Histogram selection in non Gaussian regression

Marie Sauvé

ESAIM: Probability and Statistics Will be set by the publisher
URL: http://www.emath.fr/ps/

HISTOGRAM SELECTION IN NON GAUSSIAN REGRESSION

Marie Sauvé[1]

Abstract. We deal with the problem of choosing a piecewise constant estimator of a regression function s mapping $\mathcal{X}$ into $\mathbb{R}$. We consider a non Gaussian regression framework with deterministic design points, and we adopt the non asymptotic approach of model selection via penalization developed by Birgé and Massart. Given a collection of partitions of $\mathcal{X}$, with possibly exponential complexity, and the corresponding collection of piecewise constant 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.

1991 Mathematics Subject Classification. 62G08, 62G05.

.

1. Introduction

We consider the fixed design regression framework. We observe n pairs $(x_i, Y_i)_{1\leq i\leq n}$, where the x_i's are fixed points belonging to some set $\mathcal{X}$ and the Y_i's are real valued random variables. We suppose that:

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

where s is an unknown function mapping $\mathcal{X}$ into $\mathbb{R}$, and $(\varepsilon_i)_{1\leq i\leq n}$ are centered, independent and identically distributed random perturbations. Our aim is to get informations on s from the observations $(x_i, Y_i)_{1\leq i\leq n}$.

In order to get a simple estimator of s, we consider a partition M_0 of $\mathcal{X}$ with a large number of small cells containing at least one point x_i, and we minimize the least squares contrast over the class S_{M_0} of piecewise constant functions defined on the partition M_0. The resulting estimator is denoted by $\hat{s}_{M_0}$ and is called the least squares estimator over S_{M_0}. S_{M_0} is called the histogram model associated with M_0. It is a linear space with finite dimension $D_{M_0} = |M_0|$, where $|M_0|$ is the number of cells of the partition M_0. Denoting $\|.\|_n$ the Euclidean norm on $\mathbb{R}^n$ scaled by a factor $n^{-1/2}$, and denoting for any function $u \in \mathbb{R}^{\mathcal{X}}$ $\|u\|_n = \|(u(x_i))_{1\leq i\leq n}\|_n$, 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:

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

Keywords and phrases: CART, change-points detection, deviation inequalities, model selection, oracle inequalities, regression

[1] Laboratoire de mathématiques - Bâtiment 425, Université Paris Sud, 91405 Orsay cedex, France
E-mail: marie.sauve@math.u-psud.fr
Tel: (+33) (1) 69 15 57 82
Fax: (+33) (1) 69 15 72 34

This provisional PDF is the accepted version. The article should be cited as: ESAIM: PS, doi: 10.1051/ps:2008002

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 too large number of cells and that we should rather choose a partition M built from M_0 which makes a better trade-off between the bias $\inf_{u \in S_M} \|s-u\|_n^2$ and the variance $\frac{\tau^2}{n}|M|$.

In this paper, we propose a procedure to select a partition $\hat{M}$ (or equivalently a histogram model $S_{\hat{M}}$) according to the data, and we estimate s by $\hat{s}_{\hat{M}}$ the least squares estimator over $S_{\hat{M}}$. There are already many works on data-driven histogram selection not only for regression [7,9,10,14] but also for classification [7,11] and density estimation [5,8,9,11]. We focus here on the regression framework and we adopt the non asymptotic approach of model selection via penalization developed by Birgé and Massart.
Our estimation procedure is as follows. We consider a collection $\mathcal{M}_n$ of partitions of $\mathcal{X}$ all built from an initial partition M_0, and the corresponding collection $(S_M)_{M \in \mathcal{M}_n}$ of histogram models. Denoting $\hat{s}_M$ the least squares estimator over the model S_M, the best model S_{M^*} is the one which minimizes $\mathbb{E}\left(\|s-\hat{s}_M\|_n^2\right)$ among $(S_M)_{M \in \mathcal{M}_n}$. Unfortunately this model depends on the unknown function s, and $\hat{s}_{M^*}$ can not be used as an estimator of s. Our aim is to find a data driven criterion whose minimizer $S_{\hat{M}}$ is an approximately best model, and to estimate s by $\hat{s}_{\hat{M}}$. We select a model $S_{\hat{M}}$ by minimizing over $\mathcal{M}_n$ a penalized least squares criterion $\mathrm{crit}(M) = \|Y-\hat{s}_M\|_n^2 + \mathrm{pen}(M)$:

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

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

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

The inequality (2) 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 penalty pen for which the associated penalized least squares estimator satisfies an oracle type inequality.

One of the first penalized least squares criterion is Mallows' C_p criterion [12] which corresponds to $\mathrm{pen}(M) = 2\frac{\tau^2}{n}D_M$ It leads to an oracle type inequality when the number of models with a given dimension D is a polynomial function of D (as proved by [3,4] in the Gaussian case, and by [1] in a very general case), but it can lead to very bad results when the collection of models is more complex (see [4]).
In this paper, we determine a penalty which allows to deal with large collections of histogram models. The proposed penalty $\mathrm{pen}(M)$ is the sum of two terms: the first one is proportional to $\frac{D_M}{n} = \frac{|M|}{n}$ and the second one depends on the complexity of the collection $\mathcal{M}_n$. It has the same form as the penalty obtained by Birgé and Massart [4] in the Gaussian case and those obtained by Baraud, Comte and Viennet [2] in the sub-Gaussian case, for any collections of models (not only for histogram models). But we do not assume the $(\varepsilon_i)_{1 \leq i \leq n}$ to be Gaussian nor sub-Gaussian, we only suppose that they have exponential moments around 0.

