Bernoulli **12**(4), 2006, 633–661

Penalized projection estimators of the Aalen multiplicative intensity

PATRICIA REYNAUD-BOURET

Département de Mathématiques et Applications, École Normale Supérieure, 45 Rue d'Ulm, 75230 Paris Cedex 05, France. E-mail: Patricia.Reynaud-Bouret@ens.fr

We study the problem of nonparametric, completely data-driven estimation of the intensity of counting processes satisfying the Aalen multiplicative intensity model. To do so, we use model selection techniques and, specifically, penalized projection estimators for a random inner product. For histogram estimators, under some assumptions on the process, we obtain adaptive results for the minimax risk. In general, for more intricate (predictable) models, we only obtain oracle inequalities. The study is complemented by some simulations in the right-censoring model.

Keywords: adaptive estimation; counting processes; model selection; multiplicative intensity model; penalized projection estimators

1. Introduction

1.1. The bibliographical context

Counting processes with Aalen multiplicative intensity are a generalization of temporal Poisson processes. They can model a large variety of situations (especially in biology and medicine). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and $(\mathcal{F}_t, t \geqslant 0)$ be a filtration. A counting process $N = (N_t)_{t \geqslant 0}$ satisfies the Aalen multiplicative intensity model with predictable process $Y = (Y_t)_{t \geqslant 0}$ (see Andersen *et al.* 1993) if

$$\mathrm{d}\Lambda_t = Y_t s(t) \mathrm{d}t, \tag{1.1}$$

where $(\Lambda_t)_{t \geqslant 0}$ is the compensator of $(N_t)_{t \geqslant 0}$ with respect to $(\mathcal{F}_t, t \geqslant 0)$, $(Y_t)_{t \geqslant 0}$ a non-negative predictable process and s a deterministic function. When the process $(Y_t)_{t \geqslant 0}$ is constant, $(N_t)_{t \geqslant 0}$ is a temporal Poisson process with intensity s with respect to the measure $Y\,\mathrm{d}t$.

Let us give some other examples. Let $(N_t = \mathbb{1}_{X \leqslant t})_{t \geqslant 0}$ where X is a positive random variable with density f. This process satisfies (1.1) with $Y_t = \mathbb{1}_{X \geqslant t}$ and with $s(t) = f(t)/\mathbb{P}(X \geqslant t)$, the *hazard rate* of X; if X represents a patient's lifetime, $s(t)$ represents the probability that the patient dies just after t given that he is alive at time t.

Observations of lifetimes may sometimes be censored. This is the case when a patient drops out of a hospital study. The time of death is not observed, but we know that the patient was still alive when he left the study. This situation is modelled by some other positive random variable U which is independent of X and the observations are the

variables $T = X \wedge U$ and $D = \mathbb{1}_{T=X}$. This model is known as the *right-censoring model* with independent censorship. Then the process $N_t = D\mathbb{1}_{T \leqslant t}$ has an Aalen multiplicative intensity (1.1) where $Y_t = \mathbb{1}_{T \geqslant t}$ and s is the hazard rate of X.

We may also have an n-sample of counting processes, $N^1, \ldots, N^n$, satisfying (1.1) (corresponding to n different patients, for instance). Their predictable processes are denoted by $Y^1, \ldots, Y^n$. They have the same intensity s. Then we can define the *aggregated process* N with predictable process Y by

$$N_t = \sum_{i=1}^{n} N_t^i \quad \text{and} \quad Y_t = \sum_{i=1}^{n} Y_t^i, \qquad \text{for all } t \geqslant 0. \tag{1.2}$$

This aggregated process also satisfies (1.1) with the same s.

For instance, in the right-censoring model, the process Y is a non-increasing process with integer values and with $Y_0 = n$, n being the number of observations. The number Y_t represents the number of events which will happen after t, whether these events are real observed deaths or departures.

Many other examples of processes with multiplicative intensity are mentioned in Andersen *et al.* (1993). For instance, if $(X_t)_{t \geqslant 0}$ is a Markov process with finite state space, the counting process $(N_t^{hj}, t \geqslant 0)$, where N_t^{hj} represents the number of transitions from h to j by time t, has a multiplicative intensity of the form (1.1) where s is the transition intensity from h to j and where Y is defined by $(Y_t = \mathbb{1}_{X(t)=h})_{t \geqslant 0}$. We may have an n-sample of independent and identically distributed counting processes corresponding to each individual Markov process. In this situation, we can look at the aggregated processes (1.2) where Y is still integer-valued and upper-bounded by n: at time t, Y_t represents the number of individuals in state h. This situation models, for instance, the transition from healthy to diseased (Andersen *et al.* 1993: Example I.3.10).

There are also cases where the process cannot be divided into individual processes, and so cannot be written as in (1.2). This is the case for the model of the number of matings of *Drosophila* flies (Andersen *et al.* 1993: Example III.1.10). However, this model satisfies the multiplicative intensity property (1.1) with a Y which still corresponds to a bounded number of events which may happen after time t.

The purpose of this paper is to estimate s on $[0, \tau]$ using observations of $(N_t)_{0 \leqslant t \leqslant \tau}$ and $(Y_t)_{0 \leqslant t \leqslant \tau}$. Our aim is to do so in a nonparametric adaptive way with as few assumptions on s as possible. We also try to stay within the most general framework, but, as mentioned later, we need some extra assumptions on the process itself (aggregated or not, for instance) depending on the type of estimator (piecewise constant or predictable).

Many papers consider the problem of the estimation of s in the general Aalen multiplicative intensity model. Ramlau-Hansen (1983) proved consistency and asymptotic normality results for some kernel estimators with fixed bandwidth. Grégoire (1993) gives a data-driven criterion for choosing the bandwidth of the Ramlau-Hansen estimators by cross-validation. He also proves consistency and asymptotic normality results. Other possible estimators are maximum likelihood estimators on certain sieves, whose rates of convergence were studied by van de Geer (1995). Antoniadis (1989) chooses the sieve by penalization,

proving consistency and asymptotic normality for penalized maximum likelihood estimators, the penalization depending on the regularity of the functions.

We go further in this direction by providing adaptive estimators which do not depend on previous knowledge of the regularity of s. These estimators still have good properties of convergence (in the minimax sense, for instance). The adaptive properties are also non-asymptotic. This is useful here since the data generally come from medical surveys where only a few patients can be observed.

There are already adaptive estimators of the function s but in situations where N is specified. For Poisson processes, Cavalier and Koo (2002) provide thresholding procedures and Reynaud-Bouret (2003) proposes a penalized model selection procedure. For the right-censoring model, Antoniadis *et al.* (1999) propose a family of wavelet estimators, prove their consistency under regularity assumptions and propose an adaptive method by cross-validation. In the same framework, Döhler and Rüschendorf (2002) prove adaptivity for penalized model selection.

We also use the penalized model selection method. Our results actually give a possible extension of the results of Döhler and Rüschendorf (2002) and of Reynaud-Bouret (2003) to the general Aalen multiplicative intensity model.

1.2. General tools

The problem can be reduced by means of changes of scale to the estimation of s on $[0, 1]$ and to the observation of the processes $(N_t)_{t\geqslant 0}$ and $(Y_t)_{t\geqslant 0}$ on $[0, 1]$. We make the following assumption throughout:

Assumption 1. *Y is upper-bounded by a known positive constant A.*

For instance, in the right-censoring model or in the Markovian model, $A = n$ represents the number of patients.

We need, of course, to measure the performance of the estimators, and for this purpose we need a distance between s and the estimator $\hat{s}$. For instance, Antoniadis *et al.* (1999) use the classical $\mathbb{L}^2$ norm on the entire observed interval in the right-censoring model. But in practice they reduce this interval because of the scarcity of the observations at its right-hand end. To take this into account, we have decided in this paper to use a random norm, weighted by Y. This random norm is defined for all f in $\mathbb{L}^2([0, 1], \mathrm{d}t)$ by

$$\|f\|_{\mathrm{rand}}^2 = \int_0^1 f^2(t) Y'_t \,\mathrm{d}t, \qquad \text{where } Y'_t = \frac{Y_t}{A}. \tag{1.3}$$

Now we follow Birgé and Massart (1997) whose framework is density estimation. We need to introduce a contrast. The following definition is very similar to theirs: for all f in $\mathbb{L}^2([0, 1], \mathrm{d}t)$, let us define

$$\gamma_A(f) = -2\int_0^1 f(t)\frac{\mathrm{d}N_t}{A} + \int_0^1 f^2(t) Y'_t \,\mathrm{d}t. \tag{1.4}$$

Following Birgé and Massart (1997), we call this contrast a least-squares contrast. This is not the contrast used by Döhler and Rüschendorf (2002): they use a log-likelihood contrast which is much more intricate to deal with than the least-squares one, although it gives good results.

Still following Birgé and Massart (1997), the projection estimator of s on a finite-dimensional linear subspace S is defined by

$$\hat{s} = \operatorname*{argmin}_{f \in S} \gamma_A(f). \tag{1.5}$$

If $\{h_\lambda, \lambda \in \Gamma\}$ is an orthonormal basis of S for the random norm, we can write

$$\hat{s} = \sum_{\lambda \in \Gamma} \left[\int_0^1 h_\lambda(t) \frac{\mathrm{d}N_t}{A} \right] h_\lambda. \tag{1.6}$$

If we wish to estimate s with a projection estimator, we have to choose the subspace S. If we wish to do some adaptive estimation, this finite-dimensional subspace or *model* must be chosen via a data-driven criterion. To achieve this goal, we introduce a family of models $\{S_m, m \in \mathcal{M}_A\}$ and associate with each S_m the projection estimator $\hat{s}_m$ of s on it. Let us take a penalty denoted by 'pen', which is a positive function on $\mathcal{M}_A$, independent of s and, if necessary, random. We choose the model by minimizing the following data-driven criterion:

$$\hat{m} = \operatorname*{argmin}_{m \in \mathcal{M}_A} (\gamma_A(\hat{s}_m) + \mathrm{pen}(m)), \tag{1.7}$$

and the *penalized projection estimator*, $\tilde{s}$, is defined by $\hat{s}_{\hat{m}}$.

Intuitively, we use this criterion to have $\|s - \tilde{s}\|^2$ close to $\inf_{m \in \mathcal{M}_A} \|s - \hat{s}_m\|^2$. One way to check that we have found a good penalty is to prove an *oracle-type inequality*, that is, an inequality such as

$$\|s - \tilde{s}\|^2 \leqslant C \inf_{m \in \mathcal{M}_A} (\|s - s_m\|^2 + \mathrm{pen}(m)), \tag{1.8}$$

for some positive constant C, where s_m is the projection of s on S_m for $\|\cdot\|$. This inequality should hold either in probability or in expectation, with, if necessary, the addition of some negligible term.

The norm $\|\cdot\|$ in (1.8) may be either the random norm $\|\cdot\|_{\mathrm{rand}}$ or the deterministic norm defined for all f in $\mathbb{L}^2([0, 1], \mathrm{d}t)$ by

$$\|f\|_{\mathrm{det}}^2 = \int_0^1 f^2(t) \mathbb{E}(Y'_t) \, \mathrm{d}t. \tag{1.9}$$

If $\|s - s_m\|^2 + \mathrm{pen}(m) \simeq \|s - \hat{s}_m\|$, we obtain a true oracle inequality: the penalized projection estimator $\tilde{s}$ performs as well as the best possible estimator in the family $\{\hat{s}_m, m \in \mathcal{M}_A\}$ up to a multiplicative constant, and does so without knowing s. This proves the adaptivity of the penalized projection estimator in the family $\{\hat{s}_m, m \in \mathcal{M}_A\}$.

The choice of the family of models is also very important. For technical reasons, it is easier to know an orthonormal basis of the models for the random norm. We therefore deal with two cases.

The first is the histogram case. The basis is clear, but we must narrow down our approach to aggregated processes (1.2) to be able to control the variance term of the

estimator. This is done in Section 2. Under these assumptions, we are able to prove oracle-type inequalities for some well-chosen penalties and families of models. We are also able to prove some minimax results.

