UNIVERSITE PARIS SUD
Centre d'Orsay

Mémoire

présenté pour obtenir le diplôme d'

Habilitation à diriger les recherches

Spécialité : Mathématiques

par

Yannick Baraud

Estimation par sélection de modèle et décision par tests multiples en régression non-parametrique

Soutenue le 12 Décembre 2002 devant le jury composé de:

M. Lucien Birgé	Président
M. Oleg Lespki	
M. Pascal Massart	
M. Alexander Tsybakov	
M. Aad van der Vaart	Rapporteur

Second rapporteur: M. Vladimir Spokoiny.

À ma famille, Tom, Eva et leur jolie maman.

Remerciements

Je voudrais commencer par remercier Lucien Birgé pour l'intérêt qu'il a toujours porté à mon travail. Merci pour ses conseils. Dire qu'il a influencé ma manière d'aborder les statistiques et de rédiger les mathématiques est un doux euphémisme !

Merci à Pascal pour sa sollicitude et sa générosité à mon égard. J'aimerais lui exprimer toute ma reconnaissance pour son soutien tout au long de ces années. J'espère qu'il me fera encore profiter longtemps, en avant-première, de ces inégalités. Je crois que son "congélateur" ne sera jamais vide.

J'ai été très honoré que Aad van der Vaart et Vladimir Spokoiny acceptent de rapporter mon travail. Je tiens également à remercier Oleg Lepski, Alexandre Tsybakov et Aad van der Vaart d'avoir accepté de faire partie de mon jury. Leur présence m'honore.

Je voudrais remercier tous mes co-auteurs, Béatrice, Fabienne, Gabrielle et Sylvie, pour nos collaborations riches en émotions. Merci pour tous ces moments mémorables passés ensemble.

Je crois avoir été honteusement gâté à l'École Normale Supérieure. Je tiens donc à remercier Gérard Ben Arous et Arnaud Beauville de m'avoir accueilli dans leur laboratoire. Merci également a Marc Rosso et Jean-François Le Gall pour avoir continué à promouvoir les statistiques et de m'avoir fait confiance. Merci à Jacques Beigbeder et Catherine Le Bihan pour leurs dépannages informatiques, ainsi qu'à Bénédicte Auffray et Laurence Vincent pour leur efficacité à gérer les problèmes logistiques. Merci enfin à tous les membres du DMA, présents et passés, pour leur compagnie sympathique et chaleureuse. Un merci tout particulier à Nalini Anantharaman, Vincent Beffara, Gilles Blanchard, et Patricia Reynaud pour toutes les discussions que nous avons pu avoir.

J'aimerais aussi rendre hommage à ceux qui par leurs cours de mathématiques m'ont donné envie de faire ce métier. Je pense entre autres à Alano Ancona, Annick Auzimour, Michel Bismut, Martine Dubreuil, Henri Queffelec, Fréderique Petit, Claudine Picarrony et Michel Raynaud.

Enfin, je voudrais remercier ceux qui me sont chers. Merci à Jacqueline et Michel Belluc pour leur sollicitude. Merci à ma famille, qu'ils sachent que je les aime. Merci enfin à Barbara, l'ange sans qui rien ne serait possible.

Contents

Foreword

This document aims at providing some perspective on our pieces of research in Statistics. These have mainly addressed the problems of estimation and hypothesis testing in regression models. We include in our definition of "regression models", the linear model, the functional regression model on deterministic or random design points, and the autoregression model, among others...

This document is organized as follows. In Chapter 1, we consider the problem of estimation by model selection. The aims and scope of model selection procedures are presented in Section 1 and the reader will find there illustrative examples. In Sections 2 and 3, we give an account of our contribution on the basis of the papers B. [**3, 2**] and B., Comte and Viennet [**6, 7**].

In Chapter 2, we consider the problem of hypothesis testing and describe some of the results obtained in the series of papers B., Huet and Laurent [**11, 8, 10**] and B. [**4, 5**]. We explain the role of tests in Statistics in Section 1 and provide some connections between hypothesis testing and model selection in Section 2. A description of our tests in collaboration with S. Huet and B. Laurent can be found in Section 3. Section 4 is more theoretically oriented and gives some perspective on B. [**4**] which deals with the description of minimax rates of testing in the regression and Gaussian sequence models. Finally, an application of hypothesis testing to the problem of building nonasymptotic confidence balls is presented in Section 5.

Before turning to the main part of this document, we present two regression models to which we shall repeatedly refer in view of motivating or merely illustrating our approach. We also hope that the description of these will be helpful to understand the importance of regression models in experimental sciences.

MODEL 1 (The linear regression model). Consider the statistical model given by

$$Y = f + \varepsilon, \quad \text{where } f = \Phi\theta. \tag{1}$$

In (1), Φ is a nonrandom $n \times N$-matrix ($1 \leq N \leq n$), θ is an unknown parameter in $\mathbb{R}^N$ and ε a centered random vector in $\mathbb{R}^n$ with i.i.d components of common variance σ^2. The data consist of the observation of the random vector

$Y = (Y_1, \ldots, Y_n)^T$. Model 1 arises in studying the average relationship between quantitative explanatory variables, say ϕ_j's for $j = 1, \ldots, N$, which we can vary and a quantitative dependent variable, y, which we observe. The index i varying among $\{1, \ldots, n\}$ corresponds to a time of observation or an experiment, and the values of y and ϕ_j at time or experiment i are respectively given by Y_i and $\Phi_{i,j}$. A nice feature of this model is that the dependency of the mean of y on the explanatory variables is linear. The difficulty lies in the number N of these variables which can be large, especially when the parameters influencing the issue of the experiment are not clearly identified. In this case, many explanatory variables are usually introduced in view of a reliable approximation of the truth.

MODEL 2 (The functional regression model).

$$Y_i = f(x_i) + \varepsilon_i, \ i = 1, \ldots, n. \tag{2}$$

Here, f denotes an unknown real-valued function, the x_i's are distinct deterministic points of the interval $[0, 1]$ and the ε_i's are i.i.d. centered random variables with common variance σ^2. The data consist of the observation of the pair of random variables $Z_i = (Y_i, x_i)$ for $i = 1, \ldots, n$. Model 2 arises in studying the correlation between the mean of a quantitative variable y and another quantitative variable x which we can vary in $[0, 1]$. Then, Equality (2) accounts for the relation between the quantitative variables y and x at time i, that is respectively Y_i and x_i. Unlike Model 1, the difficulty with Model 2 is that the kind of dependency between x and the mean of y is unspecified.

CHAPTER 1

Estimation via model selection

1. Introduction

1.1. What is model selection about? The choice of a good parameter set is one of the main problems a statistician faces. The parameter set must be both large enough to provide a reliable approximation of the truth and also small enough to limit the errors due to statistical estimation. Solving this dilemma is the main concern of model selection. The problem can often be settled as follows. Let $\{\mathbb{P}_f,\ f \in \mathcal{F}\}$ be a family of probabilities on some space $\mathcal{Z}$ where $\mathcal{F}$ is a subspace of a metric space $(\mathcal{S}, \mathrm{d})$. Consider the problem of estimating the unknown parameter f on the basis of the observation of a random variable $Z \in \mathcal{Z}$ of law $\mathbb{P}_f$ and a collection of spaces $\{\mathcal{F}_m,\ m \in \mathcal{M}_n\} \subset \mathcal{S}$. All along, these spaces will be called *models*. For each $m \in \mathcal{M}_n$, associate some estimator $\hat{f}_m$ of f in $\mathcal{F}_m$. The problem of model selection is to build from Z some $\hat{m}$ among $\mathcal{M}_n$ for which the distance between f and the estimator $\hat{f}_{\hat{m}}$ is as close as possible to the minimal one among the family of estimators $\left\{\hat{f}_m,\ m \in \mathcal{M}_n\right\}$. In the sequel, we offer two examples for which a model selection procedure can be relevant.

EXAMPLE 1.1 (Selecting the degree of an expansion).
Let us consider Model 2 and assume that f belongs to $\mathbb{L}^2([0,1], dx)$. Given an Hilbert basis $\{\phi_j,\ j \geq 1\}$ of $\mathbb{L}^2([0,1], dx)$, f can be expanded as

$$f = \sum_{j \geq 1} \theta_j \phi_j. \tag{3}$$

As (3) involves infinitely many unknown coefficients, one usually decides to truncate the expansion at some degree J before estimating f. The question of what the ideal J should be naturally arises. Intuitively, one must face the following dilemma. On the one hand, the integer J must be large enough to ensure that the truncated expansion contains sufficiently many coefficients to approximate f well. On the other hand, J must not be too large in order to limit the estimation errors of these coefficients. More formally, the problem of the selection of J can be settled as follows. For each integer $J \geq 1$, let us introduce $\mathcal{F}_J$ the linear space generated by the functions ϕ_j for $j = 1, \ldots, J$.

We can associate to each J the least-squares estimator of f onto $\mathcal{F}_J$, that is the function $\hat{f}_J$ which minimizes among those t in $\mathcal{F}_J$ the quantity

$$\gamma_n(Z,t) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - t(x_i)\right)^2. \tag{4}$$

The problem of selecting some best J amounts then to selecting some best estimator (in a suitable sense) among the family $\left\{\hat{f}_J,\ J \geq 1\right\}$. A natural way to compare these estimators is to look at their risks, that is the average distance between these and the target function. By choosing the (pseudo) distance $\mathrm{d}_n(.,.)$ defined for all functions s,t on $[0,1]$ by

$$\mathrm{d}_n^2(t,s) = \frac{1}{n}\sum_{i=1}^{n}\left(t(x_i) - s(x_i)\right)^2, \tag{5}$$

for each J, the risk of the estimator $\hat{f}_J$ is then defined as the quantity $\mathbb{E}\left[\mathrm{d}_n^2(f,\hat{f}_J)\right]$. Denoting by D_J the dimension of the linear subspace of $\mathbb{R}^n$

$$\left\{(t(x_1),\ldots,t(x_n))^T,\ t \in \mathcal{F}_J\right\},$$

the risk of $\hat{f}_J$ splits in two terms as follows:

$$\mathbb{E}\left[\mathrm{d}_n^2(f,\hat{f}_J)\right] = \mathrm{d}_n^2(f,\mathcal{F}_J) + \frac{D_J}{n}\sigma^2. \tag{6}$$

The first term is called the *bias* term and measures the discrepancy between f and $\mathcal{F}_J$. The second one is the *variance* term and would quantify the estimation error if f were belonging to $\mathcal{F}_J$. As expected, these two terms are monotonous functions of J: the former decreases as the latter increases. Thus, an ideal J is one which realizes the best compromise between these two terms, that is between approximation and complexity. Unfortunately, such a J is unknown since it depends on f. A natural purpose for a model selection procedure is therefore to provide an automatic choice of J, solely based on the data, in such a way that the selected J is as close as possible to the ideal one. Note that if the design is suitable, for those $J \geq n$ it is possible to find a function f_J in $\mathcal{F}_J$ whose values at the x_i's coincide with these of f. Thus, for those values of J, the least-squares estimators $\hat{f}_J$ and $\hat{f}_n$ have both the same risk, namely σ^2, and consequently, it becomes sufficient to look for the best least-squares estimator of f among those $\hat{f}_J$ for which $J \in \{1,\ldots,n\}$. Then, the natural collection of models to deal with is given by $\{\mathcal{F}_J,\ J \in \mathcal{M}_n\}$ where $\mathcal{M}_n = \{1,\ldots,n\}$.

EXAMPLE 1.2 (Variable selection). The problem of variable selection in Model 1 is to determine, from the observation of Y, which of the explanatory variables ϕ_j's for $j = 1,\ldots,N$ are influential. The word influential refers to an implicit trade-off. On the one hand, the number of explanatory variables which are retained must be small enough to warrant a reliable estimation of the

unknown parameters. On the other hand, if this number is too small, there is a chance to omit some important explanatory variables. More precisely, for each non void subset m of $\{1,\dots,N\}$, let $\mathcal{F}_m$ be the linear span generated by the column-vectors $\Phi_{.,j}$ for j in m and let $\hat{f}_m$ be the least-squares estimator of f on $\mathcal{F}_m$. For each $m \in \mathcal{M}_n$, $\hat{f}_m$ is defined as the orthogonal projection of Y onto $\mathcal{F}_m$ and its risk is given by

$$R_n(m) = \mathbb{E}\left[\mathrm{d}_n(f,\hat{f}_m)\right] = \mathrm{d}_n^2(f,\mathcal{F}_m) + \frac{|m|}{n}\sigma^2, \tag{7}$$

where $|m|$ denotes the cardinality of the set m. Thus, as in Example 1.1, the ideal m for estimating f realizes the best compromise between approximation and complexity. For this problem, the natural collection of models is $\{\mathcal{F}_m,\ m \in \mathcal{M}_n\}$ where $\mathcal{M}_n$ is the class of the non void subsets of $\{1,\dots,N\}$. However, the cardinality of this collection is very large for large values of N and some theoretical or computational difficulties may arise to solve the problem. Sometimes, when the explanatory variables can be ordered according to their presumed importance, it is easier to deal with a sub-collection of models, say $\{\mathcal{F}_m,\ m \in \mathcal{M}'_n\}$, where the set $\mathcal{M}'_n$ is totally ordered for the inclusion. In that case, we shall speak of *ordered variable selection*.

1.2. Parametric versus nonparametric. To solve these problems and others related to the choice of a proper model, many selection criteria were proposed. To our knowledge, the first ones originated from the papers by Mallows [**50**] and Akaike [**1**]. Given some collection of models $\{\mathcal{F}_m,\ m \in \mathcal{M}_n\}$, each criterion leads to a data driven choice of a $\hat{m}$ among $\mathcal{M}_n$. Originally, these criteria mainly relied on heuristics, one of the most famous being that associated to Mallows' C_p criterion (see Chapter 1 Section 2.1). Consequently, many efforts were made to justify these heuristics. In the literature, model selection criteria have mostly been studied for parametric and nonparametric inferences. Let us give an account of these.

