Hazard regression with non compactly supported bases

Elodie Brunel, Fabienne Comte

HAZARD REGRESSION WITH NON COMPACTLY SUPPORTED BASES

ELODIE BRUNEL[(1)] & FABIENNE COMTE[(2)]

ABSTRACT. In this paper, we consider the problem of nonparametric hazard rate estimation in presence of right-censored observations. We provide a generalized risk bound for a regression type nonparametric estimator of the hazard function of interest. Under adequate integrability conditions, our bound is a generalization to non necessarily compactly supported bases, of strategies which were specific to compact support of estimation. We show that it encompasses those previous compact-support results. We discuss the model selection method which comes out from the new terms of the risk bounds, and compare the performance of the new estimator to previous ones, when using a non compact Laguerre basis. A real data example is also presented.

Keywords. Hazard rate. Laguerre basis. Least squares regression. Nonparametric estimation. Projection estimator.

1. INTRODUCTION

Consider the model where the observations are

$$Z_i = X_i \wedge C_i, \quad \delta_i = \mathbf{1}_{\{X_i \leq C_i\}}, \tag{1}$$

where the sequences $(X_i)_i$ and $(C_i)_i$ are two independent sequences of i.i.d. nonnegative random variables. The function of interest is $\lambda = f/S$ where f is the density of X_1 and S its survival function, called hazard rate. The Z_i's are called right-censored observations, and the δ_i's are non-censoring indicators. This type of model is most commonly used in reliability or survival analysis: more precisely, we consider here lifetimes (or failure times) of some individuals in presence of right-censoring. This occurs for instance when some of the individuals are not observed until the end (death, remission, recovery) of the study; only a lower bound on their lifetime is observed.

There are different nonparametric methods used in the literature to estimate hazard rate, most of them rely on quotient strategies. Indeed, let S_C and S_Z denote the survival function of the C_i's and Z_i's: $S_C(x) = \mathbb{P}(C_1 > x)$, $S_Z(x) = \mathbb{P}(Z_1 > x)$. Then, the hazard rate can be written as $\lambda = fS_C/S_Z$, where the function fS_C is often called subdensity. This function can be estimated using censored observations, and S_Z has an obvious empirical counterpart, namely the empirical survival function of all Z_i's. This idea is used in Blum and Susarla (1980), Mielniczuk (1986), Diehl and Stute (1988), Lo et al.(1989), Uzunoḡullari and Wang (1992), who propose kernel estimators of the numerator. Note that bandwidth selection is an important issue in this context and practical

(1): IMAG, Univ Montpellier, CNRS, Montpellier, France, email: elodie.brunel-piccinini@umontpellier.fr
(2): Université de Paris, CNRS, MAP5 UMR 8145, F-75006 Paris, France, email: fabienne.comte@parisdescartes.fr .

methods are suggested. Antoniadis et al. (1999) consider both subdensity and hazard estimators via wavelet methods, and the optimal wavelet resolution depends on the unknown function. Brunel and Comte (2005) build projection estimators based on these ideas, and propose model selection methods to determine in a data driven way, the relevant dimension for the projection space; they prove that their strategy ensures an automatic squared-bias/variance tradeoff.

Nonparametric estimators of the hazard rate have also been constructed by convolving with a kernel some cumulative hazard estimator such as the Nelson-Aalen or the Kaplan-Meier estimators, see Tanner and Wong (1983), Ramlau-Hansen (1983) and Yandell (1983). Müller and Wang (1994) propose estimators with varying kernels and data-adaptive bandwidths and more recently Bouezmarni *et al.* (2011) study a Gamma kernel estimator. Later on, Wu and Wells (2003), proposed a wavelet-type estimator also based on the transform of a Nelson-Aalen cumulative hazard estimator. Kooperberg et al. (1995) study the $\mathbb{L}^2$ convergence rate of a hazard rate estimator in a context of tensor product splines. Dölher and Rüschendorf (2002) introduce an adaptive sieved maximum likelihood method. Reynaud-Bouret (2002) obtains adaptive results and minimax rates for penalized projection estimators of the Aalen multiplicative intensity process. Lastly, Brunel and Comte (2005) consider penalized contrast estimator using the Kaplan-Meier cumulative hazard estimator and a large variety of models.

In this work, we consider a direct regression strategy already described in Plancade (2011), or which can be obtained as a particular case of Comte *et al.* (2011). The generalization comes from the fact that we no longer assume that the estimation support is a compact set: this assumption was crucial in previous works, and we no longer require it. The ideas are inspired by those of Cohen *et al.* (2013, 2019) and Comte and Genon-Catalot (2019) for standard regression, but hazard rate regression has specificities, both in theory (e.g. the stability constraint given by (9) in section 2.2 is different from standard regression) and in practice. We have in mind that survival analysis is a context where the Laguerre basis (see Section 2.4), which is $\mathbb{R}^+$-supported, is specifically well suited for estimation: the resulting estimators are general combinations of Gamma-type distributions. As many parametric models involve Gamma densities, projection estimators in the Laguerre basis are a relevant generalization of these densities and allow a lot of flexibility.

The plan of the paper is as follows. The estimator and the assumptions are given in section 2, bounds for empirical and $\mathbb{L}^2$-risk are then stated, and examples of compact and non compact settings are given. Then, Section 3 describes a model selection procedure; the new risk bound suggests an easy-to-compute penalty function, which avoids to estimate unconvenient quantities, like upper or lower bounds of unknown functions. The method is applied through simulation experiments to previous examples of the literature, for comparison. A real data example is considered in section 4, and illustrates the relevance and the flexibility of our procedure. A short concluding section 5 ends the presentation. Proofs are postponed in Section 6.

