the number of bandlets in the square. Noticing that $(2^{-j})^2$ is a rough bound of the number of bandlets in any subspaces of V_j yields the existence of C_K such that $2^{-j(p+5)} \leq K_{(2^{-j})^2} \leq C_K 2^{-j(p+5)}$.

Approximation in bandlet dictionaries Bandlet basis dictionary provides an asymptotically optimal representation of C^α geometrically regular functions through the following theorem[17].

Theorem 2. *For any f $\mathbf{C}^\alpha$ geometrically regular function, it exists a real number C such that for any $T > 0$ and $2^j \leq T$*

$$\min_{\mathcal{B} \in \mathcal{D}_{(2^{-j})^2}} \|f - P_{\mathfrak{M}_{\mathcal{B},T}} f\|^2 + MT^2 \leqslant CT^{2\alpha/(\alpha+1)} \tag{4}$$

where M is the dimension of the subspace $\mathfrak{M}_{\mathcal{B},T}$ which corresponds to the space $\mathcal{M}_{\mathcal{B},T}$ spanned by the vectors of $\mathcal{B}$ whose inner products with $\mathbf{f}$ is larger than T.

To find the best basis that minimizes $\|\mathbf{f} - P_{\mathcal{M}_{\mathcal{B},T}}\mathbf{f}\|^2 + MT^2$, the fast algorithm proposed by Coifman and Wickerhauser [7], and Donoho [8] is used. This algorithm uses the additive structure of $\|\mathbf{f} - P_{\mathcal{M}_{\mathcal{B},T}}\mathbf{f}\|^2 + MT^2$ to conduct the optimization of the basis of a given detail space with a simple bottom-up procedure. It is based on two observations. Firstly, the best partition of a given dyadic square is either itself or the union of the best partitions of its four dyadic subsquares. This hierarchical tree structure of the partitioning process leads to a bottom up optimization algorithm once the best flow has been found for every dyadic squares. Secondly, the limited number of possible flows in a square is such that the best flow in a given dyadic square can be obtained with a simple brute force exploration.

More precisely, the brute force search of the best flow can be conducted independently over all dyadic squares and all detail spaces with a total complexity of order $O(2^{-j(p+5)})$ and yields a value of the penalized criterion for each dyadic squares. It remains now to find the best partition. We proceed in a bottom up fashion. The best partition with squares of width smaller than 2^{j+1} is obtained from the best partition with squares of width smaller than 2^j: inside each dyadic square of width 2^{j+1} the best partition is either the partition obtained so far or the considered square. This choice is made according to

the penalized criterion. The initialization of this process is straightforward as the best partition with square of size 1 is obviously the full partition. The complexity of this best partition search is of order $O(2^{-2j})$ and thus the complexity of the best basis is driven by the best flow search whose complexity is of order $O(2^{-j(p+5)})$, which nevertheless remains polynomial in 2^{-j}.

Bandlet estimators Estimating the edges is a complex task on blurred function and becomes even much harder in presence of noise. Fortunately, the bandlet estimator do not rely on such a detection process. The chosen geometry is obtain with the best basis selection of the previous section. This allows to select an efficient basis even in the noisy setting.

Indeed, combining the bandlet approximation result of Theorem 2 with the model selection results of Theorem 1 proves that the thresholding estimator in a best bandlet basis is quasi asymptotically minimax for $\mathbf{C}^\alpha$ geometrically regular images.

For any $N = (2^{-j})^2$ the observed process X is projected on the space $\mathbf{V}_N = V_j$ which yields the observations $\mathbf{X}$. These noisy data are decomposed in a dictionary of bandlet orthogonal bases. The best basis algorithm selects the bandlet basis $\widehat{\mathcal{B}}$ amongst $\mathcal{D}_N = \mathcal{D}_{(2^{-j})^2}$ that minimizes

$$\|\mathbf{X} - P_{\mathcal{M}_{\mathcal{B},T}}\mathbf{X}\|^2 + T^2 M$$

with $T = \lambda\sqrt{\log(K_{(2^{-j})^2})}\sigma$. The resulting bandlet estimator is $F = P_{\mathfrak{M}_{\widehat{\mathcal{B}},T}}X$. For the sake of simplicity, as we consider an asymptotic behavior, we assume that $\sigma \leq \frac{1}{2}$. This implies that it exists $j < 0$ such that $\sigma \in (2^{j-1}, 2^j]$ The following theorem proves that choosing $N = 2^{-2j}$ and λ large enough yields a quasi minimax estimator.