The second case deals with random predictable models, constructed as follows. If $\{\varphi_\lambda, \lambda \in m\}$ is a classical deterministic orthonormal basis of $\mathbb{L}^2([0, 1], \mathrm{d}t)$, typically a part of a Fourier basis, then $\{h_\lambda(t) = \varphi_\lambda(t)/\sqrt{Y'_t}, \lambda \in m\}$ becomes, when Y' is positive, an orthonormal basis for the random product. The model $S_m = \mathrm{Span}\{h_\lambda, \lambda \in m\}$ is consequently a random predictable subspace. Note that the resulting estimator is not smooth but piecewise continuous because of the $\sqrt{Y'_t}$ in the denominator of the h_λ. One of the advantages of these models is that they enable us to remove the aggregation assumption among many other technical assumptions. We prove an oracle-type inequality for this case in Section 3, but as the models are non-smooth and random we are not able to obtain any minimax results for this strategy.

Section 4 contains simulation studies of these two strategies in the right-censoring model and compares these results with adaptive estimators that already exist.

The main results of the paper are proved in Section 5.

2. Histogram quasi-least-squares estimators

The purpose of this section is to deal with deterministic piecewise constant models. We thus make the following assumptions:

Assumption 2.

(i) *N is an aggregated process (see* (1.2)*), with predictable process Y and with individual processes $N^1, \ldots, N^n$ and $Y^1, \ldots, Y^n$.*
(ii) *Each Y^i is bounded by* 1.
(iii) *The number of individual jumps of the N^i is bounded by a known positive constant K.*

For instance, the right-censoring model satisfies (i)–(iii) with $K = 1$ but the Markovian models and the Poisson process do not satisfy (iii).

Under these assumptions, A defined by Assumption 1 is taken to be equal to n. We also assume that there exists an unknown constant R such that s is bounded by R.

If the Y^i are just bounded by a known B, it is sufficient to divide the Y^i by B and to estimate Bs to satisfy Assumption 2.

Let us compare these assumptions with those of Grégoire (1993). He assumes N to be aggregated and Y to be bounded. He does not assume a bound on N. Thus he can manage Markovian models but he assumes that n/Y is bounded by a quantity independent of n. This is not required here.

2.1. Study on one model

Under Assumption 2, the least-squares contrast (1.4) becomes

$$\gamma_A(f) = -\frac{2}{n}\int_0^1 f(t)\,\mathrm{d}N_t + \frac{1}{n}\int_0^1 f^2(t)Y_t\,\mathrm{d}t.$$

Let us now construct the projection histogram estimator. Let m be a partition of $[0, 1]$. For all intervals I of m, let $b_I = (1/n)\int_0^1 \mathbb{1}_I Y_t\,\mathrm{d}t$: b_I depends only on the observations. For the random norm $\|\cdot\|_{\mathrm{rand}}$, the family $\{\mathbb{1}_I/\sqrt{b_I},\ I \in m\}$ is an orthonormal basis of the subspace of piecewise constant functions on m. Let $|m|$ be the number of intervals in the partition m. Let $\beta_I = \mathbb{E}(b_I)$ and let N_I be the number of points of N lying in I.

Let $\mathcal{I}_m$ be the set of intervals I of m such that the b_I are larger than $1/n^2$. Let S_m be the space of the piecewise constant functions on m and S'_m the set of piecewise constant functions on m, null outside $\mathcal{I}_m$. The *quasi-least-squares* histogram estimator on S_m is the projection estimator of s defined by (1.5) on S'_m. Using (1.6), this estimator can be rewritten as:

$$\hat{s}_m = \sum_{I\in\mathcal{I}_m} \frac{N_I}{nb_I}\mathbb{1}_I. \tag{2.1}$$

It is more convenient to deal with this quasi-least-squares estimator (i.e. the projection estimator of s on S'_m) than with the projection estimator of s on S_m, because $\hat{s}_m$ is bounded.

Turning now to the risk of the quasi-least-squares estimator, let $s_m = \sum_{I\in m}(a_I/b_I)\mathbb{1}_I$ be the projection of s on S_m for the random scalar product, where $a_I = (1/n)\int_0^1 \mathbb{1}_I Y_t s(t)\mathrm{d}t$. Note that if $b_I = 0$ then $a_I = 0$ and the corresponding coefficient of s_m is zero. Let $\alpha_I = \mathbb{E}(a_I)$ and let $s'_m = \sum_{I\in\mathcal{I}_m}(a_I/b_I)\mathbb{1}_I$ be the projection of s on S'_m. Finally, we denote by $s_m^{\mathrm{det}} = \sum_{I\in m}(\alpha_I/\beta_I)\mathbb{1}_I$ the projection of s on S_m for the deterministic scalar product. The distance between s and $\hat{s}_m$ can be split in the following way:

$$\|s - \hat{s}_m\|^2_{\mathrm{rand}} = \|s - s'_m\|^2_{\mathrm{rand}} + \|s'_m - \hat{s}_m\|^2_{\mathrm{rand}}. \tag{2.2}$$

The first term is a bias term, which is random here. We can bound it by

$$\begin{aligned}\|s - s'_m\|^2_{\mathrm{rand}} &= \|s - s_m\|^2_{\mathrm{rand}} + \|s_m - s'_m\|^2_{\mathrm{rand}} \qquad (2.3)\\ &\leqslant \inf_{t\in S_m}\|s - t\|^2_{\mathrm{rand}} + \frac{R^2}{n},\end{aligned}$$

since $\|s_m - s'_m\|^2_{\mathrm{rand}} = \sum_{I\in\mathcal{I}_m^c} a_I^2/b_I \leqslant R^2\sum_{I\in\mathcal{I}_m^c} b_I \leqslant R^2|m|/n^2$. It is sufficient to assume that there are no more than n intervals, that is, $|m| \leqslant n$. The expectation of the bias term is therefore bounded by

$$\mathbb{E}(\|s - s'_m\|^2_{\mathrm{rand}}) \leqslant \inf_{s\in S_m}\mathbb{E}(\|s - t\|^2_{\mathrm{rand}}) + \frac{R^2}{n} = \|s - s_m^{\mathrm{det}}\|^2_{\mathrm{det}} + \frac{R^2}{n},$$

which decreases when the intervals of the partition become small.

The behaviour of the second term in (2.2) is very different. Its expectation is classically called the variance term. For a set of intervals $\mathcal{T}$, let us set

$$\chi^2_{\mathcal{I}} = \sum_{I \in \mathcal{I}} \frac{(N_I/n - a_I)^2}{b_I}. \tag{2.4}$$

Then the second term in (2.2) is exactly $\chi^2_{\mathcal{I}_m}$. If we assume that the b_I are close to their expectation, denoted by β_I, and assumed to be non-zero, then $\chi^2_{\mathcal{I}_m}$, χ^2_m and

$$Z^2_m = \sum_{I \in m} \frac{(N_I/n - a_I)^2}{\beta_I} \tag{2.5}$$

are also really close to each other. Let us also assume that Z^2_m is close to $\mathbb{E}(Z^2_m) = \sum_{I \in m}(\alpha_I/(n\beta_I))$. Then this expectation lies between $r|m|/n$ and $R|m|/n$, if s is upper-bounded by R and lower-bounded by r.

If all the previous approximations are accurate, the variance term must grow like the dimension of the model S_m when the bias term decreases.

2.2. Penalized least-squares histograms

If we wish to find a good model, we must balance the bias term and the variance term, but we must do this through a data-driven criterion, without some previous knowledge of s.

Let $\{S'_m, m \in \mathcal{M}_A\}$ be the family of models corresponding to the family of partitions $\mathcal{M}_A$ of [0, 1].

The best partition or the best model, the one which we would choose if we knew s, is called the *oracle* and is defined by

$$\begin{aligned}
\overline{m} &= \operatorname*{argmin}_{m \in \mathcal{M}_A} \mathbb{E}(\|s - \hat{s}_m\|^2_{\text{rand}}) \\
&= \operatorname*{argmin}_{m \in \mathcal{M}_A} \mathbb{E}(-\|s'_m\|^2_{\text{rand}} + \|s'_m - \hat{s}_m\|^2_{\text{rand}}) \\
&\simeq \operatorname*{argmin}_{m \in \mathcal{M}_A} \mathbb{E}(-\|\hat{s}_m\|^2_{\text{rand}} + 2\|s'_m - \hat{s}_m\|^2_{\text{rand}}) \\
&\simeq \operatorname*{argmin}_{m \in \mathcal{M}_A} \mathbb{E}(-\|\hat{s}_m\|^2_{\text{rand}} + 2\chi^2_{\mathcal{I}_m}).
\end{aligned} \tag{2.6}$$

The symbol $\simeq$ indicates that the expectations are not equal but that if the coefficients of $\hat{s}_m - s'_m$ are close to zero, the expectations are close to each other.

Moreover, by simply using the definitions (1.3), (1.4) and (2.1), it is straightforward to see that $\|\hat{s}_m\|^2_{\text{rand}} = -\gamma_A(\hat{s}_m)$. Hence, in estimating the previous quantities, we will choose the model $\hat{m}$ given by (1.7) with the penalty that satisfies the requirement that pen(m) is an estimate of twice the variance term.

Equation (2.6) corresponds to the minimization of the integrated squared error for kernel estimators done by Grégoire (1993).

Here we choose the partition by the general penalized data-driven criterion given in (1.7). But in order to prove that a penalty is well chosen we have to prove an inequality of type (1.8). Hence, we need to understand how far away $\hat{m}$ can be from the oracle. More

precisely, we have to understand the behaviour of $\chi^2_{\mathcal{I}_m}$ and see whether or not the penalty overestimates it.

2.3. Control of the chi-square statistic

The behaviour of the $\chi^2_{\mathcal{I}_m}$ is, however, very difficult to control. Thus we bound them by $Z_m^2 V_m$, for all m in $\mathcal{M}_A$, where Z_m^2 is given by (2.5) and $V_m = \sup_{I\in m}(\beta_I/b_I)$. Moreover, the square root of Z_m^2 can be seen to be

$$Z_m = \sup_{\delta=(\delta_I)_{I\in m}:\Sigma_{I\in m}\delta_I^2\beta_I=1} \left\{ \frac{1}{n}\sum_{i=1}^{n} \int_0^1 \left[\sum_{I\in m} \delta_I \mathbb{1}_I(t)\right] [\mathrm{d}N_t^i - Y_t^i s(t)\,\mathrm{d}t] \right\}.$$

We can apply the recent version of Talagrand's inequality obtained by Rio (2002).

Proposition 1. *Under Assumption 2, for all* ε, $x > 0$,

$$\mathbb{P}\left[Z_m \geqslant (1+\varepsilon)\sqrt{\sum_{I\in m}\frac{\alpha_I}{n\beta_I}} + \sqrt{2v_m\frac{x}{n}} + (1/2+\varepsilon^{-1})b\frac{K+R}{n}x \right] \leqslant \mathrm{e}^{-x},$$

where $b = \sup_{I\in m} 1/\sqrt{\beta_I}$, $R \geqslant \|s\|_\infty$, *and*

$$v_m = \sup_{\delta=(\delta_I)_{I\in m}:\Sigma_{I\in m}\delta_I^2\beta_I=1} \left\{ \int_0^1 \sum_{I\in m} \delta_I^2 \mathbb{1}_I(t)\mathbb{E}(Y_t^1)s(t)\,\mathrm{d}t \right\}.$$

Proof. Applying Theorem 1.4 of Massart (2000) to

$$X_{i,\delta} = \frac{1}{n}\int_0^1 \left[\sum_{I\in m} \delta_I \mathbb{1}_I(t)\right] [\mathrm{d}N_t^i - Y_t^i s(t)\,\mathrm{d}t]$$

which are centred variables, it is very easy to derive the previous bound, knowing that the number of jumps of the N^i is bounded by K and that Y^i is bounded by 1. We can restrict the supremum to a countable dense family of δ in order to carefully apply the result of Rio (2002), which has better constants. But by density, we obtain the present result. □

We can also find a large event on which the behaviour of Z_m is sub-Gaussian, as Massart (2005) does for estimating the density of an n-sample.