From a technical point of view, we follow the same ideas as [1,4]. In this paper like in [1,4], in order to get an adequate penalty, the main point is to control the random variables $\chi_M^2 = \|\varepsilon_M\|_n^2$ where $\varepsilon_M = \arg \min_{u \in S_M} \|\varepsilon - u\|_n^2$ for all $M \in \mathcal{M}_n$ simultaneously. The difference between our work, [1], and [4], is located in the construction of sharp deviation inequalities for the $(\chi_M^2)_{M \in \mathcal{M}_n}$, which remain sharp when summing them over all $M \in \mathcal{M}_n$. In the Gaussian case (handled in [4]), the $\frac{n}{\tau^2}\chi_M^2$'s are χ^2 distributed. But in the non Gaussian case, it is much more difficult to study the deviations of these variables around their expectations. Under a mild integrability condition on the $(\varepsilon_i)_{1 \leq i \leq n}$ (assuming that $\mathbb{E}\left(|\varepsilon_i|^p\right) < +\infty$ for some $p \geq 2$), Baraud [1] gives a polynomial deviation inequality for the χ_M^2's. This inequality allows him to prove that penalties $\mathrm{pen}(M) = K'\frac{\tau^2}{n}D_M$, with $K' > 1$, lead to oracle type inequalities when the number of models with a given dimension D is a polynomial function of D. In order to deal with bigger collections of models, we need exponential deviation inequalities

for the χ^2_M's. By writing $\chi_M = \sup_{u \in B_M} < \varepsilon, u >_n$, with $B_M = \{u \in S_M;\ \|u\|_n \leq 1\}$, we can apply Bousquet's exponential concentration inequality for a supremum of an empirical process [6]. Unfortunately this general result is not sufficient here. Instead of viewing χ_M as a supremum, we can view χ^2_M as a χ^2 like random variable and write it as a sum of squares. For histogram models, we can then build adequate exponential deviation inequalities by hand, using only Bernstein's inequality. This is the reason why we determine a penalty which we prove to lead to an oracle inequality only for histogram models.

Thanks to this penalty, given a collection $\mathcal{M}_n$ of partitions, we get an estimator $\hat{s}_{\hat{M}}$ which is simple, easy to interpret and close to the optimal one among the collection of piecewise constant estimators $(\hat{s}_M)_{M \in \mathcal{M}_n}$. Unfortunately, since the estimators $(\hat{s}_M)_{M \in \mathcal{M}_n}$ are sharply discontinuous, even the best one may not provide an accurate estimation of s.

Histogram model selection may not lead to an accurate estimation of the regression function s, but it has important applications to clustering and it enables to detect the change-points of s. In the framework (1) with $\mathcal{X} = [0,1]$ and $x_i = \frac{i}{n}$, in order to detect the change-points of s, Lebarbier [10, chapter 2] considers the collection $\mathcal{M}_n$ of all partitions with endpoints belonging to the grid $(x_i)_{1 \leq i \leq n}$, and the corresponding collection $(S_M)_{M \in \mathcal{M}_n}$ of histogram models. Then, she selects a partition $\hat{M}$ among $\mathcal{M}_n$ by minimizing a penalized least squares criterion. For this collection $\mathcal{M}_n$, $|\{M \in \mathcal{M}_n;\ |M| = D\}| = \binom{n-1}{D-1}$, therefore Baraud's result [1] does not apply to this case, and penalties proportional to $\frac{|M|}{n}$ are not sufficient. Assuming the perturbations ε_i to be Gaussian, Lebarbier applies the model selection result of Birgé and Massart [3] to the collection $(S_M)_{M \in \mathcal{M}_n}$. She gets a penalty defined up to two multiplicative constants. Then she proposes a method to calibrate them according to the data and therefore gives a procedure to automatically detect the change-points of a Gaussian signal according to the data. Thanks to our result, this procedure can be extended to detect the change-points of a regression function without assuming the perturbations ε_i to be Gaussian.

One of the most famous statistical issues is variable 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 $\left(x^1, \ldots, x^p\right)$ and Y, we can use histogram models. First, we associate one or more histogram models to each subset of variables. Then, we get a large collection of models (at least 2^p models), among which we select one by minimizing a penalized least squares criterion. And finally, we keep the subset of variables which is associated with the selected model. This idea is used by Sauvé and Tuleau [16], who propose a variable selection procedure based on histogram model selection.

The CART algorithm (Classification And Regression Trees), proposed by Breiman *et al.* [7], involves histogram models. Our result allows to validate the pruning step of CART in a non Gaussian regression framework.

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 random variable. The section 4 is more technical, it exposes a deviation inequality for a χ^2 like random variable and explains why the existing deviation 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 (1) and we look for a best or approximately best piecewise constant 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 (1), then we define the piecewise constant 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 random perturbations $(\varepsilon_i)_{1\leq i\leq n}$ have finite exponential moments around 0. This assumption is equivalent to the existence of two constants $b\in\mathbb{R}_+$ and $\sigma\in\mathbb{R}_+^*$ such that

$$\forall\ \lambda\in(-1/b,1/b)\qquad \log\mathbb{E}\left(e^{\lambda\varepsilon_i}\right)\leq\frac{\sigma^2\lambda^2}{2(1-b|\lambda|)} \tag{3}$$

σ^2 is necessarily 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 2.1. Under assumption (3), 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 (3) to this last inequality because with the last one we loose a factor 2 in the variance term.

