Histogram selection in non gaussian regression

Marie Sauvé

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Histogram selection in non gaussian regression

Marie Sauvé

N° 5911

Mai 2006

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--5911--FR+ENG

INRIA
FUTURS

Histogram selection in non gaussian regression

Marie Sauvé*

Thème COG — Systèmes cognitifs
Projet SELECT

Rapport de recherche n° 5911 — Mai 2006 — 16 pages

Abstract: We deal with the problem of choosing an histogram estimator of a regression function s mapping $\mathcal{X}$ into $\mathbb{R}$. We adopt the non asymptotic approach of model selection via penalization developped by Birgé and Massart, but we do not assume that the observations are gaussian variables. We consider a collection of partitions of $\mathcal{X}$, with possibly exponential complexity, and the corresponding collection of histogram estimators. We propose a penalized least squares criterion which selects a partition whose associated estimator performs approximately as well as the best one, in the sense that its quadratic risk is close to the infimum of the risks. The risk bound we provide is non asymptotic.

Key-words: model selection, regression, CART

* Département de Mathématiques, Université Paris-Sud, 91405 Orsay Cedex, email : marie.sauve@math.u-psud.fr

Unité de recherche INRIA Futurs
Parc Club Orsay Université, ZAC des Vignes,
4, rue Jacques Monod, 91893 ORSAY Cedex (France)
Téléphone : +33 1 72 92 59 00 — Télécopie : +33 1 72 92 59 ??

Sélection d'un histogramme en régression non gaussienne

Résumé : Nous nous intéressons ici au problème du choix d'un estimateur de type histogramme d'une fonction de régression s définie sur $\mathcal{X}$ et à valeurs dans $\mathbb{R}$. Nous adoptons l'approche non asymptotique de la sélection de modèle par pénalisation développée par Birgé et Massart, mais nous ne supposons pas que les observations sont des variables gaussiennes. Nous considérons une collection de partitions de $\mathcal{X}$, de complexité éventuellement exponentielle, et les estimateurs de type histogramme correspondants. Nous proposons un critère des moindres carrés pénalisés qui sélectionne une partition dont l'estimateur associé est proche du meilleur, au sens où son risque quadratique est comparable au risque minimal. La majoration du risque obtenue est non asymptotique.

Mots-clés : sélection de modèles, régression, CART

1 Introduction

We study here regression frameworks. In these frameworks, we observe a real variable Y at n different times or for n different individuals. For the i^{th} time or individual, the variable Y_i can be written

$$Y_i = \mu_i + \varepsilon_i, \qquad 1 \leq i \leq n, \tag{1}$$

with μ_i an unknown parameter and ε_i a random perturbation. The variables $(\varepsilon_i)_{1\leq i\leq n}$ are supposed to be centered, independent and identically distributed (i.i.d.). Most of time μ_i is the value of an unknown signal $s : \mathcal{X} \to \mathbb{R}$, called regression function, at a point $x_i \in \mathcal{X}$, and then

$$Y_i = s(x_i) + \varepsilon_i, \qquad 1 \leq i \leq n. \tag{2}$$

The $(x_i)_{1\leq i\leq n}$ can be times and in this case $\mathcal{X} = \mathbb{N}$, $\mathbb{R}$ or $[0;1]$. For $\mathcal{X} = \mathbb{N}$ and $x_i = i$, $1 \leq i \leq n$, we find expression (1) again. The $(x_i)_{1\leq i\leq n}$ can also be vectors of p real components $x_i^1, \ldots, x_i^p$ corresponding to p characteristics of the i^{th} individual and in this case $\mathcal{X} = \mathbb{R}^p$. Our aim is to get informations on μ or s from the observations $(Y_i)_{1\leq i\leq n}$.

Many statistical issues can be rewritten in terms of model selection, where we call model a linear space in $\mathbb{R}^n$ (resp. $\mathbb{R}^{\mathcal{X}}$) if we observe $(Y_i)_{1\leq i\leq n}$ as defined in (1) (resp. (2)). We focus here on histogram models. An histogram model in $\mathbb{R}^n$ is the linear span of a system $\{\sum_{i\in J} e_i; J \in M\}$ where M is a partition of $\{1, \ldots, n\}$ and $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$. An histogram model in $\mathbb{R}^{\mathcal{X}}$ is a space of piecewise constant functions defined on a partition of $\mathcal{X}$. This kind of models has already been widely used in order to get partial informations on s as well as to estimate s.
The statistical problem which is obviously solved by histogram model selection is the problem of determining the groups of individuals with same means μ_i. But this is far from being the only use of histogram model selection.
One of the most famous statistical issue is variables selection. In the classical linear regression framework,

$$Y_i = \sum_{j=1}^{p} \beta_j x_i^j + \varepsilon_i, \qquad 1 \leq i \leq n,$$

selecting a small subset of variables $V \subset \{x^1, \ldots, x^p\}$ which explain "at best" the response Y is equivalent to choosing the "best" model S_V of functions linear in $\{x^j \in V\}$. Instead of considering linear interaction between $(x^1, \ldots, x^p)$ and Y, we work here with histogram models. In ([9]), Sauvé and Tuleau propose a variables selection procedure based on histogram model selection.
Histogram model selection is also used to estimate s. One looks for an estimator $\hat{s}$ of s such that $\hat{s}$ is close to s in the sense that its quadratic risk is small when the number n of observations is large. Let consider a partition M_0 of $\mathcal{X}$ with a large number of small cells. A classical method of estimating s consists in minimizing the quadratic contrast over the class S_{M_0} of piecewise constant functions defined on the partition M_0. The resulting estimator is denoted $\hat{s}_{M_0}$ and is called the least squares estimator over the histogram model S_{M_0}. Denoting $\|.\|_n$ the Euclidean norm on $\mathbb{R}^n$ scaled by a factor $n^{-1/2}$ and denoting the same way a function $u \in \mathbb{R}^{\mathcal{X}}$ and the corresponding vector $(u(x_i))_{1\leq i\leq n} \in \mathbb{R}^n$, the quadratic risk of $\hat{s}_{M_0}$, $\mathbb{E}\left(\|s - \hat{s}_{M_0}\|_n^2\right)$, is the sum of two terms, respectively called bias and variance:

$$\mathbb{E}\left(\|s - \hat{s}_{M_0}\|_n^2\right) = \inf_{u\in S_{M_0}} \|s - u\|_n^2 + \frac{\tau^2}{n}|M_0| \quad \text{where } \tau^2 = \mathbb{E}(\varepsilon_i^2)$$

We see in this expression of the risk of $\hat{s}_{M_0}$ that $\hat{s}_{M_0}$ behaves poorly when M_0 has a large number of cells and that we should rather choose a partition M built from M_0 (or equivalently an histogram model $S_M \subset S_{M_0}$) which makes a better trade-off between the bias $\inf_{u\in S_M} \|s - u\|_n^2$ and the variance $\frac{\tau^2}{n}|M|$. The CART algorithm, proposed by Breiman *et al.* [5], is based on this idea. This algorithm

first builds a partition M_0 with a large number of cells containing only one point x_i, by splitting recursively the set $\mathcal{X}$ and the subset obtained in two parts. This construction is naturally represented by a tree of maximal depth, called T_{max} and whose leaves are the cells of the partition M_0. Then it prunes T_{max} and considers the trees $\hat{T}(\alpha) = \arg\min \left\{ \|Y - \hat{s}_T\|_n^2 + \alpha \frac{|T|}{n} \right\}$, where T denotes a tree as well as the partition associated to its leaves, $|T|$ denotes the number of leaves of T, and the minimum is taken over all subtrees of T_{max} with same root. This amounts to choosing the partition which minimizes the penalized least squares criterion with the penalty term $\text{pen}(M) = \alpha \frac{|M|}{n}$ among a collection of partitions built from M_0. In the CART algorithm, there is a last step which consists in determining the right parameter α by cross-validation.

