MAXISETS FOR MODEL SELECTION ESTIMATORS

FLORENT AUTIN (1), JEAN-MICHEL LOUBES (2) AND VINCENT RIVOIRARD (2)

Abstract. The aim of this paper is to analyze the performances of penalized estimation methods. For this, we propose to describe the maximal sets where these methods attain a special rate of convergence. We deal both with regularity-type penalty, leading to penalized splines estimators, an penalty over the dimension, leading to so-called model selection estimators.

1. Introduction

In nonparametric estimation, there exist several different ways of constructing estimators of a regression function: kernel methods, thresholded estimators, projection estimators or model selection estimators. Choosing a method rather than an other, if not a question of belief, implies comparing the asymptotic performance of the different estimators. But the choice of such a criterion that enables a clear comparison is not clear.

A classical setting to compare procedures is the minimax point of view. Given a class of functional spaces, we compare two procedures by comparing the maximal rate achieved by these procedures on each member of this class. And to check that a procedure is optimal from the minimax point of view (said to be minimax), we establish that it achieves the best rate achieved by any procedure on each space. This minimax approach is widely used and many methods cited above are proved to be minimax in different statistical frameworks. However, this minimax approach has undoubtedly two drawbacks: the choice for the function class is quite subjective and providing an estimator well adapted to the worst functions of this class seems too pessimistic for practical purposes. More problematical in practice, several minimax procedures are proposed and the practitioner has no way to decide his experiment. To answer these problems, a new setting has been proposed: the maxiset point of view. It consists in deciding the accuracy of the procedure by fixing a target rate α_n and to point out all the functions that can be estimated at this rate. Of course the larger the maxiset, the better the procedure. The set of these functions is called the *maxiset* of the procedure. The maxiset point of view brings answers to the previous questions. There is no a priori functional assumption and we do not need to restrict our attention to the study on an arbitrary functional space. The practitioner states the desired accuracy and then knows the quality of the used procedure. Obviously, he chooses the procedure with the largest maxiset. For instance, in the white noise setting, the maxiset theory has been investigated in [Kerkyacharian and Picard, 2002], [Cohen et al., 2001] for kernel and thresholding estimates and in [Rivoirard, 2003] and [Rivoirard, 2004] for Bayesian and linear estimators. [Autin, 2005] investigated maxisets in the density model. For a large review of maxiset results, we refer the reader to [Autin and Rivoirard, 2005]. For instance, it has been established that the maxisets of linear methods are in fact Besov spaces, whereas the maxisets of thresholding

estimates are Lorentz spaces reflecting extremely well the practical observation that wavelet thresholding performs well when the number of wavelet coefficients is small. It has also been observed in [Kerkyacharian and Picard, 2002] that there is a deep connection between oracle inequalities and maxisets.

Our aim is here to determine the maxisets associated with model selection procedure. To estimate an unknown function s observed with observation errors, we choose to approximate the data by a function in a chosen subspace called model. For this, we define an empirical criterion $\gamma_n(.)$, depending on the observations, which will determine the quality of the approximation and minimize this criterion over the model, which will determine the asymptotic behavior of the estimator. The heuristics of this estimation strategy is that, for a large number of data, the empirical criterion $\gamma_n(t)$ behaves like a pseudo-distance between the true unknown function and the candidate t. Hence the main problem that arises in M-estimation is the choice of a proper model on which the minimum contrast estimator is to be defined. On the one hand, the model has to be close to the true function in order to guarantee a small bias errors, but on the other hand, choosing the model as large as possible, increases the variance error, leading to non consistent estimators. Hence we introduce a collection of models S_m, $m \in \mathcal{M}_n$ and consider penalized estimators as follows:

$$\hat{s} = \arg \min_{m \in \mathcal{M}_n,\, t \in S_m} \left(\gamma_n(t) + \mathrm{pen}(m,t)\right). \tag{1.1}$$

The penalty aims at restricting the choice of models by imposing either a constraint over the dimension of the model, or over the smoothness of the solution. In the first case, $\mathrm{pen}(m,t) = \mathrm{pen}(m)$ depends on the dimension of the model S_m. This case corresponds to the standard model selection procedure widely described in the literature, see for instance [Barron et al., 1999], [Birgé and Massart, 2001] or [Massart, 2005]. In the second case, mostly used in the regression framework, the penalty induces smoothness restriction over the estimator, leading to a regularized estimator. For general references, we refer to [Silverman, 1985] or [van de Geer, 2000]. In both cases, the penalty has to be chosen such that the estimator constructed over the whole range of model has a rate of convergence smaller than the rate obtained for the best model. We also point out that throughout the paper we assume that there always exists a solution to the minimization issue (1.1). If not, we consider the following version of the estimator

$$\hat{s} = \arg \min_{m \in \mathcal{M}_n,\, t \in S_m} \left(\gamma_n(t) + \mathrm{pen}(m,t) + \epsilon_n\right).$$

where $\epsilon_n \to 0$. If $\epsilon_n = O(\frac{1}{n})$, the asymptotic behavior of the estimator remains unchanged.

The article falls into the following part. Section 2 is devoted to the definition of the Besov spaces that will be discussed in this paper. Section 3 deals with maxisets for model selection estimator in the white noise model. In Section 4, we consider the regression framework. We lack smoothness property over the estimator to provide maxisets, hence in Section 5 we study the issue of finding maxisets for the regression problem using splines.

