Bandlet Image Estimation with Model Selection

Charles H Dossal, Erwan Le Pennec, Stéphane Mallat

Bandlet Image Estimation with Model Selection

Ch. Dossal[a,*], E. Le Pennec[b], S. Mallat[c]

[a] *IMB / Université de Bordeaux 1*
351, cours de la Libération
33405 Talence Cedex / France
[b] *LPMA / Université Paris Diderot*
175 rue du Chevaleret
75013 Paris / France
[c] *CMAP / Ecole Polytechnique*
91128 Palaiseau Cedex France

Abstract

To estimate geometrically regular images in the white noise model and obtain an adaptive near asymptotic minimaxity result, we consider a model selection based bandlet estimator. This bandlet estimator combines the best basis selection behaviour of the model selection and the approximation properties of the bandlet dictionary. We derive its near asymptotic minimaxity for geometrically regular images as an example of model selection with general dictionary of orthogonal bases. This paper is thus also a self contained tutorial on model selection with orthogonal bases dictionary.

Key words:
model selection, white noise model, image estimation, geometrically regular functions, bandlets
2000 MSC: 62G05

1. Introduction

A model selection based bandlet estimator has been introduced by Peyré et al. [20] 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. This paper shows that the risk of this estimator is nearly asymptotically minimax for geometrically regular images. It is also a tutorial on estimation with general dictionary of orthogonal bases, through model selection. It explains with details how to build a thresholding estimator in an adaptively chosen "best" basis and analyzes its performance with the model selection approach of Barron et al. [2].

Section 2 describes the statistical setting of the white noise model, and introduces the model of $\mathbf{C}^\alpha$ geometrically regular images. Images in this class, originally pro-

*Corresponding author

posed by Korostelev and Tsybakov [14], are, roughly, $\mathbf{C}^\alpha$ (Hölder regularity α) outside a set of $\mathbf{C}^\alpha$ curves in $[0,1]^2$. Korostelev and Tsybakov [14] prove that the minimax quadratic risk over this class, for a Gaussian white noise of variance σ^2, has an asymptotic decay of the order of $\sigma^{2\alpha/(\alpha+1)}$. They show that the risk of any possible estimator cannot decay faster than this rate uniformly for all functions of this class and exhibit an estimator that achieves this rate. Their estimator relies on the knowledge of the regularity exponent α and on an explicit detection of the contours, and is not stable relatively to any image blurring. Later, Donoho [10] 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.

The model selection based bandlet estimator, which can also be described as a thresholding estimator in a best bandlet basis, does not have this restriction. It does not rely on the detection of the precise localization of an edge but only of a looser local direction of regularity. Furthermore, these directions of regularity are not estimated directly but indirectly through a best orthogonal basis search algorithm which does not require to know the regularity parameter α. Section 3 gives a tutorial introduction of this type of estimators for arbitrary dictionary. This generic class of thresholding estimators in a best basis selected in a dictionary of orthonormal bases has been already studied by Donoho and Johnstone [11] and fit into the framework of Barron et al. [2], [3] and [18] This (self contained) section recalls the framework of these estimators and their theoretical performance. For the sake of completeness, a simplified proof of the main model selection result is given in Appendix.

Section 4 returns to the specific setting of image processing and applies the results of the previous section to geometric image estimation. The choice of the representation (the choice of the dictionary of orthogonal bases) becomes crucial and, after a short description of the bandlet bases, their use is justified. The paper is concluded with Theorem 3 which states the adaptive near asymptotic minimaxity of the selection model based bandlet estimator for geometrically regular images.

2. Image estimation

2.1. 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 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. This equation means that one is able to observe a Gaussian field X_g indexed by functions $g \in L^2$ of mean $E(X_g) = \langle f, g\rangle$ and covariance $E\left[X_g X_{g'}\right] = \langle g, g'\rangle$.

This model allows to consider asymptotics over σ of a discrete camera measurements process. Indeed, the measurement of a camera with N pixels can be modelled

as the measure of X_{ϕ_n} over a family ϕ_n of N impulse responses of the photo-sensors. Those measurements,

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

where W_g is a Gaussian field of zero mean and covariance $E\left[W_g W_{g'}\right] = \langle g, g' \rangle$, define a "projection" of our observation dX on the space V_N, spanned by the ϕ_n, that we denote $P_{V_N} X$. The white noise model allows to modify the resolution of the camera depending on the noise level. To simplify explanations, in the following we suppose that $\{\phi_n\}_{0 \leq n < N}$ is an orthogonal basis, with no loss of generality, and thus that

$$P_{V_N} X = \sum_{i=0}^{N} X_{\phi_n} \phi_n.$$

2.2. Minimax risk and geometrically regular images

We study the maximum risk of estimators for images f in a given class with respect to σ. 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 [14] in their seminal work on image estimation. It is used as a benchmark to estimate or approximate images having some kind of geometric regularity (Donoho [10], Shukla et al. [21],...). An extension of this model that incorporates a blurring kernel h has been proposed by Le Pennec and Mallat [16] 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 $\Lambda = [0,1]^2 - \{\mathscr{C}_\gamma\}_{1 \leq \gamma \leq G}$,*
- *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 $\mathscr{C}_\gamma$ are $\mathbf{C}^\alpha$ and do not intersect tangentially if $\alpha > 1$.*