Proposition 2. *Let* ε *be a positive number and let* $\Omega_m(\varepsilon)$ *be the event*

$$\Omega_m(\varepsilon) = \left\{ \forall I \in m, \left| \frac{N_I}{n} - a_I \right| \leqslant \left[\frac{2\varepsilon}{(K+R)(1/2+\varepsilon^{-1})} \right] \beta_I \right\}.$$

Then under Assumption 2, for all positive x,

$$\mathbb{P}\left[Z_m \mathbb{1}_{\Omega_m(\varepsilon)} \leqslant (1+\varepsilon)\left(\sqrt{\sum_{I\in m} \frac{\alpha_I}{n\beta_I}} + \sqrt{\frac{2R_m x}{n}} \right) \right] \leqslant \mathrm{e}^{-x},$$

where $R_m = \sup_{I\in m}(\alpha_I/\beta_I)$.

Proof. We know that Z_m is attained at $\hat{\delta}$ such that for all I, $\hat{\delta}_I = [N_I/n - a_I]/[\beta_I Z_m]$. Hence on $\Omega_m(\varepsilon) \cap \{Z_m \geqslant z\}$,

$$Z_m = \sup_{\delta=(\delta_I)_{I\in m}:\begin{cases} \sum_{I\in m}\delta_I^2\beta_I = 1, \\ \sup_{I\in m}\delta_I \leqslant \frac{2\varepsilon}{(K+R)(1/2+\varepsilon^{-1})z} \end{cases}} \left\{ \frac{1}{n}\int_0^1 \left[\sum_{I\in m} \delta_I \mathbb{1}_I(t) \right] \left[\mathrm{d}N_t^i - Y_t^i s(t)\,\mathrm{d}t \right] \right\}.$$

If we apply Talagrand's inequality (Rio 2002) to this last supremum with $z = (2R_m x/n)^{1/2}$, we obtain precisely the previous result. □

We can obtain the same kind of result by replacing R_m by every upper bound on R_m.

2.4. Oracle inequalities

We can now construct oracle-type inequalities. The first is a bound in probability on a large event, for the random norm. The second is an expectation bound for the deterministic norm (1.9).

Theorem 1. *Let N be a counting process with multiplicative intensity* $Y_t s(t)$ *(see* (1.1)*) satisfying Assumption 2. Assume that s is bounded by an unknown positive R. Let* Γ *be a fixed regular partition of* [0, 1] *(i.e. constructed on equally spaced points). Let* $\mathcal{M}_A$ *be a family of partitions which are constructed with unions of intervals of* Γ. *For a given penalty* pen *on* $\mathcal{M}_A$, *let* $\tilde{s}$ *be the associated penalized projection estimator (see* (1.5)*). Assume that:*

(i) *there exist positive constants* μ *and* ρ *such that* $\inf_{I\in\Gamma}(|\Gamma|\alpha_I) \geqslant \mu$ *and* $\inf_{I\in\Gamma}(|\Gamma|\beta_I) \geqslant \rho$;

(ii) *there exists a finite family of positive weights on* $\mathcal{M}_A$, $(L_m)_{m\in\mathcal{M}_A}$, *such that*

$$\sum_{m\in\mathcal{M}_A} \exp(-L_m|m|) \leqslant \Sigma, \qquad \text{for some } \Sigma \text{ independent of } n;$$

(iii) $|\Gamma|$ *is less than* $n/\ln^2 n$.

Let $d > 1$. *Set, for all m in* $\mathcal{M}_A$,

$$\mathrm{pen}(m) = d\tilde{R}_\Gamma \frac{|m|}{n}\left(1+\sqrt{2L_m}\right)^2, \qquad \text{where } \tilde{R}_\Gamma = \sup_{I\in\Gamma}\frac{N_I}{nb_I}.$$

Then there exists a large event $\Omega(d)$ *such that, for all* η *positive, there exist positive continuous functions* C, C' *and* C'' *such that*

$$\mathbb{P}[\Omega(d)^c] \leqslant \frac{C''(d, K, R, \rho, \mu)}{n^\eta}$$

and such that on $\Omega(d)$*, for all* $\xi > 0$ *with probability larger than* $1 - \Sigma e^{-\xi}$,

$$\|s - \tilde{s}\|^2_{\text{rand}} \leqslant C(d) \inf_{m \in \mathcal{M}_A} \left\{ \|s - s'_m\|^2_{\text{rand}} + \frac{|m| L_m}{n} R_\Gamma \right\} + C'(d) R_\Gamma \frac{\xi}{n},$$

where $R_\Gamma = \sup_{I \in \Gamma}(\alpha_I / \beta_I)$.

Corollary 1. *Under the previous assumptions and notation, there exist positive continuous functions* H *and* H' *such that*

$$\mathbb{E}(\|s - \tilde{s}\|^2_{\text{det}}) \leqslant H(d) \inf_{m \in \mathcal{M}_A} \left\{ \|s - s^{\text{det}}_m\|^2_{\text{det}} + R_\Gamma \frac{|m| L_m}{n} \right\} + \frac{H'(d, R, K, \rho, \mu, \Sigma)}{n}.$$

The weights L_m can be constant if the family of partitions has, for instance, at most one model per dimension. Then these oracle-type inequalities become true oracle inequalities and the penalized projection estimator is adaptive in the family $\{\hat{s}_m, m \in \mathcal{M}_A\}$.

The oracle-type inequality of Theorem 1 is a probability bound. It is therefore a stronger result than that of Corollary 1. But for the minimax risk, it is better to have an oracle inequality for a deterministic loss function (here $\|\cdot\|^2_{\text{det}}$ (1.9)).

We can compare this penalized model selection to the model selection constructed by Döhler and Rüschendorf (2002) for the right-censoring case. Their penalty is very large (with a factor exp[exp(R)]) and depends on the knowledge of a bound on s. Here the penalty is linear in R and as we are dealing with histogram estimators, we can estimate the bound on s by $\tilde{R}_\Gamma$. We see in the simulations reported in Section 4 that when the penalty is too large, the estimator behaves poorly and $C(d)$ becomes very large. However, the estimators constructed by Döhler and Rüschendorf (2002) apply to various types of models, in particular to smooth estimators, while we can only prove oracle-type inequalities for histogram estimators.

The weights L_m are needed to take into account the complexity of the family of models. We refer to Birgé and Massart (2001) for an extensive list of applications of these weights.

2.5. Minimax risk

The oracle inequalities imply that the penalized projection estimator is adaptive in its family $\{\hat{s}_m, m \in \mathcal{M}_A\}$: without knowledge of s, the best possible estimator in the family is found up to some multiplicative constant for the risk. But we may also wish to compare it with all other possible estimators. This is the aim of this minimax study.

We know that the histograms have good approximation properties for α-Hölderian functions with $0 < \alpha < 1$. Hence we hope that the penalized projection estimator given in Theorem 1 also has good minimax properties for such a set of functions.

Let L and r be positive constants and let $H_{L,\alpha,r}$ be

$$\{f \in \mathbb{L}^2([0,1], \mathrm{d}t) : \forall x, y \in [0,1], |f(x) - f(y)| \leqslant L|x-y|^\alpha \text{ and } r + L \geqslant f(x) \geqslant r\}.$$

Let the minimax risk on $\mathcal{H}_{L,\alpha,r}$ be defined by $\mathcal{R}(\mathcal{H}_{L,\alpha,r}) = \inf_{\hat{s}} \sup_{s \in \mathcal{H}_{L,\alpha,r}} \mathbb{E}(\|s - \hat{s}\|_{\mathrm{det}}^2)$, where $\hat{s}$ describes all possible estimators in $\mathbb{L}^2([0,1], \mathrm{d}t)$. The minimax risk on $\mathcal{H}_{L,\alpha,r}$ represents the risk of the best estimator for the toughest target function s to estimate in the family $\mathcal{H}_{L,\alpha,r}$.

Proposition 3. *If there exist μ and M such that, for all s in $\mathcal{H}_{L,\alpha,r}$, $\mu \leqslant \mathbb{E}(Y_t^1) \leqslant M$, then there exists a positive continuous function c such that*

$$\mathcal{R}(\mathcal{H}_{L,\alpha,r}) \geqslant c(\alpha) n^{-2\alpha/(2\alpha+1)} L^{2/(2\alpha+1)} r^{2\alpha/(2\alpha+1)} \mu M^{-2\alpha/(2\alpha+1)}.$$

The above assumptions are true in many situations. For instance, in the right-censoring model, retaining the notation given in the Introduction, $\mathbb{E}(Y_t^1) = \mathbb{P}(X \geqslant t)\mathbb{P}(U \geqslant t)$. Hence, it is less than 1 and larger than $\mathbb{P}(X \geqslant 1)\mathbb{P}(U \geqslant 1)$. The above assumptions are then equivalent to saying that the death time (X) and the departure time (U) may happen after the end of the observation interval.

We also remark that the exponent in n is the rate of convergence of the classical regression problem.

We now wish to compare the risk of $\tilde{s}$ constructed in Theorem 1 with the minimax risk. Let us look at the following classical strategy: $|\Gamma| = 2^J$ is of order $n/\ln^2 n$ and we take the partitions, m, constructed with union of intervals of Γ which are also regular with 2^j intervals and with j less than J. There is one model per dimension at most. Hence, we can take constant weights ($L_m = 1$, for instance) to construct the penalty. We call this strategy the *nested histogram strategy*. Now let us apply Corollary 1.

If s is in $\mathcal{H}_{L,\alpha,r}$, the bias $\|s - s_m^{\mathrm{det}}\|_{\mathrm{det}}^2$ is bounded by $L^2|m|^{-2\alpha}\varsigma$, where $\varsigma = \int_0^1 \mathbb{E}(Y_t^1)\,\mathrm{d}t$. When n tends to infinity, we obtain, taking m such that $|m|$ is of order $(n\varsigma L^2/R)^{1/(2\alpha+1)}$ (which is less than $|\Gamma|$ for n large enough),

$$\mathbb{E}(\|s - \tilde{s}\|_{\mathrm{det}}^2) = O(n^{-2\alpha/(2\alpha+1)} L^{2/(2\alpha+1)} R^{2\alpha/(2\alpha+1)} \varsigma^{1/(2\alpha+1)}).$$

We can compare this bound to the lower bound found in Proposition 3. One obtains the same power of n and L. The bound R on s replaces the infimum r of s. The quantity ς replaces $\mu^{2\alpha+1} M^{-2\alpha}$ and represents the order of magnitude of $\mathbb{E}(Y_t^1)$.

This means that without knowing α and L (depending on s), $\tilde{s}$ does as well as the best possible estimator which knows these quantities. In this sense, $\tilde{s}$ is an *adaptive estimator* for the α-Hölderian functions with $0 < \alpha < 1$.

3. Predictable models

We have seen what can easily be done for aggregated processes. Let us remove this assumption and deal with predictable models. We retain the notation introduced in (1.3) and (1.4). We now suppose that Assumption 1 and the following hold:

Assumption 3. *There exists c positive such that if $Y_t < c$, for some $t > 0$, then $Y_t = 0$.*

For a Poisson process, one has $A = c$. For the other examples, Y is an integer-valued function and $c = 1$ works.

The aggregated case leads us to think that A plays the same role as n. Consequently, the asymptotic point of view in this framework is that A tends to infinity. On the other hand, c is considered to be a fixed constant, independent of A.