2. The maxiset point of view

Let us now give the precise definition of maxisets. For this purpose, let us consider a very general sequence of statistical models: $\{\Omega_n, \mathcal{A}_n, \mathbb{P}_\theta^n, \theta \in \Theta\}$, where the $\mathbb{P}_\theta^n$'s are probability

distributions on Ω_n, and Θ is the set of parameters. Let us consider a sequence of estimates $\hat{q}_n$ of a quantity $q(\theta)$ associated with this sequence of models, a loss function ρ and a rate of convergence α_n tending to 0. The maxiset of $\hat{q}_n$ of radius R for the rate α_n associated with the loss ρ is the following set:

$$MS(\hat{q}_n, \rho, \alpha_n) = \{\theta \in \Theta : \sup_n \alpha_n^{-1} \mathbb{E}_\theta^n \rho(\hat{q}_n, q(\theta)) \leq R\}.$$

In this paper we only consider functional estimation and we only use $\rho = \|.\|_{\mathbb{L}_2}^2$, the $\mathbb{L}_2$-loss and we note:

$$MS(\hat{q}_n, \alpha_n)(R) = MS(\hat{q}_n, \|.\|_{\mathbb{L}_2}^2, \alpha_n)(R).$$

In the sequel, the following equality

$$MS(\hat{q}_n, \alpha_n) = \mathcal{B},$$

where $\mathcal{B}$ is a given space will mean that

$$\forall R,\ \exists R',\ MS(\hat{q}_n, \alpha_n)(R) \subset \mathcal{B}(R') \quad \text{and} \quad \forall R',\ \exists R,\ \mathcal{B}(R') \subset MS(\hat{q}_n, \alpha_n)(R),$$

where $R, R' > 0$ respectively denote the radii of balls of $MS(\hat{q}_n, \alpha_n)$ and $\mathcal{B}$. As we said, this paper deals only with functional estimation in different statistical models. Let us now give precise examples of maxiset results. We consider the white noise framework

$$dY_t = s_0(t)dt + \frac{1}{\sqrt{n}} dW_t, \quad t \in [0,1], \tag{2.1}$$

where s_0 is the signal to be estimated, $\frac{1}{\sqrt{n}}$ is the noise level and W is a Wiener process. Under mild conditions, [Kerkyacharian and Picard, 2000] proved that for classical wavelet thresholding estimates $\tilde{s}_n$ with thresholds of the form $\kappa\sqrt{\log(n)/n}$, where κ is a constant, the maxiset is:

$$MS(\tilde{s}_n, (\log(n)/n)^{2\alpha/(1+2\alpha)}) = W(2/(1+2\alpha)),$$

where a function f belongs to the so-called weak Besov space $W(r)$ if and only if:

$$\sup_{\lambda>0} \lambda^{r-2} \sum_{j\geq -1} \sum_k \beta_{jk}^2 I\{|\beta_{jk}| \leq \lambda\} < \infty$$

(see [Kerkyacharian and Picard, 2000]). Note that no maxiset results for thresholding estimates have been achievable when rates are polynomial. In the same model Rivoirard [Rivoirard, 2004] studied maxisets of linear estimators $\hat{s}_n$ for polynomial rates and, roughly speaking, proved that

$$MS(\hat{s}_n, n^{-2\alpha/(1+2\alpha)}) = \mathcal{B}_{2,\infty}^\alpha.$$

We recall that a function f belongs to the Besov space $\mathcal{B}_{p,\infty}^\alpha$, if and only if:

$$\sup_{j\geq -1} 2^{j(\alpha+\frac{1}{2}-\frac{1}{p})p} \sum_k |\beta_{jk}|^p < \infty.$$

(see [DeVore and Lorentz, 1993]). Note that, when $p = 2$, f belongs to $\mathcal{B}_{2,\infty}^s$ if and only if:

$$\sup_{J\geq -1} 2^{2Js} \sum_{j\geq J} \sum_k \beta_{jk}^2 \ < +\infty.$$

3. Maxisets for model selection in the white noise model and polynomial rates

The goal of this section is to point out maxisets for penalized estimators developed by Birgé and Massart in the classical Gaussian white noise model. In this framework and for an appropriate choice of models, Birgé and Massart [Birgé and Massart, 2001] proved that adaptive penalized rules achieve polynomial minimax rates on classical functional spaces, such as Besov spaces $\mathcal{B}^{\alpha}_{p,\infty}$.
So, we consider Model (2.1). We recall that it means that for any function $\phi \in \mathbb{L}_2([0,1])$,

$$\int \phi(t)dY_t = \int s_0(t)\phi(t)dt + \frac{1}{\sqrt{n}}\int \phi(t)dW_t$$

is observable. In this section, we consider a compactly supported wavelet basis denoted $(\psi_{jk})_{j\geq -1,k\in\mathbb{Z}}$, with

$$\psi_{-1k}(t) = \phi(t-k), \quad \psi_{jk}(t) = 2^{j/2}\psi(2^j t - k), \quad j \geq 0, k \in \mathbb{Z}$$

and ϕ and ψ are respectively the father and mother wavelets. So the function s_0 can be decomposed as follows:

$$s_0(t) = \sum_{j\geq -1}\sum_{k} \beta_{jk}\psi_{jk}(t),$$

where $\beta_{jk} = \int s_0(t)\psi_{jk}(t)dt$ is the wavelet coefficient of s_0 at level j and location $k2^{-j}$. If the support of ϕ and ψ is included into $[A_\psi, B_\psi]$, observe that $\beta_{jk} \neq 0$ if and only if $k \in \mathcal{I}_j := \{-B_\psi + 1, \ldots, 2^j - A_\psi - 1\}$ and the number of unknown wavelet coefficients at level j is $|\mathcal{I}_j|$. For further details on the theory of wavelets, we refer the reader to [Meyer, 1990], [Mallat, 1998] and [**?**]. Using the wavelet basis, the Gaussian white noise model is reduced to the following sequence model:

$$\hat{\beta}_{jk} = \beta_{jk} + \frac{1}{\sqrt{n}}z_{jk}, \quad z_{jk} \overset{iid}{\sim} \mathcal{N}(0,1), \quad j \geq -1, k \in \mathbb{Z},$$

and we estimate the wavelet coefficients by using observations $\hat{\beta}_{jk}$. The estimator is built as follows. We assume we are given a collection of finite-dimensional linear spaces $\{S_m : \quad m \in \mathcal{M}_n\}$, where $\mathcal{M}_n$ is a collection of subsets of $\{(j,k) : \quad j \geq -1, k \in \mathcal{I}_j\}$. The spaces S_m are called models. Actually, the models will be generated by wavelet bases. More precisely, we take:

$$S_m = \text{span}\left\{\psi_{j,k} : \quad (j,k) \in m\right\}.$$

We do not assume that s_0 belongs to $\cup_{m\in\mathcal{M}_n} S_m$. For each $m \in \mathcal{M}_n$, we denote by D_m the dimension of S_m. The proofs of our results in this section will need the following assumption:

$$\exists\, \kappa > 0, \text{ such that } \sum_{m\in\mathcal{M}_n} \exp(-\kappa D_m) < \infty. \tag{3.1}$$