2.3. Edge based estimation

Korostelev and Tsybakov [14] have built an estimator that is asymptotically minimax for geometrically regular functions f, as long as there 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

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

This rate of convergence achieves the asymptotic 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 asymptotic minimax risk. 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 [10] and Shukla et al. [21] 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 the straight edge segment in each dyadic square by minimizing a global function. This process leads to an asymptotically 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 [14] as well as Donoho [10] and [21] 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. We now consider estimators that do not have this restriction: they project the observation on adaptive subspaces in which blurred as well as sharp edges are well represented. They rely on two ingredients: the existence of bases in which geometrical images can be efficiently approximated and the existence of a mechanism to select, from the observation, a good basis and a good subset of coefficients onto which it suffices to project the observation to obtain a good estimator. We focus first on the second issue.

3. Projection Estimator and Model Selection

The projection estimators we study 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 [8]. As recalled by Candès [4], their risks have already been studied by Donoho and Johnstone [11] and are a special case of the general framework of model selection proposed by Birgé and Massart [3]. Note that Kolaczyk and Nowak [13] have studied a similar problem in a slightly different setting. We recall in this section the framework of model selection and state a selection model theorem (Theorem 1) that is the main statistical tool to prove the performance on the model selection based bandlet estimator. This section is intended as a self contained tutorial presentation of these best basis estimators and their resulting risk upper bounds and contains no new results. Nevertheless, a simple (novel) proof of the (simplified) main result is given in Appendix.

3.1. Approximation space V_N and further projection

The first step of our estimators is a projection in a finite dimension space V_N spanned by an orthonormal family $\{\phi_n\}_{0\leq n<N}$. The choice of the dimension N and of the space V_N depends on the noise level σ but should not depend on the function f to be estimated. Assume for now that V_N is fixed and thus that we observe $P_{V_N}X$. This observation can be decomposed into $P_{V_N}f + \sigma W_{V_N}$ where W_{V_N} is a finite dimensional white noise on V_N.

Our final estimator is a reprojection of this observation $P_{V_N}X$ onto a subspace $\mathscr{M} \subset V_N$ which may (and will) depend on the observation: the projection based estimator $P_{\mathscr{M}}P_{V_N}X = P_{\mathscr{M}}X$. The overall quadratic error can be decomposed in three terms:

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

The first term is a bias term corresponding to the first linear approximation error due to the projection on V_N, the second term is also a bias term which corresponds to the non linear approximation of $P_{V_N}f$ on $\mathscr{M}$ while the third term is a "variance" term corresponding to the contribution of the noise on $\mathscr{M}$.

The dimension N of V_N has to be chosen large enough so that with high probability, for reasonable $\mathscr{M}$, $\|f - P_{V_N}f\|^2 \leq \|P_{V_N}f - P_{\mathscr{M}}f\|^2 + \|P_{\mathscr{M}}W\|^2$. From the practical point of view, this means that the acquisition device resolution is set so that the first linear approximation error due to discretization is smaller than the second non linear noise related error. Engineers often set N so that both terms are of the same order of magnitude, to limit the cost in terms of storage and computations. In our white noise setting, we will explain how to chose N depending on σ.

For a fixed V_N, in order to obtain a small error, we need to balance between the two remaining terms. A space $\mathscr{M}$ of large dimension may reduce the second bias term but will increase the variance term, a space $\mathscr{M}$ of small dimension does the opposite. It is thus necessary to find a trade-off between these two trends, and select a space $\mathscr{M}$ to minimize the sum of those two terms.

3.2. Model Selection in a Dictionary of orthonormal bases

We consider a (not that) specific situation in which the space $\mathscr{M}$ is spanned by some vectors from some orthonormal bases of V_N. More precisely, let $\mathscr{B} = \{g_n\}_{0\leq n<N}$ be an orthonormal basis of V_N, that may be different from $\{\phi_n\}$, we consider space $\mathscr{M}$ spanned by a sub-family $\{g_{n_k}\}_{1\leq k\leq M}$ of M vectors and the projections of our observation on those spaces

$$P_{\mathscr{M}}X = \sum_{k=1}^{M} X_{g_{n_k}}\, g_{n_k}.$$

Note that this projection, or more precisely its decomposition in the basis $\{\phi_n\}$, can be computed easily from the decomposition of $P_{\mathscr{M}}X$ in the same basis.

Instead of choosing a specific single orthonormal basis $\mathscr{B}$, we define a dictionary $\mathscr{D}_N$ which is a collection of orthonormal bases in which we choose adaptively the basis used. Note that some bases of $\mathscr{D}_N$ may have vectors in common. This dictionary can thus also be viewed as set $\{g_n\}$ of $K_N \geq N$ different vectors, that are regrouped to form many different orthonormal bases. Any collection of M vectors from the same

orthogonal basis $\mathcal{B} \in \mathcal{D}_N$ generates a space $\mathcal{M}$ that defines a possible estimator $P_{\mathcal{M}}X$ of f. Let $\mathcal{C}_N = \{\mathcal{M}_\gamma\}_{\Gamma_N}$ be the family of all such projection spaces. Ideally we would like to find the space $\mathcal{M} \in \mathcal{C}_N$ which minimizes $\|f - P_{\mathcal{M}}X\|$. We want thus to choose a "best" model $\mathcal{M}$ amongst a collection that is we want to perform a model selection task.