3.1. Construction and risk for one model

Let $J_t = \mathbb{1}_{Y_t \neq 0}$. The family of models is then constructed as follows. Let $\{\varphi_\lambda, \lambda \in \Gamma\}$ be a classical orthonormal basis of $\mathbb{L}^2([0, 1], \mathrm{d}t)$; let $\mathcal{M}_A$ be a family of subsets of Γ. Then for m in $\mathcal{M}_A$, we set $S_m = \mathrm{Span}\{h_\lambda(\cdot) = [\varphi_\lambda(\cdot)/\sqrt{Y'_\cdot}]J_\cdot,\ \lambda \in m\}$. Let $|m|$ be the cardinality of m. Let $\hat{s}_m$ be the projection estimator associated with S_m and defined by (1.5). We remark that the $h_\lambda(\cdot)$ are not continuous, only piecewise continuous if Y is piecewise continuous, as in the right-censoring model, for instance. Consequently, there is no reason for $\hat{s}_m$ to be continuous.

Let us also define the following observable event:

$$\Omega = \{\forall t \geqslant 0,\ Y_t \neq 0\}. \tag{3.1}$$

We will see later that in many situations, Ω has a very large probability of happening when A is large enough.

On Ω, the h_λ form an orthonormal basis of S_m for the random scalar product, and consequently $\hat{s}_m$ is of the form (1.6).

We now turn to the risk of the projection estimator. On Ω, the projection s_m of s on S_m for the random inner product is given by

$$s_m(\cdot) = \sum_{\lambda \in m} \left[\int_0^1 \varphi_\lambda(t) s(t) \sqrt{Y'_t}\, \mathrm{d}t\right] \frac{\varphi_\lambda(\cdot)}{\sqrt{Y'_\cdot}}.$$

Hence, we can write $\|s - \hat{s}_m\|^2_{\mathrm{rand}} = \|s - s_m\|^2_{\mathrm{rand}} + \|s_m - \hat{s}_m\|^2_{\mathrm{rand}}$. The first term corresponds to a bias term and the expectation of the second term is a variance term.

The bias term is random as in the histogram case. We can write

$$\|s - s_m\|^2_{\mathrm{rand}} = \int_0^1 \left[s(t)\sqrt{Y'_t} - \sum_{\lambda \in m} \left[\int_0^1 \varphi_\lambda(t) s(t) \sqrt{Y'_t}\, \mathrm{d}t\right] \varphi_\lambda(t)\right]^2 \mathrm{d}t.$$

Thus, the bias term corresponds to the classical $\mathbb{L}^2([0, 1], \mathrm{d}t)$ error when one projects $s\sqrt{Y'}$ on $\mathrm{Span}\{\varphi_\lambda, \lambda \in m\}$. If m grows, this term should generally decrease.

The second term corresponds to a χ^2-type statistic and behaves quite differently: on Ω, it is $\chi(m)^2_1$ where the process $(\chi(m)^2_t)_{t \geqslant 0}$ is defined by

$$\chi(m)^2_t = \sum_{\lambda \in m} \left[\int_0^t \frac{\varphi_\lambda(u)}{\sqrt{Y'_u}} J_u \frac{\mathrm{d}M_u}{A}\right]^2, \qquad \text{for all } t \geqslant 0. \tag{3.2}$$

It has a compensator $(C(m)_t)_{t \geqslant 0}$ defined by $C(m)_t = \sum_{\lambda \in m} \int_0^t \varphi^2_\lambda(u) s(u) J_u\, \mathrm{d}u/A$, for all positive t. But on Ω, $C(m)_1$ is constant and, moreover, if $r \leqslant s \leqslant R$, then

$(r|m|/A) \leqslant C(m)_1 \leqslant (R|m|/A)$. Hence, if $\chi(m)_1^2$ is close to $C(m)_1$, it increases as the dimension of the model.

3.2. Penalized projection estimator

In the same way, if we wish to find a good model, we must balance the bias term and the variance term, but we must adaptively do this through a data-driven criterion, without extra knowledge of s. Therefore we use (1.7) and obtain $\tilde{s}$, the penalized projection estimator for the family of models $\{S_m,\ m \in \mathcal{M}_A\}$.

There also is a heuristic argument. We can define an oracle, the best model, which we could choose if we knew s:

$$\begin{aligned}\overline{m} &= \operatorname*{argmin}_{m\in\mathcal{M}_A} \|s - \hat{s}_m\|_{\text{rand}}^2 \qquad (3.3)\\ &= \operatorname*{argmin}_{m\in\mathcal{M}_A} \left(-\|s_m\|_{\text{rand}}^2 + \|s_m - \hat{s}_m\|_{\text{rand}}^2\right)\\ &\simeq \operatorname*{argmin}_{m\in\mathcal{M}_A} \left(-\|\hat{s}_m\|_{\text{rand}}^2 + \|2s_m - \hat{s}_m\|_{\text{rand}}^2\right)\\ &\simeq \operatorname*{argmin}_{m\in\mathcal{M}_A} \left(-\|\hat{s}_m\|_{\text{rand}}^2 + 2\chi(m)_1^2\right).\end{aligned}$$

The approximations ($\simeq$) are good if the coefficients of $s_m - \hat{s}_m$ are close to their expectation, which is 0, as in the histogram case. If $\chi(m)_1^2$ is close to $C(m)_1$, a penalty of the form $2c|m|/A$ would be convenient (where c is of the order of s). Again we found the factor 2 which always appears when doing this kind of heuristic and which is due to Mallows (1973) in the Gaussian framework.

The probabilistic behaviour of $\chi(m)_t$ around its compensator has already been studied (Reynaud-Bouret 2006).

3.3. Oracle inequalities

We can now derive oracle-type inequalities for predictable models.

Theorem 2. *Let N be a counting process with multiplicative intensity $Y_t s(t)$ (see* (1.1)*) satisfying Assumptions 1 and 3. Let $\{S_m,\ m \in \mathcal{M}_A\}$ be a family of predictable models constructed as previously from the deterministic classical orthonormal family $\{\varphi_\lambda,\ \lambda \in \Gamma\}$. For a given penalty* pen *on $\mathcal{M}_A$, let $\tilde{s}$ be the associated penalized projection estimator (see* (1.7)*). Assume that:*

(i) *there exists a positive constant Φ, such that for all m in $\mathcal{M}_A$, $\|\sum_{\lambda\in m} \varphi_\lambda^2\|_\infty \leqslant \Phi|m|$;*
(ii) *there exists a finite family of positive weights on $\mathcal{M}_A$, $(L_m)_{m\in\mathcal{M}_A}$, such that*

$$\sum_{m\in\mathcal{M}_A} |m|^2 \exp(-L_m) \leqslant \Sigma.$$

Moreover, assume that we know a bound on s denoted by R. Let $d > 1$. *Set, for all m in* $\mathcal{M}_A$,

$$\mathrm{pen}(m) = d\frac{|m|}{A}\left[\sqrt{R}\left(1+3\sqrt{2L_m}\right)+\sqrt{\frac{\Phi}{c}}L_m\right]^2.$$

Then there exist positive continuous functions C and C′ such that, on Ω *defined by* (3.1),

$$\mathbb{E}(\|s-\tilde{s}\|^2_{\mathrm{rand}}\mathbb{1}_\Omega) \leqslant C(d)\inf_{m\in\mathcal{M}_A}\{\mathbb{E}(\|s-s_m\|^2_{\mathrm{rand}})+\mathrm{pen}(m)\}+\frac{C'(d,\,R,\,\Phi,\,c,\,\Sigma)}{A}.$$

As the models are random, we can only derive oracle-type inequalities for the random norm. Probability bounds exist but are much more intricate than in Theorem 1 (see Section 5).

The classical case is when $\{\varphi_\lambda,\,\lambda\in\Gamma\}$ is a Fourier basis $\{\exp(-2\mathrm{i}k\pi x),\,k\in\mathbb{Z}\}$ with $\mathcal{M}_A=\{m_k=\{-k,\,k\},\,k\geqslant 0\}$. Then one has $|m_k|=2k+1$ and $L_{m_k}=4\ln k$. The constant Φ in the theorem is then equal to 1. In practice we must take a finite family of models, for instance by setting $k\leqslant A$.

We can also consider a wavelet basis $\{\varphi_{j,k},\,j\geqslant 0,\,k\geqslant 0\}$ with regularity h and $\mathcal{M}_A=\{m_l,\,l\geqslant 0\}$ where $m_j=\{(l,\,k),\,l\leqslant j\}$. If the wavelet has finite support, Φ defined in Theorem 2 depends only on the choice of the basis.

As the family of models is nested in both previous cases, the penalty is of order $|m|R\log(|m|)/A$. Thus we recover an oracle inequality up to a logarithmic factor, since the variance term is of order $|m|/A$. We can think of more complex families of models (i.e. more models with the same dimension). If the number of models with dimension D in the family is of the order of a power of D, we can have the same kind of penalty and we also recover an oracle inequality up to a logarithmic factor. If the number of models with the same dimension D is of order e^D, the penalty must be of order $R|m|^\gamma/A$, for $\gamma>1$. It is really larger than the variance term and this cannot lead to an oracle inequality.

When one has an oracle inequality, one can also say that the penalized projection estimator is adaptive in the family $\{\hat{s}_m,\,m\in\mathcal{M}_A\}$. But as we do not know the approximation properties of the random spaces S_m, we cannot, in general, consider the adaptive properties in the minimax sense.

However, if N is a Poisson process, $Y_t=A$ is deterministic. This implies that all the norms are deterministic. In this case, let us assume that s belongs to

$$\mathcal{B}(\rho,\,L,\,B^\alpha_{2,2})=\left\{t=\rho+u:\,t\geqslant 0,\,\int_0^1 u\,\mathrm{d}x=0,\,u\in B^\alpha_{2,2},\,\|u\|^\alpha_{2,2}\leqslant L\right\},$$

where ρ and L are positive constants and $B^\alpha_{2,2}$ is the classical Besov space with regularity α $(1/2\leqslant\alpha\leqslant h)$ and with $\mathbb{L}^2$ norm. Let us consider the last strategy with a wavelet family of regularity h. Then compromising between the penalty and the bias in the oracle-type inequality, we obtain, as A tends to infinity,

$$\mathbb{E}(\|s-\tilde{s}\|^2) = O\left(L^{2/(2\alpha+1)} R^{2\alpha/(2\alpha+1)} \left(\frac{A}{\ln^2 A}\right)^{-2\alpha/(2\alpha+1)}\right).$$

This is the minimax rate of Reynaud-Bouret (2003) up to the logarithmic factor and the replacement of $\int_0^1 s$ by R. Therefore, the resulting penalized projection estimator is adaptive in the minimax sense for all Besov balls with regularity less than h, up to a logarithmic factor.

This logarithmic factor is actually not necessary in the Poisson case: Reynaud-Bouret (2003) proved that penalties of the type $R|m|/A$ with the same previous families of models lead to an oracle inequality without logarithmic factor, which leads in turn to the minimax rate without logarithmic factor. If we apply Theorem 2, which is valid for more general processes, the weights L_m are constant and Σ grows too fast for complex families of models. Theorem 2 no longer implies an oracle inequality for these penalties.

The same kind of remark can be made if we wish to use a more complex family of models (i.e. more models with the same dimension in the family of models). In the Poisson framework, there exist penalties of the type $R|m|(\log A)/A$ which are proved to lead to an oracle inequality up to some logarithmic factor. Applying Theorem 2 to the same type of strategies gives an explosive last term. However, the general counting processes with Aalen multiplicative intensity are very well adapted to biomedical data. In such cases, the number of observations $n \simeq A$ is not very large and if we also take a small number of models, there is no longer an explosive phenomenon. This justifies the interest in having non-asymptotic results.

3.4. Improvements

3.4.1. Estimation of R

The fact that the penalty depends on the knowledge of a bound on s can be a nuisance. In some cases, we can estimate this bound.