Let us now describe our estimation procedure in details. We consider a collection $(S_M)_{M \in \mathcal{M}_n}$ of histogram models. Denoting $\hat{s}_M$ the least squares estimator over S_M, the best model is the one which minimizes $\mathbb{E}\left(\|s - \hat{s}_M\|_n^2\right)$. Unfortunately this model depends on s. The aim of model selection is to propose a data driven criterion, whose minimizer among $(S_M)_{M \in \mathcal{M}_n}$ is an approximately best model. We select a model $S_{\hat{M}}$ by minimizing over $\mathcal{M}_n$ a penalized least squares criterion $\text{crit}(M) = \|Y - \hat{s}_M\|_n^2 + \text{pen}(M)$.

$$\hat{M} = \arg\min_{M \in \mathcal{M}_n} \left\{ \|Y - \hat{s}_M\|_n^2 + \text{pen}(M) \right\}$$

The estimator $\hat{s}_{\hat{M}}$ is called the penalized least squares estimator (PLSE). The penalty pen has to be chosen such that the model $S_{\hat{M}}$ is close to the optimal model, more precisely such that

$$\mathbb{E}\left(\|s - \hat{s}_{\hat{M}}\|_n^2\right) \leq C \inf_{M \in \mathcal{M}_n} \mathbb{E}\left(\|s - \hat{s}_M\|_n^2\right) \tag{3}$$

The inequality (3) will be referred to as the oracle inequality. It bounds the risk of the penalized least squares estimator by the infimum of the risks on a given model up to a constant C. The main result of this paper determines a form of penalty pen which leads to an oracle type inequality.

This work is already done in ([2], [3]) when the i.i.d. random perturbations ε_i are $\mathcal{N}(0, \sigma^2)$ distributed. In this paper, we want to generalize their result in the case where the ε_i are only supposed to have finite exponential moments around 0. We use the same ideas and techniques as Birgé and Massart. The main point is to control a statistic which is the square of the supremum of an empirical process. In the gaussian case, this statistic is χ^2 distributed. But in the non gaussian case, it is much more difficult to study the deviation of this statistic around its expectation. We give more details about this difficulty in section 4. In particular, Bousquet's concentration inequality for the supremum of an empirical process is not sufficient. We explain how it should be improved and we give an other method based on Bernstein inequality to control our statistic without looking first at its square root.

Baraud in [1] determines a form of penalty which leads to an oracle type inequality with an even milder integrability condition on the $(\varepsilon_i)_{1 \leq i \leq n}$. He assumes only that the $(\varepsilon_i)_{1 \leq i \leq n}$ have a finite absolute moment of order p for some positiv integer p. Unfortunately his risk bound of the PLSE is not good when the number of models with a given dimension D is exponential in D. We call complexity of a collection of models the number of models with a given dimension. Our paper deals with collections of histogram models whose complexity may be exponential. We take stronger integrability condition on the $(\varepsilon_i)_{1 \leq i \leq n}$ but weaker condition on the complexity of the collection of models.

The paper is organized as follows. The section 2 presents the statistical framework and some notations. The section 3 gives the main result. To get this result, we have to control a χ^2 like statistic. The section 4 is more technical, it exposes a concentration inequality for a χ^2 like statistic and explains why the existing concentration inequality, due to Bousquet, is not sufficient. Sections 5 and 6 are devoted to the proofs.

2 The statistical framework

In this paper, we consider the regression framework defined by (2) and we look for a best or approximately best histogram estimator of s. In this section, we precise the integrability condition that should satisfy the random perturbations $(\varepsilon_i)_{1\leq i\leq n}$ involved in (2), then we define the histogram estimators of s and their risk. We give here some notations needed in the rest of the paper.

2.1 The random perturbations

As noted above in the introduction, we assume that the i.i.d. random perturbations $(\varepsilon_i)_{1\leq i\leq n}$ have finite exponential moments around 0. This assumption can be expressed by the existence of two positive constants b and σ such that

$$\forall\ \lambda \in (-1/b, 1/b) \qquad \log \mathbb{E}\left(e^{\lambda\varepsilon_i}\right) \leq \frac{\sigma^2\lambda^2}{2(1-b|\lambda|)} \tag{4}$$

σ^2 is necessarly greater than $\mathbb{E}(\varepsilon_i^2)$ and can be chosen as close to $\mathbb{E}(\varepsilon_i^2)$ as we want, but at the price of a larger b.

Remark 1
To get inequality (4), we show that:
if there exists $b_0 \in \mathbb{R}_+$ *satisfying* $\mathbb{E}\left(e^{\lambda\varepsilon_i}\right) < +\infty$ *for any* $\lambda \in (-1/b_0, 1/b_0)$
then, for any $b > b_0$, *we have*

$$\forall\ k \geq 2 \qquad \mathbb{E}\left(|\varepsilon_i|^k\right) \leq \frac{k!}{2}\sigma^2(b)b^{k-2}$$

with $\sigma^2(b) \geq \mathbb{E}\left(\varepsilon_i^2\right)$ *and* $\sigma^2(b) \underset{b\to+\infty}{\longrightarrow} \mathbb{E}\left(\varepsilon_i^2\right)$.

Remark 2
Under assumption (4), we have

$$\forall\ \lambda \in (-1/2b, 1/2b) \qquad \log \mathbb{E}\left(e^{\lambda\varepsilon_i}\right) \leq \sigma^2\lambda^2$$

but we prefer inequality (4) to this last inequality because with the last one we loose a factor 2 *in the variance term.*

2.2 The histogram estimators

For a given partition M of $\mathcal{X}$, we denote S_M the space of piecewise constant functions defined on the partition M and $\hat{s}_M$ the least squares estimator over S_M.

$$\hat{s}_M = \arg\min_{u\in S_M} \gamma_n(u) \text{ with } \gamma_n(u) = \|Y-u\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - u(x_i))^2$$

where $\|.\|_n$ denotes the Euclidean norm on $\mathbb{R}^n$ scaled by a factor $n^{-1/2}$ and, for $u \in \mathbb{R}^{\mathcal{X}}$, the vector $(u(x_i))_{1\leq i\leq n} \in \mathbb{R}^n$ is denoted u too. S_M is the histogram model associated with M and $\hat{s}_M$ is the histogram estimator belonging to S_M which plays the role of benchmark among all the estimators in S_M.