3.3. Oracle Model

As a projection estimator yields an estimation error

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

the expected error of such an estimator is given by

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

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

This expected error cannot be computed in practise since we have a single realization of dX (or of $P_{V_N}X$) . To (re)derive the classical model selection procedure of Birgé and Massart [3], we first slightly modify our problem by searching for a subspace $\mathcal{M}$ such that the estimation error obtained by projecting $P_{V_N}X$ on this subspace is small with only an overwhelming probability. As in all model selection papers, we use an upper bound of the estimation error obtained from an upper bound of the energy of the noise projected on $\mathcal{M}$. Each of the K_N projections of the noise on the K_N different vectors in the bases of the dictionary $\mathcal{D}_N$ is thus $W_{g_k}g_k$. Its law is a Gaussian random variable of variance σ^2 along the vector g_k. A standard large deviation result proves that the norms of 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 $\dim(\mathcal{M})$ squared dictionary noise coefficients, we get $\|P_{\mathcal{M}}W\|^2 \leq \dim(\mathcal{M})\, T^2$. It results that

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

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 $\dim(\mathcal{M})$ which yields a small approximation error $\|f - P_{\mathcal{M}}f\|$. We denote by $\mathcal{M}_O \in \mathcal{C}_N$ the space that minimizes the estimation error upper bound (1)

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

Note that this optimal space cannot be determined from the observation X since f is unknown. It is called the oracle space , hence the O in the notation, to remind this fact.

3.4. Penalized empirical error

To obtain an estimator, it is thus necessary to replace this oracle space by a "best" space obtained only from the observation $P_{V_N}X$ that yields (hopefully) a small estimation error. A first step toward this goal is to notice that since all the spaces $\mathcal{M}$ are included into V_N, minimizing

$$\|f - P_{\mathcal{M}} f\|^2 + \dim(\mathcal{M})\, T^2$$

is equivalent to minimizing

$$\|P_{V_N} f - P_{\mathcal{M}} f\|^2 + \dim(\mathcal{M})\, T^2$$

. A second step is to consider the crude estimation of $\|P_{V_N} f - P_{\mathcal{M}} f\|^2$ given by the empirical norm

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

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

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

whose expected value is $(N - \dim(\mathcal{M}))\sigma^2$, with typically $\dim(\mathcal{M}) \ll N$. 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 $\dim(\mathcal{M})$. 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:

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M} \in \mathcal{C}_N} \|P_{V_N}X - P_{\mathcal{M}}X\|^2 + \dim(\mathcal{M})\, T^2 \tag{2}$$

3.5. 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 a subset of vectors from a basis $\mathcal{B} \in \mathcal{D}_N$. The following (classical) 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 generated by a subset of vectors of an orthonormal basis $\mathcal{B} = \{g_n\}_{0 \le n < N}$ of V_N, the estimator which minimizes $\|P_{V_N}X - P_{\mathcal{M}}X\|^2 + \dim(\mathcal{M})\, T^2$ is the thresholding estimator*

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

Proof. Let $\mathscr{M} = \text{Span}\{g_n\}_{n\in I}$ with $I \subset [0,N)$, as $\mathscr{B}$ is an orthonormal basis,

$$\|X - P_{\mathscr{M}}X\|^2 + \dim(\mathscr{M})\, T^2 = \sum_{n\notin I} |\langle X, g_m\rangle|^2 + \sum_{n\in I} T^2$$

which is minimal if $I = \{n, |\langle X, g_n\rangle|^2 > T^2\}$. □

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

$$\widehat{\mathscr{B}} = \arg\min_{\mathscr{B}\in\mathscr{D}_N} \|P_{V_N}X - P_{\mathscr{M}_{\mathscr{B},X,T}}X\|^2 + \dim(\mathscr{M})\, T^2.$$

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

The following theorem, similar to the one obtained first by Barron et al. [2], proves that the thresholding estimation error in the best basis is bounded by the estimation error by projecting in the oracle space $\mathscr{M}_O$, up to a multiplicative factor.

Theorem 1. *There exists an absolute function $\lambda_0(K) \geq \sqrt{2}$ and some absolute constants $\varepsilon > 0$ and $\kappa > 0$ such that if we denote $\mathscr{C}_N = \{\mathscr{M}_\gamma\}_\Gamma$ the family of projection spaces generated by some vectors in an orthogonal basis of a dictionary $\mathscr{D}_N$ and denote K_N be the number of different vectors in $\mathscr{D}_N$. Then for any $\sigma > 0$, if we let $T = \lambda\sqrt{\log(K_N)}\,\sigma$ with $\lambda \geq \lambda_0(K_N)$, then for any $f \in L^2$, the thresholding estimator $F = P_{\mathscr{M}_{\widehat{\mathscr{B}},X,T}}X$ in the best basis*

$$\widehat{\mathscr{B}} = \arg\min_{\mathscr{B}\in\mathscr{D}_N} \|P_{V_N}X - P_{\mathscr{M}_{\mathscr{B},X,T}}X\|^2 + \dim\left(\mathscr{M}_{\mathscr{B},X,T}\right) T^2$$

satisfies

$$E\left[\|f - F\|^2\right] \leq (1+\varepsilon)\left(\min_{\mathscr{M}\in\mathscr{C}_N} \|f - P_{\mathscr{M}}f\|^2 + \dim(\mathscr{M})\, T^2\right) + \frac{\kappa}{K_N}\sigma^2.$$