Let Γ be a regular partition of [0, 1]. Suppose that s is (L, α)-Hölderian, and let s_Γ be the projection of s for the random norm on S_Γ. Then $\|s-s_\Gamma\|_\infty \leqslant L|\Gamma|^{-\alpha}$. Take $|\Gamma|$ of order $A/\ln^2 A$. Then $\|s\|_\infty \leqslant \|s_\Gamma\| + o(1)$, when A goes to infinity. But $\|s_\Gamma\|_\infty = \sup_{I\in\Gamma}\left(\int_I s(t)Y'_t\,\mathrm{d}t\right)/\left(\int_I Y'_t\,\mathrm{d}t\right)$.

So we can replace R by $(1+\varepsilon)\tilde{R}_\Gamma$, where $\tilde{R}_\Gamma = \sup_{I\in\Gamma} N_I/\left(A\int_I Y'_t\,\mathrm{d}t\right)$, if we are on $\Omega(\varepsilon) = \left\{\left|\int_I \mathrm{d}M_t/A\right| \leqslant \varepsilon(1+\varepsilon)^{-1}\int_I s(t)Y'_t\,\mathrm{d}t\right\}$. The complement of this last event is very small (it has probability of order $o(A^{-\eta})$, for all $\eta > 0$) if we assume the process to be aggregated and Assumption 2 (or moment assumptions). Then we can apply Bernstein's inequality to $\int_I \mathrm{d}M/A$ and to $\int_I s(t)Y'_t\,\mathrm{d}t$. On $\Omega\cap\Omega(\varepsilon)^c$ the estimator is bounded and one can conclude as in the proof of Corollary 1.

3.4.2 Magnitude of Ω

In the aggregated cases, Ω is a very large event and we can also give an oracle-type inequality for $\mathbb{E}(\|s-\tilde{s}\|^2_{\mathrm{rand}})$.

Let us look more closely at the right-censoring model. In this case $A = n$ and $Y'_t = \sum_{i=1}^{n} \mathbb{1}_{X_i \wedge U_i \geqslant t}$, where the X_i are the lifetimes and the U_i form the censorship. Then Y'_t can be seen as $1 - \hat{F}_n(t)$, where $\hat{F}_n(t)$ is the empirical cumulative distribution function associated with the $X_i \wedge U_i$. We have,

$$\forall \lambda > 0, \quad \mathbb{P}\left(\sqrt{n} \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| \geqslant \lambda\right) \leqslant 2\mathrm{e}^{-2\lambda^2},$$

where F is the true cumulative distribution function of the $X_i \wedge U_i$ (Massart 1990).

Thus if we assume that there exists a positive μ such that $\mathbb{E}(Y'_t) \geqslant \mu > 0$ on $[0, 1]$, then $\Omega^c \subset \{\sup_{t \in \mathbb{R}} |Y'_t - \mathbb{E}(Y'_t)| \geqslant \mu/2\}$, and $\mathbb{P}(\Omega^c) \leqslant 2\exp(-n\mu^2/2)$.

Hence, we can define the estimators on the whole probability space by

$$\hat{s}_m(\cdot) = \sum_{\lambda \in m} \left[\int_0^1 \frac{\varphi_\lambda(t)}{\sqrt{Y'_t}} J_t \frac{\mathrm{d}N_t}{A}\right] \frac{\varphi_\lambda(\cdot)}{\sqrt{Y'_\cdot}} J_\cdot,$$

even if we are not in Ω. This estimator is a projection estimator only on Ω. We do the model selection as in Theorem 2. As these estimators are still bounded, we proceed as in Corollary 1 and we can bound $\mathbb{E}(\|s - \tilde{s}\|^2_{\mathrm{rand}})$ (on the whole probability space) by the same kind of bound as in Theorem 2.

4. Simulations

The aim of this section is to illustrate the previous methods. In Section 2 we proposed a piecewise constant selected estimator which is adaptive in its family of estimators and in the minimax sense. In Section 3 we obtained a predictable estimator which is adaptive in its family of estimators. As we cannot prove minimax properties for this last estimator, we would like to check whether this estimator (which even in the Fourier case is not continuous) behaves poorly or not in practice. One way to do this is to compare the predictable estimator with the piecewise constant estimator of Section 2. But since the estimator of Section 3 looks smoother (at least visually), one may want to compare it with smoother estimators such as the wavelet-based estimator proposed by Antoniadis *et al.* (1999), which is completely data-driven. A common probabilistic set-up in which these comparisons are meaningful is the right-censoring model.

4.1. Five different strategies

The lifetimes $X_1, \ldots, X_n$ are generated for a given hazard rate s on $[0, 1]$. The censorship variables $U_1, \ldots, U_n$ are generated as uniform variables on $[0, 2]$. We observe $T_i = X_i \wedge U_i$ and $D_i = \mathbb{1}_{T_i = X_i}$, for all i less than n. Some of the Ts will be outside $[0, 1]$: this is a good case since it ensures that the event Ω introduced in the previous section happens. The random norm (1.3) is denoted by 'Risk' in the figures.

The first three strategies are histogram strategies whose adaptive properties are proved in

Section 2. The fourth strategy is a predictable estimator introduced in Section 3. The fifth strategy is a data-driven choice between the four previous strategies.

The *regular histogram strategy* consists in taking the quasi-least-squares estimators (2.1) for all the regular partitions of [0, 1] up to a certain number of intervals which is the minimum of the number of observations and 20. The factor 20 is used so that computing times are short. As there is one model per dimension and as there exists a big partition Γ (even if we do not know its precise form) such that all these partitions are subpartitions of Γ, Theorem 1 enables us to have constant weights. It is therefore convenient to take a penalty of the form

$$\text{pen}(m) = \frac{d\tilde{R}|m|}{n},$$

where $\tilde{R} = \sup_{I \in \Lambda} N_I/(nb_I)$, in which Λ is just the thinnest partition of our family of models replacing Γ in order to simplify the computations. The resulting penalized projection estimator given by (1.7) is denoted by RHS. The estimator $\tilde{R}$ of the supremum of s is used for all the other strategies in the penalty.

The *exhaustive histogram strategy* consists in taking the quasi-least-squares estimators (2.1) for all the partitions which are constructed with unions of intervals of Γ, where Γ is a regular partition with d intervals and where d is the minimum of 8 and the integer part of $[n/\log(n)^2]$. Again, the factor 8 is present so that we have short computing times. The penalty is of the form

$$\text{pen}(m) = \frac{d\tilde{R}|m|}{n}\left[1 + \sqrt{2\log\left(\frac{n}{|m|}\right)}\right]^2,$$

that is, the weights L_m of Theorem 1 are of the $\log(n/|m|)$ type to ensure the convergence of Σ. The resulting penalized projection estimator given by (1.7) is denoted by EHS.

The *progressive histogram strategy* is specially devised to take into account the fact that we have a poor estimation near 1. It consists in taking the quasi-least-squares estimators (2.1) for the partitions whose intervals are small near 0 and large near 1. Specifically, in addition to the family of regular partitions $(\{0, 1/N, 2/N, \ldots, (N-1)/N, 1\})$, we have partitions which progress polynomially $(\{0, 1^k/N^k, 2^k/N^k, \ldots, (N-1)^k/N^k, 1\})$, and also partitions which progress exponentially $(\{0, k^1/k^N, k^2/k^N, \ldots, k^{N-1}/k^N, 1\})$. We take this for all integers $k \leqslant 3$ and all integers N less than the minimum of 20 and one-third of the number of observations. Once more, the factors 3 and 20 give us small computing times. As for the RHS, Theorem 1 allows us to have constant weights, and we therefore use a penalty of the form

$$\text{pen}(m) = \frac{d\tilde{R}|m|}{n}.$$

The resulting penalized projection estimator given by (1.7) is denoted by PHS.

The *Fourier strategy* is the strategy described in the previous section. The φ_λ form the Fourier basis and we consider the nested models described in Section 3.3 (see (1.6) for the

form of the projection estimator with h_λ defined in Section 3.1). By Theorem 2, a convenient penalty is of the form

$$\text{pen}(m) = \frac{d|m|}{n}\left[\sqrt{\tilde{R}} + \log|m|\right]^2.$$

In order to give simpler formulae, we have omitted the second term of the penalty, which is smaller than the other terms. The resulting penalized projection estimator given by (1.7) is denoted by FS. This estimator is piecewise continuous.

If d is well chosen in all the previous strategies, the penalized criteria $\{\gamma_A(\tilde{s}) + \text{pen}(\hat{m})\}$ must estimate the risks of each projection estimator, $\{\|s - \tilde{s}\|^2_{\text{rand}} - \|s\|^2_{\text{rand}}\}$. The last strategy, then, is the *minimal criteria strategy* (MCS) which chooses from the four previous estimators that with the smallest penalized criterion, that is, the smallest $\{\gamma_A(\tilde{s}) + \text{pen}(\hat{m})\}$. Of course, before computing this last strategy, we have to find good parameters d for the previous four strategies, which ensure that the penalized criteria are close to the risks of the projection estimators.

First let us remark that Theorems 1 and 2 tell us that for all $d > 1$ the resulting strategies are adaptive. This is completely different from the selection of a bandwidth for kernel estimators where this choice depends on the regularity of the function s and on the number of observations. Here the choice of d is less fundamental and does not depend on the function or on the number of observations but only on the strategy we use. Intuitively, the Mallows heuristic (see (2.6) and (3.3)) suggests that $d = 2$ is a good choice. In practice, we studied a lot of examples (see Figure 3). What typically happens is that there is a very large range of possible d (see Figure 1 for the RHS, for instance) which lead to good estimators (i.e. in most cases they find the oracle model). Finding the best possible d is not our purpose here. Lebarbier (2002) shows how to choose the best optimal d in the Gaussian framework. This is a long and complete work which leads in practice to amazing results (her estimators can even sometimes beat the oracle!). Here we only aim to take reasonable values for d. In all the following simulations we hence set: for the RHS $d = 2$, for the EHS $d = 0.4$, for the PHS $d = 2.5$ and for the FS $d = 1$.

We have computed the risk for the four methods on various sets of functions. Figure 2 presents what happens for two particular examples where we clearly see the differences between the three kinds of estimators by histograms. In both cases the MCS method gives the estimator with minimum risk (i.e. the PHS for Figure 2(a) and FS for Figure 2(b)).

Figure 3 presents the hazard rate functions and Table 1 gives the risk of the different estimators. More precisely, for each kind of hazard rate with uniform censorship on [0, 2], we simulate either 200 or 500 observations. All the results are given as averages over 200 iterations.

Some of the simulated observations are greater than 1 and that is the reason why we indicate the number of data which are strictly between 0 and 1 and also the number of uncensored data. We also give the most frequent choice of the MCS to see if it corresponds to the minimum of the risk. Of course, the risk of the MCS is not exactly that of the most frequent choice because sometimes the MCS chooses something different. When two strategies are chosen with approximately the same frequency by the MCS, we give both.

We first remark that, for a fixed hazard and a fixed method, the risk seems to decrease

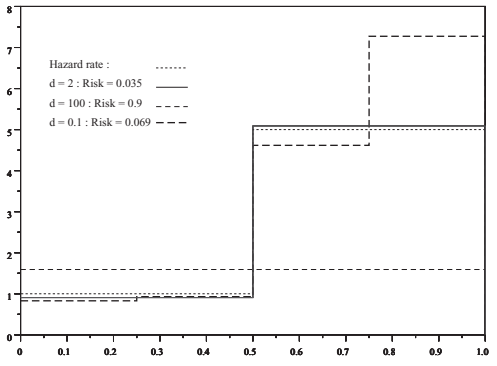

Figure 1. Example of the influence of d for the regular histogram strategy (98 observations in [0, 1], 65 uncensored observations).

with the number of variables. Moreover, this risk is proportional to s, as we can see for functions 1 and 2. This has to be taken into account when comparing the results for different types of functions.

The risk is larger when the function is not in the family of models we are using, and this happens even when the function is piecewise constant but its partition does not belong to the family of partitions we have taken. For instance, it explains why for function 4 the risk of the EHS is bigger for 500 observations than for 200: the way we have built the biggest partition Γ gives a regular partition with 8 intervals for 500 observations and with 7 intervals for 200 observations. In this last partition there is one point close to 0.3, which no longer exists when we take 500 observations.