Consider Model 1 where N is fixed and n tends to infinity. This setting is usually called *parametric* since f depends on a fixed number of unknown parameters as the number of observations tends to infinity. The problem is to determine the "exact model" for f, that is the linear space $\mathcal{F}_{m^*}$ corresponding to the subset m^* of $\{1,\dots,N\}$ gathering those indices j for which $\theta_j \neq 0$. For the problem at hand, the questions of interest are typically the following ones: what is the asymptotic probability to select m^* or a subset of $\{1,\dots,N\}$ containing it? Does the probability of selecting a subset which does not contain m^* tend to 0? and if so, how fast? These questions were, for example, addressed in the paper by Nishii [**53**] for numbers of model selection criteria including Mallows' C_p. Unfortunately, the problem of selecting m^* is different from that of variable selection presented in Example 1.2. In fact, the model which achieves

the best compromise between approximation and estimation error may coincide with $\mathcal{F}_{m^*}$ for none of the values of n. Assume for example that $N = 2$, $\Phi_{.,1} = (1, 0, \ldots, 0)^T$, $\Phi_{.,2} = (0, 1, 0, \ldots, 0)^T$, and $f = (\theta_1, \theta_2, 0, \ldots, 0)^T$ with $\theta_1 \neq 0$ and $0 < |\theta_2| < \sigma$, then $m^* = \{1, 2\}$ but the inequality

$$nR_n(\{1\}) = \theta_2^2 + \sigma^2 < 2\sigma^2 = nR_n(m^*)$$

shows that for all n the least-squares estimator performs better on the linear space $\mathcal{F}_{\{1\}}$ rather than on the exact model. The results obtained by Nishii (and others) provide thus little perspective on the problem of variable selection raised in Example 1.2.

In the opposite direction, Polyak and Tsybakov [**54**] addressed the problem of selecting the order of an expansion as described in Example 1.1. For this problem, Mallows' C_p criterion was put to the test (see also the earlier results of Shibata [**57**] and Li [**49**] for related problems). As the function to estimate was depending on possibly infinitely many parameters, the setting was said to be nonparametric. The point of view of Polyak and Tsybakov was also asymptotic, but in contrast with the parametric setting, the cardinality of their collection of models was increasing with the number of observations n. They showed that under suitable assumptions, the selected order $\hat{J} = \hat{J}_n$ was asymptotically allowing to balance at best the bias and the variance terms. Unfortunately, their results (as those of Shibata and Li) were requiring the unpleasant condition that the expansion (3) of f did not involve a finite number of non zero θ_j's. Thus, their setting was excluding the parametric one.

As a consequence, the results that had been established on the performance of Mallows' C_p were unsatisfactory in the sense that they were not providing a unified perspective reconciling both the parametric and nonparametric settings. Besides, the question of how Mallows' C_p behaved for fixed values of n was remaining open.

1.3. A new approach to model selection. One way to reconcile the parametric and nonparametric settings is to adopt a nonasymptotic point of view. Then, the distinction between these two settings vanishes. This point of view is ours and has been developed in various regression models (including the autoregression model) in the papers B. [**3, 2**], B., Comte and Viennet [**6, 7**]. It was earlier the one considered in the papers by Birgé and Massart [**15**] and Barron, Birgé and Massart [**12**]. These papers had been influenced by the notion of complexity regularization introduced in Barron and Cover [**14**] and extended in Barron [**13**]. Barron and Cover considered the problem of estimating a density function by selecting among a family Γ_n of countable candidates one which minimized some penalized likelihood criterion. Their approach was information-theoretically oriented. Some summability condition on Γ_n ensured that the

family of candidates could be encoded by a instantaneously decodable binary code (no codeword is the prefix of any other code word). For fixed n, the estimator was proven to realize, under some suitable assumptions, a trade-off among those $q \in \Gamma_n$ between accuracy (for the Kullback-Leiber distance between q and the true) and complexity (the length of the code for q renormalized by the number of observation n). These inequalities were, in a certain sense, the ancestors of those oracle inequalities established in [**12**] and described in the next section. Barron, Birgé and Massart gave an new impulse to model selection in their attempt to establish a general theory for minimum penalized contrast estimators. Such a theory provided a new perspective on model selection in various statistical frameworks including density estimation and regression. In particular, it allowed to relax the assumption imposed in both Barron and Cover [**14**] and Barron [**13**] that the number of candidate functions for estimating the "true" was countable. Our approach to model selection is based on their theory.

2. Model selection in $\mathbb{R}^n$

Models 1 and 2 can be handled simultaneously by studying the regression model

$$Y_i = f_i + \varepsilon_i,\ i = 1, \ldots, n, \tag{8}$$

where $f = (f_1, \ldots, f_n)^T$ is an unknown $\mathbb{R}^n$-vector and the ε_i's i.i.d. centered random variables of common variance σ^2. One recovers Model 1 by taking $f = \Phi\theta$. In the functional regression model, we shall identify the functions t on $[0,1]$ with the vectors $(t(x_1), \ldots, t(x_n))^T$. In particular, we shall not distinguish between the regression function f and the vector $(f(x_1), \ldots, f(x_n))^T$. We denote by $\mathbb{P}_f$ the law of the vector $(Y_1, \ldots, Y_n)^T$.

For the sake of simplicity, we assume here that σ^2 is known. Throughout this section, $\sqrt{n}\mathrm{d}_n(.,.)$ denotes the Euclidean distance in $\mathbb{R}^n$ and $\gamma_n(Y,.)$ the least-squares contrast function given for $t \in \mathbb{R}^n$ by

$$\gamma_n(Y,t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - t_i)^2\,. \tag{9}$$

2.1. Mallows' heuristic. The problems raised in Examples 1.1 and 1.2 can be translated in $\mathbb{R}^n$ in the following way. Given some collection $\{\mathcal{F}_m,\ m \in \mathcal{M}_n\}$ of linear subspaces of $\mathbb{R}^n$ and the corresponding sequence of least-squares estimators $\left\{\hat{f}_m,\ m \in \mathcal{M}_n\right\}$, how can we select *from the data* some $\hat{m}$ among $\mathcal{M}_n$ in such a way that the risk of $\hat{f}_{\hat{m}}$ is as close as possible to the minimal one among these estimators? In a more formal way, we would like the following

inequality to hold true for some universal constant C as close as possible to 1

$$\mathbb{E}\left[\mathrm{d}_n^2(f,\hat{f}_{\hat{m}})\right] \leq C \inf_{m\in\mathcal{M}_n} \mathbb{E}\left[\mathrm{d}_n^2(f,\hat{f}_m)\right]. \tag{10}$$

To solve the problem, the heuristic proposed by Mallows and known as Mallows' C_p relied on the following analysis of the risk. For each $m \in \mathcal{M}_n$, the risk of $\hat{f}_m$ is given by the formula

$$R_n(m) = \mathbb{E}\left[\mathrm{d}_n^2(f,\hat{f}_m)\right] = \mathrm{d}_n^2(f,\mathcal{F}_m) + \frac{D_m}{n}\sigma^2, \tag{11}$$

where $D_m = \dim(\mathcal{F}_m)$. On the other hand, by replacing f by Y in the left-hand side of (11) we have that

$$\mathbb{E}\left[\mathrm{d}_n^2(Y,\hat{f}_m)\right] = \mathbb{E}\left[\gamma_n(Y,\hat{f}_m)\right] = \mathrm{d}_n^2(f,\mathcal{F}_m) + \sigma^2 - \frac{D_m}{n}\sigma^2.$$

Consequently, by adding the term $2\sigma^2 D_m/n$ to the value $\gamma_n(Y,\hat{f}_m)$ we derive that

$$\mathbb{E}\left[\gamma_n(Y,\hat{f}_m) + 2\frac{D_m}{n}\sigma^2\right] - \sigma^2 = \mathbb{E}\left[\mathrm{d}_n^2(f,\hat{f}_m)\right],$$

and therefore, the index m which minimizes $R_n(m)$ also minimizes the expectation of the quantity

$$\mathrm{Crit}(m) = \gamma_n(Y,\hat{f}_m) + 2\frac{D_m}{n}\sigma^2. \tag{12}$$

Unlike $\mathrm{d}_n^2(f,\hat{f}_m)$, $\mathrm{Crit}(m)$ does not depend on f but only on the data. Thus, the idea of Mallows is to select $\hat{m}$ as the minimizer of $\mathrm{Crit}(m)$ among those $m \in \mathcal{M}_n$. This choice is of course reasonable if for all $m \in \mathcal{M}_n$, $\mathrm{Crit}(m)$ is close to its expectation.

The procedure we propose in B. [**3**] generalizes that by Mallows. Given some $\eta > 0$, we study the performance of the criterion given by

$$\mathrm{Crit}(m) = \gamma_n(Y,\hat{f}_m) + (1+\eta)\frac{D_m}{n}\sigma^2, \tag{13}$$

the choice $\eta = 1$ corresponding thus to Mallows' C_p. As expected, we select $\hat{m}$ as

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \mathrm{Crit}(m). \tag{14}$$

2.2. From model selection to oracle inequalities. Inequalities such as (10) are called *oracle inequalities* as defined by Donoho and Johnstone [**26**]. From a general point of view, they show that the risk of the estimator of interest is comparable to the minimal one among a family of estimators given beforehand. Unfortunately, it is not always possible to obtain an oracle inequality without any condition on the collection of estimators at hand. In the regression model

given by (8), assume for example that the ε_i's are Gaussian, so that the probability distributions $\{\mathbb{P}_f,\ f \in \mathbb{R}^n\}$ are equivalent and consider a collection of models $\{\mathcal{F}_m,\ m \in \mathcal{M}_n\}$ containing $\{0\}$. The least-squares estimator of f on this particular model is necessarily 0 whatever the values of the data. If the oracle inequality (10) were true, then $\tilde{f}$ and 0 would be equal a.s. under $\mathbb{P}_0$, and therefore, under any probability $\mathbb{P}_f$ with $f \in \mathbb{R}^n$. In other words, the selection criterion would always select 0!

The "oracle inequality" we get in B. [**3**] has the following form.

THEOREM 1. *Assume that $\tau_p = \mathbb{E}\left[|\varepsilon_1|^p\right]$ is finite for some $p > 4$. Selecting $\hat{m}$ as in (14) results in the estimator $\tilde{f} = \hat{f}_{\hat{m}}$ which satisfies for some constant κ the inequality*

$$\mathbb{E}\left[\mathrm{d}_n^2(f, \tilde{f})\right] \leq \kappa \left[\inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\mathrm{d}_n^2(f, \hat{f}_m)\right] + \Delta_n \frac{\sigma^2}{n}\right]. \tag{15}$$

where

$$\Delta_n = \frac{\tau_p}{\sigma^p}\left(1 + \sum_{\substack{m \in \mathcal{M}_n, \\ D_m \geq 1}} D_m^{-(p/2-2)}\right).$$

The constant κ only depends on η and p.

The difference between Inequalities (10) and (15) lies in the presence of the remaining term $\Delta_n \sigma^2/n$ in the latter. The quantity Δ_n measures in some sense the complexity of the collection of models with respect to the integrability of the errors. If Δ_n can be bounded for all values of n by some positive constant Δ, the remaining term is of order $\Delta\sigma^2/n$ and is thus small if $\Delta\sigma^2$ is small or n large enough. In this case, an oracle inequality can be deduced, assuming (as expected) that for all $m \in \mathcal{M}_n$ $\mathcal{F}_m \neq \{0\}$, since then the inequality $\mathbb{E}\left[\mathrm{d}_n^2(f, \hat{f}_m)\right] \geq \sigma^2/n$ (from (11)) ensures that the remaining term is dominated (up to a constant depending on Δ) by the quantity $\inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\mathrm{d}_n^2(f, \hat{f}_m)\right]$. Then, we deduce form (15) the oracle inequality

$$\mathbb{E}\left[\mathrm{d}_n^2(f, \tilde{f})\right] \leq C \inf_{m \in \mathcal{M}_n} \mathbb{E}\left[\mathrm{d}_n^2(f, \hat{f}_m)\right] \tag{16}$$

$$= C \inf_{m \in \mathcal{M}_n} \left[\mathrm{d}_n^2(f, \mathcal{F}_m) + \frac{D_m}{n}\sigma^2\right] \tag{17}$$

where $C = \kappa(1 + \Delta)$.

As an application of this result, let us consider the problem of selecting the order of an expansion described in Example 1.1 (doing the identification between functions and vectors). The collection of models of interest contains

here at most one model per dimension. Assuming that τ_p is finite for some $p > 6$, we have

$$\begin{aligned}\Delta_n &= \frac{\tau_p}{\sigma^p}\left(1+\sum_{J=1}^{n} D_J^{-(p/2-2)}\right)\\ &\leq \frac{\tau_p}{\sigma^p}\left(1+\sum_{D\geq 1} D^{-(p/2-2)}\right) = \Delta < +\infty,\end{aligned}$$

which leads to the oracle inequality (16).

For the problem of variable selection described in Example 1.2, the collection of models is too complex to obtain a uniform upper bound on Δ_n. In fact, for this problem oracle inequalities are impossible to establish (see Donoho and Johnstone [**26**]). Nevertheless, the problem of ordered variable selection can be handled by arguing as previously.

2.3. Model selection, adaptive estimation and approximation theory. The aim of this section is to shed light on some connections between model selection, adaptive estimation and approximation theory.

Throughout this section, we consider Model 2 and assume that f belongs to some class $\mathcal{F}$ of functions. A classical way to compare the performances of two estimators of f is to compare the supremum of their risks (say with respect to pseudo distance $\mathrm{d}_n(.,.)$ defined by (5)) when f varies among $\mathcal{F}$. This point of view is called *minimax*. For an estimator $\hat{f}$ of f, the maximal risk of $\hat{f}$ over $\mathcal{F}$ is thus defined by

$$R_{\max}(n,\mathcal{F},\hat{f}) = \sup_{f\in\mathcal{F}} \mathbb{E}\left[\mathrm{d}_n^2(f,\hat{f})\right].$$

The infimum, $R_{\min}(n,\mathcal{F})$, of $R_{\max}(n,\mathcal{F},\hat{f})$ when $\hat{f}$ varies among the whole set of estimators is called the minimax risk over $\mathcal{F}$. An estimator is usually said to be minimax if its maximal risk over $\mathcal{F}$ is comparable (up to a constant depending on $\mathcal{F}$) to the minimax one as n becomes large.