For the sake of completion, we propose in Appendix a simple proof of Theorem 1, inspired by Birgé and Massart [3], which requires only a concentration lemma for the norm of the noise in all the subspaces spanned by the K_N generators of $\mathscr{D}_N$ but with worse constants: $\lambda_0(K) = \sqrt{32 + \frac{8}{\log(K)}}$, $\varepsilon = 3$ and $\kappa = 64$. Note this Theorem can be deduced from Massart [18] with different (better) constant (and for roughly $\lambda_0(K) > \sqrt{2}$) using a more complex proof based on subtle Talagrand's inequalities. It results that any bound on $\min_{\mathscr{M}\in\mathscr{C}_N} \|f - P_{\mathscr{M}}f\|^2 + \dim(\mathscr{M})\, T^2$, gives a bound on the risk of the best basis estimator F.

To obtain a computational estimator, the minimization

$$\widehat{\mathscr{B}} = \arg\min_{\mathscr{B}\in\mathscr{D}_N} \|P_{V_N}X - P_{\mathscr{M}_{\mathscr{B},X,T}}X\|^2 + \dim\left(\mathscr{M}_{\mathscr{B},X,T}\right) T^2 \quad ,$$

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 [8] with wavelet packets or by Coifman and Meyer [7] with local cosine bases for signals having localized time-frequency structures. Next section reviews some possible dictionaries for images and recall the construction of the dictionary of bandlet orthogonal bases that is adapted to the estimation of geometrically regular images.

4. Best basis image estimation and bandlets

4.1. Thresholding in a single basis

When the dictionary $\mathscr{D}_N$ is reduced to a single basis $\mathscr{B}$, and there is thus no basis choice, Theorem 1 clearly applies and reduces to the classical thresholding Theorem of Donoho and Johnstone [12]. The corresponding estimator is thus the classical thresholding estimator which quadratic risk satisfies

$$E\left[\|f - P_{\mathscr{M}_{\mathscr{B},X,T}} X\|^2\right] \leq (1+\varepsilon)\left(\min_{\mathscr{M}\in\mathscr{C}_N} \|f - P_{\mathscr{M}} f\|^2 + \dim(\mathscr{M})\, T^2\right) + \frac{\kappa}{N}\sigma^2$$

It remains "only" to choose which basis to use and how to define the space 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, each space

$$V_j = \text{Span}\{\phi_{j,k_1}(x_1)\phi_{j,k_2}(x_2)\}_{k_1,k_2},$$

called approximation space of scale 2^j, admits $\{\psi^o_{l,k}\}_{o,l\geq j,k_1,k_2}$ as an orthogonal basis. The approximation space V_N of the previous section coincides with the classical wavelet approximation space V_j when $N = 2^{-j/2}$.

A classical approximation result ensures that for any function f $\mathbf{C}^\alpha$, as soon as the wavelet has more than $\lfloor\alpha\rfloor + 1$ vanishing moments, there is a constant C such

that, for any T, $\min_{\mathcal{M}\in\mathscr{C}_N} \|P_{V_N} f - P_{\mathcal{M}} f\|^2 + \dim(\mathcal{M})\, T^2 \leq C(T^2)^{\frac{\alpha}{\alpha+1}}$, 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 remain large. As explained in Mallat [17], by noticing that those edges contribute at scale 2^j to $O(2^{-j})$ coefficients of order $O(2^{j/2})$, one verifies that the rate of convergence in a wavelet basis decays like $(|\log(\sigma)|\sigma^2)^{1/2}$, which is far from the asymptotical minimax rate.

A remarkably efficient representation was introduced by Candès and Donoho [5]. 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 $\mathscr{C} = \{c_n\}_n$ is not a basis of $L^2([0,1]^2)$ but a tight frame of $L^2(\mathbb{R}^2)$. This implies, nevertheless, that for any $f \in L^2([0,1]^2)$

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

Although this is not an orthonormal basis, the results of Section 3 can be extended to this setting. Projecting the data on the first $N = \sigma^{-1/2}$ curvelets with significant intersection with the unit square 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.

4.2. 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[15, 16, 19] 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 neighbourhood of an edge, these wavelets coefficients are regular along the direction of the edge as illustrated by Fig 1.

To capture this geometric regularity, the key tool is a local orthogonal transform, inspired by the work of Alpert [1], that combines locally the wavelets along the direction of regularity, represented by arrows in the rightmost image of Fig 1), to produce a new orthogonal basis, a bandlet basis. By construction, the bandlets are elongated along the direction of regularity and have the vanishing moments along this direction. The