Finally, we note $\hat{s}_m$ the least-squares estimator of s_0 in S_m and the penalty function from $\mathcal{M}_n$ into $\mathbb{R}_+$ will be denoted pen(.). We select $\hat{m} \in \mathcal{M}_n$ as follows:

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n}\left[-\|\hat{s}_m\|_2^2 + \text{pen}(m)\right],$$

where $\|.\|_2$ denotes the $\mathbb{L}_2$-norm with respect to the Lebesgue measure μ on $[0,1]$. Now, let us investigate different choices for the models to obtain maxisets as large as possible. The most natural choice would consists in considering all the possible models: $\mathcal{M}_n$ is the collection of all subsets of $\{(j,k): \quad k \in \mathcal{I}_j, j \in \mathbb{N}\}$. But in this case, Assumption (3.1) is not checked. So, our choice of models will be more restrictive. Another natural choice of models, checking Assumption (3.1), is the following:

$$\mathcal{M}_n = \{m: \quad m = \{(j,k): \quad k \in \mathcal{I}_j, -1 \leq j \leq J_m\}\},$$

with $J_m \in \mathbb{N}$. Models are then nested and for any m, D_m verifies:

$$c_1 2^{J_m} \leq D_m \leq c_2 2^{J_m}, \tag{3.2}$$

where c_1 and c_2 only depend on the wavelet basis. For this poor class, maxisets of penalized estimators are the following:

Theorem 3.1. *Let $\alpha > 0$ and*

$$\mathrm{pen}(m) = KD_m/n,$$

where $K > (1+\sqrt{2\kappa})^2$. If $\hat{s}_{\hat{m}}$ is the associated penalized estimator,

$$MS(\hat{s}_{\hat{m}}, n^{-2\alpha/(2\alpha+1)}) = \mathcal{B}^{\alpha}_{2,\infty}.$$

Proof: Using (3.1) and applying Theorem 47 of Massart [Massart, 2005], we have:

$$\mathcal{B}^{\alpha}_{2,\infty} \subset MS(\hat{s}_{\hat{m}}, n^{-2\alpha/(2\alpha+1)}).$$

To prove the other inclusion, we use that for any m,

$$\mathbb{E}\|\hat{s}_m - s_0\|_2^2 = \|s_0 - s_m\|_2^2 + D_m/n,$$

where s_m is the $\mathbb{L}_2$-projection of s_0 onto S_m. Then,

$$\begin{aligned} \mathbb{E}\|\hat{s}_{\hat{m}} - s_0\|_2^2 &\geq \inf_{m \in \mathcal{M}_n} \mathbb{E}\|\hat{s}_m - s_0\|_2^2 \\ &\geq \inf_{m \in \mathcal{M}_n} [\|s_0 - s_m\|_2^2 + D_m/n] \end{aligned}$$

So, using (3.2), if s_0 belongs to $MS(\hat{s}_{\hat{m}}, n^{-2\alpha/(2\alpha+1)})$, for any $n \in \mathbb{N}^*$, there exists $m = m(n)$ such that

- $2^{J_m} n^{-1} \leq C n^{-2\alpha/(2\alpha+1)}$,
- $\sum_{j \geq J_m} \sum_k \beta_{jk}^2 \leq C n^{-2\alpha/(2\alpha+1)}$,

where C is a constant. So,

$$\sum_{j \geq J_m} \sum_k \beta_{jk}^2 \leq C^{1+4\alpha/(2\alpha+1)} 2^{-2\alpha J_m}.$$

If there exists j_0 such that

$$\forall\, j \geq j_0, \quad \forall\, k \in \mathcal{I}_j, \quad \beta_{jk} = 0,$$

then s_0 obviously belongs to $\mathcal{B}^{\alpha}_{p,\infty}$. Otherwise $J_{m(n)}$ tends to $+\infty$ when n tends to $+\infty$. This implies that

$$\sup_{J \geq -1} 2^{2J\alpha} \sum_{j \geq J} \sum_{k \in \mathcal{I}_j} \beta_{jk}^2 < \infty.$$

Using the characterization of Besov spaces by using wavelet coefficients, this shows that s_0 belongs to $\mathcal{B}^{\alpha}_{2,\infty}$, and

$$MS(\hat{s}_{\hat{m}}, n^{-2\alpha/(2\alpha+1)}) \subset \mathcal{B}^{\alpha}_{2,\infty}.$$

The theorem is proved. □

From this result, we can draw following conclusions. First of all, we conclude that in the maxiset framework, these penalized estimators achieve exactly the same performance as linear ones (see [Rivoirard, 2004]). Secondly, note that this procedure was built to achieve optimal rates on spaces $\mathcal{B}^{\alpha}_{2,\infty}$. And we show that it cannot estimate other functions at the rate $n^{-2\alpha/(2\alpha+1)}$. In particular, functions of $\mathcal{B}^{\alpha}_{p,\infty} \setminus \mathcal{B}^{\alpha}_{2,\infty}$ cannot be estimated at this rate with this penalized estimator.
To overcome this drawback, we can consider the procedure built by Massart [Massart, 2005] that achieves minimax rates on $\mathcal{B}^{\alpha}_{p,\infty}$, when $\alpha > 1/p - 1/2$. For this purpose, we define for any $J_m \in \mathbb{N}$ and any $j \geq J_m$,

$$A(j, J_m) = \lfloor 2^{J_m}(j - J_m + 1)^{\theta} \rfloor,$$

where $\theta > 2$. We take:

$$\mathcal{M}_n = \{m : \quad m = \{(j,k), \quad k \in \mathcal{I}_j, -1 \leq j \leq J_m - 1\} \cup \{(j,k), \quad k \in K(j, J_m), j \geq J_m\}\},$$

with $J_m \in \mathbb{N}$, and $K(j, J_m) \subset \mathcal{I}_j$, such that

$$|K(j, J_m)| = A(j, J_m).$$

Note that one more time, for any m, D_m verifies:

$$c_1 2^{J_m} \leq D_m \leq c_2 2^{J_m},$$

where c_1 and c_2 only depend on the wavelet basis and θ. Furthermore, this collection of models can be view as a compromise between too rich and too poor collections considered before. So, we can hope to obtain satisfying maxiset results. Indeed, we have the following theorem:

Theorem 3.2. *Let $\alpha > 0$ and*

$$\mathrm{pen}(m) = K D_m / n,$$

where $K > (1 + \sqrt{2\kappa})^2$. If $\hat{s}_{\hat{m}}$ is the associated penalized estimator,

$$MS(\hat{s}_{\hat{m}}, n^{-2\alpha/(2\alpha+1)}) = \mathcal{W}^{\alpha}_{2,\infty},$$

where

$$\mathcal{W}^{\alpha}_{2,\infty} = \left\{ s : \quad \sup_{J \geq -1} 2^{2\alpha J} \sum_{j \geq J} \sum_{k > A(j,J)} \beta^2_{j(k)} < \infty \right\},$$

and

$$\beta_{j(1)} \geq \beta_{j(2)} \geq \cdots \geq \beta_{j(|\mathcal{I}_j|)}$$

are the reordered wavelet coefficients of s at level j.

Proof: Using Theorem 47 of Massart [Massart, 2005] and elements of the proof of Theorem 3.1, we prove that the risk of $\hat{s}_{\hat{m}}$ is of the same order as

$$\inf_{m\in\mathcal{M}_n}[\|s_0 - s_m\|_2^2 + D_m/n].$$

So, one more time, s_0 belongs to $MS(\hat{s}_{\hat{m}}, n^{-2\alpha/(2\alpha+1)})$ if and only if for any $n \in \mathbb{N}^*$, there exists $m = m(n)$ such that

- $2^{J_m} n^{-1} \leq C n^{-2\alpha/(2\alpha+1)}$,
- $\sum_{(j,k)\notin m} \beta_{jk}^2 \leq C n^{-2\alpha/(2\alpha+1)}$,

where C is a constant. And as previously, we conclude that s_0 belongs to $MS(\hat{s}_{\hat{m}}, n^{-2\alpha/(2\alpha+1)})$ if and only if there exists a constant C such that for any J, there exists m with $J_m = J$ and

$$2^{2J_m s} \sum_{(j,k)\notin m} \beta_{jk}^2 \leq C,$$

which means that $s_0 \in \mathcal{W}^{\alpha}_{2,\infty}$. The theorem is proved. □

There is no doubt that $\mathcal{B}^{\alpha}_{p,\infty} \subsetneq \mathcal{W}^{\alpha}_{2,\infty}$, when $\alpha > 1/p - 1/2$. For instance, dealing with the case $p \geq 2$, let $\alpha > 0$ and $s \in \mathbb{L}_2([0,1])$ be such that

$$s = \sum_{j\geq -1}\sum_k \beta_{jk}\psi_{jk} \text{ with } \beta_{jk} = \frac{2^{-\alpha j}}{|k|+1} \text{ if } |k| > j^{\theta}, \quad \beta_{jk} = 2^{-\alpha j} \text{ otherwise .}$$

One gets that $s \in \mathcal{W}^{\alpha}_{2,\infty} \setminus \mathcal{B}^{\alpha}_{p,\infty}$. On the one hand, to prove that $s \notin \mathcal{B}^{\alpha}_{p,\infty}$, it suffices to observe that, for any $j \in \mathbb{N}$

$$2^{j(\alpha-\frac{1}{p}+\frac{1}{2})p} \sum_k |\beta_{jk}|^p \geq (2j^{\theta}+1)\, 2^{j(\frac{p}{2}-1)} \geq j^{\theta}\, 2^{j(\frac{p}{2}-1)}.$$

So

$$\sup_{j\geq -1} 2^{j(\alpha-\frac{1}{p}+\frac{1}{2})p} \sum_k |\beta_{jk}|^p = \infty,$$

which implies that $s \notin \mathcal{B}^{\alpha}_{p,\infty}$. On the other hand, since $\sum_{j\geq j_0}\sum_{k>A(j,j_0)} \beta_{j(k)}^2 \leq \sum_{j\geq j_0}\sum_{|k|\geq \frac{A(j,j_0)}{2}} \beta_{jk}^2$, to prove that the function $s \in \mathcal{W}^{\alpha}_{2,\infty}$, it is sufficient to show that for any j_0 large enough,

$$\sum_{j\geq j_0}\sum_{|k|\geq \frac{A(j,j_0)}{2}} \beta_{jk}^2 \leq C\, 2^{-2\alpha j_0},$$

where C is a constant which does not depend on j_0. So let j_0 be such that $\frac{\ln(3j_0)}{j_0} < \frac{\ln(2)}{\theta}$. Then, $3j_0^{\theta} < 2^{j_0}$ which implies that for any $j \geq j_0$, $2^{j_0}(1-\frac{j_0-1}{j})^{\theta} > 3j^{\theta}$, $\lfloor 2^{j_0}(1-\frac{j_0-1}{j})^{\theta}\rfloor > 2j^{\theta}$ and that

$$\frac{A(j,j_0)}{2} > j^{\theta}.$$

So

$$\begin{aligned}\sum_{j\geq j_0}\sum_{|k|\geq\frac{A(j,j_0)}{2}}\beta_{jk}^2 &\leq \sum_{j\geq j_0}\sum_{|k|>j^\theta}\frac{2^{-2\alpha j}}{((|k|+1)^2}\\ &\leq \sum_{j\geq j_0}2^{-2\alpha j}\sum_k\frac{1}{(|k|+1)^2}\\ &= C\,2^{-2\alpha j_0}.\end{aligned}$$

Finally one gets that $s\in\mathcal{W}^\alpha_{2,\infty}$.