Remark 2.2. Thanks to assumption (3) and Cramer-Chernoff method (see [13, section 2.1]), we can easily get deviation inequalities for any linear combination of the $(\varepsilon_i)_{1\leq i\leq n}$.
First, since the $(\varepsilon_i)_{1\leq i\leq n}$ are independent, we get from inequality (3) similar inequalities for any linear combination $\sum_{i=1}^n\alpha_i\varepsilon_i$. Denoting $\|\alpha\|_\infty=\max_{1\leq i\leq n}|\alpha_i|$ and $v=\sigma^2\left(\sum_{i=1}^n\alpha_i^2\right)$,

$$\forall\ \lambda\in\left(0,\frac{1}{b\|\alpha\|_\infty}\right)\qquad \log\mathbb{E}\left(e^{\lambda\sum_{i=1}^n\alpha_i\varepsilon_i}\right)\leq\frac{v\lambda^2}{2(1-b\|\alpha\|_\infty\lambda)}. \tag{4}$$

We denote by $\psi(\lambda)$ the right term of (4) and by $h(u)=1+u-\sqrt{1+2u}$ for any $u\in\mathbb{R}_+^*$.
Then, applying Cramer-Chernoff method, since for any $x>0$

$$\sup_{0<\lambda<\frac{1}{b\|\alpha\|_\infty}}\{\lambda x-\psi(\lambda)\}=\frac{v}{b^2\|\alpha\|_\infty^2}h\left(\frac{b\|\alpha\|_\infty x}{v}\right),$$

we get for any $x>0$:

$$\mathbb{P}\left(\sum_{i=1}^n\alpha_i\varepsilon_i\geq x\right)\quad\leq\quad\exp\left(-\frac{v}{b^2\|\alpha\|_\infty^2}h\left(\frac{b\|\alpha\|_\infty x}{v}\right)\right).$$

Finally we deduce the two following inequalities:

- Since h is inversible with $h^{-1}(u)=u+\sqrt{2u}$,

$$\mathbb{P}\left(\sum_{i=1}^n\alpha_i\varepsilon_i\geq\sqrt{2vx}+b\|\alpha\|_\infty x\right)\quad\leq\quad e^{-x}.$$

- Since $h(u) \geq \frac{u^2}{2(1+u)}$,

$$\mathbb{P}\left(\sum_{i=1}^{n} \alpha_i \varepsilon_i \geq x\right) \quad \leq \quad \exp\left(\frac{-x^2}{2\left(v + b\|\alpha\|_\infty x\right)}\right).$$

2.2. **The piecewise constant estimators**

For a given partition M of $\mathcal{X}$ whose cells contain at least one point x_i, 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 $n^{-1/2}$, $Y = (Y_i)_{1 \leq i \leq n}$, and for $u \in S_M$, the vector $(u(x_i))_{1 \leq i \leq n} \in \mathbb{R}^n$ is denoted by u too. S_M is the histogram model associated with M and $\hat{s}_M$ is the piecewise constant estimator belonging to S_M which plays the role of benchmark among all the estimators in S_M.

Let now calculate the quadratic risk of $\hat{s}_M$: $\mathbb{E}\left(\|s - \hat{s}_M\|_n^2\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(s(x_i) - \hat{s}_M(x_i)\right)^2\right)$.

To this end, we denote by

$s_M = \arg\min_{u \in S_M} \|s - u\|_n^2$,

$\varepsilon_M = \arg\min_{u \in S_M} \|\varepsilon - u\|_n^2$ where $\varepsilon = (\varepsilon_i)_{1 \leq i \leq n}$,

$|M|$ the number of elements of the partition M.

$\hat{s}_M$, s_M and ε_M are respectively the orthogonal projections of Y, s and ε on the space S_M according to $\|.\|_n$. Thanks to Pythagore's equality, we get that:

$$\mathbb{E}\left(\|s - \hat{s}_M\|_n^2\right) \quad = \quad \|s - s_M\|_n^2 + \mathbb{E}\left(\|\varepsilon_M\|_n^2\right).$$

For any element J of the partition M, we denote by $|J| = |\{1 \leq i \leq n;\ x_i \in J\}|$ and by $\mathbb{1}_J : x \to 1$ if $x \in J$ and 0 if $x \notin J$. Since $\left(\sqrt{\frac{n}{|J|}}\mathbb{1}_J\right)_{J \in M}$ is an orthonormal basis of $(S_M, \|.\|_n)$, we have

$$\|\varepsilon_M\|_n^2 \quad = \quad \sum_{J \in M}\left\langle \varepsilon, \sqrt{\frac{n}{|J|}}\mathbb{1}_J \right\rangle_n^2 \quad = \quad \frac{1}{n}\sum_{J \in M}\frac{\left(\sum_{x_i \in J}\varepsilon_i\right)^2}{|J|}. \tag{5}$$