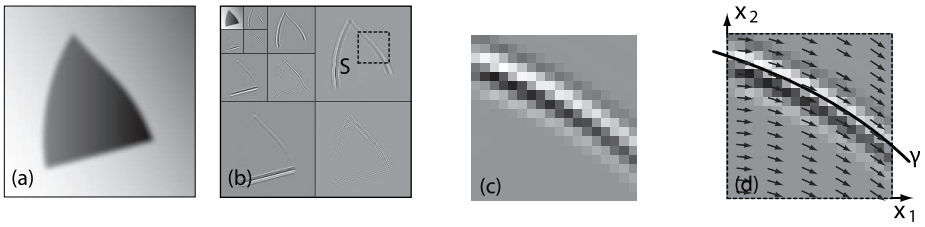

Figure 1: a) a geometrically regular image, b) the associated wavelet coefficients, c) a close-up of wavelet coefficients in a detail space W_j^o that shows their remaining regularity, d) the geometrical flow adapted to this square of coefficients, here it is vertically constant and parametrized by a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

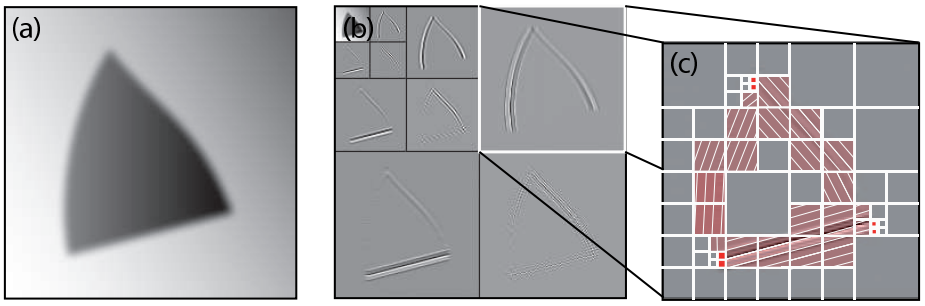

Figure 2: a) a geometrically regular image b) the corresponding wavelet coefficients c) the quadtree associated to the segmentation of a detail space W_j^o. In each square where the image is not uniformly regular, the flow is shown.

(possibly large) wavelets coefficients are thus locally recombined along this direction, yielding more coefficients of small amplitudes than before.

More precisely, the construction of a bandlet 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 $(2^{-l})^2$. Its coefficients are recombined using the Alpert transform induced by some directions of regularity. This geometry is specified by a local geometric flow, a vector field meant to follow the geometric direction of regularity. This geometric flow is further constraint to have a specific structure as illustrated in Fig. 2, It is structured by a partition into dyadic squares in which the flow, if there exists, is vertically or horizontally constant. In each square of the partition, the flow being thus easily parametrized by its tangent.

For each choice of geometric flow, a specific orthogonalization process [19] yields an orthogonal basis of bandlets that have vanishing moments along 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 is the one of the best basis.

The first step is to define a dictionary $\mathscr{D}_{(2^{-j})^2}$ of orthogonal bandlet bases of V_j or equivalently a dictionary of possible geometric flows. Obviously this dictionary should be finite and this require a discretization of the geometry. As proved by Peyré and Mallat [19], this is not an issue: the flow does not have to follow exactly the direction of regularity but only up to a sufficient known precision. It is indeed sufficient to parametrize the flow in any dyadic square by the tangent of a polynomial of degree p (the number of vanishing moments of the wavelets). The coefficients of this polynomial can be further quantized. The resulting family of geometric flow in a square is of size $O(2^{-jp})$.

A basis of the dictionary $\mathscr{D}_{(2^{-j})^2}$ is thus 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 $O(2^{-jp})$ polynomials. The number of bases in the dictionary $\mathscr{D}_{(2^{-j})^2}$ grows exponentially with 2^{-j}, but the total number of different bandlets $K_{(2^{-j})^2}$ grows only polynomially like $O(2^{-j(p+4)})$. 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 $O(2^{-jp)})$ 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, we obtain the existence of a constant C_K such that $2^{-j(p+4)} \leq K_{(2^{-j})^2} \leq C_K 2^{-j(p+4)}$.

4.3. Approximation in bandlet dictionaries

The key property of the bandlet basis dictionary is that it provides an asymptotically optimal representation of C^α geometrically regular functions. Indeed Peyré and Mallat

[19] prove

Theorem 2. *Let $\alpha < p$ where p in the number of wavelet vanishing moments, for any f $\mathbf{C}^\alpha$ geometrically regular function, there exists a real number C such that for any $T > 0$ and $2^j \leq T$*

$$\min_{\mathscr{B} \in \mathscr{D}_{(2^{-j})^2}} \|f - P_{\mathscr{M}_{\mathscr{B},f,T}} f\|^2 + \dim\left(\mathscr{M}_{\mathscr{B},f,T}\right) T^2 \leqslant C T^{2\alpha/(\alpha+1)} \tag{4}$$

where the subspace $\mathscr{M}_{\mathscr{B},f,T}$ is the space spanned by the vectors of $\mathscr{B}$ whose inner product with f is larger than T.

This Theorem gives the kind of control we require in Theorem 1.