For illustration, let us consider the class of Hölderian functions $\mathcal{F} = \mathcal{H}_s(R)$ defined for $s \in]0,1]$ and $R > 0$ by

$$\mathcal{H}_s(R) = \{g/\ |g(x)-g(y)| \leq R|x-y|^s,\ \forall x,y\in[0,1]\}. \tag{18}$$

For this functional class, the minimax risk is known to be of order $n^{-2s/(1+2s)}$ (when the design is regular in $[0,1]$) and a minimax estimator can easily be obtained by arguing as follows. Let $\mathcal{M}_n = \{1,\ldots,n\}$ and define for each m in $\mathcal{M}_n$ the space $\mathcal{F}_m$ as the linear span generated by the piecewise constant functions over the regular partition of $[0,1]$ into m pieces. For each $m \in \mathcal{M}_n$,

the risk of the least-squares estimator, $\hat{f}_m$, of f onto $\mathcal{F}_m$ satisfies

$$\mathbb{E}\left[\mathrm{d}_n^2(f,\hat{f}_m)\right] \leq \mathrm{d}_n^2(f,\mathcal{F}_m) + \frac{m}{n}\sigma^2. \tag{19}$$

Some simple calculation shows that the distance in sup-norm, and therefore with respect to $\mathrm{d}_n(.,.)$, between f and $\mathcal{F}_m$ is not larger than Rm^{-s}. Consequently, we deduce from (19) that the risk of $\hat{f}_m$ satisfies for all f in $\mathcal{H}_s(R)$

$$\mathbb{E}\left[\mathrm{d}_n^2(f,\hat{f}_m)\right] \leq C(R,\sigma^2)\left(m^{-2s} + \frac{m}{n}\right).$$

As the right-hand side of this inequality is now free of f, it becomes possible to obtain an explicit value of $m = m^*$ for which the bias and the variance terms are balanced. This value is given by m^* of order $n^{1/(1+2s)}$ and thus, when s is known, the estimator $\hat{f}_{m^*}$ is proven to converge at minimax rate $n^{-2s/(1+2s)}$. This method was known from Grenander [**34**] as *the method of sieves* and has been studied by Stone [**60**], Shen and Wong [**56**] and van de Geer [**63**] among other references.

This estimation procedure has the drawback to depend, via the choice of m^*, on the prior information that f belongs to some known class of smooth functions. By using a model selection procedure, we can take advantage of an automatic choice of m among $\mathcal{M}_n$. More precisely, by considering the collection of models $\{\mathcal{F}_m,\ m \in \mathcal{M}_n\}$ described above and by using Inequality (17) (which holds true as soon as ε_1 admits a moment of order larger than 6) the optimality of $\tilde{f}$ can be deduced in the same way by bounding the quantity

$$a_n(f) = \inf_{m\in\mathcal{M}_n}\left[\mathrm{d}_n^2(f,\mathcal{F}_m) + \frac{D_m}{n}\sigma^2\right]. \tag{20}$$

It is worth emphasizing the fact that the optimality of $\tilde{f}$ is now obtained without any prior knowledge of s! Such a property is known as *adaptation in the minimax sense*. To our knowledge, the first result of adaptation originated from the paper by Efroimovich and Pinsker [**28**]. Adaptation in the minimax sense was intensively studied in the 90's. Let us mention the work of Lepski [**47, 48**], Donoho and Johnstone [**24**] and refer to Donoho and Johnstone [**25**] and Barron, Birgé and Massart [**12**] (Section 5) for more references on this topic.

The quantity $a_n(f)$ is usually called the *accuracy index*. A nice feature of Inequality (17) is that it allows an immediate connection between the approximation properties of the collection and the rate of convergence of the estimator. As a consequence, the problem of adaptation can easily by solved by considering collections of models with suitable approximation properties. The problem of building adaptive estimators over Besov balls is among the most challenging ones in statistics. These balls, usually denoted by $\mathcal{B}_{s,p,\infty}(R)$ with $s > \max\{1/p - 1/2, 0\}$, $p \in [1,\infty]$ and $R > 0$, generalize the Hölderian and Sobolev balls often encountered in functional analysis. For a precise definition

of these balls we refer to the book of DeVore and Lorentz [**23**]. The approximation of Besov balls (with respect to the $\mathbb{L}^2([0,1],dx)$-distance $\mathrm{d}(.,.)$) by linear spaces based on trigonometric polynomials, wavelets or splines is available in the literature. Various results can indeed be found in the books by Lorentz and DeVore [**23**] or Meyer [**51**] (for wavelets). This makes it possible to deduce uniform rates of convergence of $\tilde{f}$ over these Besov balls whenever the distances $\mathrm{d}(.,.)$ and $\mathrm{d}_n(.,.)$ are comparable.

2.4. Model selection and empirical process theory. In Barron, Birgé and Massart [**12**], the construction of inequalities such as (15) relied on techniques of empirical processes. In our work as in others based on their approach, the use of the empirical process theory is at the heart of the proofs. The aim of this section is to describe how empirical processes get involved and what are the probabilistic tools allowing their stochastic control. For convenience, we shall restrict ourself to the functional regression case described in Model 2 and use the notations introduced there.

Let us note that the least-squares contrast function (4) has some nice properties. First, it relates to the distance $\mathrm{d}_n(.,.)$ (defined by (5)) in the following way. For all functions s,t on $[0,1]$

$$\mathbb{E}_f\left[\gamma_n(Z,t)-\gamma_n(Z,s)\right]=\mathrm{d}_n^2(t,f)-\mathrm{d}_n^2(s,f).$$

Moreover, the centered process

$$\gamma_n(Z,t)-\gamma_n(Z,s)-\mathbb{E}_f\left[\gamma_n(Z,t)-\gamma_n(Z,s)\right]=-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(t(x_i)-s(x_i))$$

is a linear function of $t-s$. We denote it by $-\nu_n(Z,t-s)$. Consequently, we have that for all s,t

$$\mathrm{d}_n^2(f,t)=\mathrm{d}_n^2(f,s)+\nu_n(Z,t-s)+\gamma_n(Z,t)-\gamma_n(Z,s). \tag{21}$$

This decomposition was obtained in van de Geer [**62**]. In the functional setting, she used it to study the rate of convergence of the least-squares estimator $\hat{f}$ of the function f on a given functional class $\mathcal{F}$. She showed how the entropy structure of the space $\mathcal{F}$ could be related to the rate of convergence of the estimator $\hat{f}$ towards f. By substituting s for f and t for $\hat{f}$ and using that $\hat{f}$ satisfies $\gamma_n(Z,\hat{f})\leq\gamma_n(Z,f)$, one derives from (21) that $\mathrm{d}_n^2(f,\hat{f})\leq\nu_n(Z,\hat{f}-f)$. This inequality shows that the distance between f and its estimator is bounded from above by the value of a centered empirical process on $\mathcal{F}$, namely $g\to\nu_n(Z,g-f)$, at the random point $g=\hat{f}$. For a suitable normalization $N(.)$, she controlled this quantity thanks to a deviation inequality on the supremum of the empirical process $g\to\nu_n(Z,g-f)/N(g)$ over a small ball around f. She obtained this inequality via chaining arguments.

One of the differences between her computations and ours is that we deal with a collection of $\mathcal{F}_m$'s and not only one. Let us fix some arbitrary m in $\mathcal{M}_n$ and denote by f_m the closest element to f in $\mathcal{F}_m$ (with respect to $\mathrm{d}_n(.,.)$). By definition of $\hat{f}_{\hat{m}}$, we have that

$$\gamma_n(Z, \hat{f}_{\hat{m}}) - \gamma_n(Z, f_m) \leq \mathrm{pen}(\hat{m}) - \mathrm{pen}(m)$$

and therefore, we deduce from (21) that

$$\mathrm{d}_n^2(f, \hat{f}_{\hat{m}}) \leq \mathrm{d}_n^2(f, f_m) + \mathrm{pen}(m) + N_m(\hat{f}_{\hat{m}})\mathcal{U}_{\hat{m}} - \mathrm{pen}(\hat{m}),$$

where the $N_{m'}(.)$'s are suitable normalizations and the $\mathcal{U}_{m'}$'s, random variables given by

$$\mathcal{U}_{m'} = \sup_{g \in \mathcal{F}_{m'}} \nu_n \left(Z, \frac{g - f_m}{N_m(g)} \right). \tag{22}$$

The control of the random variable $\mathcal{U}_{\hat{m}}$ is obtained by controlling $\mathcal{U}_{m'}$ for all $m' \in \mathcal{M}_n$ simultaneously. This control influences the choice of the penalty term since ideally we would like to have $N_m(\hat{f}_{\hat{m}})\mathcal{U}_{\hat{m}} - \mathrm{pen}(\hat{m}) \leq 0$.

Suprema of empirical processes not only arose in our work but also in that of Barron, Birgé and Massart [**12**] and in Birgé and Massart [**15**]. These two papers differ, however, in the way they controlled these suprema. In the former, an inequality based on chaining arguments was established whereas the latter used, for the first time in Statistics to our knowledge, a concentration inequality. This inequality had been established by Talagrand [**61**] (Theorem 1.4) and applied for the control the fluctuations of suprema of *bounded* empirical processes around their means. Since then, the use of concentration inequalities has become very popular in Statistics and has contributed to the development of model selection, see the work of Castellan [**19**] in density estimation and that of Reynaud [**55**] in the estimation of the intensity of an inhomogeneous Poisson process.

Conversely, the development of model selection in Statistics has allowed the emergence of new problems connected with the control of suprema of empirical processes around their means and thus, has simulated in turn research in empirical process theory. An illustrative example can be found in B. [**3**]. In fact, the normalizations $N_m(.)$ chosen in this paper are merely given by $\mathrm{d}_n(., f_m)$ and consequently, the random variables $\mathcal{U}_m$ take the simple form

$$\mathcal{U} = \frac{2}{n} \sup_{t \in \mathcal{B}} \sum_{i=1}^{n} h_t(Z_i) \tag{23}$$

where for all t, $h_t(Z_i) = \varepsilon_i t(x_i)$ and $\mathcal{B}$ is the set defined by

$$\mathcal{B} = \left\{ t \in S,\ \frac{1}{n} \sum_{i=1}^{n} t^2(x_i) \leq 1 \right\},$$

for some linear space S. Unless one assumes that the ε_i's take their values inside a compact interval, the random variable $\mathcal{U}$ is the supremum of an empirical process over an *unbounded* class of functions and consequently, its fluctuations around its expectation cannot be controlled by using Talagrand's Bennett-type inequality. To overcome this problem, we established the following inequality that derived from Talagrand's.

THEOREM 2. *Let $Z_1, \ldots, Z_n$ be independent random variables with values in some measurable space $\mathcal{E}$ and $\mathcal{H}$ be some countable class of real-valued measurable functions on $\mathcal{E}$. Let us set*

$$\mathcal{U} = \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} h(Z_i) \quad \text{or} \quad \mathcal{U} = \sup_{h \in \mathcal{H}} \left| \sum_{i=1}^{n} h(Z_i) \right|.$$

Then, in both cases, for all $p \geq 2$

$$C^{-1}\mathbb{E}\left[|\mathcal{U} - \mathbb{E}[\mathcal{U}]|^p\right] \leq \mathbb{E}\left[\max_{i=1,\ldots,n} \sup_{h \in \mathcal{H}} |h(Z_i)|^p\right] + \mathbb{E}^{p/2}\left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} h^2(Z_i)\right],$$

where C is a positive constant depending on p only.

This inequality generalizes to suprema of empirical processes an inequality on sums of centered independent random variables known as Rosenthal's inequality. The latter can be recovered by reducing $\mathcal{H}$ to a singleton. We derive from Theorem 2 and Markov's inequality a concentration inequality of the random variable $\mathcal{U}$ around its expectation. More generally, this inequality can be applied to suprema of empirical processes over classes of functions possessing an envelope in $\mathbb{L}_p$.

3. Regression with random designs

In experimental sciences, Model 2 usually arises when the design of the experiment is decided before it is actually carried out. For example, in studying the growth y of a plant as a function of the concentration x of a chemical substance in the soil, the values of the x_i's can naturally be predetermined by the experimenter. In other situations, the values of the response y and the quantitative variable x are known simultaneously, the experimenter having no control on the latter. This is the case in the study of the correlation between height and weight in a population for example. For these situations, the following regression model becomes more appropriate.

MODEL 3 (Regression with independent random design points).

$$Y_i = f(X_i) + \varepsilon_i, \ i = 1, \ldots, n \tag{24}$$

where the X_i's are independent real-valued random variables and the ε_i i.i.d. centered random variables. The sequences of X_i's and ε_i's are independent. The data consists of the pairs $Z_i = (Y_i, X_i)$ for $i = 1, \ldots, n$.

This model is considered in B. [**2**] and we propose to give an account of the results therein. Throughout this section we assume, for the sake of simplicity, that the X_i's are i.i.d. uniformly distributed in $[0,1]$ and we denote by $\mathrm{d}(.,.)$ and $\| \ \|$ respectively the distance and the norm associated to the Hilbert space $\mathbb{L}^2([0,1], dx)$. All along we shall assume that f belongs to $\mathbb{L}^2([0,1], dx)$.

3.1. The problem at hand. The problem we consider in this section is that of estimating f and to establish risk bounds with respect to the distance $\mathrm{d}(.,.)$ for the proposed estimator. Thus, if $\hat{f}$ is an estimator of f, we measure its performance by looking at its risk $\mathbb{E}\left[\mathrm{d}^2(f, \hat{f})\right]$.