Denoting $s_M = \arg\min_{u \in S_M} \|s-u\|_n^2$, $\varepsilon_M = \arg\min_{u \in S_M} \|\varepsilon - u\|_n^2$ and $|M|$ the number of elements of the partition M, the quadratic risk of the estimator $\hat{s}_M$ is

$$\begin{aligned}\mathbb{E}\left(\|s-\hat{s}_M\|_n^2\right) &= \|s-s_M\|_n^2 + \mathbb{E}\left(\|\varepsilon_M\|_n^2\right) = \|s-s_M\|_n^2 + \mathbb{E}(\varepsilon_1^2)\frac{|M|}{n} \\ &\leq \|s-s_M\|_n^2 + \sigma^2\frac{|M|}{n}\end{aligned}$$

3 The main theorem

Let M_0 a partition of $\mathcal{X}$ and $\mathcal{M}_n$ a family of partitions of $\mathcal{X}$ built from M_0, i.e. for any $M \in \mathcal{M}_n$ and any element J of M, J is the union of elements of M_0. In the following theorem, we assume that the initial partition M_0 is not too fine in the sense that the elements of the partition M_0 contain a minimal number of points x_i. We measure the fineness of the partition M_0 by the number $N_{min} = \inf_{J \in M_0} |J|$ where $|J| = |\{1 \leq i \leq n;\ x_i \in J\}|$.

The ideal partition M^* minimizes the quadratic risk $\mathbb{E}\left(\|s-\hat{s}_M\|_n^2\right)$ over all the partitions $M \in \mathcal{M}_n$. Unfortunately M^* depends on the unknown regression function s and $\hat{s}_{M^*}$ can not be used as an estimator of s. The purpose of model selection is to propose a data driven criterion which selects a partition $\hat{M}$ whose associated histogram estimator $\hat{s}_{\hat{M}}$ performs approximately as well as $\hat{s}_{M^*}$ in terms of risks. We select a partition $\hat{M}$ by minimizing a penalized least squares criterion $\text{crit}(M) = \|Y-\hat{s}_M\|_n^2 + \text{pen}(M)$ over $\mathcal{M}_n$.

$$\hat{M} = \arg\min_{M \in \mathcal{M}_n} \left\{\|Y-\hat{s}_M\|_n^2 + \text{pen}(M)\right\}$$

It remains to provide a penalty pen such that the partition $\hat{M}$ is close to the optimal partition, in the sense that the PLSE $\hat{s}_{\hat{M}}$ satisfies an oracle inequality like (3). The following theorem determines a general form of penalty pen which leads to an oracle type inequality for any family of partitions built from a partition M_0 not too fine. We compare our result to those of Birgé and Massart and those of Baraud, and we study in more details two particular families of partitions.

Example 1: The partition M_0 is built by splitting recursively $\mathcal{X}$ and the subsets obtained in two different parts as long as each subset contains at least N_{min} points x_i. A usefull representation of this construction is a tree of maximal depth, called maximal tree and denoted T_{max}. The leaves of the maximal tree are the elements of the partition M_0. Every pruned subtree of the maximal tree gives a partition of $\mathcal{X}$ built from M_0. This first family corresponds to the one proposed by CART.

Example 2: We consider $\mathcal{X} = [0,1]$ and a grid on $[0,1]$, and we take the family of all partitions of $[0,1]$ with endpoints belonging to the grid.

Theorem 1
Let $b \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}_+^*$ such that inequality (4) holds.
Let M_0 a partition of $\mathcal{X}$ such that $N_{min} = \inf_{J \in M_0} |J|$ satisfies $N_{min} \geq 12\frac{b^2}{\sigma^2}\log n$.
Let $\mathcal{M}_n$ a family of partitions of $\mathcal{X}$ built from M_0 and $(x_M)_{M \in \mathcal{M}_n}$ a family of weights such that

$$\sum_{M \in \mathcal{M}_n} e^{-x_M} \leq \Sigma \in \mathbb{R}_+^*$$

Assume $\|s\|_\infty \leq R$, with R a positiv constant.
Let $\theta \in (0,1)$ and $K > 2-\theta$ two numbers.

Taking a penalty satisfying

$$pen(M) \geq K\frac{\sigma^2}{n}|M| + 8\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left[\left(4(2-\theta)+\frac{2}{\theta}\right)\frac{\sigma^2}{n} + \frac{4Rb}{n}\right]x_M \tag{5}$$

we have

$$\begin{aligned}\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\right) &\leq \frac{2}{1-\theta}\inf_M\left\{\|s-s_M\|_n^2 + pen(M)\right\} \\ &\quad + \frac{1}{1-\theta}\left(8(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{4}{\theta}+2\right)\frac{\sigma^2}{n}\Sigma + \frac{12}{1-\theta}\frac{Rb}{n}\Sigma \\ &\quad + C(b,\sigma^2,R)\frac{\mathbb{1}_{b\neq 0}}{n(\log n)^{3/2}}\end{aligned}$$

where $C(b,\sigma^2,R)$ is a positive constant which depends only on b, σ^2 and R.

This theorem gives the general form of the penalty function

$$\mathrm{pen}(M) = K\frac{\sigma^2}{n}|M| + \left(\kappa_1(\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left[\kappa_2(\theta)\frac{\sigma^2}{n} + \frac{4Rb}{n}\right]x_M\right)$$

The penalty is the sum of two terms: the first one is proportional to $\frac{|M|}{n}$ and the second one depends on the complexity of the family $\mathcal{M}_n$ via the weights $(x_M)_{M\in\mathcal{M}_n}$. For $\theta \in (0,1)$ and $K > 2-\theta$, the PLSE $\hat{s}_{\hat{M}}$ satisfies an oracle type inequality with an additional term tending to 0 like $1/n$ when $n \to +\infty$.

$$\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\right) \leq C_1\inf_M\left\{\|s-s_M\|_n^2 + \mathrm{pen}(M)\right\} + \frac{C_2}{n}$$

where the constant C_1 only depends on θ, whereas C_2 depends on s (via R), on the family of partitions (via Σ) and on the integrability condition of $(\varepsilon_i)_{1\leq i\leq n}$ (via σ^2 and b).

For the two particular families $\mathcal{M}_n$ quoted above, we calculate adequate weights $(x_M)_{M\in\mathcal{M}_n}$ and we get a simpler form of penalty. Before studying these two examples, we compare the general result with those of Birgé and Massart [3] and those of Baraud [1].