Being able to perform efficiently the minimization of the previous Theorem is very important to exploit numerically this property. It turns out that a fast algorithm can be used to find the best basis that minimizes $\|f - P_{\mathscr{M}_{\mathscr{B},f,T}} f\|^2 + \dim\left(\mathscr{M}_{\mathscr{B},f,T}\right) T^2$ or equivalently $\|P_{V_j} f - P_{\mathscr{M}_{\mathscr{B},f,T}} f\|^2 + \dim\left(\mathscr{M}_{\mathscr{B},f,T}\right) T^2$. We use first the additive structure with respect to the subband W_l^o of this "cost" $\|P_{V_j} f - P_{\mathscr{M}_{\mathscr{B},f,T}} f\|^2 + \dim\left(\mathscr{M}_{\mathscr{B},f,T}\right) T^2$ to split the minimization into several independent minimizations on each subbands. A bottom-top fast optimization of the geometry (partition and flow) similar to the one proposed by Coifman and Wickerhauser [8], and Donoho [9] can be performed on each subband thanks to two observations. Firstly, for a given dyadic square, the limited number of possible flows is such that the best flow can be obtained with a simple brute force exploration. Secondly, the hierarchical tree structure of the partition and the additivity of the cost function with respect to the partition implies that the best partition of a given dyadic square is either itself or the union of the best partitions of its four dyadic subsquares. This leads to a bottom up optimization algorithm once the best flow has been found for every dyadic squares. Note that this algorithm is adaptive with respect to α: it does not require the knowledge of the regularity parameter to be performed.

More precisely, the optimization algorithm goes as follows. The brute force search of the best flow is conducted independently over all dyadic squares and all detail spaces with a total complexity of order $O(2^{-j(p+4)})$. This 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 cost computed so far. Remark that the initialization 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 $O(2^{-j(p+4)})$, which nevertheless remains polynomial in 2^{-j}.

4.4. 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 proposed by Peyré et al.

[20] do not rely on such a detection process. The chosen geometry is obtained with the best basis selection of the previous section. This allows one 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 selection model based bandlet estimator is near asymptotically minimax for $\mathbf{C}^\alpha$ geometrically regular images.

For a given noise level σ, one has to select a dimension $N = (2^{-j})^2$ and a threshold T. The best basis algorithm selects then the bandlet basis $\widehat{\mathcal{B}}$ amongst $\mathcal{D}_N = \mathcal{D}_{(2^{-j})^2}$ that minimizes

$$\|P_{V_N}X - P_{\mathcal{M}_{\mathcal{B},X,T}}X\|^2 + T^2 \dim\left(\mathcal{M}_{\mathcal{B},X,T}\right)$$

and the model selection based estimate is $F = P_{\mathcal{M}_{\widehat{\mathcal{B}},X,T}}X$. We should now specify the choice of $N = (2^{-j})^2$ and T in order to be able to use Theorem 1 and Theorem 2 to obtain the near asymptotic minimaxity of the estimator. On the one hand, the dimension N should be chosen large enough so that the unknown linear approximation error $\|f - P_{V_N}\|^2$ is small. One the other hand, the dimension N should not be too large so that the total number of bandlets K_N, which satisfies $\sqrt{N}^{(p+4)} \leq K_N \leq C_K\sqrt{N}^{(p+4)}$, imposing a lower bound on the value of the threshold remains small. For the sake of simplicity, as we consider an asymptotic behaviour, we assume that σ is smaller than $1/4$. 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 $T = \widetilde{\lambda}\sqrt{|\log\sigma|}\sigma$ with $\widetilde{\lambda}$ large enough yields a nearly asymptotically minimax estimator.

Theorem 3. *Let $\alpha < p$ where p in the number of wavelet vanishing moments and let $K_0 \in \mathbb{N}^*$ and $\widetilde{\lambda} \geq \sqrt{2(p+4)}\sup_{K\geq K_0}\lambda_0(K)$. For any $\mathbf{C}^\alpha$ geometrically regular function f, there exists $C > 0$ such that for any*

$$\sigma \leq \min(\frac{1}{4}, \max(C_K, K_0/2)^{-1/(p+4)}),$$

if we let $N = 2^{-2j}$ with j such that $\sigma \in (2^{j-1}, 2^j]$ and $T = \widetilde{\lambda}\sqrt{|\log\sigma|}\sigma$, the estimator $F = P_{\mathcal{M}_{\widehat{\mathcal{B}},X,T}}X$ obtained by thresholding $P_{V_N}X$ with a threshold T in the basis $\widehat{\mathcal{B}}$ of $\mathcal{D}_N$ that minimizes

$$\|P_{V_N}X - P_{\mathcal{M}_{\mathcal{B},X,T}}X\|^2 + T^2 \dim\left(\mathcal{M}_{\mathcal{B},X,T}\right)$$

satisfies

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

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

Proof. For any $\sigma \in (2^{j-1}, 2^j]$, observe that $2^{-j(p+4)} \leq K_N = K_{(2^{-j})^2} \leq C_K 2^{-j(p+4)}$ and thus $(2\sigma)^{-(p+4)} \leq K_N \leq C_K\sigma^{-(p+4)}$. The restriction on σ further implies then that $K_N \geq K_0$ and $K_N \leq \sigma^{-2(p+4)}$. As $\widetilde{\lambda} \geq \sqrt{2(p+4)}\sup_{K\geq K_0}\lambda_0(K)$, $T = \widetilde{\lambda}\sqrt{|\log\sigma|}\sigma \geq \lambda\sqrt{\log(K_N)}\sigma$ with $\lambda \geq \lambda_0(K_N)$ so that Theorem 1 applies. This yields

$$E\left[\|f - F\|^2\right] \leq (1+\varepsilon)\min_{\mathcal{M}\in\mathcal{C}_N}\left(\|f - P_{\mathcal{M}}f\|^2 + T^2\dim(\mathcal{M})\right) + \frac{\kappa}{K_N}\sigma^2 \quad . \qquad (5)$$

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

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

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

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

Although Theorem 3 applies only to $\mathbf{C}^\alpha$ geometrically regular function, one can use the bandlet estimator for any type of images. Figure 3 illustrates the good behaviour of the bandlet estimator for natural images already shown in [20]. Each line presents the original image, the degraded noisy image and two estimations, one using classical translation invariant estimator [6]. and the other using the bandlet estimator. The bandlet improvement with respect to the classical wavelet estimator 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 g_k, of $\mathcal{D}_N$.