Since $(\varepsilon_i)_{1 \leq i \leq n}$ are centered, independent and identically distributed random variables with $\mathbb{E}(\varepsilon_i^2) \leq \sigma^2$, we get that

$$\mathbb{E}\left(\|\varepsilon_M\|_n^2\right) \quad = \quad \mathbb{E}(\varepsilon_1^2)\frac{|M|}{n} \quad \leq \quad \sigma^2\frac{|M|}{n}.$$

Therefore

$$\mathbb{E}\left(\|s - \hat{s}_M\|_n^2\right) \quad = \quad \|s - s_M\|_n^2 + \mathbb{E}(\varepsilon_1^2)\frac{|M|}{n} \quad \leq \quad \|s - s_M\|_n^2 + \sigma^2\frac{|M|}{n}.$$

Remark 2.3. In the following, the random variable $\|\varepsilon_M\|_n^2$ is denoted by χ_M^2. Thanks to the decomposition (5), we can see χ_M^2 as a χ^2 like random variable. If the $(\varepsilon_i)_{1 \leq i \leq n}$ were Gaussian variables with variance τ^2, then the variables $\frac{n}{\tau^2}\chi_M^2$ would be $\chi^2(|M|)$-distributed.

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 piecewise constant 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 $\mathrm{crit}(M) = \|Y - \hat{s}_M\|_n^2 + \mathrm{pen}(M)$ over $\mathcal{M}_n$:

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

It remains to provide a penalty pen such that the partition $\hat{M}$ is close to the optimal partition, in the sense that the penalized least squares estiamtor $\hat{s}_{\hat{M}}$ satisfies an oracle inequality like (2). 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 [4] and those of Baraud [1], and we study in more detail two particular families of partitions which are too large to apply Baraud's result.

Example 1: We consider $\mathcal{X} = [0,1]$ and a grid on $[0,1]$ such that there are at least N_{min} points x_i between two consecutive grid points. For example, we can take the grid $(v_j)_{1 \leq j \leq [n/N_{min}]-1}$ with $v_j = x_{jN_{min}}$. We define M_0 as the partition associated with the whole grid, and $\mathcal{M}_n^1$ as the family of all partitions of $[0,1]$ with endpoints belonging to the grid. $\mathcal{M}_n^1$ corresponds to the collection of partitions used by Lebarbier [10] to detect the change-points of a Gaussian signal $s : [0,1] \longrightarrow \mathbb{R}$.

Example 2: We consider some set $\mathcal{X}$. We build a partition M_0 by splitting recursively $\mathcal{X}$ and the obtained subsets in two different parts as long as each subset contains at least N_{min} points x_i. A useful representation of this construction is a tree of maximal depth, called maximal tree and denoted by T_{max}. The leaves of T_{max} are the elements of the partition M_0. Every pruned subtree of T_{max} gives a partition of $\mathcal{X}$ built from M_0. We denote by $\mathcal{M}_n^2$ this second family of partitions. $\mathcal{M}_n^2$ corresponds to the family of partitions obtained via the first step of the CART algorithm.

Theorem 3.1. *Let $b \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}_+^*$ such that inequality (3) holds.*
Let M_0 a partition of $\mathcal{X}$ such that $N_{min} = \inf_{J \in M_0} |J|$ satisfies: $N_{min} \geq (\log n)^2$ if $b \neq 0$ and $N_{min} \geq 1$ if $b = 0$.
Let $\mathcal{M}_n$ a family of partitions of $\mathcal{X}$ built from M_0 and $(w_M)_{M \in \mathcal{M}_n}$ a family of weights such that

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

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

Taking a penalty satisfying

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

we have

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

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

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

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 $(w_M)_{M\in\mathcal{M}_n}$. For $\theta\in(0,1)$ and $K>2-\theta$, the penalized least squares estimator $\hat{s}_{\hat{M}}$ 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 the integrability condition of $(\varepsilon_i)_{1\leq i\leq n}$ (via σ^2 and b), on s (via R) unless $b=0$, and on the family of partitions (via Σ).

For the two particular families $\mathcal{M}_n$ quoted above, we calculate adequate weights $(w_M)_{M\in\mathcal{M}_n}$ and we get a simpler form of penalty. Before studying these two examples, we compare the general result with those of Birgé and Massart [4], those of Baraud, Comte and Viennet [2] and those of Baraud [1].

If b can be taken equal to zero in (3), then the variables $(\varepsilon_i)_{1\leq i\leq n}$ are said to be sub-Gaussian. In this case, we do not need any assumption on the regression function s and the minimal number N_{min} of observations in each element of the partition M_0 is only supposed to be positive. And taking a penalty satisfying

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

we have

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

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. Using the inequality $2\sqrt{|M|w_M}\leq|M|+w_M$, we recover the result obtained by Baraud, Comte and Viennet [2] in the sub-Gaussian case.

Baraud [1] studies the non Gaussian regression framework as defined in (1) 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 $\mathrm{pen}(M)$ 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 $|\{M \in \mathcal{M}_n;\ |M| = D\}| \leq \Gamma D^r$ for some constants $\Gamma \in \mathbb{R}_+^*$ and $r \in \mathbb{N}$, one only needs to assume that the random perturbations have a finite absolute moment of order $p > 2r + 6$. The minimal admissible value of p increases with the degree r 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 valid 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, and therefore Baraud's result does not apply to this case.