More precisely,

Theorem 3. *Suppose that $\alpha \leq p$ where p in the number of wavelet vanishing moments.*

Suppose that $\widetilde{\lambda} \geq \sqrt{32\,(p+5+\log C_K)+8}$. *For any* $\mathbf{C}^\alpha$ *geometrically regular function* f, *there exists* $C>0$ *such that for any* $\sigma \leq \frac{1}{2}$, *if we let* $N = 2^{-2j}$ *with* j *such that* $\sigma \in (2^{j-1}, 2^j]$ *and* $T = \widetilde{\lambda}\sqrt{|\log\sigma|}\sigma$, *if* $F = P_{\mathfrak{M}_{\widehat{\mathcal{B}},T}} X$ *is the estimator in the best bandlet basis* $\widehat{\mathcal{B}} \in \mathcal{D}_N$ *which minimizes*

$$\|\mathbf{X} - P_{\mathcal{M}_{\mathcal{B},T}}\mathbf{X}\|^2 + T^2 M$$

then

$$E\left[\|f-F\|^2\right] \leq C(|\log\sigma|\sigma^2)^{\frac{\alpha}{\alpha+1}}.$$

Theorm 3 is a direct consequence of Theorem 1 and Theorem 2,

Proof. For any σ, observe that $2^{-j(p+5)} \leq K_N = K_{(2^{-j})^2} \leq C_K 2^{-j(p+5)}$ so that Theorem 1 applies as $T = \widetilde{\lambda}\sqrt{|\log\sigma|}\sigma \geq \lambda\sqrt{\log(K_N)}\sigma$ with a large enough λ. This yields

$$E\left[\|f-F\|^2\right] \leq 4 \min_{\mathfrak{M}\in\mathfrak{C}} \left(\|f - P_{\mathfrak{M}} f\|^2 + T^2 M\right) + \frac{64}{K_N}\sigma^2 \quad . \tag{5}$$

Now as $T \geq 2^j$, Theorem 2 applies and there is a constant C independant of T such that

$$\min_{\mathfrak{M}\in\mathfrak{C}} \left(\|f - P_{\mathfrak{M}} f\|^2 + T^2 M\right) \leq C(T^2)^{\alpha/(\alpha+1)} \quad .$$

Pluging this bound into (5) gives the result. □

The estimate $F = P_{\mathfrak{M}_{\widehat{\mathcal{B}},T}} X$ is computed efficiently by the same fast algorithm used in the approximation setting without requiring the knowledge of the regularity parameter α. The bandlet estimator is thus a tractable adaptive estimator that provides, up to the logarithmic term, the best possible minimax rate of convergence for $\mathbf{C}^\alpha$ geometrically regular function.

Theorem 3 applies only to $\mathbf{C}^\alpha$ geometrically regular function but the bandlet estimator can be applied to any images[16]. Figure 3 illustrates the good behavior of the

Missing Figure. See `http://www.math.jussieu.fr/~lepennec/` for the complete version.

Figure 3: Comparison between the translation invariant wavelet estimator and the bandlet estimator. The number within parenthesis is the PSNR defined by $-10 \log \left(\frac{\|f-F\|^2}{\|f\|_\infty^2} \right)$ (the larger the better).

bandlet estimator for natural images. Each line presents the original image, the degraded noisy image and two estimations, one using classical translation invariant estimator[5]. and the other using the bandlet estimator. The bandlet improvement can be seen numerically as well as visually. The quadratic error is smaller with the bandlet estimator and the bandlets preserve much more geometric structures in the images.

A Proof of Theorem 1

Concentration inequalities are at the core of all the selection model estimators. Essentially, the penalty should dominate the random fluctuation of the minimized quantity. The key lemma, Lemma 2, uses a concentration inequality for gaussian variable to ensure, with high probability, that the noise energy is small simultaneously in all the subspaces $\mathcal{M}_I$ spanned by a subset I of the K_N different vectors, denoted by $\mathbf{h}_k$, of $\mathcal{D}_N$.

Lemma 2. *For all $u \geq 0$, with a probability greater than $1 - 2/K_N e^{-u}$,*

$$\forall I \subset [1, K_N] \text{ and } \mathcal{M}_I = \mathrm{Span}\{\mathbf{h}_k\}_{k \in I}, \quad \|P_{\mathcal{M}_I} \mathbf{W}\| \leq \sqrt{M_I} + \sqrt{4 \log(K_N) M_I + 2u}$$

where M_I is the dimension of $\mathcal{M}_I$.