This problem was considered by several authors among which Barron, Birgé and Massart [**12**], Kohler [**42**], Wegkamp [**65**], Yang [**66**] and Catoni [**20**]. The common idea of these authors is to build an estimator which achieves (up to a constant) the minimal risk among a family of estimators given beforehand. Besides assuming strong integrability conditions on the errors, a common feature of these papers was that the proposed estimators were depending on a known upper bound B on the sup-norm, $\|f\|_\infty$, of the regression function. Furthermore, the risk bounds established for these estimators involve constants that are increasing functions of B. Consequently, these constants become large when one chooses B large enough to satisfy the condition $\|f\|_\infty \leq B$. These estimators were proved to possess minimax properties. However, because of the dependency of the estimation procedure on the sup-norm of the target function, these results were unfortunately restricted to those functional classes which were uniformly bounded in sup-norm. This is for example the case for those Besov balls $\mathcal{B}_{s,p,\infty}(R)$ with $s > 1/p$ and $p \in [1, +\infty]$. Surprisingly, the minimax rates over the Besov balls $\mathcal{B}_{s,p,\infty}(R)$ for which $s < 1/p$ (which are not uniformly bounded in sup-norm) had, to our knowledge, never been described in the regression framework with random design points. For each $s < 1/p$, only the lower bound $n^{-2s/(1+2s)}$ was available from the papers by Korostelev and Tsybakov [**43**] or Yang and Barron [**67**].

In B. [**2**], we relaxed the assumption that the regression function was bounded in sup-norm and thus, overcame the problems encountered earlier by the previous authors.

3.2. The estimation procedure. We start with a collection of finite dimensional subspaces $\{\mathcal{F}_m, \ m \in \mathcal{M}_n\}$ of $\mathbb{L}^2([0,1], dx)$ and associate to each

$m \in \mathcal{M}_n$ the least squares estimator $\hat{f}_m$ of f onto $\mathcal{F}_m$, that is the minimizer among those $t \in \mathcal{F}_m$ of the contrast

$$\gamma_n(Z,t) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - t(X_i)\right)^2. \tag{25}$$

Given some positive η, we select $\hat{m}$ in $\mathcal{M}_n$ as the minimizer of the penalized criterion

$$\mathrm{Crit}(m) = \gamma_n(Z,\hat{f}_m) + (1+\eta)\frac{D_m}{n}\sigma^2. \tag{26}$$

By setting $k_n = 2\exp\left(\ln^2(n)\right)$, we define our estimator $\tilde{f}$ as follows

$$\tilde{f} = \hat{f}_{\hat{m}} \text{ if } \|\hat{f}_{\hat{m}}\| \leq k_n \text{ and } \tilde{f} = 0 \text{ otherwise.}$$

3.3. Oracle types inequalities. Under the condition that $\|f\|^2 \leq K\exp\left(2\ln^2(n)\right)\sigma^2/n$ (where K is a positive constant) and suitable assumptions (that we shall discuss later) on the collection of models and the distribution of the errors, we obtain the following inequality

$$\mathbb{E}\left[\mathrm{d}^2(f,\tilde{f})\right] \leq C \inf_{m\in\mathcal{M}_n}\left[\mathrm{d}^2(f,\mathcal{F}_m) + \frac{D_m}{n}\sigma^2\right], \tag{27}$$

where C is a constant which does not depend on f and n. This inequality relates to (17). However, it is not an oracle inequality as the quantity $\mathrm{d}^2(f,\mathcal{F}_m) + D_m\sigma^2/n$ is not necessarily of the same order as the risk of $\hat{f}_m$, $\mathbb{E}\left[\mathrm{d}^2(f,\hat{f}_m)\right]$. In fact, there are situations for which the former quantity is finite and the latter infinite. Therefore, we call Inequality (27) an oracle-type inequality.

Let us now give some comments on this result. First, the estimator we consider is not the estimator $\hat{f}_{\hat{m}}$ as in the previous section, but rather a truncated version. By doing so, the risk of our estimator remains bounded even for an unlucky draw of the X_i's. However, we want, of course, $\tilde{f}$ and 0 to coincide only on an exceptional set and this is the reason why the constraint that $\|f\|$ is not too large is required. Note that this constraint holds true at least for large values of n as f is assumed to belong to $\mathbb{L}^2([0,1],dx)$.

Second, we have established this result under both Gaussian and weak moment conditions on the errors. In the former case, the condition on the collection of models is less restrictive than in the former, as expected. More precisely, in the second case the collection is assumed to contain only a few models of the same dimension which is actually enough to handle the problem of selecting the order of an expansion as already seen in Section 2.2. On the other hand, when the errors are Gaussian, the collection is allowed to contain, for each $D \geq 1$, $e^{L(\eta)D}$ models of dimension D where $L(\eta)$ is an increasing function of η. For a suitable choice of η, we take advantage of this possibility to deal with the

collections of models described in Birgé and Massart [**16**] which possess remarkable approximation properties with respect to those Besov balls $\mathcal{B}_{s,p,\infty}(R)$ with $p \geq 1$, $R > 0$ and $s > \max\{1/p - 1/2, 0\}$. By arguing as in Section 2.3, these properties allow us to deduce that the estimator $\tilde{f}$ achieves the rate $n^{-2s/(1+2s)}$ over each Besov ball $\mathcal{B}_{s,p,\infty}(R)$ for which $p \geq 1$ and $s > s_p$ where s_p is a nonnegative number (made explicit in B. [**2**]) satisfying

$$\max\left\{0, \frac{1}{p} - \frac{1}{2}\right\} \leq s_p < \frac{1}{p}.$$

This result combined with the lower bounds we mentioned previously shows that the minimax rate of estimation over the Besov ball $\mathcal{B}_{s,p,\infty}(R)$ is still of order $n^{-2s/(1+2s)}$ for those $s \in]s_p, 1/p]$. We conjecture that the minimax rates are different for $s < s_p$. Besides describing the minimax rates, our estimation procedure provides the first adaptive estimator over these Besov spaces.

Finally, let us say a few words of the proof of (27). The difficulty mainly lies in the control of the random variable $\mathcal{U}$ defined by (23). This control can no longer be deduced from Theorem 2 since the set $\mathcal{B}$ is now random (for all $i = 1, \ldots, n$, $x_i = X_i$). We overcame this problem by establishing some connections between the distances $\mathrm{d}(.,.)$ and $\mathrm{d}_n(.,.)$ (defined by (5) with X_i in place of x_i). Note that the latter is thus random. On each linear space $\mathcal{F}_m$, we show that these two distances are close to one another provided that some conditions on the dimension of $\mathcal{F}_m$ and the structure of its $\mathbb{L}^2(dx)$-orthonormal bases are fulfilled. The technical problems we have met in this framework were also met in those considered in B., Comte, Viennet [**6, 7**].

3.4. The autoregression framework. In time series, the index i represents a time of experiment, the result of the experiment at time i being an influential factor for the result at time $i + 1$. The evolution of temperature on earth from day to day provides a simple illustrative example. For such a phenomenon, one of the simplest models is given by the autoregression model of order 1 described as follows.

MODEL 4 (The AR(1) model).

$$X_0,\ Y_i = X_i = f(X_{i-1}) + \varepsilon_i,\ i = 1, \ldots, n. \tag{28}$$

The X_i's are real-valued random variables and X_0 is independent of ε_1. The ε_i's are i.i.d. centered random variables. The observations consist of the sequence of random variables X_i for $i = 1, \ldots, n$. For this model, the difficulty, as compared with Model 3, lies in the dependency between the data.

Estimation by model selection in this model (and others) is considered in B., Comte and Viennet [**6, 7**].

In B., Comte and Viennet [**6**], we assume that the X_i's are geometrically β-mixing. This property is usually met under the same assumptions on f and the distribution of the ε_i's which ensure the existence of a stationary law for the Markovian process defined by (28). Under weak moment assumptions on the errors we establish oracle-type inequalities for the model selection criterion given by (26). As before, we deduce from these optimality properties in the minimax sense for our estimator. As the proof of our results only relies on the β-mixing properties of the data, it can be generalized to other regression models among which regression models such as Model 3 in which both the X_i's and the ε_i's are β-mixing. In this case, we show that our model selection procedure is still relevant provided that η in (13) is large enough. This result shows thus the robustness of our estimation procedure with respect to a possible dependency of the ε_i's.

The proof of our oracle-type inequality is based on a Rosenthal-type inequality for suprema of empirical processes based on β-mixing data. It is established by combining the results of B. [**3, 2**] with those of Viennet [**64**] on β-mixing random variables.

Because of the weak moment assumptions on the errors, in B., Comte, Viennet [**6**], the collection of models considered there only contain few models of the same dimension. More general models are considered in B., Comte, Viennet [**7**] at the price of a more restrictive assumption on the law of the ε_i's. It is assumed there to be sub-Gaussian, that is to satisfy for some $s > 0$ the inequality

$$\mathbb{E}\left[\exp(u\varepsilon_1)\right] \leq \exp\left(\frac{u^2 s^2}{2}\right) \ \forall u \in \mathbb{R}.$$

This conditions is fulfilled for bounded or Gaussian ε_i's. Then, it becomes possible to consider a collection of models for which our estimator is proven to be adaptive over a large family of Besov spaces as in B. [**2**]. In contrast to B., Comte, Viennet [**6**], the techniques allow to relax the geometrical β-mixing assumption on the design. The proof relies on the martingale structure of the random variable $\sum_{i=1}^{n} \varepsilon_i t(X_i)$.

CHAPTER 2

Hypothesis testing

1. Introduction

1.1. What is a test? In the sequel we shall call *test* any measurable function ϕ of the data which takes its values in $\{0,1\}$. Tests are used in Statistics to validate some "intuition" about the probability distribution $\mathbb{P}$ of the data. More precisely, let α be some number in $]0,1[$ and let us consider two disjoint families of probabilities $\mathcal{P}_0$ and $\mathcal{P}_1$ such that $\mathbb{P}$ belongs to $\mathcal{P}_0 \cup \mathcal{P}_1$. A test ϕ aims at deciding which of these two families contains $\mathbb{P}$ under the constraint that the probability of error must not exceed α if $\mathbb{P}$ actually belongs to $\mathcal{P}_0$. We then say that ϕ is a level α test of the null hypothesis "$\mathbb{P} \in \mathcal{P}_0$" against the alternative "$\mathbb{P} \in \mathcal{P}_1$". By convention, we shall say that the test ϕ accepts the null (hypothesis) when $\phi = 0$.

In this section we consider the regression model given by

MODEL 5 (The Gaussian regression model).

$$Y_i = f_i + \sigma\varepsilon_i,\ i = 1, \dots, n,$$

where the f_i's are unknown real numbers, the ε_i's are i.i.d. standard Gaussian random variables and σ some unknown positive quantity. The data is given by the observation of the vector $Y = (Y_1, \dots, Y_n)^T$ which is distributed as a Gaussian vector of mean $f = (f_1, \dots, f_n)^T$ and covariance matrix $\sigma^2 I_n$ (where I_n denotes the identity matrix in $\mathbb{R}^n$). We shall denote by $\mathbb{P}_{f,\sigma^2}$ this distribution. This regression model corresponds to both Model 1 and Model 2 when the errors in these models are assumed to be Gaussian. Throughout this chapter, when considering Model 2, we shall denote by F (in place of f) the regression function.

Given some subset $\mathcal{C}$ of $\mathbb{R}^n$, we consider the problem of testing $\mathcal{P}_0$ against $\mathcal{P}_1$ where $\mathcal{P}_0$ and $\mathcal{P}_1$ are respectively given by

$$\mathcal{P}_0 = \{\mathbb{P}_{f,\tau},\ f \in \mathcal{C}, \tau > 0\} \quad \text{and } \mathcal{P}_1 = \{\mathbb{P}_{f,\tau},\ f \in \mathbb{R}^n \setminus \mathcal{C}, \tau > 0\}\,.$$

For simplicity, we shall merely say that we test the null hypothesis "$f \in \mathcal{C}$" against "$f \notin \mathcal{C}$" and call *alternative* any subset of $\mathbb{R}^n$ disjoint from $\mathcal{C}$.

The comparison of two tests of a same level, say α, can be done in various ways. For each $f \notin \mathcal{C}$, one way is the compare under $\mathbb{P}_{f,\sigma^2}$ the probabilities of rejection, also called *powers*, of these tests. A test is all the more powerful that its power is close to one over a larger class of alternatives. Let us now fix some $\beta \in]0, 1-\alpha[$. Given some alternative $\mathcal{A}$ and some distance $\delta(.,.)$ in $\mathbb{R}^n$, another way is to look at the smallest positive number ρ such that the test rejects the null with probability larger than $1-\beta$ for all those f in $\mathcal{A}$ at distance larger than ρ from $\mathcal{C}$. We shall call this quantity the δ-separation rate of the test over $\mathcal{A}$ and for a given test ϕ denote it by $\rho_n(\phi, \mathcal{A})$. To keep the notation as simple as possible the dependency with respect to α, β and $\mathcal{C}$ are omitted. Comparing the performances of two tests over an alternative $\mathcal{A}$ amounts then to comparing their δ-separation rates. An interesting quantity is the infimum of those δ-separation rates over all possible level-α tests. It describes the optimal separation rate over a given alternative. We shall call it the δ-minimax separation rate over $\mathcal{A}$, or shortly the minimax separation rate when δ is clearly specified in the text. We denote it by $\rho_n(\mathcal{A})$.

1.2. Case of a linear hypothesis. When $\mathcal{C} = V$ is a linear subspace of $\mathbb{R}^n$, one of the most famous test is the Fisher test. More precisely, the Fisher test is classically used to test the null hypothesis "$f \in V$" against the alternative "$f \in W \setminus V$" where W is a linear subspace of $\mathbb{R}^n$ satisfying $V \subsetneq W \subsetneq \mathbb{R}^n$. Denoting by Π_W and Π_V the orthogonal projectors onto W and V, the Fisher test consists in rejecting the null when the ratio $\hat{T} = \mathrm{d}_n^2(\Pi_W Y, \Pi_V Y)/\mathrm{d}_n^2(Y, \Pi_W Y)$ is large enough. The advantages of this test are twofold. First, it is easy to implement in practice. The law of the statistic $\hat{T}$ under the null only depends on n and the dimensions of the linear spaces W and V. In particular, it does not depend on the unknown variance of the errors. This makes it possible to establish a table of quantiles for $\hat{T}$ under the null. Finally, when $\dim(W)/n$ is small compared to one, the d_n-separation rate of the Fisher over $W \setminus V$ turns to be small, making thus the testing procedure powerful over a large subset of $W \setminus V$.