Note that the strict inclusion between the spaces $\mathcal{B}^\alpha_{p,\infty}$ and $\mathcal{W}^\alpha_{2,\infty}$ shows that this procedure strictly improves the previous one from the maxiset point of view. □

4. Maxisets for model selection in regression model

The goal of this section is to find maxisets in the regression model. Unfortunately, we will show that the maxiset point of view is not well adapted to this problem since the smoothness conditions for the maxiset are imposed by the discretization issue. Here, we consider the following model:

$$Y_i = s_0(t_i)+\frac{\sigma}{\sqrt{n}}w_i,\ i=1,\dots,n,\quad w_i\overset{iid}{\sim}\mathcal{N}(0,1), \tag{4.1}$$

where $(t_i)_{i=1,\dots,n}\in[0,1]^n$ are discrete fixed observation times. Along this paper, we assume that σ is kwown and $\sigma=1$. Under conditions on the t_i's, Brown and Low (1996) showed the asymptotic equivalence of this model and Model (2.1). For each function s, define the corresponding empirical norm as $\|s\|_n^2=\frac{1}{n}\sum_{i=1}^n s^2(t_i)$ to be compared with the integrated norm $\|s\|_2^2$. On one hand, the first norm is well adapted to the estimation issue since, in the sequence space, the only available values of the functions are to be taken at the observation points. In the literature, asymptotic results are often given with respect to the empirical norm, see for example [Kohler, 1999], [van de Geer, 2000] or [Loubes and van de Geer, 2002]. On the other hand, in the maxiset point of view, the sets are to be characterized as smoothness functional spaces, regardless of the discretization grid. Hence we aim at finding a control of approximation ! properties of the unknown function, with respect to the integrated norm. In model selection, the functions are characterized by their projection onto a basis, so in the regression framework, we introduce a bias between the real coefficients and their discrete corresponding values. The gap between a real observation model and a discretized model has been stressed by a large number of authors, see for instance [Donoho and Johnstone, 1995], [Donoho and Johnstone, 1994], [Antoniadis et al., 1997] or [Antoniadis and Pham, 1998].
As in Section 3, consider $\mathcal{S}_m, m\in\mathcal{M}_n$ a collection of finite linear dimensional sets of $\mathbb{L}^2([0,1])\cap\mathbb{L}_\infty([0,1])$ satisfying the condition that there exists some positive number R_n such that

$$\sup_{s\in\mathcal{S}_m}\frac{\|s\|_n}{\|s\|_2}\leq R_n.$$

Set D_m the dimension of $\mathcal{S}_m$. A well-known method to estimate the regression function s_0 is to use a least-squares estimator over the sieves $\mathcal{S}_m$, $m\in\mathcal{M}_n$, once $\mathcal{M}_n$ has been chosen. It

means

$$\hat{s}_m = \arg\min_{s\in\mathcal{S}_m} \sum_{i=1}^{n} (Y_i - s(t_i))^2,$$

If $\|.\|_n$ is the (empirical) norm associated with $\mathbb{L}_2(\mu_n)$, it is easy to check that

$$\mathbb{E}\|\hat{s}_m - s_0\|_n^2 = \sigma^2 \frac{D_m}{n} + \inf_{s\in\mathcal{S}_m} \|s - s_0\|_n^2. \tag{4.2}$$

In (4.2), there exists an optimal choice of m among the indexes in $\mathcal{M}_n$ which achieves the best trade-off between the biased term $\inf_{s\in\mathcal{S}_m} \|s - s_0\|_n^2$ and the approximation term $\frac{D_m}{n}$. But, the drawback of this previous approach lies in the fact that this best parameter heavily relies on the knowledge of the regularity α of the space of the true function s_0. Thus, the aim of model selection is to provide a methodology able to construct an adaptive estimator. Using the method, originally described by Birgé and Massart in [Birgé and Massart, 1997] or [Birgé and Massart, 1998] and developed by Barraud in [Baraud, 2000] for the regression model, we define the model selection estimator as follows.

$$\tilde{s}_n = \arg\min_{m\in\mathcal{M}_n,\, s\in\mathcal{S}_m} \left(\|Y - s\|_n^2 + \mathrm{pen}(m)\right) \tag{4.3}$$

with $\mathrm{pen}(m)$ is of the form $\mathrm{pen}(m) = (1+c)\frac{D_m L_m}{n}\sigma^2$, for some weights L_m and c is a positive number.

Such estimators in the regression scheme have been studied by Baraud in [Baraud, 2000], providing rates of convergence. The rate of convergence of the estimator is proved to be less than the best rate for all the models. More precisely the following theorem gives a rate of convergence under regularity conditions

Theorem 4.1. *Assume that $s_0 \in \mathcal{B}^{\alpha}_{l,\infty}$, for $l \geq 2$ and that S_m is the space of trigonometric polynomials of degree less or equal 2^m and $\mathcal{M}_n = [0, \ldots, m_n]$. Assume also that s_0 belongs to $\mathbb{L}_2([0,1]) \cap \mathbb{L}_\infty([0,1])$. Hence for $\alpha > 1/l$, the estimator $\tilde{s}_n$ defined by* (4.3) *satisfies*

$$\mathbb{E}\|\tilde{s}_n - s_0\|_2^2 \leq C n^{-\frac{2\alpha}{2\alpha+1}}.$$

In order to get maxisets, we saw in Section 2 that we both need the rate of convergence of the estimation procedure as well as a way to characterize a smoothness set by the approximation provided by the estimator. This implies that we must characterize the set for which it is possible to build an estimator $\tilde{s}_n$ such that $\|\tilde{s}_n - s_0\|_2^2 \leq n^{-\frac{2\alpha}{2\alpha+1}}$. The norm related to smoothness approximation property is the integrated norm since the set cannot depend on the discretization scheme. But, in model selection theory, approximation property and estimation property are related through the empirical norm by relation (4.2). Hence we obtain

$$\mathbb{E}\|\tilde{s}_n - s_0\|_n^2 \geq \inf_{m\in\mathcal{M}_n} \left(\sigma^2 \frac{D_m}{n} + \inf_{s\in\mathcal{S}_m} \|s - s_0\|_n^2\right).$$

So the search for maxisets implies comparing the empirical norm to the integrated norm for functions $s - s_0$, $s \in S_m$. Relationships between empirical norms and integrated norms have been investigated in the estimation literature by several authors. It is achieved under regularity conditions for the set to whom the functions $s - s_0$ belongs. In the finite dimensional case, the following lemma states the equivalence of both norms.