Example 1: Since $\left|\left\{M \in \mathcal{M}_n^1; |M| = D\right\}\right| = \binom{D_0 - 1}{D - 1} \leq \left(\frac{eD_0}{D}\right)^D$, where $D_0 - 1$ is the number of grid points, taking $w_M = |M| \left(a + \log \frac{D_0}{|M|}\right)$ with $a > 1$ leads to $\sum_{M \in \mathcal{M}_n^1} e^{-w_M} \leq \left(e^{a-1} - 1\right)^{-1} \in \mathbb{R}_+^*$. We deduce from the above theorem that:
taking a penalty

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

with α and β big enough, we have

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

Since σ^2, b and R are unknown, we consider penalties of the form $\mathrm{pen}(M) = \frac{|M|}{n} \left(\alpha' \log \frac{|M_0|}{|M|} + \beta'\right)$ and we determine the right constants α' and β' according to the data by using, for example, the same technique as Lebarbier [10]. We get a data driven criterion which selects a close to optimal partition $\hat{M}$. The endpoints of the partition $\hat{M}$ provide estimators of the change points of the signal s.

Example 2: Thanks to Catalan inequality, $\left|\left\{M \in \mathcal{M}_n^2; |M| = D\right\}\right| \leq \frac{1}{D} \binom{2(D-1)}{D-1} \leq \frac{2^{2D}}{D}$. Thus taking $w_M = a|M|$ with $a > 2 \log 2$, we get $\sum_{M \in \mathcal{M}_n^2} e^{-w_M} \leq -\log\left(1 - e^{-(a - 2\log 2)}\right) \in \mathbb{R}_+^*$. We deduce from the above theorem that:
taking a penalty

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

with α big enough, we have

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

For this second example, we recommend a penalty $\mathrm{pen}(M)$ 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 $n \to +\infty$. This result

validates the CART pruning step which involves a penalized least squares criterion with $\text{pen}(M) = \alpha' \frac{|M|}{n}$. The last step of CART consists in choosing the right parameter α' via cross-validation or test sample.

Remark 3.2. Theorem 3.1 do not determine completely the right penalty to be used to get an approximately best estimator. It only gives a form of penalty. The multiplicative constants depend on absolute constants $\theta \in (0;1)$ and $K > 2 - \theta$ (whose optimal values are unknown), and on unknown parameters σ^2, b and R.
In practice the multiplicative constants (α' and β' in example 1, or α' in example 2) are determined according to the data using for example one of the methods developped in [10] or in [15, sections 4.4 and 4.5]

Remark 3.3. In theorem 3.1, the assumption "all partitions $M \in \mathcal{M}_n$ are built from some initial partition M_0" allows to deal with large collections of partitions $\mathcal{M}_n$, but is useless for collections $\mathcal{M}_n$ satisfying $|\{M \in \mathcal{M}_n;\ |M| = D\}| \Gamma D^r$ with $\Gamma \in \mathbb{R}_+^*$ and $r \in \mathbb{N}$. (For more details see remark 6.2 at the end of the proof of theorem 3.1.) Such collections have already been handled by Baraud [1] under a mild integrability condition on the ε_i and without assuming $\|s\|_\infty \leq R$. Our result allows to recover Baraud's result (under stronger assumptions), but is only interesting for larger collections.

Remark 3.4. In this paper, the points $(x_i)_{1 \leq i \leq n}$ of the design are deterministic. If the points of the design were 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 would get a similar result. This work is done in a second paper [16].

4. Two deviation inequalities

As will be seen in the proof of theorem 3.1 (via equality (10)), in order that the penalized least squares estimator $\hat{s}_{\hat{M}}$ satisfies an oracle type inequality, the penalty $\text{pen}(M)$ has to compensate the deviations of the random variables:

$$\chi_M^2 = \|\varepsilon_M\|_n^2 \qquad \text{and} \qquad \langle \varepsilon, s - s_M \rangle_n$$

for all partitions $M \in \mathcal{M}_n$ simultaneously, without being too large. Therefore we need sharp deviation inequalities for these random variables.
This section is composed of two lemmas, which give appropriate deviation inequalities for respectively $\langle \varepsilon, s - s_M \rangle_n$ and χ_M^2. First, in lemma 4.1, we give the deviation inequality for $\langle \varepsilon, s - s_M \rangle_n$, which is the most easy to obtain. Then we recall two known deviation inequalities for χ_M^2. The first one is only valid if the ε_i's are Gaussian variables. The second one is very general and is due to Bousquet [6], but unfortunately it is not sharp enough. Finally, in lemma 4.2, we give our own deviation inequality for χ_M^2. The proof of lemma 4.2 is quite long and thus postponed to section 5.

Let begin with the deviation inequality for $\langle \varepsilon, s - s_M \rangle_n$:

Lemma 4.1. *Let $b \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}_+^*$ such that inequality (3) holds.*
For any partition M and for any $x > 0$

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

and if $\|s\|_\infty \leq R$ then

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

Proof. The first inequality is obtained by applying Cramer-Chernoff method as explained in remark 2.2.
If $\|s\|_\infty \leq R$ then $\max_{1 \leq i \leq n} |s(x_i) - s_M(x_i)| \leq 2R$, and thus we get the second inequality. □