If b can be taken equal to zero in (4), then the variables $(\varepsilon_i)_{1\leq i\leq n}$ are said to be sub-gaussian. In this case, we do not need any assumptions neither on N_{min} the minimal number of observations in each element of the partition M_0 nor on s the regression function. And taking a penalty satisfying

$$\mathrm{pen}(M) \geq K\frac{\sigma^2}{n}|M| + 8\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left[4(2-\theta)+\frac{2}{\theta}\right]\frac{\sigma^2}{n}x_M$$

we have

$$\begin{aligned}\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\right) &\leq \frac{2}{1-\theta}\inf_M\left\{\|s-s_M\|_n^2 + \mathrm{pen}(M)\right\} \\ &\quad + \frac{1}{1-\theta}\left(8(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{4}{\theta}+2\right)\frac{\sigma^2}{n}\Sigma\end{aligned}$$

Up to some small differences in the constants (which can be improved by looking more precisely at the proof), this is the result obtained by Birgé and Massart in the gaussian case.

In [1], Baraud studies the non gaussian regression framework as defined in (2) with a milder integrability condition on the random perturbations than ours. For a collection of histogram models $(S_M)_{M\in\mathcal{M}_n}$

whose complexity is polynomial, our theorem and those of Baraud both validate penalties $\text{pen}(M)$ proportional to $|M|/n$ through an oracle type inequality with an additional term tending to 0 like $1/n$ when $n \to +\infty$. Thanks to Baraud's result, if $|\{M \in \mathcal{M}_n;\ |M| = D\}| \leq \Gamma D^a$ for some constants $\Gamma \in \mathbb{R}_+^*$ and $a \in \mathbb{N}$, one only needs to assume that the random perturbations have a finite absolute moment of order $p > 2a + 6$. The minimal admissible value of p increases with the degree a of the polynomial complexity. And, whatever p, having a finite absolute moment of order p seems to be not enough to deal with collections of exponential complexity. Our assumption on the exponential moments is too strong when the complexity is polynomial, but it allows us to propose a general form of penalty which is still valide when the complexity is exponential.

Let now see which form of penalty is adapted to the two collections of partitions quoted above. The complexity of the two corresponding collections of histogram models is exponential, thus Baraud's result is not available.

EXAMPLE 1: Let T_{max} a binary tree of maximal depth built on $\mathcal{X}$ such that each leaf contains at least N_{min} points x_i, with $N_{min} \geq 12\frac{b^2}{\sigma^2}\log n$. Let M_0 the partition of $\mathcal{X}$ whose elements are the leaves of the maximal tree T_{max}. Let $\mathcal{M}_n$ the collection of partitions corrresponding to the pruned subtrees of T_{max}. Thanks to Catalan inequality, $|\{M \in \mathcal{M}_n; |M| = D\}| \leq \frac{1}{D}\binom{2(D-1)}{D-1} \leq \frac{2^{2D}}{D}$. Thus taking $x_M = a|M|$ with $a > 2\log 2$, we get $\sum_{M\in\mathcal{M}_n} e^{-x_M} \leq -\log\left(1 - e^{-(a-2\log 2)}\right) \in \mathbb{R}_+^*$. We deduce from the above theorem that:
taking a penalty

$$\text{pen}(M) = \alpha\frac{\sigma^2 + Rb}{n}|M|$$

with α big enough, we have

$$\mathbb{E}\left(\|s - \hat{s}_{\hat{M}}\|_n^2\right) \leq C_1(\alpha)\inf_M\left\{\|s - s_M\|_n^2 + \frac{\sigma^2 + Rb}{n}|M|\right\} + C_2(\alpha)\frac{\sigma^2 + Rb}{n} + C(b, \sigma^2, R)\frac{1\!\!1_{b\neq 0}}{n(\log n)^{3/2}}$$

For this first example, we recommend a penalty $\text{pen}(M)$ proportional to $\frac{|M|}{n}$. For such a penalty, the selected model satisfies an oracle inequality with an additional term tending to 0 like $1/n$ when $n \to +\infty$.

In practice, as σ^2, b and R are unknown, we consider penalties of the form $\text{pen}(M) = \gamma\frac{|M|}{n}$ and we determine the right constant γ by using, for example, the slope heuristic of Massart [8], section 8.5.2.

EXAMPLE 2: Let $\mathcal{X} = [0, 1]$, M_0 a partition of $[0, 1]$ composed by D_0 segments such that $N_{min} = \inf_{J\in M_0}|J| \geq 12\frac{b^2}{\sigma^2}\log n$, and $\mathcal{M}_n$ the collection of all partitions of $[0, 1]$ in segments built from those of M_0. As $|\{M \in \mathcal{M}_n; |M| = D\}| \leq \binom{D_0 - 1}{D - 1} \leq \left(\frac{eD_0}{D}\right)^D$, taking $x_M = |M|\left(a + \log\frac{D_0}{|M|}\right)$ with $a > 1$ leads to $\sum_{M\in\mathcal{M}_n} e^{-x_M} \leq \left(e^{a-1} - 1\right)^{-1} \in \mathbb{R}_+^*$. We deduce from the above theorem that:
taking a penalty

$$\text{pen}(M) = \frac{\sigma^2 + Rb}{n}|M|\left(\alpha + \beta\log\left(\frac{|M_0|}{|M|}\right)\right)$$

with α and β big enough, we have

$$\begin{aligned}\mathbb{E}\left(\|s - \hat{s}_{\hat{M}}\|_n^2\right) \leq\ & C_1(\alpha,\beta)\inf_M\left\{\|s - s_M\|_n^2 + \frac{\sigma^2 + Rb}{n}|M|\left(1 + \log\left(\frac{|M_0|}{|M|}\right)\right)\right\} + C_2(\alpha,\beta)\frac{\sigma^2 + Rb}{n} \\ & + C(b, \sigma^2, R)\frac{1\!\!1_{b\neq 0}}{n(\log n)^{3/2}}\end{aligned}$$

In this second example, the penalty differs from the preceding one on a factor $\log\left(\frac{|M_0|}{|M|}\right)$. This additional factor allows us to get nearly the same oracle inequality as before despite a larger complexity of $\mathcal{M}_n$.

Like in the first example, as σ^2, b and R are unknown, we consider penalties of the form $\text{pen}(M) = \left(\gamma_1 + \gamma_2 \log\left(\frac{|M_0|}{|M|}\right)\right)\frac{|M|}{n}$ and we determine the right constants γ_1 and γ_2 by using, for example, the same technique as Lebarbier in [7].

Remark 3
If the points $(x_i)_{1\leq i\leq n}$ of the design are random points $(X_i)_{1\leq i\leq n}$, then with the same approach, working first conditionnally to $(X_i)_{1\leq i\leq n}$,we get a similar result. For more details see ([9]).

4 The key to determine an adequate form of penalty: a concentration inequality for a χ^2 like statistic

This section is more technical. First we give an expression of $\|s - \hat{s}_{\hat{M}}\|_n^2$, which allows us to see that the penalty $\text{pen}(M)$ has to compensate the deviation of a χ^2 like statistic, denoted χ^2_M, in order that the PLSE $\hat{s}_{\hat{M}}$ satisfies an oracle type inequality. The square root of this statistic is the supremum of a random process. Then we explain why Bousquet's concentration inequality for the supremum of a random process is not convenient. And finally lemma 1 gives a self-made concentration inequality for χ^2_M. This concentration inequality is the main point of the proof of theorem 1, the remaining of the proof only consists in technical details.