In general, for piecewise constant hazard rates, the histograms family are better, and this is the most frequent choice of the MCS. When the function is smoother, the FS sometimes seems better even if it is not smooth, and this is also the most frequent choice of the MCS.

The PHS seems to work well even for smooth functions. This is probably because it is very well adapted to finding differences in behaviour near the origin: for instance, it detects more easily the bump in function 11 (see also function 13) than the FS which is not localized and which tends to oscillate when the hazard rate remains flat (see also Figure 2(b)).

The EHS does not seem so useful at first sight, but in some cases (function 6 with 500 observations, for instance), it is in fact similar to the PHS and therefore has the same risk. In terms of criteria, as there is a logarithmic factor for the penalty of the EHS, the MCS always prefers the PHS. Another explanation for the fact that the PHS seems better than the EHS is that we cannot trust any estimator at the end of the interval. The PHS, which already takes this fact into account and which has a simpler family of models, gives better estimation. This phenomenon is due to the fact that we have right-censored data; for other types of counting processes, this would probably not happen.

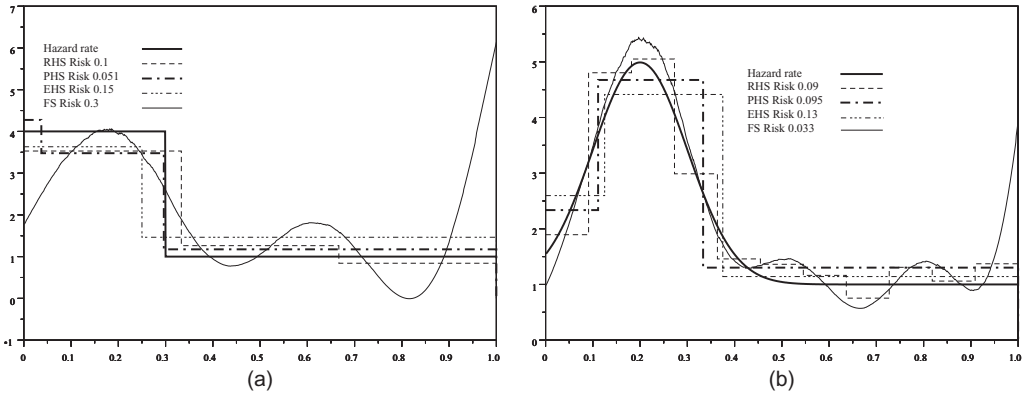

Figure 2. Results for (a) a piecewise constant function (460 observations in [0, 1], 370 uncensored observations, MCS = PHS) and (b) a smooth function (470 observations in [0, 1], 390 uncensored observations, MCS = FS).

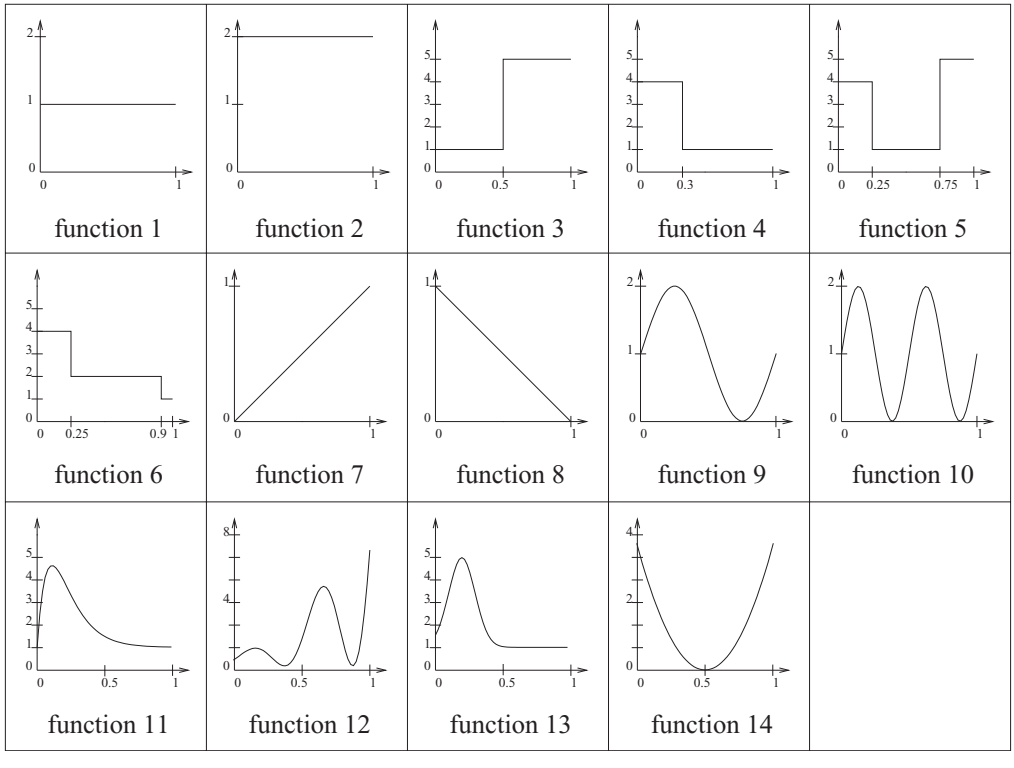

Figure 3. Hazard rates to be estimated.

Table 1. Risks for the different estimators for 200 and 500 simulated observations: the most frequent choice of the MCS is in bold type; HR stands for 'hazard rate'

	Observations in [0, 1] (uncensored)		RHS		PHS		EHS		FS		MCS	
HR	200	500	200	500	200	500	200	500	200	500	200	500
1	163 (99)	408 (249)	**0.007**	**0.003**	0.008	0.004	0.004	0.002	0.032	0.020	0.014	0.005
2	186 (142)	467 (357)	**0.013**	**0.005**	0.015	0.006	0.010	0.004	0.119	0.072	0.023	0.008
3	195 (144)	487 (360)	**0.03**	**0.01**	0.03	0.01	0.40	0.01	0.28	0.16	0.04	0.01
4	185 (152)	462 (380)	0.107	0.08	**0.09**	**0.04**	0.06	0.13	0.31	0.21	0.09	0.04
5	193 (159)	484 (400)	**0.07**	**0.02**	0.13	0.02	0.43	0.02	0.40	0.23	0.12	0.02
6	190 (160)	477 (402)	0.10	0.03	**0.07**	**0.02**	0.14	0.02	0.33	0.22	0.07	0.02
7	140 (54)	350 (134)	**0.02**	**0.01**	0.02	0.01	0.04	0.02	**0.03**	**0.01**	0.03	0.01
8	140 (66)	349 (168)	**0.02**	**0.01**	0.02	0.01	0.03	0.01	0.04	0.03	0.02	0.01
9	163 (108)	407 (271)	0.06	0.04	0.06	0.04	0.06	0.05	**0.03**	**0.02**	0.04	0.03
10	163 (104)	408 (261)	0.06	0.05	0.09	0.05	0.20	0.05	**0.03**	**0.02**	0.04	0.02
11	188 (157)	469 (391)	0.22	0.17	**0.17**	**0.10**	0.23	0.19	0.21	0.10	0.18	0.10
12	186 (131)	465 (327)	0.25	0.17	0.26	0.19	0.34	0.21	**0.10**	**0.07**	0.12	0.07
13	186 (152)	465 (380)	0.26	0.12	**0.16**	**0.11**	0.27	0.15	**0.10**	**0.06**	0.14	0.07
14	171 (115)	428 (289)	**0.15**	**0.08**	0.21	0.10	0.21	0.12	**0.17**	**0.10**	0.18	0.10

Globally the MCS is a good way to choose among all the strategies: even if it does not achieve the minimal risk, its risk is always of the same order as the minimum of the risks.

4.2. Comparison with other existing results

In this subsection, we wish to compare our estimators with existing ones. Of course, there are a lot of nonparametric estimators for the hazard rate, but actually Antoniadis *et al.* (1999) are the only authors we know who propose adaptive procedures and apply them. Their estimator is a wavelet estimator and they choose the coefficients to keep by a cross-validation criterion. Therefore, their estimator has the same quality as ours: this is a completely data-driven nonparametric estimator.

As their estimator is constructed on $[0, \tau]$, where τ is the last observation, we do the following rescaling: we divide the observations by τ to obtain a new set of observations in [0, 1], and as the last point is always 1, we are always in Ω. This new set of observations has an intensity of the form $\bar{s}(t) = \tau s(\tau t)$ (if τ is deterministic). We estimate it on [0, 1] by $\tilde{s}$ coming either from the RHS ($d = 2$), the PHS ($d = 2.5$), the EHS ($d = 0.4$), the FS ($d = 1$) or finally the MCS. Then the resulting estimator for s on $[0, \tau]$ is $\hat{s}(x) = \tilde{s}(x/\tau)/\tau$.

In the first set of simulations, the X_i follow a gamma distribution with shape parameter 5 and scale 1 and the U_i follow an exponential distribution with mean 6. The results are displayed in Figure 4(a).

In the second set of simulations, the X_i have a bimodal density defined by

$$f = 0.8g + 0.2h,$$

where g is the density of $\exp(Z/2)$ with Z having a standard normal distribution and where h is the density of $0.17Z + 2$. The U_i have an exponential distribution with mean 2.5. The results are displayed in Figure 4(b).

In both cases, we see that all the estimators (and especially the FS) are very inefficient at the end of the interval since by construction one has few observations towards the end of the interval.

We can compare our estimators with theirs by computing the same error on a lot of simulations. If one takes K regularly spaced points in $[0, \tau]$, denoted by t_k, the AMSE error is defined by

$$\text{AMSE} = \frac{1}{K}\sum_{k=1}^{K}(\hat{s}(t_k) - s(t_k))^2.$$

The AMSE2 error is defined for the first simulation by the same kind of mean squared error but only for $t_k < 6$. This is done in order to remove the effect of scarcity of the observations. One has $\mathbb{P}(X > 6) = 0.25$.

For the second simulations, after discussion with G. Grégoire, the AMSE2 is done for $t_k < 2$. One has here that $\mathbb{P}(X > 2) = 0.16$. (There is a small misprint in Antoniadis *et al.* (1999), in which they should have written 2 instead of 2.5 which is inadequate since $\mathbb{P}(X > 2.5) = 0.02$.)

All the errors are computed over 200 simulations.

The results of Antoniadis *et al.* (1999) are presented in Table 2. As their procedure of estimation depends on the t_k, there are three possible choices for the partitions.

We give AMSE, AMSE2 and the risk for our estimators in Table 3. As our procedures do not depend on the choice of the t_k, we find the same order of magnitude for each possibility. The results presented here are given with 64 points regularly spaced.

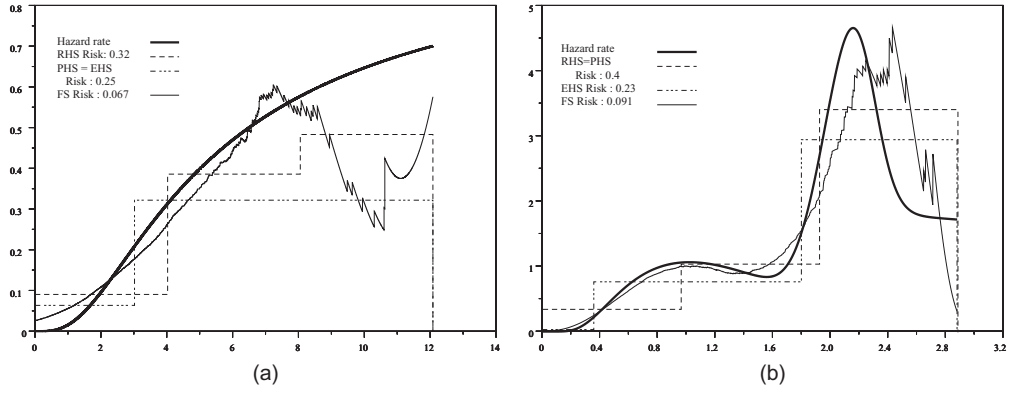

Figure 4. Estimation for (a) a gamma distribution (230 uncensored observations) and (b) a bimodal distribution (280 uncensored observations). In both cases MCS = FS.