It remains to study the deviations of the random variable χ^2_M around its expectation.
If the perturbations $(\varepsilon_i)_{1\leq i\leq n}$ were Gaussian variables with variance τ^2, then the variable $\frac{n}{\tau^2}\chi^2_M$ would be $\chi^2(|M|)$-distributed (see remark 2.3). Thus, according to [4, lemma 1], χ^2_M would satisfy for any $x>0$ the following deviation inequality:

$$\mathbb{P}\left(\chi^2_M \geq \frac{\tau^2}{n}|M| + 2\frac{\tau^2}{n}\sqrt{|M|x} + 2\frac{\tau^2}{n}x\right) \leq e^{-x}. \tag{6}$$

Let remark that the term which is linear in x: $2\frac{\tau^2}{n}x$, does not depend on M. (We will see in the proof of theorem 3.1 that this property is necessary to define a penalty which compensates the deviations of the χ^2_M for all $M \in \mathcal{M}_n$ simultaneously.)
The square root of χ^2_M: χ_M, would then satisfy

$$\mathbb{P}\left(\chi_M \geq \frac{\tau}{\sqrt{n}}\sqrt{|M|} + \frac{\tau}{\sqrt{n}}\sqrt{2x}\right) \leq e^{-x}. \tag{7}$$

Recall that $\chi_M = \|\varepsilon_M\|_n$ where ε_M is the orthogonal projection of $\varepsilon = (\varepsilon_i)_{1\leq i\leq n}$ on S_M (more pecisely on $\{u(x_i)_{1\leq i\leq n};\ u \in S_M\}$). According to Cauchy-Schwarz inequality, we can write χ_M as the supremum of a random process:

$$\chi_M = \|\varepsilon_M\|_n = \sup_{\substack{u\in S_M \\ \|u\|_n=1}} <\varepsilon, u>_n = \frac{1}{n}\sup_{\substack{u\in S_M \\ \|u\|_n=1}} \sum_{i=1}^n u_i\varepsilon_i. \tag{8}$$

Therefore, if the $(\varepsilon_i)_{1\leq i\leq n}$ were Gaussian variables, we could apply the concentration inequality for the supremum of a Gaussian process due to Cirel'son, Ibragimov and Sudakov (see [13, chapter 3]). Then we would recover inequality (7), which is a little bit less sharp than inequality (6).

Here, the $(\varepsilon_i)_{1\leq i\leq n}$ are not Gaussian variables, they are only supposed to have exponential moments around 0, but the expression (8) of χ_M is still valid. Our first idea was to consider expression (8) and use Bousquet's concentration inequality for the supremum of an empirical process instead of the Gaussian result of Cirel'son, Ibragimov and Sudakov. Thanks to Bousquet's result [6], 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 $v = \sum_{i=1}^n \sup_{\substack{u\in S_M \\ \|u\|_n=1}} \mathrm{Var}(u_i\varepsilon_i)$.
$c = \sqrt{\frac{n}{\inf_{J\in M}|J|}}$, $v = n|M|\tau^2$ with $\tau^2 = \mathbb{E}(\varepsilon_i^2) \leq \sigma^2$, and $\mathbb{E}(\chi_M) \leq \sqrt{\mathbb{E}(\chi^2_M)} = \frac{\tau}{\sqrt{n}}\sqrt{|M|}$. Thus we get that:

$$\mathbb{P}\left(\chi_M \geq (1+\gamma)\frac{\tau}{\sqrt{n}}\sqrt{|M|} + \frac{\tau}{\sqrt{n}}\sqrt{2|M|x} + \frac{1}{n}(2+\gamma^{-1})bcx\right) \leq e^{-x}. \tag{9}$$

If the ε_i's are Gaussian variables, then inequality (9) is sub-optimal. Indeed, if we compare inequality (7) (obtained in the case $b=0$ and ε_i Gaussian) with inequality (9), we see that the variance term $v = \sum_{i=1}^n \sup_u \mathrm{Var}(u_i\varepsilon_i)$ is much too large. We should have $v = \sup_u \sum_{i=1}^n \mathrm{Var}(u_i\varepsilon_i)$ instead of $v = \sum_{i=1}^n \sup_u \mathrm{Var}(u_i\varepsilon_i)$.
With such a refinement, we would obtain here $v = n\tau^2$ instead of $v = n|M|\tau^2$ and the term $\frac{\tau}{\sqrt{n}}\sqrt{2|M|x}$ in (9) would be replaced by $\frac{\tau}{\sqrt{n}}\sqrt{2x}$ like in (7). Moreover we would be able to get rid of the term which is linear in x: $\frac{1}{n}(2+\gamma^{-1})bcx$, by truncating χ_M. More precisely, since the supremum in (8) is achieved with $u = \frac{\varepsilon_M}{\|\varepsilon_M\|_n}$,

denoting $\Omega_\delta = \left\{\forall J \in M_0;\ \left|\sum_{x_i \in J} \varepsilon_i\right| \leq \delta \tau^2 |J|\right\}$ and truncating χ_M with $\Omega_\delta \bigcap \left\{\chi_M \geq \frac{\tau}{\sqrt{n}}\sqrt{2x}\right\}$, we would get:

$$\mathbb{P}\left(\chi_M 1\!\!1_{\Omega_\delta} \geq (1+\gamma)\mathbb{E}(\chi_M) + \frac{\tau}{\sqrt{n}}\sqrt{2x} + \frac{(2+\gamma^{-1})\delta b}{\sqrt{2}}\frac{\tau}{\sqrt{n}}\sqrt{x}\right) \leq e^{-x}.$$

Bousquet's result leads to a deviation inequality for χ_M which is not sharp enough for our problem (the term "$\frac{\tau}{\sqrt{n}}\sqrt{2|M|x}$ " in (9) is in particular too large and its square becomes a linear term in x which depends on M in the deviation inequality for χ_M^2). Since the refinement of Bousquet's general result seems difficult to obtain, we build our own deviation inequality for χ_M^2 by hand. Instead of considering expression (8) where χ_M is written as a supremum, we wiew $\chi_M^2 = \|\varepsilon_M\|_n^2$ as a χ^2 like random variable and we write it as a sum of squares (see expression (5)). Then, by truncating χ_M^2 and applying Bernstein inequality [13, section 2.2.3], we get the following lemma.