Proof of Lemma 2. The key ingredient of this proof is a concentration inequality. Tsirelson's Lemma[19] implies that for any 1-Lipschitz function $\phi : \mathbb{C}^n \to \mathbb{C}$ ($|\phi(x) - \phi(y)| \leq \|x - y\|$) if $\mathbf{W}$ is a gaussian standard white noise in $\mathbb{C}^n$ then

$$\mathbb{P}\left\{\phi(\mathbf{W}) \geq E\left[\phi(\mathbf{W})\right] + t\right\} \leq e^{-t^2/2} \quad .$$

For any space $\mathcal{M}$, $\mathbf{f} \mapsto \|P_{\mathcal{M}} \mathbf{f}\|$ is 1-Lipschitz. Applying Tsirelson's Lemma with

$t = \sqrt{4\log(K_N)M + 2u}$ yields

$$\mathbb{P}\left\{\|P_{\mathcal{M}}\mathbf{W}\| \geq E\left[\|P_{\mathcal{M}}\mathbf{W}\|\right] + \sqrt{4\log(K_N)M + 2u}\right\} \leq K_N^{-2M}e^{-u} \quad .$$

Now as $E\left[\|P_{\mathcal{M}}\mathbf{W}\|\right] \leq \left(E\left[\|P_{\mathcal{M}}\mathbf{W}\|^2\right]\right)^{1/2} = \sqrt{M}$, one derives

$$\mathbb{P}\left\{\|P_{\mathcal{M}}\mathbf{W}\| \geq \sqrt{M} + \sqrt{4\log(K_N)M + 2u}\right\} \leq K_N^{-2M}e^{-u} \quad .$$

Now

$$\begin{aligned}
&\mathbb{P}\left\{\exists I \subset [1, K_n], \|P_{\mathcal{M}_I}\mathbf{W}\| \geq \sqrt{M_I} + \sqrt{4\log(K_N)M_I + 2u}\right\} \\
&\qquad \leq \sum_{I\subset[1,K_n]} \mathbb{P}\left\{\|P_{\mathcal{M}_I}\mathbf{W}\| \geq \sqrt{M_I} + \sqrt{4\log(K_N)M_I + 2u}\right\} \\
&\qquad \leq \sum_{I\subset[1,K_n]} K_N^{-2M_I}e^{-u} \\
&\qquad \leq \sum_{n=1}^{K_N} \binom{n}{K_N} K_N^{-2n}e^{-u} \leq \sum_{n=1}^{K_N} K_N^{-n}e^{-u} \\
&\qquad \leq \frac{K_N^{-1}}{1-K_N^{-1}}e^{-u}
\end{aligned}$$

and thus

$$\begin{aligned}
&\mathbb{P}\left\{\exists I \subset [1, K_n], \|P_{\mathcal{M}_I}\mathbf{W}\| \geq \sqrt{M_I} + \sqrt{4\log(K_N)M_I + 2u}\right\} \\
&\qquad \leq \frac{2}{K_N}e^{-u}
\end{aligned}$$

□

The proof of Theorem 1 follows from the definition of the best basis, the oracle subspace and the previous Lemma.

Proof of Theorem 1. Recall, that $\mathbf{X} = \mathbf{f} + \sigma\mathbf{W} \in \mathbb{C}^N$ with $\mathbf{W}$ a gaussian white noise.

By construction, the thresholding estimate is $P_{\mathcal{M}_{\widehat{\mathcal{B}},T}}\mathbf{X}$ where

$$\widehat{\mathcal{B}} = \arg\min_{\mathcal{B}\in\mathcal{D}_N} \|\mathbf{X} - P_{\mathcal{M}_{\mathcal{B},T}}\mathbf{X}\|^2 + M\,T^2 \quad .$$

To simplify the notation, we denote by $\widehat{\mathcal{M}}$ and $\widehat{M}$ the corresponding space and its dimension.

Denote now M_O the dimension of the oracle subspace $\mathcal{M}_O$ that has been defined as the minimizer of

$$\|\mathbf{f} - P_{\mathcal{M}}\mathbf{f}\|^2 + M\,T^2 \quad .$$

By construction,

$$\|\mathbf{X} - P_{\widehat{\mathcal{M}}}\mathbf{X}\|^2 + \lambda^2\log(K_N)\sigma^2\widehat{M} \leq \|\mathbf{X} - P_{\mathcal{M}_O}\mathbf{f}\|^2 + \lambda^2\log(K_N)\sigma^2 M_O$$