Unfortunately, the use of the Fisher test requires that some alternative set for f is specified. Moreover, the choice of this alternative has a great influence on the issue of the test. Consider, for example, two possible alternatives , $W \setminus V$ and $W' \setminus V$, for which the dimension of W is very small compared to that of W'. By choosing the first alternative, the power of the resulting Fisher test over $W' \cap W^{\perp}$ will not exceed its level. By choosing the second alternative, the Fisher test will behave very poorly over $W \setminus V$ compared to the former. In other words, f being unknown, the problem of choosing among a sequence of alternatives the one over which the Fisher test achieves its greatest power is not an easy task. The alternative must not be too large for the test to be powerful but also not too small for f to be, at least, close to this alternative. It is worth

mentioning that this dilemma closely relates to that which we mentioned for the problem of estimation. Thus, it is no wonder that procedures based on model selection naturally arose in hypothesis testing.

This is the case for the procedure proposed by Eubank and Hart [**30**]. Let us give an account of their approach in the particular case where $V = \{0\}$. Their statistical model is given by Model 2 where the regression function, F, is assumed to belong to $\mathbb{L}^2([0,1], dx)$. They consider the expansion of F onto some suitable basis, say the sine and cosine system, and by means of a model selection criterion select the order of the truncated Fourier series estimator for F. The model selection criterion is supposed to choose an order close to the optimal one, that is, as explained in Example 1.1, one which realizes the best trade-off between approximation and estimation error. The selecting order $\hat{J}$ has the possibility to be 0 which then leads to estimate F by 0. This is of course the best estimation of F under the null, the procedure rejects thus when $\hat{J} \geq 1$. One of the main advantage of this approach is that it gives the opportunity to revise the modelisation when the hypothesis is rejected, suggesting that some terms in the expansion have not been taken into account. Unfortunately, little is known on the optimality of this procedure in the minimax sense...

For the problem considered by Eubank and Hart, natural alternative sets for f take the form

$$\mathcal{A} = \left\{(F(x_1), \ldots, F(x_n))^T,\ F \in \mathcal{K}\right\} \setminus \{0\},$$

where $\mathcal{K}$ is a set of smooth functions. Let $\sqrt{n}\mathrm{d}_n(.,.)$ be the Euclidean distance in $\mathbb{R}^n$. When $\mathcal{K}$ is a linear space of dimension 1, the d_n-separation rate over $\mathcal{A}$ is of order $1/\sqrt{n}$. A test which achieves this rate over any of those $\mathcal{A}$ is said to be $(1/\sqrt{n})$-consistent over directional alternatives. The procedure proposed by Eubank and Hart possesses such a property. However, nothing is known on this d_n-separation rate when $\mathcal{K}$ is a more general functional set as, for example, a Hölder or Sobolev ball. In fact, the literature on the topic, Staniswalis and Severini [**59**], Müller [**52**], Härdle and Mammen [**36**], Hart [**37**], Chen [**21**], Eubank and LaRiccia [**31**], Dette and Munk [**22**] among other references, never addressed such an issue. The only exception we are aware of is the paper by Horowitz and Spokoiny [**38**]. Their procedure applies for more general sets $\mathcal{C}$ than linear spaces but, in the particular case we consider here, has the drawback to be difficult to implement.

When $\mathcal{C}$ is a linear subspace of $\mathbb{R}^n$, the procedure we propose has the property to be optimal or nearly optimal in the minimax sense and also to be $(1/\sqrt{n})$-consistent over directional alternatives. Besides, it is easy to implement and do not require any estimation of the nuisance parameter σ (which is the case for the procedure proposed by Eubank and Hart). In B., Huet and Laurent [**11,**

9], simulation studies show that the procedure is very powerful. Further details on this procedure is given in Section 3.

1.3. Case of a qualitative hypothesis. Given some estimator $\hat{f}$ of f, a natural idea to test the hypothesis "$f \in \mathcal{C}$" is to reject the null when $\mathrm{d}_n(\hat{f}, \mathcal{C})$ is large. When $\hat{f}$ and $\mathcal{C}$ satisfy the condition

$$\kappa^2 = \sup_{f \in \mathcal{C}} E\left[\mathrm{d}_n^2(\hat{f}, f)\right] < +\infty, \tag{29}$$

this approach results (by Markov's inequality) in a level-α test by taking $\kappa/\sqrt{\alpha}$ as the threshold. Such a test is usually called a *plug-in procedure*. Unfortunately, Condition (29) cannot be satisfied when $\mathcal{C}$ is large, as for those sets

$$\mathcal{C}_{\geq 0} = \{f \in \mathbb{R}^n,\ \forall i \in \{1, \dots, n\},\ f_i \geq 0\}$$

$$\mathcal{C}_{\nearrow} = \{f \in \mathbb{R}^n,\ f_1 \leq f_2 \leq \dots \leq f_n\},$$

making thus the use of plug-in procedures impossible. We shall call such hypotheses "qualitative". As suggested by the forms of the sets $\mathcal{C}_{\geq 0}$ and $\mathcal{C}_{\nearrow}$, tests of qualitative hypotheses usually arise in Model 2 to test a specific feature of the regression function such as positivity or monotonicity.

In the literature, the problem of testing a qualitative hypothesis was addressed in view of detecting a local discrepancy to the null hypothesis. Let us mention the work of Bowman et al. [**18**], Gijbels et al. [**33**], Hall and Heckman [**35**], Ghosal et al. [**32**] for the problem of testing the monotonicity of a regression function. However, as for those tests of linear hypotheses, little was known on their performances from the minimax point of view. The only exception we know are the tests of non-monotonicity, non-negativity and non-convexity proposed by Dümbgen and Spokoiny [**27**] in the Gaussian white noise model. These were proved to be asymptotically optimal (as the noise level tends to zero) when the signal was belonging to a class of Lipschitz function. The drawback of this test is that it does not apply to the regression framework where the variance σ is unknown and its optimality is restricted to a particular class of alternatives.

Let us now make two comments. First, the problem of detecting a "global" discrepancy to the null hypothesis, say with respect to the Euclidean distance, had never been addressed in the literature. However, such a test can be of interest at least for testing monotonicity. For example, if the index i in Model 5 represents a time of experiment, it may be more appropriate to reject the null hypothesis if the values of the f_i's are not monotonous but oscillating with time rather than detecting a gap between two of their consecutive values. Second, as pointed out by Dümbgen and Spokoiny [**27**], the problem of building tests of qualitative hypotheses which are rate optimal over Hölderian balls was remaining

open. These problems have been solved in the papers B., Huet and Laurent [**8, 10**].

In the sequel, we denote by $\|\ \|$ the Euclidean norm in $\mathbb{R}^n$ and for each linear subspace V of $\mathbb{R}^n$, Π_V the orthogonal projector onto V. Finally, $(e_1, \ldots, e_n)$ denotes the canonical basis of $\mathbb{R}^n$.

2. Multiple testing and model selection

The aim of this section is twofold. First, to give an account of the basic ideas that underline our testing procedures and second, to establish some connections between these and model selection.

2.1. A description of our tests. For particular subsets $\mathcal{C}$ of $\mathbb{R}^n$, we address the problem of testing at some level α the null hypothesis "$f \in \mathcal{C}$" against the alternative "$f \notin \mathcal{C}$". The sets $\mathcal{C}$ we consider have the particularity to contain 0. The tests we propose are based on the following principle.

We start with a family of tests $\{\phi_{m,\alpha},\ m \in \mathcal{M}_n\}$ of the form $\phi_{m,\alpha}(Y) = \mathbb{1}\left\{\hat{T}_m - q_m(\alpha) > 0\right\}$ where $\hat{T}_m$ is a test statistic and $q_m(\alpha)$ the $1-\alpha$ quantile of $\hat{T}_m$ under $\mathbb{P}_{0,1} = \mathcal{N}(0, I_n)$ (the reason for this choice will become clearer later). Typically, we construct the collection of tests $\{\phi_{m,\alpha}, m \in \mathcal{M}_n\}$ in such a way that for each m, $\phi_{m,\alpha}$ is a "good" test of the hypothesis "$f \in \mathcal{C}$" against a specific alternative of interest. For a suitable choice of positive numbers $\{\alpha_m,\ m \in \mathcal{M}_n\}$ in $]0,1[$, we reject the null when the test statistic

$$\hat{T}_\alpha = \sup_{m \in \mathcal{M}_n} \left(\hat{T}_m - q_m(\alpha_m)\right), \tag{30}$$

is positive. Equivalently, our test, say ϕ_α, rejects the null if one of the tests ϕ_{m,α_m}'s does. Thus, our test is based on a multiple testing procedure.

The numbers α_m are chosen for the test ϕ_α to be of level α, that is to satisfy

$$\sup_{\tau>0} \sup_{f \in \mathcal{C}} \mathbb{P}_{f,\tau}\left[\sup_{m \in \mathcal{M}_n} \left(\hat{T}_m - q_m(\alpha_m)\right) > 0\right] \leq \alpha. \tag{31}$$

At first glance, such a condition may seem difficult to fulfill because of the presence of the two suprema. In fact, this will be easy. Indeed, the test statistics $\hat{T}_m$ are chosen to satisfy the following conditions:

- for any $\tau > 0$, the supremum

$$\sup_{f \in \mathcal{C}} \mathbb{P}_{f,\tau}\left[\sup_{m \in \mathcal{M}_n} \left(\hat{T}_m - q_m(\alpha_m)\right) > 0\right]$$

  is achieved for $f = 0$.

- for each $m \in \mathcal{M}_n$ and $\tau > 0$, the distribution of $\hat{T}_m$ under $\mathbb{P}_{0,\tau}$ is free from τ.

As a consequence, to satisfy Inequality (31), it is enough to ensure that

$$\mathbb{P}_{0,1}\left[\sup_{m\in\mathcal{M}_n}\left(\hat{T}_m - q_m(\alpha_m)\right) > 0\right] \leq \alpha \tag{32}$$

holds true. Then, we suggest two ways of choosing the α_m's. The first one is to take $\alpha_m = \alpha'$ for all $m \in \mathcal{M}_n$ where α' satisfies

$$\mathbb{P}_{0,1}\left[\sup_{m\in\mathcal{M}_n}\left(\hat{T}_m - q_m(\alpha')\right) > 0\right] = \alpha. \tag{33}$$

The quantity α' is obtained by carrying out a simulation study. The second way is to choose the α_m's constraint by the summability condition $\sum_{m\in\mathcal{M}_n} \alpha_m = \alpha$, since the following inequalities lead to Condition (32):

$$\begin{aligned}\mathbb{P}_{0,1}\left[\sup_{m\in\mathcal{M}_n}\left(\hat{T}_m - q_m(\alpha_m)\right) > 0\right] &\leq \sum_{m\in\mathcal{M}_n} \mathbb{P}_{0,1}\left[\hat{T}_m - q_m(\alpha_m) > 0\right] \\ &= \sum_{m\in\mathcal{M}_n} \alpha_m = \alpha.\end{aligned}$$

This second choice of α_m's leads to a conservative procedure as the size of the test is then smaller than α. However, when the $q_m(.)$'s are known, this choice allows to avoid simulations and makes thus the procedure easier to implement in practice.

2.2. Connection with model selection. In this section, we propose to establish some connections between the ideas underlying our approach to hypothesis testing and model selection. To simplify our task, we restrict ourself to the problem of testing "$f = 0$".

As already mentioned in Section 1.2, the major drawback of the Fisher test is that an alternative must be specified to implement it. Of course, such a choice is a difficult task when no prior information on f is available. For the problems considered here, the idea of our tests is to propose several alternatives, $\{\mathcal{F}_m \setminus \{0\},\ m \in \mathcal{M}_n\}$, rather than a single one. For each m, $\mathcal{F}_m$ is a linear subspace of $\mathbb{R}^n$ ($\mathcal{F}_m \neq \mathbb{R}^n$) to which we associate the Fisher test, ϕ_m, of level α_m where the α_m's for $m \in \mathcal{M}_n$ are suitably chosen. Finally, we decide to reject the null hypothesis if one of these Fisher tests rejects. Equivalently, our test can be seen as a model selection procedure among the family $\Lambda = \{\mathcal{F}_m,\ m \in \mathcal{M}_n\} \cup \{0\}$ and let us explain how. Assume for convenience that the index 0 does not belong to $\mathcal{M}_n$ and set $\mathcal{F}_0 = \{0\}$, $\hat{T}_0 = q_0(u) = 0$ for all $u \in]0,1[$. For the particular choice $\mathrm{pen}(m) = q_m(\alpha_m)$ for all $m \in \mathcal{M}_n \cup \{0\}$,

consider the model selection criterion which selects among Λ the model $\mathcal{F}_{\hat{m}}$ for which $\hat{m}$ is the maximizer among $\mathcal{M}_n \cup \{0\}$ of the penalized criterion

$$\mathrm{Crit}(m) = \hat{T}_m - \mathrm{pen}(m). \tag{34}$$

Then, it is easy to see that our test rejects if and only if the selected model $\mathcal{F}_{\hat{m}}$ is not $\mathcal{F}_0 = \{0\}$. This means that, equivalently, we could have built our procedure on a model selection criterion (that given by (34)) and decide to reject the null if the selected model is different from the $\{0\}$. This idea was earlier proposed by Eubank and Hart [**30**]. In their paper $\hat{T}_m$ is defined as $\|\Pi_{\mathcal{F}_m} Y\|^2/n$ and $\mathrm{pen}(m)$ is taken as $C\sigma^2 D_m/n$ for some suitable positive constant C. Note that when $C = 2$, this criterion corresponds to Mallows' C_p! In fact, their constant C is chosen for the so-defined test to be of level α. Consequently, our approach mainly differs from theirs in the choice of the penalty term. It is interesting to notice that our choice of the penalty leads to a model selection criterion which is analogous to that proposed by Laurent and Massart [**44**] for estimating the quadratic functional $\|f\|^2$.