Let us recall that $\hat{M} = \underset{M\in\mathcal{M}_n}{\arg\min}\ \left\{\|Y - \hat{s}_M\|_n^2 + \text{pen}(M)\right\}$
with the penalty pen to be chosen such that

$$\mathbb{E}\left(\|s - \hat{s}_{\hat{M}}\|_n^2\right) \leq C' \inf_M \left\{\|s - s_M\|_n^2 + \sigma^2\frac{|M|}{n}\right\}$$

According to the definition of $\hat{M}$, we have

$$\|s - \hat{s}_{\hat{M}}\|_n^2 = -2\left\langle\varepsilon, s - \hat{s}_{\hat{M}}\right\rangle_n - \text{pen}(\hat{M}) + \inf_{M\in\mathcal{M}_n}\left\{\|s - \hat{s}_M\|_n^2 + 2\left\langle\varepsilon, s - \hat{s}_M\right\rangle_n + \text{pen}(M)\right\}$$

Since $\hat{s}_M = s_M + \varepsilon_M$,

$$\left\langle\varepsilon, s - \hat{s}_M\right\rangle_n = \left\langle\varepsilon, s - s_M\right\rangle_n - \|\varepsilon_M\|_n^2 \text{ and } \|s - \hat{s}_M\|_n^2 = \|s - s_M\|_n^2 + \|\varepsilon_M\|_n^2$$

Thus

$$\|s - \hat{s}_{\hat{M}}\|_n^2 = 2\|\varepsilon_{\hat{M}}\|_n^2 - 2\left\langle\varepsilon, s - s_{\hat{M}}\right\rangle_n - \text{pen}(\hat{M}) + \inf_{M\in\mathcal{M}_n}\left\{\|s - s_M\|_n^2 - \|\varepsilon_M\|_n^2 + 2\left\langle\varepsilon, s - s_M\right\rangle_n + \text{pen}(M)\right\}$$

and

$$\|s - \hat{s}_{\hat{M}}\|_n^2 = \|s - s_{\hat{M}}\|_n^2 + \|\varepsilon_{\hat{M}}\|_n^2$$

We deduce from these two last equalities that for any $\theta \in (0,1)$,

$$\begin{aligned}(1-\theta)\|s - \hat{s}_{\hat{M}}\|_n^2 = {} & (2-\theta)\|\varepsilon_{\hat{M}}\|_n^2 - 2\left\langle\varepsilon, s - s_{\hat{M}}\right\rangle_n - \theta\|s - s_{\hat{M}}\|_n^2 - \text{pen}(\hat{M}) \\ & + \inf_{M\in\mathcal{M}_n}\left\{\|s - s_M\|_n^2 - \|\varepsilon_M\|_n^2 + 2\left\langle\varepsilon, s - s_M\right\rangle_n + \text{pen}(M)\right\}\end{aligned} \tag{6}$$

To get an oracle type inequality, the penalty pen(M) has to compensate the deviations of the statistics

$$\chi^2_M = \|\varepsilon_M\|_n^2 = \frac{1}{n}\sum_{J\in M}\frac{\left(\sum_{x_i\in J}\varepsilon_i\right)^2}{|J|} \text{ and } \langle \varepsilon, s-s_M\rangle_n$$

for all partitions $M \in \mathcal{M}_n$ simultaneously.
Thanks to assumption (4), it is easy to obtain the following concentration inequality for $\langle \varepsilon, s-s_M\rangle_n$

$$\text{for all } x>0 \qquad \mathbb{P}\left(\pm < \varepsilon, s-s_M >_n \geq \frac{\sigma}{\sqrt{n}}\|s-s_M\|_n\sqrt{2x} + \frac{b}{n}\left(\max_{1\leq i\leq n}|s(x_i)-s_M(x_i)|\right)x\right) \leq e^{-x}$$

If $\|s\|_\infty \leq R$ then

$$\text{for all } x>0 \qquad \mathbb{P}\left(\pm < \varepsilon, s-s_M >_n \geq \frac{\sigma}{\sqrt{n}}\|s-s_M\|_n\sqrt{2x} + \frac{2Rb}{n}x\right) \leq e^{-x} \tag{7}$$

It remains to study the deviation of the χ^2 like statistic χ^2_M around its expectation.

As

$$\chi_M = \|\varepsilon_M\|_n = \sup_{\substack{u\in S_M\\ \|u\|_n=1}} < \varepsilon, u >_n = \frac{1}{n}\sup_{\substack{u\in S_M\\ \|u\|_n=1}}\sum_{i=1}^n u_i\varepsilon_i$$

where the supremum is achieved with $u = \frac{\varepsilon_M}{\|\varepsilon_M\|_n}$, we could be tempted to use Bousquet's concentration inequality for the supremum of an empirical process. Thanks to Bousquet's result [4], we have for any $x>0$ and any $\gamma>0$:

$$\mathbb{P}\left(\chi_M \geq (1+\gamma)\mathbb{E}(\chi_M) + \frac{1}{n}\sqrt{2vx} + \frac{1}{n}(2+\gamma^{-1})bcx\right) \leq e^{-x}$$

where $c = \sup_{\substack{u\in S_M\\ \|u\|_n=1}}\|u\|_\infty$ and the variance term $v = \sum_{i=1}^n \sup_{\substack{u\in S_M\\ \|u\|_n=1}}\mathrm{Var}(u_i\varepsilon_i) \leq nc^2\sigma^2$.
The variance term v should be $\sup_u\sum_{i=1}^n \mathrm{Var}(u_i\varepsilon_i)$ instead of $\sum_{i=1}^n\sup_u \mathrm{Var}(u_i\varepsilon_i)$. With such a refinement, we would obtain here $v\leq n\sigma^2$ instead of $nc^2\sigma^2$ (and the presence of c in the last term $(2+\gamma^{-1})bcx$ would be solved by truncating χ_M and using the fact that the supremum, which defines χ_M, is achieved with $u = \frac{\varepsilon_M}{\chi_M}$). As Bousquet's concentration inequality is not convenient for our problem, we build our own concentration inequality.