Lemma 4.2. *Let V_n be the approximation space spanned by $(\phi_{j_n k}(.) = 2^{j_n/2}\phi(2^{j_n}. - k))_{k\in\mathbb{Z}}$, where $j_n \in \mathbb{N}^*$, $j_n \leq n$ and ϕ is a r-regular compactly supported father wavelet. Then, for any*

$$s \in V_n \cap L^\infty \cap B_{2,\infty}^{\frac{\alpha}{2\alpha+1}},$$

$$\left| \frac{1}{n}\sum_{i=1}^{n} s^2(i/n) - \int_0^1 s^2(x)dx \right| = O(n^{-\frac{2\alpha}{2\alpha+1}}).$$

In our case $s - s_0$ are the remainder term in the approximation, hence it is not finite dimensional. So if we consider smoothness restrictions, we see that, in [Kress and Sloan, 1993], it implies that $s - s_0$ belongs to a Sobolev ball, while, in [van de Geer, 2000], the regularity condition is expressed by entropy conditions over a ball of the set $\{s - s_0,\ s \in S_m\}$. So it is necessary to restrict drastically the choice of the approximation spaces S_m, $m \in \mathcal{M}_n$, by imposing smoothness conditions over the derivative of the functions in S_m. As a consequence, maxisets can be established for the following estimation procedure: minimizing a contrast function penalized by a smoothness constraint of a derivative of the estimator, namely spline estimators.

5. Maxisets for spline regularized estimators in regression model

In this section we still consider the model (4.1) with $\sigma = 1$ and $t_i = \frac{i}{n}$, $1 \leq i \leq n$. For q and k strictly positive integers, let S_k^q be the set of spline functions defined on $[0, 1]$ of order q with k equispaced interior knots. So, consider the following penalized estimator. For a smoothing sequence $\lambda_n \geq 0$ and a positive integer m such that $1 < m < q - 1$, define

$$\hat{s}_n = \arg\min_{s\in S_k^q} \left(\|Y - s\|_n^2 + \lambda_n^2 \int_0^1 (s^{(m)}(t))^2 dt \right). \tag{5.1}$$

The space S_k^q is well adapted to the estimation problem in the regression framework since spline functions are designed to take into account the smoothness of a function with respect to the observation scheme. As a matter of fact, there exists a basis of S_k^q composed of normalized B-splines, B_j^q, $j = 1, \ldots, q + k$. We refer to [de Boor, 1978] or [DeVore and Lorentz, 1993] for general references. Write $B_q = (B_1^q, \ldots, B_{q+k}^q)'$. A spline basis satisfies the following property. For all $1 < m < q - 1$, there exists a matrix $\Delta^{(m)}$ such that

$$\forall\ s = B_q'\theta,\ s^{(m)} = B'_{q-m}\theta^{(m)} = B'_{q-m}\Delta^{(m)}\theta.$$

So, using the properties of B-splines, the estimator can be written in the following way:

$$\hat{s}_n = \sum_{j=1}^{q+k} \hat{\theta}_j B_j^q = B_q'\hat{\theta},$$

where $\hat{\theta}$ is the solution of the following minimization problem

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^{q+k}} \left(\frac{1}{n}\sum_{i=1}^{n} [Y_i - (B_q{}'\theta)(t_i)]^2 + \lambda_n^2 \| \sum_{j=1}^{q+k-m} \theta_j^{(m)} B_j^{q-m} \|_2^2 \right). \tag{5.2}$$

The following lemma gives the expression of the estimator.

Lemma 5.1. *Let A_n be the matrix with elements $B_j^q(t_i), i = 1, \dots, n,\ j = 1, \dots, q+k$, and C_{q-m} the Gram matrix with elements*

$$(C_{q-m})_{i,j} = \int_0^1 B_i^{q-m}(t) B_j^{q-m}(t) dt, \quad 1 \le i, j \le n.$$

The explicit solution of (5.2) *is given by*

$$\hat{\theta} = \left(\frac{1}{n} A_n^{'} A_n + \lambda_n^2 \Delta^{(m)^{'}} C_{q-m} \Delta^{(m)} \right)^{-1} \frac{1}{n} A_n^{'} Y. \tag{5.3}$$

Proof of Lemma 5.1: The estimator $\hat{s}_{n,\lambda_n} = B_q^{'} \hat{\theta}$ is defined as the solution of the constrained minimization issue (5.1). Since

$$\begin{aligned} \| \sum_{j=1}^{q+k-m} \theta_j^{(m)} B_j^{q-m} \|_2^2 &= \sum_{i,j=1}^{q+k-m} \theta_i^{(m)} \int_0^1 B_i^{q-m}(t) B_j^{q-m}(t) dt\, \theta_j^{(m)} \\ &= \theta^{'} \Delta^{(m)^{'}} C_{q-m} \Delta^{(m)} \theta, \end{aligned}$$

then, we can write that

$$\hat{\theta}_n = \arg \min_{\theta \in \mathbb{R}^{q+k}} \left(\frac{1}{n} (Y - B_q^{'} \theta)^{'} (Y - B_q^{'} \theta) + \lambda_n^2 \theta^{'} \Delta^{(m)^{'}} C_{q-m} \Delta^{(m)} \theta \right).$$