3. An overview of our tests

3.1. Linear hypotheses. In B., Huet and Laurent [**11**], we consider the case where $\mathcal{C} = V$ is a linear subspace of $\mathbb{R}^n$. The test we propose is based upon a choice of a collection of linear subspaces $\{\mathcal{F}_m,\ m \in \mathcal{M}_n\}$ of $\mathcal{C}^\perp$. Our test statistic is given by (30) where $\hat{T}_m$ is that of the Fisher test $\phi_{m,\alpha}$ for testing the null hypothesis "$f \in V$" against "$f \in (V + \mathcal{F}_m) \setminus V$". As already mentioned, the Fisher statistic offers the advantage to be distribution free from the nuisance parameter σ under the null and consequently, our test does not rely on any estimation of σ.

Our work differs from previous ones in that, for each $\beta \in]0,1[$, we describe a set of vectors over which the power of our test is larger than $1-\beta$. When $\mathcal{C} = \{0\}$ and $\sigma^2 = 1$, this set gathers the vectors f satisfying

$$\mathrm{d}_n^2(f,\mathcal{C}) \geq a^2_{\alpha,\beta}(f) = \kappa \inf_{m \in \mathcal{M}_n} \left[\mathrm{d}_n^2(f,\mathcal{F}_m) + \frac{\sqrt{D_m \mathcal{L}_m}}{n} + \frac{\mathcal{L}_m}{n} \right] \tag{35}$$

where $\mathcal{L}_m = -\log(\alpha_m \beta/2)$ and κ denotes some universal constant. For suitable α_m's, we show that, for each β, this set is close to the set of alternatives f which are detected with probability larger than $1-\beta$ by at least one of the Fisher tests $\phi_{m,\alpha}$'s. Our procedure is thus nearly uniformly more powerful than the family of Fisher tests $\{\phi_{m,\alpha},\ m \in \mathcal{M}_n\}$.

The role played by the quantity $a_{\alpha,\beta}(f)$ in (35) is similar to that played by the accuracy index (20) in the oracle-type inequalities. Let us provide some illustration of this fact when $\mathcal{C} = \{0\}$. Take $\mathcal{M}_n = \{1,\ldots,n-1\}$ and for each

$m \in \mathcal{M}_n$, define $\mathcal{F}_m$ as the linear span of the e_i's for $i = 1, \ldots, m$. Finally, we choose for all $m \in \mathcal{M}_n$, $\alpha_m = \alpha/(n-1)$ in order to satisfy the summability condition $\sum_{m \in \mathcal{M}_n} \alpha_m = \alpha$ which, in turn, ensures that our test is of level α. Let us now consider the ellipsoid

$$\mathcal{E}_{s,2}(1) = \left\{ g \in \mathbb{R}^n,\ \frac{1}{n} \sum_{i=1}^{n} i^{2s} g_i^2 \leq 1 \right\}$$

where s denotes some positive number. It is not hard to see that for those f in the ellipsoid, $\mathrm{d}_n^2(f, \mathcal{F}_m) \leq m^{-2s}$ for each $m \in \mathcal{M}_n$. Consequently, by bounding $a_{\alpha,\beta}(f)$ as in Chapter 1 Section 2.3, we derive from (35) that our test achieves a power larger than $1-\beta$ over those f in $\mathcal{E}_{s,2}(1)$ satisfying $\mathrm{d}_n(f, 0) \geq \rho_s$ where ρ_s^2 is of order

$$\inf_{m \in \mathcal{M}_n} \left(m^{-2s} + \frac{\sqrt{m \ln(n)}}{n} \right).$$

By optimizing this expression with respect to m, we deduce that the minimax separation rate of our procedure over $\mathcal{E}_{s,2}(1)$ is, up to a power of $\ln(n)$, of order $n^{-2s/(1+4s)}$. The rate $n^{-2s/(1+4s)}$ is known to be the minimax rate of testing over the ellipsoid $\mathcal{E}_{s,2}(1)$ (see Section 4) making thus our procedure nearly rate optimal. Since our procedure is not based on any prior knowledge of s, this rate is achieved simultaneously for all $s > 0$. In contrast with the estimation problem, it is not possible to achieve the rate $n^{-2s/(1+4s)}$ simultaneously over each ellipsoid $\mathcal{E}_{s,2}(1)$ with $s > 0$. A loss of efficiency is unavoidable and this is the reason why our rate differs from the minimax one (by a $\ln(n)$-factor). In fact, it is possible to achieve the minimax rate up to a $\ln\ln(n)$-factor by using a more appropriate collection of $\mathcal{F}_m$'s.

3.2. Qualitative hypotheses: detecting a global discrepancy. An extension of the procedure described above to the case where $\mathcal{C}$ is no longer a linear space but a cone of $\mathbb{R}^n$ is described in B., Huet and Laurent [**8**]. More precisely, we consider the case where $\mathcal{C}$ is a subset of $\mathbb{R}^n$ containing 0 and satisfying the following condition

$$\forall g, h \in \mathcal{C},\ g + h \in \mathcal{C}.$$

Note that these conditions hold true in the cases where $\mathcal{C}$ is $\mathcal{C}_{\geq 0}$ or $\mathcal{C}_{\nearrow}$.

Our testing procedure can be described as follows. Let $\{\mathcal{F}_m, m \in \mathcal{M}_n\}$ be a suitable collection of linear subspaces of $\mathbb{R}^n$. By suitable we mean that if f belongs to $\mathcal{C}$, we require that $\Pi_{\mathcal{F}_m} f$ still belongs to $\mathcal{C}$. Besides, we assume that there exists some linear subspace $\mathcal{F}$ of $\mathbb{R}^n$ such that $\mathcal{F} \neq \mathbb{R}^n$ and $\bigcup_{m \in \mathcal{M}_n} \mathcal{F}_m \subset \mathcal{F}$. Our test statistic is then given by (30) where for each $m \in \mathcal{M}_n$, $\hat{T}_m$ is the

test statistic

$$\frac{\mathrm{d}_n^2(\Pi_{\mathcal{F}_m} Y, \mathcal{C})}{\mathrm{d}_n^2(Y, \mathcal{F})/(n - \dim(\mathcal{F}))}.$$

The idea of the test is to reject the null if for some $m \in \mathcal{M}_n$, the distance between $\Pi_{\mathcal{F}_m} Y$ (which is intuitively close to $\Pi_{\mathcal{F}_m} f$) and $\mathcal{C}$ is large. By introducing the denominator $\mathrm{d}_n^2(Y, \mathcal{F})/(n - \dim(\mathcal{F}))$, the distribution of the test statistics $\hat{T}_m$ is free from σ under $\mathbb{P}_{0,\sigma^2}$ and thus the computations of the α_m's do not require any prior estimation of the variance of the ε_i's.

As for our tests of linear hypotheses, given some $\beta \in]0,1[$, we describe for each value of n a subset of $\mathbb{R}^n$ over which we prove the test to achieve a power larger than $1-\beta$. When f is the vector of the values of a regression function F at some deterministic points, we deduce uniform separation rates (with respect to $\mathrm{d}_n(.,.)$) as F varies among some Hölderian ball $\mathcal{H}_s(R)$ defined by (18). Let us emphasize that these rates are established under the a posteriori assumption that F belongs to $\mathcal{H}_s(R)$ since our testing procedure does not depend on any prior assumption on f. Over Hölderian balls, the separation rates we get are of the same order as the minimax estimation rates and we do not know whether the formers are optimal or not. For the problem of testing positivity, only lower bounds established by Juditsky and Nemirovski [**41**] in the Gaussian white noise model are available. Their bounds differ from ours by a $\log(n)$ factor.

3.3. Qualitative hypotheses: detecting a local discrepancy. In B., Huet and Laurent [**10**] we consider the case where $\mathcal{C}$ is a convex subset of $\mathbb{R}^n$ of the form

$$\mathcal{C} = \{f \in \mathbb{R}^n,\ \forall j \in 1, \ldots, p, < f, v_j > \leq 0\},$$

where $< .,. >$ denotes the Euclidean inner product of $\mathbb{R}^n$ and $\{v_j,\ j = 1, \ldots, p\}$ a family of vectors which are linearly independent in $\mathbb{R}^n$.

In the functional regression model, for a suitable choice of the set $\mathcal{C}$, the procedure allows to test that the derivative of order $r \in \mathbb{N}$ of F is positive. The case $r = 0$, respectively $r = 1, 2$ corresponds to testing the positivity, respectively the monotonicity and convexity, of F. More generally, our approach allows to test that F satisfies a differential inequality of the form

$$\frac{d^r}{dx^r}\left(R(x)F(x)\right) \geq 0, \quad \text{for all } x \in]0,1[$$

where r is an integer and R a non vanishing function on the interval $[0,1]$. For example, by taking $r = 1$ and $R(x) = -\exp(ax)$ for some positive number $a > 0$, the problem amounts to testing that the function F has an exponential decay on the interval $[0,1]$, i.e. satisfies for all $x \in [0,1]$, $F(x) \leq F(0)\exp(-ax)$.

Let us now describe our procedure. Let $\mathcal{T}$ be the set of vectors defined by

$$\mathcal{T} = \left\{ t = \sum_{j=1}^{p} \lambda_j v_j,\ \|t\| = 1,\ \lambda_j \geq 0\ \forall j = 1, \ldots, p \right\}.$$

We consider a finite subset $\mathcal{T}_n = \{t_m,\ m \in \mathcal{M}_n\}$ of $\mathcal{T}$ and a linear space V_n of dimension $\mathcal{D}_n$ smaller than n containing $\mathcal{T}_n$. Our test statistic is then given by (30) where for each $m \in \mathcal{M}_n$, $\hat{T}_m$ is the studentized statistics $\sqrt{n - \mathcal{D}_n} < Y, t_m > /\|Y - \Pi_{V_n} Y\|$. Since for all $t \in \mathcal{T}_n < f, t > \leq 0$ when f belongs to $\mathcal{C}$, the test is driven by the idea that the null hypothesis should be rejected when for some $t \in \mathcal{T}_n < Y, t >$ is large. As in the previous sections, the denominator $\|Y - \Pi_{V_n} Y\|$ ensures that the distribution of the test statistic $\hat{T}_m$ is free from σ under $\mathbb{P}_{0,\sigma^2}$. The vectors of $\mathcal{T}_n$ are typically chosen to have a small number of nonzero coordinates, giving thus a local character to the tests $\phi_{m,\alpha}$.

Let us now turn to an example by considering the problem of testing that F is nondecreasing. The set $\mathcal{C}$ is then given by

$$\mathcal{C} = \{f \in \mathbb{R}^n,\ f_{i+1} - f_i \geq 0, \forall i \in \{1, \ldots, n-1\}\},$$

which amounts to taking $p = n - 1$ and for all $j = 1, \ldots, p$, $v_j = e_j - e_{j+1}$. For different choices of sets $\mathcal{T}_n$, our testing procedure takes different features. We offer two examples below assuming that the x_i's are equispaced points in $[0, 1]$.

- For two disjoint subsets I and J, each of which consisting of consecutive integers among $\{1, \ldots, n\}$, let us define

  $$t'_{I,J} = \frac{1}{|I|} \sum_{i \in I} e_i - \frac{1}{|J|} \sum_{i \in J} e_j.$$

  When I is to the left of J, we show that $t = t'_{I,J}/\|t'_{I,J}\|$ belongs to $\mathcal{T}$. Thus, a possible choice of $\mathcal{T}_n$ is a collection of those vectors t for suitable choices of sets I and J. The resulting procedure, say TEST T1, is based on the differences of local means.

- For a subset I of consecutive integers among $\{1, \ldots, n\}$, let us define x_I (respectively $\mathbb{1}_I$) as the vector of $\mathbb{R}^n$ the i-th coordinate of which is x_i (respectively 1) if $i \in I$ and 0 otherwise. We show that the vector

  $$t_I = \frac{\left(\sum_{i \in I} x_i/|I|\right) \mathbb{1}_I - x_I}{\| \left(\sum_{i \in I} x_i/|I|\right) \mathbb{1}_I - x_I \|},$$

  belongs to $\mathcal{T}$ and therefore we can choose a collection $\mathcal{T}_n$ of such vectors t for suitable I's. In this case, the resulting procedure, say TEST T2, is based on the local slope $- < Y, t_I >$ of the regression of the Y_i's on the x_i's for $i \in I$.

Let us mention that TEST T2 is akin to that proposed by Hall and Heckman [**35**]. Our results provide thus some perspective on the optimality of their procedure.

We describe the power of our tests in the following way. For each $\beta \in]0,1[$, we show that there exists a constant κ depending on α, β such that the power of the test is larger than $1-\beta$ over those f satisfying both $\mathrm{d}_n^2(f, V_n) \leq \sigma^2$ (at least for n large enough) and

$$\begin{aligned} \delta_{\mathcal{T}_n,\mathcal{C}}(f) &= \max_{t \in \mathcal{T}_n} [< f, t > \mathbb{1}\{< f, t > \geq 0\}] \\ &\geq \kappa \sqrt{\ln(n)} \sigma. \end{aligned} \tag{36}$$

An interesting feature of this result is that it is possible to relate the quantity $\delta_{\mathcal{T}_n,\mathcal{C}}(f)$ to a more tractable distance between F and the null hypothesis under suitable *a posteriori* regularity condition on F. We use this property to establish the rate optimality of our procedure in the minimax sense over classes of smooth functions. To explain how we proceed, let us consider again the problem of testing the monotonicity of F, taking for simplicity $\sigma^2 = 1$. Let us also introduce the two distances $\delta_1(G)$ and $\delta_2(G)$ defined for all functions G on $[0,1]$ as the distance in sup-norm between G and respectively, the set of nondecreasing functions on $[0,1]$ and the set of nonnegative functions on $[0,1]$.

In the case of TEST T1, a suitable choice of the sets I and J (free from F) together with the assumption that F belongs to some Hölderian ball $\mathcal{H}_s(R)$ allows us to establish that $\delta_{\mathcal{T}_n,\mathcal{C}}(f)$ is bounded from below (up to a constant depending on R) by $\sqrt{n}\,(\delta_1(F))^{1+1/(2s)}$. Then by combining this result with (36), we obtain that TEST T1 rejects the null with probability larger than $1-\beta$ as soon as $\delta_1(F)$ is larger than $(\log(n)/n)^{s/(1+2s)}$. This rate is known to be optimal over $\mathcal{H}_s(R)$ (see Dümbgen and Spokoiny [**27**]).