Lemma 1
Let $b\in\mathbb{R}_+$ and $\sigma\in\mathbb{R}_+^*$ such that inequality (4) holds.
Let M_0 a partition of $\mathcal{X}$ and denote $N_{min} = \inf_{J\in M_0}|J|$.
Let $\delta>0$ and $\Omega_\delta = \left\{\forall J\in M_0;\ \left|\sum_{x_i\in J}\varepsilon_i\right| \leq \delta\sigma^2|J|\right\}$
For any partition M built from M_0 and for any $x>0$

$$\mathbb{P}\left(\chi^2_M \mathbb{1}_{\Omega_\delta} \geq \frac{\sigma^2}{n}|M| + 4\frac{\sigma^2}{n}(1+b\delta)\sqrt{2|M|x} + 2\frac{\sigma^2}{n}(1+b\delta)x\right) \leq e^{-x}$$

and

$$\mathbb{P}\left(\Omega_\delta^c\right) \leq 2\frac{n}{N_{min}}\exp\left(\frac{-\delta^2\sigma^2N_{min}}{2(1+b\delta)}\right)$$

If $b=0$, we do not need to truncate χ^2_M with Ω_δ and for any $x>0$

$$\mathbb{P}\left(\chi^2_M \geq \frac{\sigma^2}{n}|M| + 4\frac{\sigma^2}{n}\sqrt{2|M|x} + 2\frac{\sigma^2}{n}x\right) \leq e^{-x}$$

The concentration inequalities of the $\left(\chi^2_M\right)_{M\in\mathcal{M}_n}$ and $\left(\langle\varepsilon, s-s_M\rangle_n\right)_{M\in\mathcal{M}_n}$ are the key to determine the adequate form of penalty. $\langle\varepsilon, s-s_M\rangle_n$ is centered and the expectation of χ^2_M is upper bounded by $\frac{\sigma^2}{n}|M|$. The weights $(x_M)_{M\in\mathcal{M}_n}$ satisfying $\sum_{M\in\mathcal{M}_n} e^{-x_M} \leq \Sigma \in \mathbb{R}^*_+$ allow to control χ^2_M and $\langle\varepsilon, s-s_M\rangle_n$ for all $M \in \mathcal{M}_n$ simultaneously. This is the reason why, as told in section 3, the right penalty pen is the sum of two terms: one proportional to $\frac{|M|}{n}$ (corresponding to $\mathbb{E}(\chi^2_M)$) and a second depending on x_M.

Remark 4
This lemma is based on Bernstein inequality. We must truncate to get concentration inequalities which remain sharp when summing them over all partitions $M \in \mathcal{M}_n$. In the context of histogram density estimation, Castellan ([6]) has to control an other χ^2 like statistic. Like here, the main point is to truncate the statistic. While she concludes by applying a Talagrand inequality to the truncated statistic, we use Bernstein inequality.

Remark 5
If $N_{min} \geq 2(k+1)\frac{(1+b\delta)}{\delta^2\sigma^2}\log n$,

$$\mathbb{P}\left(\Omega^c_\delta\right) \leq \frac{1}{(k+1)}\frac{\delta^2\sigma^2}{(1+b\delta)}\frac{1}{n^k\log n}$$

5 Proof of lemma 1

Let M a partition built from M_0 and denote, for any $J \in M$,

$$Z_J = \frac{\left(\sum_{i\in J}\varepsilon_i\right)^2}{|J|} \wedge \left(\delta^2\sigma^4|J|\right)$$

$(Z_J)_{J\in M}$ are independent random variables, $\mathbb{E}\left(Z_J\right) \leq \mathbb{E}\left(\varepsilon_1^2\right) \leq \sigma^2$, and for any $k \geq 2$ we have

$$\begin{aligned}
\mathbb{E}\left(|Z_J|^k\right) &= \frac{1}{|J|^k}\mathbb{E}\left[\left(\left|\sum_{i\in J}\varepsilon_i\right| \wedge \left(\delta\sigma^2|J|\right)\right)^{2k}\right] \\
&= \frac{1}{|J|^k}\int_0^{+\infty} 2kx^{2k-1}\mathbb{P}\left(\left|\sum_{i\in J}\varepsilon_i\right| \wedge \left(\delta\sigma^2|J|\right) \geq x\right)dx \\
&= \frac{1}{|J|^k}\int_0^{\delta\sigma^2|J|} 2kx^{2k-1}\mathbb{P}\left(\left|\sum_{i\in J}\varepsilon_i\right| \geq x\right)dx
\end{aligned}$$

We deduce from assumption (4) that for any $x > 0$

$$\mathbb{P}\left(\left|\sum_{i\in J}\varepsilon_i\right| \geq x\right) \leq 2\exp\left(\frac{-x^2}{2(\sigma^2|J|+bx)}\right)$$

Thus

$$\begin{aligned}
\mathbb{E}\left(|Z_J|^k\right) &\leq \frac{1}{|J|^k}\int_0^{\delta\sigma^2|J|} 2kx^{2k-1}2\exp\left(\frac{-x^2}{2(\sigma^2|J|+bx)}\right)dx \\
&\leq \frac{4k}{|J|^k}\int_0^{+\infty} x^{2k-1}\exp\left(\frac{-x^2}{2\sigma^2|J|(1+b\delta)}\right)dx
\end{aligned}$$

Integrating part by part, we get

$$\mathbb{E}\left(|Z_J|^k\right) \leq \frac{k!}{2}\left(4\sigma^2(1+b\delta)\right)^2\left(2\sigma^2(1+b\delta)\right)^{k-2}$$

Thanks to Bernstein inequality we obtain that for any $x>0$

$$\mathbb{P}\left(\sum_{J\in M} Z_J \geq \sigma^2|M| + 4\sigma^2(1+b\delta)\sqrt{2|M|x} + 2\sigma^2(1+b\delta)x\right) \leq e^{-x}$$

Since $\frac{1}{n}\sum_{J\in M} Z_J = \chi^2_M$ on the set Ω_δ,

$$\mathbb{P}\left(\chi^2_M 1\!\!1_{\Omega_\delta} \geq \frac{\sigma^2}{n}|M| + 4\frac{\sigma^2}{n}(1+b\delta)\sqrt{2|M|x} + 2\frac{\sigma^2}{n}(1+b\delta)x\right) \leq e^{-x}$$

Thanks to assumption (4), for any $J\in M_0$, we have

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i\in J}\varepsilon_i\right| \geq \delta\sigma^2|J|\right) &\leq 2\exp\left(\frac{-\delta^2\sigma^2|J|}{2(1+b\delta)}\right)\\ &\leq 2\exp\left(\frac{-\delta^2\sigma^2 N_{min}}{2(1+b\delta)}\right)\end{aligned}$$

Summing these inequalities over $J\in M_0$, we get

$$\begin{aligned}\mathbb{P}\left(\Omega_\delta^c\right) &\leq 2|M_0|\exp\left(\frac{-\delta^2\sigma^2 N_{min}}{2(1+b\delta)}\right)\\ &\leq 2\frac{n}{N_{min}}\exp\left(\frac{-\delta^2\sigma^2 N_{min}}{2(1+b\delta)}\right)\end{aligned}$$

6 Proof of the theorem

Let $\theta\in(0,1)$ and $K>2-\theta$.