We see that the histogram strategies are better than theirs on the entire intervals in both cases. This is due to the fact that histograms do not oscillate at the end of the interval where there are fewer and fewer observations, for they are more robust. On the other hand, histograms give larger results on the shorter intervals, because they are less 'smooth' than the FS strategy. The FS strategy, which is also, in this case, the one chosen by the MCS, gives results of the same order as that of Antoniadis *et al.* (1999). The FS is better for the whole interval (especially for the bimodal hazard rate), but is worse for AMSE2 with 200

Table 2. Antoniadis *et al.*'s results

Distributions		Gamma		Bimodal	
Number of observations		200	500	200	500
AMSE	K = 16	0.0644	0.0554	3.050	3.090
	32	0.0786	0.0554	4.060	1.820
	64	0.112	0.0995	2.080	1.970
AMSE2	16	0.0058	0.0059	0.182	0.295
	32	0.0026	0.0021	0.152	0.066
	64	0.0025	0.0016	0.048	0.032

Source: Antoniadis *et al.* (1999: Table 2).

Table 3. Results of the penalized projection estimators

Distributions		Gamma		Bimodal	
Number of observations		200	500	200	500
RHS	AMSE	0.0333	0.0376	0.894	0.789
	AMSE2	0.0086	0.0048	0.255	0.152
	Risk	0.278	0.179	0.559	0.321
PHS	AMSE	0.0275	0.0224	1.107	0.862
	AMSE2	0.0069	0.0054	0.265	0.142
	Risk	0.246	0.190	0.617	0.338
EHS	AMSE	0.0431	0.0315	1.384	0.832
	AMSE2	0.0123	0.0059	0.363	0.175
	Risk	0.397	0.243	0.865	0.415
FS	AMSE	0.055	0.0579	1.259	1.122
	AMSE2	0.0032	0.0012	0.150	0.051
	Risk	0.138	0.0817	0.426	0.183
MCS	AMSE	0.055	0.0579	1.289	1.103
	AMSE2	0.0032	0.0012	0.160	0.051
	Risk	0.138	0.0817	0.437	0.185

observations. However, the same phenomenon appears: AMSE2 is much smaller than AMSE in every case.

4.3. Conclusion

It seems that the methods introduced in Sections 2 and 3 are suitable for use in practice, for they give results of the same order as other estimators and even better ones if we want to estimate the hazard rate as far as possible (i.e. until the last observation). The FS, for which we are not able to prove minimax results in the general case, seems to work quite well and gives results that look smoother than the histogram strategies even if it is not continuous. The MCS which assumes that the penalized criterion is close to the risk up to a constant, allows us to take almost the best estimator among a heterogeneous family of estimators (RHS, PHS, EHS, FS) and seems to be more robust than each individual strategy.

5. Proofs of the main results

Proof of Theorem 1. Let d be a real number larger than 1 and let ε be a positive continuous function of d which we will choose later. Define the event

$$\Omega(d) = \left\{ \forall I \in \Gamma, \left| \frac{N_I}{n} - a_I \right| \leqslant \frac{2\varepsilon}{(K+R)(1+\varepsilon^{-1})} \beta_I, \left| \frac{N_I}{n} - \alpha_I \right| \leqslant \frac{\varepsilon}{1+\varepsilon} \alpha_I, \right.$$
$$\left. |b_I - \beta_I| \leqslant \frac{\varepsilon}{1+\varepsilon} \beta_I \right\}.$$

Let us bound the probability of $\Omega(d)^c$:

$$\mathbb{P}[\Omega(d)^c] \leqslant \sum_{I \in \Gamma} \left[\mathbb{P}\left(\left| \frac{N_I}{n} - a_I \right| \geqslant \frac{2\varepsilon}{(K+R)(1+\varepsilon^{-1})} \beta_I \right) \right.$$
$$\left. + \mathbb{P}\left(\left| \frac{N_I}{n} - \alpha_I \right| \geqslant \frac{\varepsilon}{1+\varepsilon} \alpha_I \right) + \mathbb{P}\left(|b_I - \beta_I| \geqslant \frac{\varepsilon}{1+\varepsilon} \beta_I \right) \right].$$

For each of these quantities one can employ Bernstein's inequality, using the individual counting processes. All the quantities are sums of n independent and centred quantities. For the first probability, we have to deal with the sum of the $(1/n) \int_0^1 \mathbb{1}_I (\mathrm{d}N^i - Y^i s \, \mathrm{d}t)$, which are random variables with variance $(1/n^2)\alpha_I$. For the second probability, we have to deal with the sum of the $(1/n) \int_0^1 \mathbb{1}_I (\mathrm{d}N^i - \mathbb{E}(Y^i) s \, \mathrm{d}t)$, which are random variables with variance less than $(1/n^2)\alpha_I$. Each is bounded by $M = K + R$ divided by n. For the third probability, we have to deal with the sum of the $(1/n) \int_0^1 \mathbb{1}_I (Y^i - \mathbb{E}(Y_i)) \mathrm{d}t$, which are random variables bounded by $1/n$ with variance bounded by $(1/n^2)\beta_I$. Hence, we obtain

$$\mathbb{P}[\Omega(d)^c] \leqslant 2 \sum_{I \in \Gamma} \left[\mathrm{e}^{-n\beta_I h(\varepsilon, M, R_\Gamma)} + \mathrm{e}^{-n\alpha_I h'(\varepsilon, K, M)} + \mathrm{e}^{-n\beta_I h''(\varepsilon)} \right],$$

where h, h' and h'' are positive continuous functions. Finally, we obtain, for some positive continuous function f,

$$\mathbb{P}[\Omega(d)^c] \leqslant 6\frac{n}{\ln^2 n}\mathrm{e}^{-(\ln n)^2 f(\varepsilon,\rho,\mu,K,R)}$$

which is, for fixed $\eta > 0$, less than some $C''(d, \rho, \mu, K, \|s\|_\infty)/n^\eta$.

Let us now look at $\Omega(d)$. Let m be some fixed partition in $\mathcal{M}_A$. We know that by construction $\gamma_A(\tilde{s}) + \mathrm{pen}(\hat{m}) \leqslant \gamma_A(\hat{s}_m) + \mathrm{pen}(m) \leqslant \gamma_A(s'_m) + \mathrm{pen}(m)$. For any g in $\mathbb{L}^2([0, 1], \mathrm{d}t)$, let

$$\nu_n(g) = \int_0^1 g(t)\frac{\mathrm{d}N_t - Y_t s(t)\,\mathrm{d}t}{n}.$$

Using the fact that $\gamma_A(g) = \|s - g\|^2_{\mathrm{rand}} - \|s\|^2_{\mathrm{rand}} - 2\nu_n(g)$, we obtain:

$$\|s - \tilde{s}\|^2_{\mathrm{rand}} \leqslant \|s - s'_m\|^2_{\mathrm{rand}} + 2\nu_n(\tilde{s} - s'_m) - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m).$$

Now, for a partition m', we denote by $m \cup m'$ the partition built on the union of sets of points which are used to construct m and m'. We denote by $\chi_{\mathcal{T}}$ the square root of $\chi^2_{\mathcal{T}}$ defined in (2.4) for a set $\mathcal{T}$ of intervals.

Then, for all $m' \in \mathcal{M}_A$, one has

$$\sup_{f \in S'_m + S'_{m'}} \frac{\nu_n(f)}{\|f\|_{\mathrm{rand}}} \leqslant \sup_{f \in S_{m\cup m'}} \frac{\nu_n(f)}{\|f\|_{\mathrm{rand}}} = \chi_{m\cup m'}.$$

Hence,

$$\begin{aligned} 2\nu_n(\tilde{s} - s'_m) &\leqslant 2\|\tilde{s} - s'_m\|_{\mathrm{rand}}\chi_{m\cup\hat{m}} \\ &\leqslant \frac{2}{\varepsilon}\|s - s'_m\|^2_{\mathrm{rand}} + \frac{2}{2+\varepsilon}\|s - \tilde{s}\|^2_{\mathrm{rand}} + (1+\varepsilon)\chi_{m\cup\hat{m}}, \end{aligned}$$

using twice the fact that for all a, b, θ positive numbers, $2ab \leqslant \theta a^2 + b^2/\theta$. Then we obtain

$$\frac{\varepsilon}{2+\varepsilon}\|s - \tilde{s}\|^2_{\mathrm{rand}} \leqslant \left(1 + \frac{2}{\varepsilon}\right)\|s - s'_m\|^2_{\mathrm{rand}} + (1+\varepsilon)\chi^2_{m\cup\hat{m}} - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m). \tag{5.1}$$

In order to control $\chi^2_{m\cup\hat{m}}$, we have to control all the $\chi^2_{m\cup m'}$ for m' in $\mathcal{M}_A$. First we bound $\chi^2_{m\cup m'}$ by $Z^2_{m\cup m'}V_\Gamma$ since $S_{m\cup m'} \subset S_\Gamma$. We control all the $Z^2_{m\cup m'}$ using Proposition 2 with an upper bound on $R_{m\cup m'}$ that we denote by R_Γ (this is an upper bound by additivity). As we are on $\Omega(d)$, by additivity we are on $\Omega_{m\cup m'}(\varepsilon)$ defined in Proposition 2, and we can write that for all $x_{m'}$ positive, with probability larger than $1 - \exp(-x_{m'})$,

$$Z_{m\cup m'} \leqslant (1+\varepsilon)\left(\sqrt{\sum_{I\in m\cup m'}\frac{\alpha_I}{n\beta_I}} + \sqrt{\frac{2R_\Gamma x_{m'}}{n}}\right).$$

We choose $x_{m'} = L_{m'}|m'| + \xi$. With probability larger than $1 - \Sigma\mathrm{e}^{-\xi}$, we control all the $Z_{m\cup m'}$ and also $Z_{m\cup\hat{m}}$. After some easy computations, we obtain on $\Omega(d)$, with probability larger than $1 - \Sigma\mathrm{e}^{-\xi}$,

$$Z^2_{m\cup\hat{m}} \leqslant (1+\varepsilon)^3 R_\Gamma \frac{|\hat{m}|}{n}(1+\sqrt{2L_m})^2 + (1+\varepsilon)^3(1+\varepsilon^{-1})R_\Gamma \frac{|m|}{n} + (1+\varepsilon)^2(1+\varepsilon^{-1})^2 \frac{2R_\Gamma \xi}{n}.$$

We now remark that we have constructed $\Omega(d)$ in such a way that on $\Omega(d)$, $V_\Gamma \leqslant (1+\varepsilon)$ and $R_\Gamma \leqslant (1+2\varepsilon)\tilde{R}_\Gamma$. Taking ε such that $(1+\varepsilon)^5(1+2\varepsilon) = d$ fixes ε and finishes the proof. □

Proof of Corollary 1. Let us return to the proof of Theorem 1. One has $\|s-\tilde{s}\|^2_{\mathrm{det}} = \|s - s^{\mathrm{det}}_\Gamma\|^2_{\mathrm{det}} + \|s^{\mathrm{det}}_\Gamma - \tilde{s}\|^2_{\mathrm{det}}$. On $\Omega(d)$, the random norm and the deterministic norms are equivalent for functions in S_Γ. Thus one has $\|s^{\mathrm{det}}_\Gamma - \tilde{s}\|^2_{\mathrm{det}} \leqslant (1+\varepsilon)\|s^{\mathrm{det}}_\Gamma - \tilde{s}\|^2_{\mathrm{rand}}$. Then on $\Omega(d)$, we obtain

$$\|s-\tilde{s}\|^2_{\mathrm{det}} \leqslant \|s - s^{\mathrm{det}}_\Gamma\|^2_{\mathrm{det}} + 2(1+\varepsilon)\|s - s^{\mathrm{det}}_\Gamma\|^2_{\mathrm{rand}} + 2(1+\varepsilon)\|s-\tilde{s}\|^2_{\mathrm{rand}}.$$