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

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

where $\dim(\mathcal{M}_I)$ is the dimension of $\mathcal{M}_I$.

Proof. The key ingredient of this proof is a concentration inequality. Tsirelson's Lemma[22] implies that for any 1-Lipschitz function $\phi : \mathbb{C}^n \to \mathbb{C}$ ($|\phi(x) - \phi(y)| \leq \|x - y\|$) if W is a Gaussian standard white noise in $\mathbb{C}^n$ then

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

For any space $\mathcal{M}$, $f \mapsto \|P_{\mathcal{M}} f\|$ is 1-Lipschitz. Note that one can first project f into the finite dimensional space V_N without modifying the norm. We can thus apply Tsirelson's Lemma with $t = \sqrt{4 \log(K_N) \dim(\mathcal{M}) + 2u}$ and obtain

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

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

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

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

Now

$$\begin{aligned}
\mathbb{P}\left\{\exists I \subset \{1,\ldots,K_N\}, \|P_{\mathscr{M}_I}W\| \geq \sqrt{\dim(\mathscr{M}_I)} + \sqrt{4\log(K_N)\dim(\mathscr{M}_I) + 2u}\right\} & \\
&\leq \sum_{I\subset\{1,\ldots,K_N\}} \mathbb{P}\left\{\|P_{\mathscr{M}_I}W\| \geq \sqrt{\dim(\mathscr{M}_I)} + \sqrt{4\log(K_N)\dim(\mathscr{M}_I) + 2u}\right\} \\
&\leq \sum_{I\subset\{1,\ldots,K_N\}} K_N^{-2\dim(\mathscr{M}_I)}e^{-u} \\
&\leq \sum_{d=1}^{K_N} \binom{K_N}{d} K_N^{-2d}e^{-u} \leq \sum_{d=1}^{K_N} K_N^{-d}e^{-u} \\
&\leq \frac{K_N^{-1}}{1-K_N^{-1}}e^{-u}
\end{aligned}$$

and thus

$$\begin{aligned}
\mathbb{P}\left\{\exists I \subset \{1,\ldots,K_N\}, \|P_{\mathscr{M}_I}W\| \geq \sqrt{\dim(\mathscr{M}_I)} + \sqrt{4\log(K_N)\dim(\mathscr{M}_I) + 2u}\right\} & \\
&\leq \frac{2}{K_N}e^{-u}
\end{aligned}$$

□

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 $P_{V_N}X = P_{V_N}f + \sigma P_{V_N}W \in V_N$ with $P_{V_N}W$ a Gaussian white noise. By construction, the thresholding estimate is $P_{\mathscr{M}_{\widehat{\mathscr{B}},X,T}}X$ where

$$\widehat{\mathscr{B}} = \arg\min_{\mathscr{B}\in\mathscr{D}_N} \|P_{V_N}X - P_{\mathscr{M}_{\mathscr{B},X,T}}X\|^2 + \dim\left(\mathscr{M}_{\mathscr{B},X,T}\right) T^2 \quad .$$

To simplify the notation, we denote by $\widehat{\mathscr{M}}$ and $\dim\left(\widehat{\mathscr{M}}\right)$ the corresponding space and its dimension.

Denote now $\dim(\mathscr{M}_0)$ the dimension of the oracle subspace $\mathscr{M}_O$ that has been defined as the minimizer of

$$\|P_{V_N}f - P_{\mathscr{M}}f\|^2 + \dim(\mathscr{M})\, T^2 \quad .$$

By construction,

$$\|P_{V_N}X - P_{\widehat{\mathscr{M}}}X\|^2 + \lambda^2\log(K_N)\sigma^2\dim\left(\widehat{\mathscr{M}}\right) \leq \|P_{V_N}X - P_{\mathscr{M}_O}f\|^2 + \lambda^2\log(K_N)\sigma^2\dim(\mathscr{M}_0).$$

Using

$$\|P_{V_N}X - P_{\widehat{\mathscr{M}}}X\|^2 = \|P_{V_N}X - P_{V_N}f\|^2 + \|P_{V_N}f - P_{\widehat{\mathscr{M}}}X\|^2 + 2\langle P_{V_N}X - P_{V_N}f, P_{V_N}f - P_{\widehat{\mathscr{M}}}X\rangle$$

and a similar equality for $\|P_{V_N}X - P_{\mathscr{M}_O}f\|^2$, one obtains

$$\|P_{V_N}f - P_{\widehat{\mathscr{M}}}X\|^2 + \lambda^2\log(K_N)\sigma^2\dim\left(\widehat{\mathscr{M}}\right) \leq \|P_{V_N}f - P_{\mathscr{M}_O}f\|^2 + \lambda^2\log(K_N)\sigma^2\dim(\mathscr{M}_0) + 2\langle P_{V_N}X - P_{V_N}f, P_{\widehat{\mathscr{M}}}X - P_{\mathscr{M}_O}f\rangle$$