According to (6),

$$(1-\theta)\|s-\hat{s}_{\hat{M}}\|_n^2 = \Delta_{\widehat{M}} + \inf_{M\in\mathcal{M}_n} R_M \tag{8}$$

where

$$\begin{aligned}\Delta_M &= (2-\theta)\|\varepsilon_M\|_n^2 - 2<\varepsilon, s-s_M>_n -\theta\|s-s_M\|_n^2 - \mathrm{pen}(M)\\ R_M &= \|s-s_M\|_n^2 - \|\varepsilon_M\|_n^2 + 2<\varepsilon, s-s_M>_n + \mathrm{pen}(M)\end{aligned}$$

Let denote $\Omega = \left\{\forall J\in M_0;\ \left|\sum_{i\in J}\varepsilon_i\right| \leq \frac{\sigma^2}{b}|J|\right\}$

Thanks to lemma 1,

$$\mathbb{P}\left(\Omega^c\right) \leq 2\frac{n}{N_{min}}\exp\left(\frac{-\sigma^2 N_{min}}{4b^2}\right)$$

and, for any $M\in\mathcal{M}_n$ and any $x>0$,

$$\mathbb{P}\left(\|\varepsilon_M\|_n^2 1\!\!1_\Omega \geq \frac{\sigma^2}{n}|M| + 8\frac{\sigma^2}{n}\sqrt{2|M|x} + 4\frac{\sigma^2}{n}x\right) \leq e^{-x} \tag{9}$$

Thanks to (7), we have for any $M \in \mathcal{M}_n$ and any $x > 0$,

$$\mathbb{P}\Big(- < \varepsilon, s - s_M >_n \geq \frac{\sigma}{\sqrt{n}} \|s - s_M\|_n \sqrt{2x} + \frac{2Rb}{n} x \Big) \quad \leq \quad e^{-x} \tag{10}$$

Setting $x = x_M + \xi$ with $\xi > 0$, and summing all inequalities (9) and (10) with respect to $M \in \mathcal{M}_n$, we derive a set E_ξ such that:

- $\mathbb{P}\left(E_\xi^c\right) \leq e^{-\xi} 2\Sigma$
- on the set $E_\xi \bigcap \Omega$, for any M,

$$\begin{aligned} \Delta_M \quad \leq \quad & (2-\theta)\frac{\sigma^2}{n}|M| + 8(2-\theta)\frac{\sigma^2}{n}\sqrt{2|M|(x_M+\xi)} + 4(2-\theta)\frac{\sigma^2}{n}(x_M+\xi) \\ & + 2\frac{\sigma}{\sqrt{n}}\|s - s_M\|_n \sqrt{2(x_M+\xi)} + \frac{4Rb}{n}(x_M+\xi) \\ & - \theta\|s - s_M\|_n^2 - \mathrm{pen}(M) \end{aligned}$$

Using the two following inequalities

$$2\tfrac{\sigma}{\sqrt{n}}\|s - s_M\|_n \sqrt{2(x_M+\xi)} \leq \theta\|s - s_M\|_n^2 + \tfrac{2}{\theta}\tfrac{\sigma^2}{n}(x_M+\xi),$$

$$8(2-\theta)\tfrac{\sigma^2}{n}\sqrt{2|M|(x_M+\xi)} \leq 8\sqrt{2}(2-\theta)\tfrac{\sigma^2}{n}\sqrt{|M|x_M} + 4\sqrt{2}(2-\theta)\tfrac{\sigma^2}{n}\left(\eta|M| + \eta^{-1}\xi\right)$$

with $\eta = \frac{1}{4\sqrt{2}}\frac{K+\theta-2}{2-\theta} > 0$, we deduce that on the set $E_\xi \bigcap \Omega$, for any M,

$$\begin{aligned} \Delta_M \quad \leq \quad & (2-\theta)\frac{\sigma^2}{n}|M| + 8(2-\theta)\frac{\sigma^2}{n}\sqrt{2|M|(x_M+\xi)} \\ & + \left(4(2-\theta) + \frac{2}{\theta}\right)\frac{\sigma^2}{n}(x_M+\xi) + \frac{4Rb}{n}(x_M+\xi) \\ & - \mathrm{pen}(M) \\ \leq \quad & K\frac{\sigma^2}{n}|M| + 8\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left(4(2-\theta) + \frac{2}{\theta}\right)\frac{\sigma^2}{n}x_M + \frac{4Rb}{n}x_M \\ & + \left(4(2-\theta)\left(1 + \frac{8(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta}\right)\frac{\sigma^2}{n}\xi + \frac{4Rb}{n}\xi - \mathrm{pen}(M) \end{aligned}$$

Taking a penalty pen wich compensates for all the other terms in M, i.e.

$$\mathrm{pen}(M) \geq K\frac{\sigma^2}{n}|M| + 8\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left[\left(4(2-\theta) + \frac{2}{\theta}\right)\frac{\sigma^2}{n} + \frac{4Rb}{n}\right] x_M$$

we get that, on the set $E_\xi \bigcap \Omega$,

$$\Delta_{\widehat{M}} \leq \left(4(2-\theta)\left(1 + \frac{8(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta}\right)\frac{\sigma^2}{n}\xi + \frac{4Rb}{n}\xi$$

In other words, on the set E_ξ,

$$\Delta_{\widehat{M}} \mathbb{1}\!\mathrm{I}_\Omega \leq \left(4(2-\theta)\left(1 + \frac{8(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta}\right)\frac{\sigma^2}{n}\xi + \frac{4Rb}{n}\xi$$

Integrating with respect to ξ,

$$\mathbb{E}\left(\Delta_{\widehat{M}} \mathbb{1}\!\mathrm{I}_\Omega\right) \leq 2\left(4(2-\theta)\left(1 + \frac{8(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta}\right)\frac{\sigma^2}{n}\Sigma + \frac{8Rb}{n}\Sigma \tag{11}$$

We are going now to control $\mathbb{E}\left(\inf_M R_M \mathbb{1}_\Omega\right)$.
Thanks to (7), for any M and any $x > 0$

$$\mathbb{P}\Big(\langle \varepsilon, s - s_M \rangle_n \geq \frac{\sigma}{\sqrt{n}} \|s - s_M\|_n \sqrt{2x} + \frac{2Rb}{n} x \Big) \quad \leq \quad e^{-x}$$

Thus we derive a set F_ξ such that

- $\mathbb{P}\left(F_\xi^c\right) \leq e^{-\xi}\Sigma$
- on the set F_ξ, for any M,

$$< \varepsilon, s - s_M >_n \quad \leq \quad \frac{\sigma}{\sqrt{n}} \|s - s_M\|_n \sqrt{2\left(x_M + \xi\right)} + \frac{2Rb}{n}\left(x_M + \xi\right)$$

It follows from definition of R_M that on the set F_ξ, for any M,

$$\begin{aligned} R_M &\leq \|s - s_M\|_n^2 + 2\frac{\sigma}{\sqrt{n}} \|s - s_M\|_n \sqrt{2\left(x_M + \xi\right)} + \frac{4Rb}{n}\left(x_M + \xi\right) + \mathrm{pen}(M) \\ &\leq 2\|s - s_M\|_n^2 + 2\frac{\sigma^2}{n}\left(x_M + \xi\right) + \frac{4Rb}{n}\left(x_M + \xi\right) + \mathrm{pen}(M) \\ &\leq 2\|s - s_M\|_n^2 + 2\mathrm{pen}(M) + 2\frac{\sigma^2}{n}\xi + \frac{4Rb}{n}\xi \end{aligned}$$