using $\|\mathbf{X}-P_{\widehat{\mathcal{M}}}\mathbf{X}\|^2 = \|\mathbf{X}-\mathbf{f}\|^2+\|\mathbf{f}-P_{\widehat{\mathcal{M}}}\mathbf{X}\|^2+2\langle\mathbf{X}-\mathbf{f},\mathbf{f}-P_{\widehat{\mathcal{M}}}\mathbf{X}\rangle$ and a similar equality for $\|\mathbf{X} - P_{\mathcal{M}_O}f\|^2$, one obtains

$$\begin{aligned}\|\mathbf{f} - P_{\widehat{\mathcal{M}}}\mathbf{X}\|^2 + \lambda^2\log(K_N)\sigma^2\widehat{M} \leq \|\mathbf{f} - P_{\mathcal{M}_O}\mathbf{f}\|^2 + \lambda^2\log(K_N)\sigma^2 M_O \\ + 2\langle\mathbf{X} - \mathbf{f}, P_{\widehat{\mathcal{M}}}\mathbf{X} - P_{\mathcal{M}_O}\mathbf{f}\rangle\end{aligned}$$

One should now concentrate on the bound on the scalar product :

$$\begin{aligned}|2\langle\mathbf{X} - \mathbf{f}, P_{\widehat{\mathcal{M}}}\mathbf{X} - P_{\mathcal{M}_O}\mathbf{f}\rangle| &= |2\langle\sigma P_{\widehat{\mathcal{M}}+\mathcal{M}_O}\mathbf{W}, P_{\widehat{\mathcal{M}}}\mathbf{X} - P_{\mathcal{M}_O}\mathbf{f}\rangle| \\ &\leq 2\sigma\|P_{\widehat{\mathcal{M}}+\mathcal{M}_O}\mathbf{W}\|(\|P_{\widehat{\mathcal{M}}}\mathbf{X} - \mathbf{f}\| + \|\mathbf{f} - P_{\mathcal{M}_O}\mathbf{f}\|)\end{aligned}$$

and, using Lemma 2, with a probability greater than $1 - \frac{2}{K_N}e^{-u}$

$$\leq 2\sigma\left(\sqrt{\widehat{M} + M_O} + \sqrt{4\log(K_N)(\widehat{M} + M_O) + 2u}\right)$$

$$\times (\|P_{\widehat{\mathcal{M}}}\mathbf{X} - \mathbf{f}\| + \|\mathbf{f} - P_{\mathcal{M}_O}\mathbf{f}\|)$$

applying $2xy \leq \beta^{-2}x^2 + \beta^2 y^2$ successively with $\beta = \frac{1}{2}$ and $\beta = 1$ leads to

$$|2\langle \mathbf{X} - \mathbf{f}, P_{\widehat{\mathcal{M}}}\mathbf{X} - P_{\mathcal{M}_O}\mathbf{f}\rangle| \leq \left(\frac{1}{2}\right)^{-2} 2\sigma^2(\widehat{M} + M_O + 4\log(K_N)(\widehat{M} + M_O) + 2u) + \left(\frac{1}{2}\right)^2 2(\|P_{\widehat{\mathcal{M}}}\mathbf{X} - \mathbf{f}\|^2 + \|\mathbf{f} - P_{\mathcal{M}_O}\mathbf{f}\|^2) \quad .$$

Inserting this bound into

$$\|\mathbf{f} - P_{\widehat{\mathcal{M}}}\mathbf{X}\|^2 + \lambda^2 \log(K_N)\sigma^2\widehat{M} \leq \|\mathbf{f} - P_{\mathcal{M}_O}\mathbf{f}\|^2 + \lambda^2 \log(K_N)\sigma^2 M_O + |2\langle \mathbf{X} - \mathbf{f}, P_{\widehat{\mathcal{M}}}\mathbf{X} - P_{\mathcal{M}_O}\mathbf{f}\rangle|$$

yields

$$\frac{1}{2}\|\mathbf{f} - P_{\widehat{\mathcal{M}}}\mathbf{X}\|^2 \leq \frac{3}{2}\|\mathbf{f} - P_{\mathcal{M}_O}\mathbf{f}\|^2 + \sigma^2(\lambda^2 \log(K_N) + 8(1 + 4\log(K_N)))M_O + \sigma^2(8(1 + 4\log(K_N)) - \lambda^2 \log(K_N))\widehat{M} + 16\sigma^2 u$$

So that if $\lambda^2 \geq 32 + \frac{8}{\log(K_N)}$

$$\|\mathbf{f} - P_{\widehat{\mathcal{M}}}\mathbf{X}\|^2 \leq 3\|\mathbf{f} - P_{\mathcal{M}_O}\mathbf{f}\|^2 + 4\sigma^2\lambda^2 \log(K_N)M_O + 32\sigma^2 u$$

which implies

$$\|\mathbf{f} - P_{\widehat{\mathcal{M}}}\mathbf{X}\|^2 \leq 4(\|\mathbf{f} - P_{\mathcal{M}_O}\mathbf{f}\|^2 + \sigma^2\lambda^2 \log(K_N)M_O) + 32\sigma^2 u$$

where this result holds with probability greater than $1 - \frac{2}{K_N}e^{-u}$.