In the case of TEST T2, for a suitable choice of set I together with the assumption that F' belongs to some Hölderian ball $\mathcal{H}_s(R)$, we show that $\delta_{\mathcal{T}_n,\mathcal{C}}(f)$ is bounded from below (up to a constant depending on R) by

$$\left(\delta_2(F')\,(n/\ln(n))^{s/(3+2s)} - 1\right)\sqrt{\ln(n)}$$

and by arguing as before, we obtain that TEST T2 rejects the null with probability larger than $1-\beta$ as soon as $\delta_2(F)$ is larger than $(\log(n)/n)^{s/(3+2s)}$. This rate is also known to be optimal over those F for which F' belongs to $\mathcal{H}_s(R)$.

In the rates specified here, we only mention the dependency with respect to n. The dependency of these with respect to the radii R of the Hölderian balls is made explicit in B., Huet and Laurent [**10**], but for the sake of simplicity it is omitted here. Nevertheless, it turns out that our procedures provide optimal rates with respect to both parameters s and R.

4. Nonasymptotic minimax rates of testing

When $\mathcal{C} = \{0\}$, the description of the minimax rates of testing was mainly obtained in the Gaussian white noise model and in the Gaussian sequence model. In this latter framework, the case of ellipsoids was first considered in Ermakov [**29**] under some conditions on the decay of the semi-axes. In the former, other kinds of alternatives were considered in Ingster [**39**] including Hölderian functional spaces among others. Lepski and Spokoiny [**46**] described the minimax rates of testing over Besov spaces $\mathcal{B}_{s,p,q}(R)$ with $p \in]0, 2[$ (see also Ingster and Suslina [**40**]) and showed an unexpected dependence (with regard to the case $p = 2$) of the minimax rate of testing with respect to s.

The point of view of the previous authors was asymptotic and little was known on the minimax rates of testing in regression for a fixed value of n. In B. [**4**], our point of view is nonasymptotic and aims at describing the minimax rates of testing in both the regression and the Gaussian sequence model. We consider various alternatives including ellipsoids (and more generally ℓ_p-bodies). The upper and lower bounds we get are free from any assumptions on the decay of the semi-axes which is new in the literature. Let us now give an account of our work mainly restricting ourself to the Gaussian sequence model.

4.1. The problem at hand. In contrast with the other sections of this chapter, we shall assume that Y is an infinite sequence of independent Gaussian random variables Y_i of mean f_i and common variance σ^2. We shall keep the notation f for the mean of Y and assume that f belongs to

$$\ell_2(\mathbb{N}^*) = \left\{ f \in \mathbb{R}^{\mathbb{N}^*},\ \|f\|^2 = \sum_{i \geq 1} f_i^2 < +\infty \right\}$$

where $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$.

For a fixed $\beta \in]0, 1[$, we measure the performance of a level-α test, ϕ_α, by means of the smallest radius ρ for which the power of the test is larger than $1-\beta$ over the alternative "$\|f\| \geq \rho$ and $f \in \mathcal{F}$". We denote by $\rho_{\mathcal{F}}(\phi_\alpha)$ this quantity, omitting its dependency with respect to α, β and σ^2. In this section, our aim is to describe the infimum, $\rho_{\mathcal{F}}$, of $\rho_{\mathcal{F}}(\phi_\alpha)$ when ϕ_α varies among all possible level-α tests. In the literature, one usually studies the asymptotics of $\rho_{\mathcal{F}}$ when the parameter σ tends to 0. The quantity $\rho_{\mathcal{F}}$ is then called the *minimax rate of testing* (or the minimax separation rate) over $\mathcal{F}$, the word "rate" referring to the scaling parameter σ. In the sequel, we shall keep calling $\rho_{\mathcal{F}}$ this way even though σ is fixed in our setting.

4.2. Alternatives consisting of sparse sequences. For given integers N and D ($D \leq N$), B. [**4**] first considers the case where $\mathcal{F} = \mathcal{F}_{D,N}$ is the set of sequences f which have at most D nonzero coordinates located among the N first. Under the assumption that $\alpha + \beta \leq 59\%$, we obtain that

$$D \ln\left(1 + \frac{N}{D^2} \vee \sqrt{\frac{N}{D^2}}\right) \sigma^2 \leq \rho^2_{\mathcal{F}} \leq \kappa \left[D \ln\left(e\frac{N}{D}\right) \wedge \sqrt{N}\right] \sigma^2, \tag{37}$$

where $\kappa = 2(\sqrt{5}+4)\ln(2e/(\alpha\beta))$. Let us give some elementary consequences of this formula when $\sigma^2 = 1$. First, by taking $N = D$ and using the invariance of the Gaussian law with respect to orthogonal transformations, we obtain that, when $\mathcal{F}$ is a finite dimensional space of dimension D, the squared minimax rate of testing is of order $\sqrt{D}$. Note that in the context of estimation, the minimax estimation rate with respect to the quadratic loss $\| \|^2$ is of order D. Second, another interesting fact is that $\rho^2_{\mathcal{F}}$ is of order $\sqrt{N}$ whatever the values of D in $[\sqrt{N}, N]$. In this case, it is not necessary to take into account the fact that at most D coordinates among $\{f_1, \ldots, f_N\}$ are nonzero to obtain a rate optimal procedure. Only the information that $f_i = 0$ for all $i > N$ plays a role. Let us mention that for the estimation problem the minimax rate is then of order $D \ln(eN/D)$ and is thus much larger when N is large and $D \in [\sqrt{N}, N]$. Finally, let us mention that the minimax rates for both estimation and testing coincide (up to constants) for small values of D.

The classical regression framework corresponds to the case where $\mathcal{F} = \mathcal{F}_{N,N}$. This means that in this setting, the information that f belongs to $\mathcal{F}_{N,N}$ is available for free. It follows from (37) that the (squared) minimax rates of testing in regression are always bounded from above (up to a constant) by $\sqrt{N}$. This result is worth mentioning indeed as it is no longer true in other statistical frameworks such as the Gaussian white noise model or the Gaussian sequence model. For these, such information on f is no longer available.

4.3. Case of ellipsoids and ℓ_p-bodies. Considering the case where $\mathcal{F} = \mathcal{F}_{D,N}$ may seem unusual. Its interest lies in the fact that by approximation arguments we can derive lower bounds for the minimax rates of testing over various sets such as ellipsoids, Besov bodies or more generally ℓ_p-bodies. Such an approach has been already adopted by Birgé and Massart [**17**] in the estimation framework. We recall that an ℓ_p-body (in $\ell_2(\mathbb{N}^*)$) is a set of the form

$$\mathcal{E}_{a,p}(R) = \left\{ f \in \ell_2(\mathbb{N}^*),\ \sum_{i \geq 1} \left| \frac{f_i}{a_i} \right|^p \leq R^p \right\},$$

where p, R are positive numbers and the a_i's a sequence of nonnegative ones which are non-increasing towards 0. We shall use the convention that $x/0 =$

$+\infty$ for all $x > 0$ and $0/0 = 0$. Restricting our study to the values $p \in]0, 2]$, we deduce from (37) that for $\mathcal{F} = \mathcal{E}_{a,p}(R)$ and α, β small enough

$$\rho^2_{\mathcal{F}} \geq \sup_{D \in I} \left[\left(\lceil \sqrt{D} \rceil \sigma^2 \right) \wedge \left(R^2 a_D^2 \lceil \sqrt{D} \rceil^{1-2/p} \right) \right],$$

where $\lceil \sqrt{D} \rceil$ denotes the smallest integer k satisfying $k \geq \sqrt{D}$. This result is both nonasymptotic and assumption free on the decay of the a_i's and shows that a lower bound is obtained by finding some trade-off between two terms: $\lceil \sqrt{D} \rceil$ which increases with D and $R^2 a_D^2 \lceil \sqrt{D} \rceil^{1-2/p}$ which decreases with D. Typically, this trade-off is achieved for some D^* satisfying (roughly speaking)

$$\sqrt{D^*} \approx R^p a^p_{D^*} / \sigma^p.$$

When for all D, $a_D = D^{-s}$ for some $s > 0$, the lower bound is of order $R^{2/(1+4s")} \sigma^{8s"/(1+4s")}$ where $s" = s - 1/4 + 1/(2p)$. The lower bound is known to be optimal (up to a constant free from σ and R) from Lepski and Spokoiny [**46**]. In the case where $p = 2$ (case of an ellipsoid), the lower bound becomes

$$\rho^2_{a,R} = \sup_{D \in I} \left[\left(\lceil \sqrt{D} \rceil \sigma^2 \right) \wedge \left(R^2 a_D^2 \right) \right],$$

and is optimal. Indeed, we show that for all $\sigma < R$, $\rho^2_{\mathcal{F}} \leq \kappa \rho^2_{a,R}$ for some constant κ depending on α and β only.

One usually compares the minimax rates of testing obtained in the Gaussian sequence model with those obtained in the regression one by taking in the former $\sigma^2 = 1/N$ where N corresponds to the number of observations in the latter. For such a value of σ^2, we deduce from the result above that when $p = 2$ and $a_D = D^{-s}$ for all D, the (squared) minimax rate of testing over the ellipsoid $\mathcal{E}_{a,2}(R)$ is given by $R^{2/(1+4s)} N^{-4s/(1+4s)}$. Note that this rate can be slower than $\sqrt{N}$ when $s < 1/4$, which contrasts with the regression setting.

4.4. Adaptive hypothesis testing. When $\mathcal{F}$ is a union of sets, say $\bigcup_{\lambda \in \Lambda} \mathcal{F}_\lambda$, we also consider the problem of describing for each $\lambda \in \Lambda$ the minimal radius, $\bar{\rho}_{\mathcal{F}_\lambda}$, for which it is possible to find a test whose power over the union of alternatives "$\|f\| \geq \bar{\rho}_{\mathcal{F}_\lambda}$ and $f \in \mathcal{F}_\lambda$" is larger than $1 - \beta$. This problem is related to the problem of adaptation in the minimax sense. In contrast with the estimation problem, adaptation in the minimax sense without loss of efficiency is in general impossible. More precisely, the ratio $\sup_{\lambda \in \Lambda} \bar{\rho}_{\mathcal{F}_\lambda} / \rho_{\mathcal{F}_\lambda}$ cannot be in general bounded by some universal constant. In the Gaussian white noise model, this fact was mentioned by Spokoiny [**58**] for a union of Besov balls. In B. [**4**], we provide an illustration by considering the collection of ellipsoids $\{\mathcal{E}_{a,2}(\lambda), \lambda \in]0, +\infty[\}$ for which the problem of adaptation addresses here to the radius λ. For $\alpha + \beta$ small enough, we show that *at least for one* $\lambda > 0$ (not

necessarily for all)

$$\bar{\rho}^2_{\mathcal{F}_\lambda} \geq \kappa \sup_{D\geq 1} \left[\left(\sqrt{D\mathcal{L}(D)}\sigma^2 \right) \wedge \left(R^2 a_D^2 \right) \right], \tag{38}$$

where κ denotes some positive constant and $\mathcal{L}(D) = \ln\ln(D+1)$. Denoting by λ_1 a value of λ for which (38) holds and repeating the argument for the collection of ellipsoids $\{\mathcal{E}_{a,2}(\lambda), \lambda \in]0,+\infty[\setminus \{\lambda_1\}\}$, we can exhibit another value λ_2 for which (38) is satisfied an so on. This means that if (38) is not satisfied for all values of λ than it is necessarily satisfied by infinitely many of these.

By arguing as previously, when for all $D \geq 1$, $a_D = D^{-s}$ for some $s > 0$, we obtain that the lower bound in (38) differs from the minimax rate by an additional $\ln\ln(1/\sigma^2)$ factor.

5. Confidence balls

In this section we propose to give an account of the results established in B. [**5**]. In this paper, we consider the problem of building a nonasymptotic Euclidean confidence ball around the vector f with prescribed probability of coverage. Our approach is inspired by Lepski [**45**] and is based on a combination of estimation and hypothesis testing. The problem considered by Lepski is more general than the one on which we focus on here. Lepski addresses the problem of improving the accuracy of estimation in many statistical frameworks of interest. The construction of a confidence ball turns out to be a consequence of his general approach. For our specific problem, we shall use more adequate tools and different techniques.

In the sequel, let us denote by $\mathcal{B}(x,r)$ the Euclidean ball centered at $x \in \mathbb{R}^n$ of radius $r > 0$ and by $q_D(u)$, the $1-u$ quantile of a χ^2 random variable with D degrees of freedom. Hereafter, we shall fix some $\beta \in]0,1[$. To explain the basic ideas of our approach, let us assume that σ^2 is known and that f belongs to some known linear space $\mathcal{F} \subset \mathbb{R}^n$. We allow the case where $\mathcal{F} = \mathbb{R}^n$, which is the practical one for which no information on f is available. Since the random variable $\|f - \Pi_{\mathcal{F}} Y\|^2$ is distributed as a χ^2 with $N = \dim(\mathcal{F})$ degrees of freedom, the Euclidean ball $\mathcal{B}(\Pi_{\mathcal{F}} Y, R)$ with $R = q_N(\beta)$ provides a confidence ball for f with probability of coverage larger than $1-\beta$. However, this confidence ball may be very broad since for large values of N, $q_N(\beta)$ is of order N. To obtain sharper confidence balls with the same probability of coverage, the idea (inspired by Lepski) is to introduce some simpler candidate models for f, namely a collection of linear subspaces $\{\mathcal{F}_m, m \in \mathcal{M}_n\}$ of $\mathcal{F}$. Among this collection, we only retain those $\mathcal{F}_m$'s with $m \in \mathcal{M}'_n$ for which the χ^2-test based on the statistics $\|\Pi_{\mathcal{F}} Y - \Pi_{\mathcal{F}_m} Y\|^2$ accepts the null hypothesis "H_m : $f \in \mathcal{F}_m$" at some level $\alpha \in]0,1[$. If $\mathcal{M}'_n$ is empty, then our confidence

ball if merely given by $\mathcal{B}(\Pi_{\mathcal{F}}Y, R)$. Otherwise, a sequence of positive numbers $\{\rho_m,\ m \in \mathcal{M}_n\}$ computed beforehand, ensures that with probability larger than $1-\beta$, f belongs to the intersection of the balls $\mathcal{B}(\Pi_{\mathcal{F}_m}Y, \rho_m)$ for those $m \in \mathcal{M}'_n$. In particular, with probability larger than $1-\beta$, f belongs to the one for which the radius ρ_m is the smallest among $\mathcal{M}'_n$.