And

$$\begin{aligned} \mathbb{E}\left(\inf_M R_M \mathbb{1}_\Omega\right) \quad \leq \quad & 2\inf_M \left\{ \|s - s_M\|_n^2 + \mathrm{pen}(M) \right\} \\ & + 2\frac{\sigma^2}{n}\Sigma + \frac{4Rb}{n}\Sigma \end{aligned} \tag{12}$$

We conclude from (8), (11) and (12) that

$$\begin{aligned} (1-\theta)\mathbb{E}\left(\|s - \hat{s}_{\hat{M}}\|_n^2 \mathbb{1}_\Omega\right) \quad \leq \quad & 2\inf_M \left\{ \|s - s_M\|_n^2 + \mathrm{pen}(M) \right\} \\ & + \left(8(2-\theta)\left(1 + \frac{8(2-\theta)}{K + \theta - 2}\right) + \frac{4}{\theta} + 2 \right) \frac{\sigma^2}{n}\Sigma + \frac{12Rb}{n}\Sigma \end{aligned}$$

It remains to control $\mathbb{E}\left(\|s - \hat{s}_{\hat{M}}\|_n^2 \mathbb{1}_{\Omega^c}\right)$, except if $b = 0$ in which case it is finished.

$$\begin{aligned} \mathbb{E}\left(\|s - \hat{s}_{\hat{M}}\|_n^2 \mathbb{1}_{\Omega^c}\right) &= \mathbb{E}\left(\|s - s_{\widehat{M}}\|_n^2 \mathbb{1}_{\Omega^c}\right) + \mathbb{E}\left(\|\varepsilon_{\widehat{M}}\|_n^2 \mathbb{1}_{\Omega^c}\right) \\ &\leq \mathbb{E}\left(\|s\|_n^2 \mathbb{1}_{\Omega^c}\right) + \mathbb{E}\left(\|\varepsilon_{M_0}\|_n^2 \mathbb{1}_{\Omega^c}\right) \\ &\leq R^2 \mathbb{P}\left(\Omega^c\right) + \sqrt{\mathbb{E}\left(\|\varepsilon_{M_0}\|_n^4\right)} \sqrt{\mathbb{P}\left(\Omega^c\right)} \end{aligned}$$

By developping $\|\varepsilon_{M_0}\|_n^4$, since $\mathbb{E}\left(\varepsilon_i^2\right) \leq \sigma^2$ and $\mathbb{E}\left(\varepsilon_i^4\right) \leq C(b, \sigma^2)^2$, we get

$$\begin{aligned} \mathbb{E}\left(\|\varepsilon_{M_0}\|_n^4\right) &\leq \frac{\sigma^4 |M_0|^2}{n^2} + \frac{C(b,\sigma^2)^2 |M_0|}{n^2 N_{min}} + \frac{3\sigma^4 |M_0|}{n^2} \\ &\leq \frac{\sigma^4}{N_{min}^2} + \frac{C(b,\sigma^2)^2}{n N_{min}^2} + \frac{3\sigma^4}{n N_{min}} \end{aligned}$$

and thus

$$\mathbb{E}\left(\|s - \hat{s}_{\hat{M}}\|_n^2 \mathbb{1}_{\Omega^c}\right) \quad \leq \quad R^2 \mathbb{P}\left(\Omega^c\right) + \left(\frac{\sigma^2}{N_{min}} + \frac{C(b,\sigma^2)}{\sqrt{n} N_{min}} + \frac{\sqrt{3}\sigma^2}{\sqrt{n N_{min}}} \right) \sqrt{\mathbb{P}\left(\Omega^c\right)}$$

Let us recall that

$$\mathbb{P}\left(\Omega^c\right) \leq 2\frac{n}{N_{min}}\exp\left(\frac{-\sigma^2 N_{min}}{4b^2}\right)$$

For $N_{min} \geq 12\frac{b^2}{\sigma^2}\log n$,

$$\mathbb{P}\left(\Omega^c\right) \leq \frac{\sigma^2}{6b^2}\frac{1}{n^2\log n}$$

and

$$\begin{aligned}
\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2 1\!\!1_{\Omega^c}\right) &\leq \frac{R^2\sigma^2}{6b^2}\frac{1}{n^2\log n} \\
&\quad + \left(\frac{\sigma^4}{12b^2\log n} + \frac{\sigma^2 C(b,\sigma^2)}{12b^2\sqrt{n}\log n} + \frac{\sqrt{3}\sigma^3}{\sqrt{12b^2 n\log n}}\right)\frac{1}{\sqrt{6}}\frac{\sigma}{b}\frac{1}{n\sqrt{\log n}} \\
&\leq \left[\frac{R^2}{6} + \frac{1}{\sqrt{6}}\left(\frac{\sigma}{12b} + \frac{C(b,\sigma^2)}{12b\sigma} + \frac{1}{2}\right)\sigma^2\right]\frac{\sigma^2}{b^2}\frac{1}{n(\log n)^{3/2}}
\end{aligned}$$

Finally we have the following result:
Taking a penalty which satisfies for all $M \in \mathcal{M}_n$

$$\mathrm{pen}(M) \geq K\frac{\sigma^2}{n}|M| + 8\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left[\left(4(2-\theta)+\frac{2}{\theta}\right)\frac{\sigma^2}{n} + \frac{4Rb}{n}\right]x_M$$

if $N_{min} \geq 12\frac{b^2}{\sigma^2}\log n$, we have

$$\begin{aligned}
\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\right) &\leq \frac{2}{1-\theta}\inf_M\left\{\|s-s_M\|_n^2 + \mathrm{pen}(M)\right\} \\
&\quad + \frac{1}{1-\theta}\left(8(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{4}{\theta}+2\right)\frac{\sigma^2}{n}\Sigma + \frac{12}{1-\theta}\frac{Rb}{n}\Sigma \\
&\quad + \left[\frac{R^2}{6} + \frac{1}{\sqrt{6}}\left(\frac{\sigma}{12b} + \frac{C(b,\sigma^2)}{12b\sigma} + \frac{1}{2}\right)\sigma^2\right]\frac{\sigma^2}{b^2}\frac{1\!\!1_{b\neq 0}}{n(\log n)^{3/2}}
\end{aligned}$$

References

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

[2] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, (2001).

[3] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. To be published in Probability Theory and Related Fields, (2005).

[4] O. Bousquet. Concentration Inequalities for Sub-Additive Functions Using the Entropy Method. *Stochastic Inequalities and Applications*, 56:213–247, (2003).

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification And Regression Trees.* Chapman et Hall, (1984).

[6] G. Castellan. Modified Akaike's criterion for histogram density estimation. *C.R. Acad. Sci. Paris Sér. I Math.*, 330(8):729–732, (2000).

[7] E. Lebarbier. *Quelques approches pour la détection de ruptures à horizon fini.* PhD thesis, Université Paris XI Orsay, (2002).

[8] P. Massart. Notes de Saint-Flour. Lecture Notes to be published, (2003).

[9] M. Sauvé and C. Tuleau. Variables selection through CART. *Rapport de recherche INRIA*, (2006).

Unité de recherche INRIA Futurs
Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399