Recalling that this is valid for all $u \geq 0$, one has

$$\mathbb{P}\left\{\|\mathbf{f} - P_{\widehat{\mathcal{M}}}\mathbf{X}\|^2 - 4(\|\mathbf{f} - P_{\mathcal{M}_O}\mathbf{f}\|^2 + \sigma^2\lambda^2 \log(K_N)M_O) \geq 32\sigma^2 u\right\} \leq \frac{2}{K_N}e^{-u}$$

which implies by integration over u

$$E\left[\|\mathbf{f} - P_{\widehat{\mathcal{M}}}\mathbf{X}\|^2 - 4(\|\mathbf{f} - P_{\mathcal{M}_O}\mathbf{f}\|^2 + \sigma^2\lambda^2 \log(K_N)M_O)\right] \leq 32\sigma^2\frac{2}{K_N}$$

that is the bound of Theorem 1

$$E\left[\|\mathbf{f} - P_{\widehat{\mathcal{M}}}\mathbf{X}\|^2\right] \leq 4(\|\mathbf{f} - P_{\mathcal{M}_O}\mathbf{f}\|^2 + \sigma^2\lambda^2 \log(K_N)M_O) + 32\sigma^2\frac{2}{K_N}$$

□

References

[1] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Th. Rel. Fields*, 113:301–413, 1999.

[2] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. L. Yang, editors, *A Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1995.

[3] E. J. Candès. Modern statistical estimation via oracle inequalities. *Acta Numerica*, 2006.

[4] E. J. Candès and D. L. Donoho. A surprinsingly effective nonadaptive representation for objects with edges. *Curves and Surfaces*, 1999.

[5] R. R. Coifman and D. L. Donoho. Translation-invariant denoising. In A. Antoniadis and G. Oppenheim, editors, *Wavelet and Statistics*, Lecture Notes in Statistics. Springer Verlag, Berlin, 1995.

[6] R. R. Coifman and Y. Meyer. Remarques sur l'analyse de Fourier à fenêtre. *C. R. Acad. Sci. Paris Sér. I Math.*, 312(3):259–261, 1991.

[7] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718, 1992.

[8] D. L. Donoho. Cart and best-ortho-basis: A connection. *Ann. Statist.*, pages 1870–1911, 1997.

[9] D. L. Donoho. Wedgelets: Nearly-minimax estimation of edges. *Ann. Statist*, 27:353–382, 1999.

[10] D. L. Donoho and I. M. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of bases. *Comptes Rendus de l'Académie des Sciences*, Serie 1(319):1317–1322, 1994.

[11] E. D. Kolaczyk and R. D. Nowak. Multiscale likelihood analysis and complexity penalized estimation. *Annals of Statistics*, 32:500–527, 2004.

[12] A. P. Korostelev and A. B. Tsybakov. *Minimax Theory of Image Reconstruction*, volume 82. Spinger, 1993.

[13] E Le Pennec and S. Mallat. Sparse Geometrical Image Approximation with Bandelets. *IEEE Transaction on Image Processing*, 14(4):423–438, 2004.

[14] E. Le Pennec and S. Mallat. Bandlet image approximation and compression. *SIAM Multiscale Modeling and Simulation*, 4(3):992–1039, 2005.

[15] P. Massart. *Concentration Inequalities and Model Selection (Saint Flour Notes)*. Springer, 2003.

[16] G. Peyré, Ch. Dossal, E. Le Pennec, and S. Mallat. Geometric estimation with orthogonal bandlet bases. In *Proceedings of SPIE Wavelet XII*, Aug 2007.

[17] G. Peyré and S. Mallat. Orthogonal bandlets bases for geometric images approximation. *Journal of Pure and Applied Mathematics*, 2008.

[18] R. Shukla, P.L. Dragotti, M. N. Do, and M. Vetterli. Rate-distortion optimized tree structured compression algorithms for piecewise polynomial images. *IEEE Trans. on Image Processing*, 14(3):343–359, March 2005.

[19] B. S. Tsirelson, I. A. Ibragimov, and V. N. Sudakov. Norms of gaussian sample functions. In *Lecture Notes in Mathematics*, volume 550, pages 20–41. Springer, 1976.