Lemma 4.2. *Let $b \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}_+^*$ such that inequality (3) 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 $x > 0$

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

and

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

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

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

In lemma 4.2, the $(\varepsilon_i)_{1 \leq i \leq n}$ are only supposed to have exponential moments around 0. In this case, by truncating χ_M^2, we get a deviation inequality which differs from inequality (6) (corresponding to the Gaussian case) only in the multiplicative constants. The set Ω_δ on which we control the deviations of χ_M^2 is very large. More precisely, if N_{min} satisfies $N_{min} \geq (\log n)^2$ then, for all $k \in \mathbb{N}$, $\mathbb{P}\left(\Omega_\delta^c\right) = o\left(\frac{1}{n^k}\right)$ when $n \to +\infty$.

Remark 4.3. In the context of histogram density estimation, Castellan [8] has to control an other χ^2 like random variable. Like here the main point is to truncate the variable. While she concludes by applying a Talagrand inequality to the truncated variable, we use Bernstein inequality. The detailed proof of lemma 4.2 is given in the next section.

5. Proof of lemma 4.2

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 (3) and Cramer-Chernoff method (see remark 2.2) 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 [13, section 2.2.3], 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 (3), 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 the definition of $\hat{M}$: $\hat{M} = \arg\min_{M\in\mathcal{M}_n}\left\{\|Y-\hat{s}_M\|_n^2 + \mathrm{pen}(M)\right\}$, we have

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

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

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

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,

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

or equivalently,

$$(1-\theta)\|s-\hat{s}_{\hat{M}}\|_n^2 = \Delta_{\widehat{M}} + \inf_{M\in\mathcal{M}_n} R_M \tag{10}$$

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 4.2,

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

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

$$\mathbb{P}\left(\|\varepsilon_M\|_n^2 \mathbb{1}_\Omega \geq \frac{\sigma^2}{n}|M| + 8\frac{\sigma^2}{n}\sqrt{2|M|x} + 4\frac{\sigma^2}{n}x \right) \leq e^{-x} \tag{11}$$

Thanks to lemma 4.1, for any $M \in \mathcal{M}_n$ and any $x > 0$,

$$\mathbb{P}\Big(- < \varepsilon, s - s_M >_n \geq \frac{\sigma}{\sqrt{n}}\|s - s_M\|_n\sqrt{2x} + \frac{2Rb}{n}x \Big) \quad \leq \quad e^{-x} \tag{12}$$

Setting $x = w_M + \xi$ with $\xi > 0$, and summing all inequalities (11) and (12) 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,

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

Using the two following inequalities

$2\frac{\sigma}{\sqrt{n}}\|s - s_M\|_n\sqrt{2(w_M+\xi)} \leq \theta\|s - s_M\|_n^2 + \frac{2}{\theta}\frac{\sigma^2}{n}(w_M+\xi)$,
$8(2-\theta)\frac{\sigma^2}{n}\sqrt{2|M|(w_M+\xi)} \leq 8\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\sqrt{|M|w_M} + 4\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\left(\eta|M| + \eta^{-1}\xi\right)$

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,

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

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

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

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,

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

Integrating with respect to ξ,

$$\mathbb{E}\left(\Delta_{\widehat{M}} 1\!\!1_\Omega\right) \leq 2\left\{ 4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}\right\}\frac{\sigma^2}{n}\Sigma+\frac{8Rb}{n}\Sigma \tag{13}$$

We are going now to control $\mathbb{E}\left(\inf_M R_M 1\!\!1_\Omega\right)$.
Thanks to lemma 4.1, 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,

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

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

$$\begin{aligned} R_M &\leq \|s-s_M\|_n^2+2\frac{\sigma}{\sqrt{n}}\|s-s_M\|_n\sqrt{2\left(w_M+\xi\right)}+\frac{4Rb}{n}\left(w_M+\xi\right)+\mathrm{pen}(M) \\ &\leq (1+\sqrt{\theta})\|s-s_M\|_n^2+\frac{2}{\sqrt{\theta}}\frac{\sigma^2}{n}\left(w_M+\xi\right)+\frac{4Rb}{n}\left(w_M+\xi\right)+\mathrm{pen}(M) \\ &\leq (1+\sqrt{\theta})\|s-s_M\|_n^2+(1+\sqrt{\theta})\mathrm{pen}(M)+\frac{2}{\sqrt{\theta}}\frac{\sigma^2}{n}\xi+\frac{4Rb}{n}\xi \end{aligned}$$