We apply Theorem 1 to the last term and we integrate in ξ on $\Omega(d)$. We obtain after some computations

$$\mathbb{E}(\|s-\tilde{s}\|^2_{\mathrm{det}}\mathbb{1}_{\Omega(d)}) \leqslant (3+2\varepsilon)\|s - s^{\mathrm{det}}_\Gamma\|^2_{\mathrm{det}} + C(d)\mathbb{E}\left[\inf_{m\in\mathcal{M}_A}\left\{\|s-s'_m\|^2_{\mathrm{rand}} + \frac{|m|L_m}{n}R_\Gamma\right\}\right] + C'(d)R_\Gamma\frac{\Sigma}{n}.$$

Using (2.3) and exchanging the expectations and the infimum, there exist D and D' positive continuous functions such that

$$\mathbb{E}(\|s-\tilde{s}\|^2_{\mathrm{det}}\mathbb{1}_{\Omega(d)}) \leqslant D(d)\inf_{m\in\mathcal{M}_A}\left\{\mathbb{E}(\|s-s^{\mathrm{det}}_m\|^2_{\mathrm{det}}) + \frac{|m|L_m}{n}R_\Gamma\right\} + \frac{D'(d,\Sigma,R)}{n}.$$

On $\Omega(d)^c$, we use the fact that $\|s-\tilde{s}\|_\infty$ is bounded by $R + Kn^2$, and also the upper bound on $\mathbb{P}[\Omega(d)^c]$ given by Theorem 1 with $\eta = 3$, to obtain the result. □

Proof of Proposition 3. Let ψ be a positive function on $[0,1]$ symmetric about $1/2$, belonging to $\mathcal{H}_{1,\alpha,0}$ and such that $\psi(0) = 0$. Then for all positive integers D, $\psi_D(x) = LD^{-\alpha}\psi(Dx)$ belongs to $\mathcal{H}_{L,\alpha,0}$. Let us fix the regular partition Γ of $[0,1]$ with D intervals. Let m be a set of intervals of Γ and let any u_I be the left extremity of any I in Γ. Then

$$s_m = r + \sum_{I\in m}\psi_D(x-u_I)$$

belongs to $\mathcal{H}_{L,\alpha,r}$. Let $\mathcal{C}$ be a set such that for all m, m' in $\mathcal{C}$, $|m\triangle m'| \geqslant \theta D$ and $\log|\mathcal{C}| \geqslant \sigma D$, for θ and σ absolute constants. Such a set exists by application of Lemma 8 of Barron *et al.* (1999: 400). Let $\mathcal{A} = \{s_m,\ m\in\mathcal{C}\}$. Clearly, one has that

$$\mathcal{R}(\mathcal{H}_{L,\alpha,r}) \geqslant \frac{1}{4}\inf_{\hat{s}\in\mathcal{A}}\sup_{s\in\mathcal{A}}\mathbb{E}(\|s-\hat{s}\|^2_{\mathrm{det}}).$$

But for all $m \neq m'$ in $\mathcal{C}$,

$$\|s_m - s_{m'}\|_{\mathrm{det}}^2 = \int_0^1 \sum_{I \in m \triangle m'} \psi_D(t - a_I)^2 \mathbb{E}(Y_t^1)\,\mathrm{d}t \geqslant \mu |m \triangle m'| \int_0^1 \psi_D(t)^2\,\mathrm{d}t \geqslant \mu\theta L^2 D^{-2\alpha} P,$$

where $P = \int_0^1 \psi^2$ depends only on α. Hence,

$$\mathcal{R}(\mathcal{H}_{L,\alpha,r}) \geqslant \frac{1}{4}\mu\theta P L^2 D^{-2\alpha} \inf_{\hat{s} \in \mathcal{A}} \sup_{s \in \mathcal{A}} \mathbb{P}_s(\hat{s} \neq s) \geqslant \frac{1}{4}\mu\theta P L^2 D^{-2\alpha} \inf_{\hat{s} \in \mathcal{A}} \left(1 - \inf_{s \in \mathcal{A}} \mathbb{P}_s(\hat{s} = s)\right).$$

We next use a new version of Fano's lemma due to Birgé (2001): the infimum of the probabilities on the above right-hand side is bounded by an absolute constant α' if the Kullback–Leibler distance is bounded by $\alpha' \log|\mathcal{C}|$. By the combinatorial lemma previously used, it is sufficient to bound the Kullback–Leibler distance by $\alpha' \sigma D$. But by taking the expectation of the classical formula for log-likelihood for counting processes, one has (Andersen *et al.* 1993) that for all $m' \neq m \in \mathcal{C}$,

$$K(\mathbb{P}_{s_{m'}}, \mathbb{P}_{s_m}) = \int s_{m'} \phi\left(\ln \frac{s_m}{s_{m'}}\right) \mathbb{E}_{s_m}(Y_t)\,\mathrm{d}t \leqslant \int \frac{(s_m - s_{m'})^2}{s_m}(x) \mathbb{E}_{s_m}(Y_t)\,\mathrm{d}t \leqslant \frac{1}{r} nMPL^2 D^{-2\alpha}.$$

Finally, one fixes D such that $nMPL^2 D^{-2\alpha} \simeq r\alpha' \sigma D$. This leads to the result. □

Proof of Theorem 2. Let ε be a positive continuous function of d that we will choose later. On Ω, we can perform the same computations as in the histogram case to obtain

$$\|s - \tilde{s}\|_{\mathrm{rand}}^2 \leqslant \|s - s'_m\|_{\mathrm{rand}}^2 + 2\nu_A(\tilde{s} - s'_m) - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m),$$

where for all g in $\mathbb{L}^2([0, 1], \mathrm{d}t)$, $\nu_A(g) = \int_0^1 g(t)\,\mathrm{d}M_t/A$. On Ω, one can see that

$$\chi(m \cup \hat{m})_1 = \sup\{\nu_A(f) : f \in S_{m \cup \hat{m}},\ \|f\|_{\mathrm{rand}} = 1\}.$$

Therefore, using the same method as for the histograms, we obtain

$$\frac{\varepsilon}{2 + \varepsilon}\|s - \tilde{s}\|_{\mathrm{rand}}^2 \leqslant \left(1 + \frac{2}{\varepsilon}\right)\|s - s'_m\|_{\mathrm{rand}}^2 + (1 + \varepsilon)\chi(m \cup \hat{m})_1^2 - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m). \quad (5.2)$$

Moreover, one has that $\chi(m \cup \hat{m})_1^2 \leqslant \chi(m)_1^2 + \chi(\hat{m})_1^2$. But for all m' in $\mathcal{M}_A$, we can apply the exponential formula derived in Reynaud-Bouret (2006: Proposition 6): for all $x_{m'}$ positive with probability larger than $1 - 2\exp(-x_{m'})$, $\chi(m')_1 \leqslant \sqrt{C(m')_1} + 3\sqrt{2\upsilon_{m'} x_{m'}} + b_{m'} x_{m'}$, where $\upsilon_{m'}$ is a deterministic bound on $C(m')_1$ and $b_{m'}^2$ is a deterministic bound on $\sum_{\lambda \in m'} \varphi_\lambda^2/(Y' A^2)$. Under the assumptions of Theorem 2, we obtain that for all $x_{m'} > 0$, $\chi(m')_1 \leqslant \sqrt{(|m'|/A)}[\sqrt{R} + 3\sqrt{2Rx_{m'}} + \sqrt{(\Phi/c)}x_{m'}]$, with probability larger than $1 - 2\mathrm{e}^{-x_{m'}}$. Let $\xi > 0$ and let $x_{m'} = L_{m'} + \xi/|m'|$. Then we can bound $\chi(m')_1^2$ by

$$(1 + \varepsilon)\frac{|m'|}{A}\left[\sqrt{R}(1 + 3\sqrt{2L_{m'}}) + \sqrt{\frac{\Phi}{c}}L_{m'}\right] + (1 + \varepsilon^{-1})(1 + \varepsilon)\frac{18\xi}{A} + (1 + \varepsilon^{-1})^2 \frac{\xi^2}{A}.$$

Taking $d = (1 + \varepsilon)^2$, which fixes ε, it follows that, with probability larger than $1 - 2\sum_{m' \in \mathcal{M}_A} \exp(-L_{m'} - (\xi/|m'|))$,

$$\frac{\varepsilon}{2+\varepsilon}\|s-\tilde{s}\|_n^2 \leqslant \left(1+\frac{2}{\varepsilon}\right)\|s-s'_m\|_n^2 + 2\mathrm{pen}(m) + 2\left[(1+\varepsilon^{-1})(1+\varepsilon)\frac{18\xi}{A} + (1+\varepsilon^{-1})^2\frac{\xi^2}{A}\right].$$

It remains to integrate in ξ. We finally obtain the result by a change of variables and the Beppo Levi theorem. □

Acknowledgement

I would like to thank L. Birgé for the idea of the progressive histogram strategy, and C. Houdré, F. Comte and the anonymous referees for their helpful suggestions.

References

Andersen, P., Borgan, Ø., Gill, R. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Antoniadis, A. (1989) A penalty method for nonparametric estimation of the intensity function of a counting process. *Ann. Inst. Statist. Math.*, **41**, 781–807.

Antoniadis, A., Grégoire, G. and Nason, G. (1999) Density and hazard rate estimation for right-censored data by using wavelet methods. *J. R. Statist. Soc. Ser. B*, **61**, 63–84.

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

Birgé, L. (2001) A new look at an old result: Fano's lemma. Preprint PMA-632, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris VI/VII.

Birgé, L. and Massart, P. (1997) From model selection to adaptive estimation. In D. Pollard, E. Torgersen and G.L. Yang (eds), *Festschrift for Lucien Le Cam*, pp. 55–87. New York: Springer-Verlag.

Birgé, L. and Massart, P. (2001) Gaussian model selection. *J. Eur. Math. Soc.*, **3**, 203–268.

Cavalier, L. and Koo, J.-Y. (2002) Poisson intensity estimation for tomographic data using a wavelet shrinkage approach. *IEEE Trans. Inform. Theory*, **48**, 2794–2802.

Döhler, S. and Rüschendorf, L. (2002) Adaptive estimation of hazard functions. *Probab. Math. Statist.*, **22**, 355–379.

Grégoire, G. (1993) Least squares cross-validation for counting process intensities. *Scand. J. Statist.*, **20**, 343–360.

Lebarbier, E. (2002) Quelques approches pour la détection de ruptures à horizon fini. Doctoral thesis, Université Paris XI, UFR Orsay.

Mallows, C. (1973) Some comments on C_p. *Technometrics*, **15**, 661–675.

Massart, P. (1990) The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Ann. Probab.*, **18**, 1269–1283.

Massart, P. (2000) About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, **28**, 863–884.

Massart, P. (2005) *Concentration Inequalities and Model Selection: École d'Été de Probabilités de Saint-Flour XXXIII–2003* (ed. J. Picard), Lecture Notes in Math. To appear. http://www.math.u-psud.fr/∼massart/ (accessed 9 February 2006).

Ramlau-Hansen, H. (1983) Smoothing counting process intensity by means of kernel functions. *Ann. Statist.*, **11**, 453–466.

Reynaud-Bouret, P. (2003) Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields*, **126**, 103–153.

Reynaud-Bouret, P. (2006) Compensator and exponential inequalities for some suprema of counting processes: *Statist. Probab. Lett*., **76**, 1514–1521.

Rio, E. (2002) Une inégalité de Bennett pour les maxima de processus empiriques. *Ann. Inst. H. Poincaré Probab. Statist.*, **38**, 1053–1057.

van de Geer, S. (1995) Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.*, **23**, 1779–1801.

Received December 2002 and revised November 2005.