Since $G_n = \frac{1}{n} A_n^{'} A_n + \lambda_n^2 \Delta^{(m)^{'}} C_{q-m} \Delta^{(m)}$ is a symmetric matrix defining a scalar product, there exists a matrix U_n such that

$$G_n = U_n^{'} U_n.$$

We get

$$\begin{aligned} &\frac{1}{n} (Y - B_q^{'} \theta)^{'} (Y - B_q^{'} \theta) + \lambda_n^2 \theta^{'} \Delta^{(m)^{'}} C_{q-m} \Delta^{(m)} \theta \\ =&\frac{1}{n} Y^{'} Y - \frac{1}{n} (Y^{'} B_q^{'} \theta + \theta^{'} B_q Y) + \theta^{'} U_n^{'} U_n \theta \\ =&\frac{1}{n} Y^{'} Y - \frac{1}{n} (Y^{'} B_q^{'} U_n^{-1} [U_n \theta] + [U_n \theta]^{'} (U_n^{-1})^{'} B_q Y) + [U_n \theta]^{'} [U_n \theta] \\ =&C + \| U_n \theta - \frac{1}{n} [U_n^{'}]^{-1} B_q Y \|^2, \end{aligned}$$

where C does not depend on θ. As a result, we have turned the minimization program (5.1) into a least squares minimization. So, the estimator has the following expression

$$\begin{aligned} \hat{\theta} &= (U_n^{'} U_n)^{-1} U_n^{'} \left(\frac{1}{n} [U_n^{'}]^{-1} B_q Y \right) \\ &= G_n^{-1} \frac{1}{n} B_q Y. \end{aligned}$$

□

Remark : Since the design sequence is $\{t_i = \frac{i}{n}, 1 \le i \le n\}$, it ensures that G_n is invertible and hence the uniqueness of $\hat{s}_n$ provided that n is sufficiently large.

The asymptotic behavior of penalized splines estimates has been studied by several authors: we refer here to [Agarwal and Studden, 1980], [Cardot, 2002b], [Cardot, 2002a],[van de Geer, 1990] or [Wahba, 1990] for more general references. The following theorem gives the asymptotic rate of convergence of a spline estimator.

Theorem 5.2. *Let α be a strictly positive integer such that $\alpha < m < q-1$. Choosing $k = n^{\frac{1}{2\alpha+1}}$ and the optimal choice of regularizing sequence $\lambda_n = n^{-\frac{\alpha+2m}{4\alpha+2}}$, the risk of any function $s_0 \in W_{q,k}^{\alpha*}$ satisfies*

$$\sup_{n\in\mathbb{N}^*} n^{\frac{2\alpha}{2\alpha+1}} \mathbb{E}\|\hat{s}_n - s_0\|_2^2 < \infty,$$

where

$$W_{q,k}^{\alpha*} = \left\{ s;\ \inf_{t\in S_k^q} \|s-t\|_2^2 = O(k^{-2\alpha}) \right\}.$$

Remark : The rate of convergence provided in this theorem follows from an optimal choice of both the number of knots k and the smoothing sequence λ_n. So this estimation procedure is a linear non adaptive procedure, which enables us to achieve the minimax rate of convergence for functions belonging to the space $W_{2\infty}^{\alpha,*}$. Such spaces are deeply linked with spline smoothing estimators in the sense that the maxiset of a spline estimation procedure, associated with the rate of convergence $n^{-\frac{2\alpha}{2\alpha+1}}$, as we shall prove it in the next theorem. Before that, let us notice that we have the following properties :

$$\forall\, \alpha \in \mathbb{N}^*,\ \ \mathcal{B}_{2,2}^{\alpha} = W_{\alpha,k}^{\alpha*},$$
$$\forall\, \alpha < q \Longrightarrow \mathcal{B}_{2,2}^{\alpha} \subsetneq W_{q,k}^{\alpha*}.$$

The first equality is obtained by using (ii) of Theorem 2.4 of Chapter 12 in [DeVore and Lorentz, 1993]. The second equality is obtained by just observing that for any $\alpha < q$, $\mathcal{B}_{2,2}^{\alpha} \subsetneq \mathcal{B}_{2,2}^{q}$ and that $\mathcal{B}_{2,2}^{q} \subseteq W_{q,k}^{\alpha*}$.

Proof of Theorem 5.2 : Since $m > \alpha$, our estimation problem deals with functions belonging to the functional space $C^{\alpha}([0,1])$ and using same arguments of proof as in Theorem 3.1 in [Cardot, 2002a], Theorem is easily proved. □

Theorem 5.3. *Let α be a strictly positive integer such that $\alpha < m < q-1$. Considering the spline penalized estimators* (5.1) *associated with the optimal number of knots for the spline basis $k = n^{\frac{1}{2\alpha+1}}$ and the optimal choice of regularizing sequence $\lambda_n = n^{-\frac{\alpha+2m}{4\alpha+2}}$, one gets:*

$$MS(\hat{s}_n, n^{-2\alpha/(2\alpha+1)}) = W_{q,k}^{\alpha*}.$$

Proof: Theorem 5.2 proves the inclusion $W_{q,k}^{\alpha*} \subset MS(\hat{s}_n, n^{-2\alpha/(2\alpha+1)})$. To prove the other inclusion, let us consider a function $s_0 \in MS(\hat{s}_n, n^{-2\alpha/(2\alpha+1)})$. The optimal spline estimator constructed with splines in S_k^q is such that

$$\mathbb{E}\|\hat{s}_n - s_0\|_2^2 = O\left(n^{-\frac{2\alpha}{2\alpha+1}}\right).$$

First, note that the estimator itself belongs to the set of splines, so we can write that

$$\mathbb{E}\|\hat{s}_n - s_0\|_2^2 \geq \inf_{\theta\in\mathbb{R}^{q+k}} \|s_0 - B_q^{'}\theta\|_2^2,$$

leading to

$$\inf_{t\in S_k^q} \|s_0 - t\|_2^2 = O(n^{-\frac{2\alpha}{2\alpha+1}}) = O(k^{-2\alpha}). \tag{5.4}$$

Finally, the bound (5.4) yields $s_0 \in W_{q,k}^{\alpha*}$. In other words, one gets $MS(\hat{s}_n, n^{-2\alpha/(2\alpha+1)}) \subset W_{q,k}^{\alpha*}$. It ends the proof.

□

6. Conclusion

In this work, we consider the efficiency of penalized model selection estimators in a maxiset point of view. In the white noise model, we obtain for the non nested model selection procedure the maxiset $\mathcal{W}_{2,\infty}^{\alpha}$. This set is very similar to the weak Besov spaces obtained when studying wavelet thresholded estimators. Such results can be found in [Cohen et al., 2001], [Rivoirard, 2004] or [Autin, 2005] for instance. Hence, when considering the model where the coefficients of the functions are observed, both non linear procedures provide the same kind of results. They enable to build fully tractable estimators which converge at the same minimax rate of convergence (up to logarithmic terms) with maxisets of the same kind.