And

$$\begin{aligned} \mathbb{E}\left(\inf_M R_M 1\!\!1_\Omega\right) \quad \leq \quad & (1+\sqrt{\theta})\inf_M\left\{\|s-s_M\|_n^2+\mathrm{pen}(M)\right\} \\ & +\frac{2}{\sqrt{\theta}}\frac{\sigma^2}{n}\Sigma+\frac{4Rb}{n}\Sigma \end{aligned} \tag{14}$$

We conclude from (10), (13) and (14) that

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

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

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

By developing $\|\varepsilon_{M_0}\|_n^4$, 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

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

and thus

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

Let us recall that

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

For $N_{min}\geq(\log n)^2$,

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

and

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

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

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

if $N_{min}\geq(\log n)^2$, we have

$$\begin{aligned}\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\right) &\leq \frac{1+\sqrt{\theta}}{1-\theta}\inf_M\left\{\|s-s_M\|_n^2+\mathrm{pen}(M)\right\}\\ &\quad+\frac{1}{1-\theta}\left\{8(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{4}{\theta}+\frac{2}{\sqrt{\theta}}\right\}\frac{\sigma^2}{n}\Sigma+\frac{12}{1-\theta}\frac{Rb}{n}\Sigma\\ &\quad+C(b,\sigma^2,R)\frac{1\!\!1_{b\neq0}}{n(\log n)^3}\end{aligned}$$

Remark 6.1. Starting from equality (10), the deviation inequalities for the $\left(\chi^2_M\right)_{M\in\mathcal{M}_n}$ and the $\left(\langle\varepsilon, s-s_M\rangle_n\right)_{M\in\mathcal{M}_n}$ (lemmas 4.1 and 4.2) are the key to determine the adequate form of penalty and to prove theorem 3.1. The weights $(w_M)_{M\in\mathcal{M}_n}$ satisfying $\sum_{M\in\mathcal{M}_n} e^{-w_M} \leq \Sigma \in \mathbb{R}_+^*$ allow to sum all deviation inequalities for $\left(\chi^2_M\right)_{M\in\mathcal{M}_n}$ and $\left(\langle\varepsilon, s-s_M\rangle_n\right)_{M\in\mathcal{M}_n}$ with respect to $M \in \mathcal{M}_n$, and to control χ^2_M and $\langle\varepsilon, s-s_M\rangle_n$ for all $M \in \mathcal{M}_n$ simultaneously. The right penalty pen is thus the sum of two terms: one proportional to $\frac{|M|}{n}$, which upper-bounds $\mathbb{E}(\chi^2_M) = \frac{\tau^2|M|}{n}$, and a second depending on w_M, which compensates the deviations of χ^2_M around $\mathbb{E}(\chi^2_M) = \frac{\tau^2|M|}{n}$ and those of $\langle\varepsilon, s-s_M\rangle_n$ around $\mathbb{E}\left(\langle\varepsilon, s-s_M\rangle_n\right) = 0$.

Remark 6.2. The assumption "all partitions $M \in \mathcal{M}_n$ are built from some initial partition M_0" allows to define a large set Ω on which we control all variables χ^2_M, $M \in \mathcal{M}_n$, simultaneously, whatever the complexity of $\mathcal{M}_n$. If the collection $\mathcal{M}_n$ satisfies $|\{M \in \mathcal{M}_n;\ |M| = D\}| \, \Gamma D^r$ with $\Gamma \in \mathbb{R}_+^*$ and $r \in \mathbb{N}$, then one can remove the assumption "all partitions $M \in \mathcal{M}_n$ are built from some initial partition M_0" from theorem 3.1, and only suppose that $N_{min} = \inf_{M\in\mathcal{M}_n} \inf_{J\in M} |J| \geq (\log n)^2$ (except if $b = 0$). Then, in the proof of theorem 3.1, we consider $\Omega' = \bigcap_{M\in\mathcal{M}_n} \left\{\forall J \in M;\ \left|\sum_{i\in J} \varepsilon_i\right| \leq \frac{\sigma^2}{b}|J|\right\}$ (instead of $\Omega = \left\{\forall J \in M_0;\ \left|\sum_{i\in J} \varepsilon_i\right| \leq \frac{\sigma^2}{b}|J|\right\}$).
$\mathbb{P}\left(\Omega'^c\right) \leq 2\Gamma \left(\frac{n}{N_{min}}\right)^{r+2} \exp\left(\frac{-\sigma^2 N_{min}}{4b^2}\right) \leq C''(b, \sigma^2, \Gamma, r) \frac{1}{n^2(\log n)^2}$

References

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

[2] Y. Baraud, F. Comte, and G. Viennet. Model Selection for (auto-)regression with dependent data. *ESAIM Probability and Statistics*, 5:33–49, 2001.

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

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

[5] L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram. *ESAIM: Probability and Statistics*, 10:24–45, 2006.

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

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

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

[9] O. Catoni. Universal aggregation rules with sharp oracle inequalities. *Annals of Statistics*, pages 1–37, 1999.

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

[11] G. Lugosi and A. Nobel. Consistency of data-driven histogram methods for density estimation and classification. *The Annals of Statistics*, 24(2):786–706, 1996.

[12] C.L. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

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

[14] A. Nobel. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084–1105, 1996.

[15] M. Sauvé. *Sélection de modèles en régression non gaussienne. Applications la sélection de variables et aux tests de survie accélérés*. PhD thesis, Université Paris XI Orsay, 2006.

[16] M. Sauvé and C. Tuleau. Variable selection through CART. Research Report 5912, INRIA, 2006.