Our construction warrants that under the *a posteriori* information that f belongs to $\mathcal{F}_m$, the radius of the so-defined confidence ball is not larger than ρ_m with probability larger than $1-\alpha$. We show that ρ_m^2 is of the order $b_m = \max\{\sqrt{NL_m}, D_m, L_m\}$ where the L_m's are chosen in the construction to satisfy a summability condition, say $\sum_{m\in\mathcal{M}_n} e^{-L_m} \leq \beta/2$. For a fixed value of m, by choosing $L_m = \ln(4/\beta)$, b_m is of order $\max\{\sqrt{N}, D_m\}$ and this order of magnitude cannot be improved. More precisely, we show that for some positive constant κ which is made explicit in the paper and depends on α and β only, it is not possible to find random variables $\hat{f}, \hat{\rho}$ satisfying both $\mathbb{P}_f\left[f \in \mathcal{B}(\hat{f}, \hat{\rho})\right] \geq 1-\beta$ for all $f \in \mathbb{R}^n$ and $\mathbb{P}_f\left[\hat{\rho} \leq \kappa b_m\right] \geq 1-\alpha$ for all $f \in \mathcal{F}_m$.

As already mentioned, the result presented above also holds true for the choice $\mathcal{F} = \mathbb{R}^n$ which requires thus no prior information on f. Yet, we have assumed that σ^2 was known. A natural question is now what happens if this quantity is unknown and one takes $\mathcal{F} = \mathbb{R}^n$. We show that the problem is then impossible. Even if one has some piece of information on σ such as $\sigma^2 \in [(1-\eta)\tau^2, \tau^2]$ for some $\tau > 0$ and $\eta \in]0,1[$, any procedure would be almost useless unless η is very small. More precisely, we show that for any random variables $\hat{f}, \hat{\rho}$ satisfying both $\mathbb{P}_f\left[f \in \mathcal{B}(\hat{f}, \hat{\rho})\right] \geq 1-\beta$ for all $f \in \mathbb{R}^n$ and $\mathbb{P}_f\left[\hat{\rho} \leq r_m\right] \geq 1-\alpha$ for all $f \in \mathcal{F}_m$ for some positive number r_m, we necessarily have $r_m \geq \kappa'\eta n\tau^2$ for some positive constant κ'.

The procedures obtained in this paper extend to the problem of building a confidence ball around f when the ratio f/σ is known to belong to some ellipsoid in $\mathbb{R}^n$. Beside, we provide an application of these to the problem of variable selection described in (1.2) when the errors are Gaussian.

Publications et Prépublications

Les papiers et preprints sont disponibles à l'adresse

`http://www.dma.ens.fr/~baraud`

Papiers parus ou à paraître

(1) *Model selection for regression on a fixed design* Probability Theory and Related Fields 117 $467 - 493$ (2000).
(2) *Adaptive estimation in an autoregression and a geometrical β-mixing framework* Annals of Statist. 29 $839 - 875$ (2001). Collaboration avec F. Comte et G. Viennet.
(3) *Model Selection for (auto-)regression with dependent data* ESAIM Probab. Statist. 5 $33 - 49$ (2001). Collaboration avec F. Comte et G. Viennet.
(4) *"Model selection for regression on a random design"* ESAIM Probab. Statist. 6 $127 - 146$ (2002)
(5) *A new test of linear hypothesis in regression* Goodness-of-fit tests and model validity (Paris, 2000), $195 - 207$, Stat. Ind. Technol., Birkhäuser Boston, Boston, MA, (2002). Collaboration avec S. Huet et B. Laurent.
(6) *Non-asymptotic minimax rates of testing in signal detection* Bernoulli 8 $577 - 606$ (2002)
(7) *Adaptive tests of qualitative hypotheses* ESAIM Probab. Statist. (2002). Collaboration avec S. Huet et B. Laurent.
(8) *Adaptive tests of linear hypotheses by model selection* Annals of Statist. 31 (2003). Collaboration avec S. Huet et B. Laurent.

Papiers soumis ou en révision

(1) *Confidence balls in nonparametric Gaussian regression* Préprint de L'Ecole Normale Supérieure 01-28. (En révision pour les Annals of Statistics).
(2) *Tests for convex hypotheses* Préprint de L'Ecole Normale Supérieure 01-32. Collaboration avec S. Huet et B. Laurent.

Bibliography

[1] Akaike, H. Information theory and an extension of the maximum likelihood principle. *Proceedings 2nd International Symposium on Information Theory,* $267 - 281$, 1973.

[2] Baraud, Y. Model selection for regression on a random design. *Preprint 01-10, Ecole Normale Superieure*, 1997.

[3] Baraud, Y. Model selection for regression on a fixed design. *Probab. Theory Relat. Fields* **117**, $467 - 493$, 2000.

[4] Baraud, Y. Non asymptotic minimax rates of testing in signal detection. *Technical Report. 00-25, Ecole Normale Supérieure*, 2000.

[5] Baraud, Y. Confidence balls in nonparametric gaussian regression. *Préprint 01.28 Ecole Normale Supérieure, DMA, Paris.*, 2001.

[6] Baraud, Y., Comte, F., and Viennet G. Adaptive estimation in autoregression or β-mixing regression via model selection. *Ann. Statist.* **29**, $839 - 875$, 2001.

[7] Baraud, Y., Comte, F., and Viennet, G. Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist.* **5** $33 - 49$, 2001.

[8] Baraud, Y., Huet, S., and Laurent, B. Adaptive tests of qualitative hypotheses. *Préprint 2001.37 Université Paris-Sud, Orsay.*, 2001.

[9] Baraud, Y., Huet, S., and Laurent, B. Nonparametric smoothing and lack-of-fit tests. In Huber-Carol, C, Balakrishnan, N, Nikulin, M. S. and Mesbah, M., editor, *Goodness-of-fit tests and validity of models*, pages 193–204, Boston, 2001. Birkhauser.

[10] Baraud, Y., Huet, S., and Laurent, B. Tests for convex hypotheses. *Préprint 01.32 Ecole Normale Supérieure, DMA, Paris.*, 2001.

[11] Baraud, Y., Huet, S., and Laurent, B. Adaptive tests of linear hypotheses by model selection. *To appear in Ann. Statist.*, 2002.

[12] Barron, A., Birgé, L., and Massart, P. Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113**, $301 - 413$, 1999.

[13] Barron, A.R. Complexity regularization with application to artificial neural networks. *Proceedings NATO Advanced Study Institute on Nonparametric Functional estimation,* $561 - 576$. *G.Roussas, Ed., Dordrecht, The Netherlands: Kluwer*, 1991.

[14] Barron, A.R. and Cover, T.M. Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37**, $1034 - 1054$, 1991.

[15] Birgé, L. and Massart, P. From model selection to adaptative estimation. *In Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics (D.Pollard, E. Torgensen and G. Yang, eds.),* $55 - 87$. *Springer-Verlag, New York*, 1997.

[16] Birgé, L. and Massart, P. An adaptive compression algorithm in Besov spaces. *Constr. Approx.* **16**, $1 - 36$, 2000.

[17] Birgé, L. and Massart, P. Gaussian model selection. *JEMS* **3** $203 - 268$, 2001.

[18] Bowman, A.W., Jones, M.C., and Gijbels, I. Testing monotonicity of regression. *J. Comput. Graph. Statist.* **7** $489 - 500$, 1998.

[19] Castellan, G. Density estimation via exponential model selection. Technical report, 00.25 Université Paris XI, 2000.

[20] Catoni, O. Statistical learning theory and stochastic optimization. *Ecole d'été de probabilités de Saint-Flour, Springer, to appear.*

[21] Chen, J-C. Testing for no effect in nonparametric regression via spline smoothing techniques. *Ann. Inst. Statist. Math.* **46** $251-265$, 1994.

[22] Dette, H. and Munk, A. Validation of linear regression models. *Ann. Statist.* **26** $778-800$, 1998.

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

[24] Donoho, D. and Johnstone, I. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer, Statist. Assoc.* **90** $1200-1224$, 1995.

[25] Donoho, D. and Johnstone, I. Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** $879-921$, 1998.

[26] Donoho, D.L. and Johnstone, I.M. Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81** $425-455$, 1994.

[27] Dümbgen, L. and Spokoiny, V.G. Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29** $124-152$, 2001.

[28] Efroimovich, S. and Pinsker, M. Learning algorithm for nonparametric filtering. *Auto. Remote. Control* **11**, $1434-1140$, 1984.

[29] Ermakov, M.S. Minimax detection of a signal in a Gaussian white noise. *Theory Probab. Appl.* **2** $85-114$, 1991.

[30] Eubank, R.L. and Hart, J.D. Testing goodness-of-fit in regression via order selection criteria. *Ann. Statist.* **20** $1412-1425$, 1992.

[31] Eubank, R.L. and LaRiccia, V.N. Testing for no effect in nonparametric regression. *J. Statist. Plann. Inference* **36** $1-14$, 1993.

[32] Ghosal, S and Sen, A. and Van der Vaart, A. Testing monotonicity of regression. *Ann. Statist.* **28** $1054-1082$, 2001.

[33] Gijbels, I., Hall, P., Jones, M.C., and Koch, I. Tests for monotonicity of a regression mean with guaranteed level. *Biometrika,* **87** $663-673$, 2000.

[34] Grenander, U. *Abstract inference*. Wiley, New York, 1981.

[35] Hall, P. and Heckman, N.E. Testing for monotonicity of a regression mean by calibrating for linear functions. *The Annals of Statistics* **28** $20-39$, 2000.

[36] Härdle, W. and Mammen, E. Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21** $1926-1947$, 1993.

[37] J.D. Hart. *Nonparametric smoothing and lack-of-fit tests*. Springer-Verlag, New-York, 1997.

[38] Horowitz, J.L. and Spokoiny, V.G. An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* **3** $599-631$, 2001.

[39] Ingster, Yu.I. Asymptotically minimax hypothesis testing for nonparametric alternatives I-III. *Math. Methods Statist.* **2**, $85-114$, **3** $171-189$, **4** $249-268$, 1993.

[40] Ingster, Yu.I. and Suslina, I.A. Minimax detection of a signal for besov bodies and balls. *Problems Inform. Transmission* **34** $48-59$, 1998.

[41] Juditsky, A. and Nemirovski, A. On nonparametric tests of positivity/monotonicity/ convexity. *Ann. Statist.* **30**, 2002.

[42] Kohler, M. Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *J. Statist. Plann. Inference* **89** $1-23$, 2000.

[43] Korostelev, A.P. and Tsybakov, A.B. Minimax theory of image reconstruction. In *Lecture notes in statistics, New York NY*. Springer-Verlag, 1993.

[44] Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28**, $1302-1338$, 2000.

[45] Lepski, O. How to improve the accuracy of estimation. *Math. Methods Statist.* **8**, $441-486$, 1999.

[46] Lepski, O.V. and Spokoiny, V.G. Minimax nonparametric hypothesis testing: the case of inhomogeneous alternative. *Bernoulli* **5**, $333-358$, 1999.

[47] Lepskii, O. A problem of adaptive estimation in gaussian white noise. *Theory Probab. Appl.* **35** $454-466$, 1990.

[48] Lepskii, O. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** $682-697$, 1991.

[49] Li, K.C. Asymptotic optimality for C_p,C_l, cross-validation and generalized cross validation: discrete index set. *Ann. Statist.* **15**, $958-975$, 1987.

[50] Mallows, C.L. Some comments on C_p. *Technometrics* **15**, $661-675$, 1973.

[51] Meyer, Y. *Ondelettes et opérateurs 1*. Hermann, 1990.

[52] Müller, H-G. Goodness-of-fit diagnostics for regression models. *Scand. J. Statist.* **19** $157-172$, 1992.

[53] Nishii, R. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, $758-765$, 1984.

[54] Polyak, B.T. and Tsybakov, A.B. Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory Probab. Appl.* **35**, $293-306$, 1990.

[55] Reynaud-Bouret, P. Concentration inequalities for inhomogeneous Poisson processes and adaptive estimation of the intensity. Technical report, 01-18 Université Paris XI, 2001.

[56] Shen, X. and Wong, W.H. Convergence rate of sieve estimates. *Ann. Statist.* **22** $580-615$, 1994.

[57] Shibata, R. An optimal selection of regression variables. *Biometrika* **68**, $45-54$, 1981.

[58] Spokoiny, V.G. Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24**, $2477-2498$, 1996.

[59] Staniswalis, J.G. and Severini, T.A. Diagnostics for assessing regression. *J. Amer. Statist. Assoc.* **86** $684-692$, 1991.

[60] Stone, C.J. The use of polynomial spline and their tensor products in multivariate function estimation. *Ann. Statist.* **22** $118-184$, 1994.

[61] Talagrand, M. New concentration inequalities in product spaces. *Invent. Math.* **126**, $505-563$, 1996.

[62] van de Geer, S. Estimating a regression function. *Ann. Statist.* **18**, $907-924$, 1990.

[63] van de Geer, S. The method of sieves and minimum contrast estimators. *Math. Methods Statist.* **4** $20-38$, 1995.

[64] Viennet, G. Inequalities for absolutely regular sequences, application to density estimation. *Prob. Theory Relat. Fields* **107**, $467-492$, 1997.

[65] Wegkamp, M. Model selection in non-parametric regression. Technical report, Preprint Yale University, 2000.

[66] Yang, Y. Combining different procedures for adaptive regression. *Journal of Multivariate Analysis* **74** $135-161$, 2000.

[67] Yang, Y. and Barron, A. Information-Theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** $1564-1599$, 1999.

N° D'IMPRESSION 2459
4ème TRIMESTRE 2002