When dealing with observations drawn from the regression model, we face the discretization issue. Indeed the empirical norm must be compared to the integrated norm, which is often solved, in the literature, by assuming sufficient smoothness conditions over the functional set. Unfortunately, finding the maxiset of an estimation procedure implies finding the minimal regularity conditions such that the estimator converges at a given rate of convergence. In the regression scheme, it first implies to control the regularity of the estimator, preventing from choosing too irregular models. Hence the maxiset point of view is not adapted to model selection procedure in model (4.1). We point out that maxisets are not given for other type of estimation procedure in the regression model but only when assuming that the sampled model is equivalent to the white noise model and that the wavelet coefficients of the data are directly observed. This approximation is highlighted in [Donoho and Johnstone, 1999] or [Donoho and Johnstone, 1994]. Finding maxiset for the real regression model is only achieved by considering M-estimation with regularity penalties and hence so-called spline estimator. Then we note that the regularity condition to ensure the comparison between the empirical and the integrated norm is stronger than the one needed to ensure the convergence of the estimator, preventing as large maxisets as in the white noise model. Perhaps a key to more efficient estimation procedure, in the maxiset point of view, would be to construct bases adapted to the discretization scheme, warped bases taking more into account the regularity of the functions with respect to the observation points.

References

[Agarwal and Studden, 1980] Agarwal, G. and Studden, W. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statit.*, 8:1307–1325.

[Antoniadis et al., 1997] Antoniadis, A., Grgoire, G., and Vial, P. (1997). Random design wavelet curve smoothing. *Stat. Probab. Lett.*, 35(3):225–232.

[Antoniadis and Pham, 1998] Antoniadis, A. and Pham, D. T. (1998). Wavelet regression for random or irregular design. *Comput. Stat. Data Anal.*, 28(4):353–369.

[Autin, 2005] Autin, F. (2005). Ideal maxisets for 3 families of estimation procedures. *Technical Report.*

[Autin and Rivoirard, 2005] Autin, F., P. D. and Rivoirard, V. (2005). Maxiset approach for choosing priors in the bayesian setting. *Technical Report*.

[Baraud, 2000] Baraud, Y. (2000). Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493.

[Barron et al., 1999] Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.

[Birgé and Massart, 1997] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York.

[Birgé and Massart, 1998] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.

[Birgé and Massart, 2001] Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*.

[Cardot, 2002a] Cardot, H. (2002a). Local roughness penalties for regression splines. *Computational Statistics*, 17:89–102.

[Cardot, 2002b] Cardot, H. (2002b). Spatially adaptive splines for statistical linear inverse problems. *Journal of Multivariate Analysis*, 81:100–119.

[Cohen et al., 2001] Cohen, A., DeVore, R., Kerkyacharian, G., and Picard, D. (2001). Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon. Anal.*, 11(2):167–191.

[de Boor, 1978] de Boor, C. (1978). *A practical guide to Splines*. Springer, New-York.

[DeVore and Lorentz, 1993] DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation*. Springer-Verlag, Berlin.

[Donoho and Johnstone, 1994] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

[Donoho and Johnstone, 1995] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.*, 90(432):1200–1224.

[Donoho and Johnstone, 1999] Donoho, D. L. and Johnstone, I. M. (1999). Asymptotic minimaxity of wavelet estimators with sampled data. *Stat. Sin.*, 9(1):1–32.

[Kerkyacharian and Picard, 2000] Kerkyacharian, G. and Picard, D. (2000). Thresholding algorithms, maxisets and well-concentrated bases. *Test*, 9(2):283–344.

[Kerkyacharian and Picard, 2002] Kerkyacharian, G. and Picard, D. (2002). Minimax or maxisets? *Bernoulli*, 8(2):219–253.

[Kohler, 1999] Kohler, M. (1999). Nonparametric estimation of piecewise smooth regression functions. *Statist. Probab. Lett.*, 43(1):49–55.

[Kress and Sloan, 1993] Kress, R. and Sloan, I. H. (1993). On the numerical solution of a logarithmic integral equation of the first kind for the Helmholtz equation. *Numer. Math.*, 66(2):199–214.

[Loubes and van de Geer, 2002] Loubes, J.-M. and van de Geer, S. A. (2002). Adaptive estimation using thresholding type penalties. *Statistica Neerlandica*, 56:1–26.

[Mallat, 1998] Mallat, S. (1998). *A wavelet tour of signal processing*. Academic Press Inc., San Diego, CA.

[Massart, 2005] Massart, P. (2005). *Concentration inequalities and model selection*. Cours de Saint-Flour.

[Meyer, 1990] Meyer, Y. (1990). *Ondelettes et Opérateurs*. Hermann.

[Rivoirard, 2003] Rivoirard, V. (2003). Bayesian modelization of sparse sequences and maxisets for bayes rules. *Technical Report. Submitted to Math. Methods Statist.*

[Rivoirard, 2004] Rivoirard, V. (2004). Maxisets for linear procedures. *Statist. Probab. Lett.*, 67(3):267–275.

[Silverman, 1985] Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B*, 47(1):1–52. With discussion.

[van de Geer, 1990] van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.*, 18(2):907–924.

[van de Geer, 2000] van de Geer, S. A. (2000). *Applications of empirical process theory*. Cambridge University Press, Cambridge.

[Wahba, 1990] Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

1: Laboratoire de Probabilités et Modèles Aléatoires , UMR 7599, Université Paris VII, 118 Route de Narbonne, 2 place Jussieu, 75251 Paris Cedes 05

2: CNRS - Laboratoire de Mathématiques, UMR 8628, Université Paris-Sud, Bât 425, 91405 Orday cedex

E-mail address: autin@math.jussieu.fr

E-mail address: Jean-Michel.Loubes@math.u-psud.fr

E-mail address: Vicent.Rivoirar@math.u-psud.fr