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

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

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

$$\begin{aligned}&\leq 2\sigma\left(\sqrt{\dim\left(\widehat{\mathscr{M}}\right) + \dim(\mathscr{M}_0)} + \sqrt{4\log(K_N)(\dim\left(\widehat{\mathscr{M}}\right) + \dim(\mathscr{M}_0)) + 2u}\right) \\ &\quad\times(\|P_{\widehat{\mathscr{M}}}X - P_{V_N}f\| + \|P_{V_N}f - P_{\mathscr{M}_O}f\|)\end{aligned}$$

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

$$\begin{aligned}&|2\langle P_{V_N}X - P_{V_N}f, P_{\widehat{\mathscr{M}}}X - P_{\mathscr{M}_O}f\rangle| \\ &\quad\leq \left(\frac{1}{2}\right)^{-2}2\sigma^2(\dim\left(\widehat{\mathscr{M}}\right) + \dim(\mathscr{M}_0) + 4\log(K_N)(\dim\left(\widehat{\mathscr{M}}\right) + \dim(\mathscr{M}_0)) + 2u) \\ &\qquad + \left(\frac{1}{2}\right)^{2}2(\|P_{\widehat{\mathscr{M}}}X - P_{V_N}f\|^2 + \|P_{V_N}f - P_{\mathscr{M}_O}f\|^2) \quad .\end{aligned}$$

Inserting this bound into

$$\|P_{V_N}f - P_{\widehat{\mathscr{M}}}X\|^2 + \lambda^2\log(K_N)\sigma^2\dim\left(\widehat{\mathscr{M}}\right) \leq \|P_{V_N}f - P_{\mathscr{M}_O}f\|^2 + \lambda^2\log(K_N)\sigma^2\dim(\mathscr{M}_0) + |2\langle P_{V_N}X - P_{V_N}f, P_{\widehat{\mathscr{M}}}X - P_{\mathscr{M}_O}f\rangle|$$

yields

$$\begin{aligned}\frac{1}{2}\|P_{V_N}f - P_{\widehat{\mathscr{M}}}X\|^2 \leq \frac{3}{2}\|P_{V_N}f - P_{\mathscr{M}_O}f\|^2 &+ \sigma^2(\lambda^2\log(K_N) + 8(1+4\log(K_N)))\dim(\mathscr{M}_0) \\ &+ \sigma^2(8(1+4\log(K_N)) - \lambda^2\log(K_N))\dim\left(\widehat{\mathscr{M}}\right) + 16\sigma^2u\end{aligned}$$

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

$$\|P_{V_N}f - P_{\widehat{\mathscr{M}}}X\|^2 \leq 3\|P_{V_N}f - P_{\mathscr{M}_O}f\|^2 + 4\sigma^2\lambda^2\log(K_N)\dim(\mathscr{M}_0) + 32\sigma^2u$$

which implies

$$\|P_{V_N}f - P_{\widehat{\mathscr{M}}}X\|^2 \leq 4(\|P_{V_N}f - P_{\mathscr{M}_O}f\|^2 + \sigma^2\lambda^2\log(K_N)\dim(\mathscr{M}_0)) + 32\sigma^2u$$

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

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

$$\mathbb{P}\left\{\|P_{V_N}f - P_{\widehat{\mathscr{M}}}X\|^2 - 4(\|P_{V_N}f - P_{\mathscr{M}_O}f\|^2 + \sigma^2\lambda^2\log(K_N)\dim(\mathscr{M}_0)) \geq 32\sigma^2 u\right\} \leq \frac{2}{K_N}e^{-u}$$

which implies by integration over u

$$E\left[\|P_{V_N}f - P_{\widehat{\mathscr{M}}}X\|^2 - 4(\|P_{V_N}f - P_{\mathscr{M}_O}f\|^2 + \sigma^2\lambda^2\log(K_N)\dim(\mathscr{M}_0))\right] \leq 32\sigma^2\frac{2}{K_N}$$

that is the bound of Theorem 1

$$E\left[\|P_{V_N}f - P_{\widehat{\mathscr{M}}}X\|^2\right] \leq 4(\|P_{V_N}f - P_{\mathscr{M}_O}f\|^2 + \sigma^2\lambda^2\log(K_N)\dim(\mathscr{M}_0)) + 32\sigma^2\frac{2}{K_N}$$

up to $\|f - P_{V_N}f\|^2$ which can be added on both size of the inequality. □

Bibliography

References

[1] B. K. Alpert. *Wavelets and Other Bases for Fast Numerical Linear Algebra*, pages 181–216. C. K. Chui, editor, Academic Press, San Diego, CA, USA, 1992.

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

[3] 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.

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

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

[6] 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.

[7] 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. ISSN 0764-4442.

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

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

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

[11] 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.

[12] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, Dec 1994.

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

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

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

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

[17] S. Mallat. *A Wavelet Tour of Signal Processing, 3rd ed., Third Edition: The Sparse Way*. Academic Press, 3rd edition, 2008.

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

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

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

[21] 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.

[22] 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.