2. Hazard rate estimation in presence of right censoring

Let us start with preliminary notations. For a function u, we denote by $\|u\|^2 := \int u^2(x)dx$, by $\|u\|^2_{S_Z} := \int u^2(x)S_Z(x)dx$ and for two square integrable functions u_1 and u_2, by $\langle u_1, u_2\rangle$ and $\langle u_1, u_2\rangle_{S_Z}$ the associated scalar products. The corresponding spaces for

square-integrable A-supported functions are denoted by $\mathbb{L}^2(A, dx)$ and $\mathbb{L}^2(A, S_Z(x)dx)$. For a m-dimensional vector $\vec{v}$ with coordinates $(v_1, \ldots, v_m)$, we denote by $\|\vec{v}\|_{2,m}^2 := \sum_{j=1}^m v_j^2$ its euclidean norm.

For a matrix M, we define the operator norm $\|M\|_{\text{op}}$ as the square-root of the largest (nonnegative) eigenvalue of $M\,{}^tM$, where ${}^tM$ is the transpose of M. When M is symmetric, it coincides with its largest eigenvalue in absolute value. The so-called Frobenius norm is also defined by $\|M\|_{\text{F}}^2 = \sum_{i,j}[M]_{i,j}^2 = \text{Tr}(M\,{}^tM)$ where Tr denotes the trace.

2.1. **Definition of the estimator.** The following contrast has been considered in Comte *et al.* (2011) and in Plancade (2011). Let $s, t : A \mapsto \mathbb{R}$ be two square integrable functions from $A \subseteq \mathbb{R}^+$ into $\mathbb{R}$ and

$$\gamma_n(t) = \|t\|_n^2 - \frac{2}{n}\sum_{i=1}^n \delta_i t(Z_i), \tag{2}$$

where the empirical scalar product and its associated empirical norm are defined by

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^n \int t^2(x)\mathbf{1}_{\{Z_i>x\}}dx\,,\ \langle s,t\rangle_n = \frac{1}{n}\sum_{i=1}^n \int s(x)t(x)\mathbf{1}_{\{Z_i>x\}}dx. \tag{3}$$

Let us explain why this contrast is related to our hazard rate estimation problem. First note that $\mathbb{E}(\mathbf{1}_{\{Z_1>x\}}) = \mathbb{P}(Z_1 > x) = S_Z(x) = S_C(x)S(x)$ with S_Z and S_C denoting the survival functions of Z_1 and C_1. Secondly, we have

$$\mathbb{E}(\delta_1 t(Z_1)) = \mathbb{E}(\mathbf{1}_{\{X_1\le C_1\}}t(X_1)) = \mathbb{E}(S_C(X_1)f(X_1)).$$

Therefore, we find that

$$\begin{aligned}\mathbb{E}(\gamma_n(t)) &= \int t^2(x)S_Z(x)dx - 2\int t(x)S_C(x)f(x)dx\\ &= \int (t(x)-\lambda(x))^2 S_Z(x)dx - \int \lambda^2(x)S_Z(x)dx.\end{aligned}$$

Thus, minimizing γ_n for large n, should provide a function t minimizing $\int(t(x)-\lambda(x))^2S_Z(x)dx$, that is a weighted $\mathbb{L}^2$-distance to λ. Therefore, we should estimate the $\mathbb{L}^2$ orthogonal projection of λ w.r.t the S_Z-weighted scalar product on a subspace S_m of functions over which the minimization is performed.

Let $A \subseteq \mathbb{R}^+$ and let $(\varphi_j, j = 0, \cdots, m-1)$ be an orthonormal system of functions supported on A belonging to $\mathbb{L}^2(A, dx)$, i.e. such that $\langle\varphi_j, \varphi_k\rangle = \delta_{j,k}$, $0 \le j, k \le m-1$. We define S_m as the space linearly spanned by the functions φ_j: $S_m = \text{span}(\varphi_0, \cdots, \varphi_{m-1})$. The space S_m has thus finite dimension m.
We define the matrix

$$\widehat{\Psi}_{m,Z} = (\langle\varphi_j, \varphi_k\rangle_n)_{0\le j,k\le m-1} = \left(\int \varphi_j(x)\varphi_k(x)\widehat{S}_{Z,n}(x)dx\right)_{0\le j,k\le m-1}$$

where

$$\widehat{S}_{Z,n}(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Z_i>x\}}$$

and the matrix

$$\Psi_{m,Z} := \left(\int \varphi_j(x)\varphi_k(x) S_Z(x) dx \right)_{0\leq j,k\leq m-1}.$$

Note that the matrix $\Psi_{m,Z} = \mathbb{E}\left(\widehat{\Psi}_{m,Z}\right)$ is the matrix of the scalar products $\langle \varphi_j, \varphi_k \rangle_{S_Z}$ (with associated $\mathbb{L}^2$-weighted norm $\|.\|_{S_Z}$) and $\widehat{\Psi}_{m,Z}$ is its empirical counterpart with $\langle \varphi_j, \varphi_k \rangle_n$ for $1 \leq j, k \leq m$. Then we define

$$\widehat{\lambda}_m = \arg\min_{t \in S_m} \gamma_n(t).$$

Setting the gradient of $\gamma_n(t)$ to zero and standard algebra calculations give, provided that $\widehat{\Psi}_{m,Z}$ is a.s. invertible,

$$\widehat{\lambda}_m = \sum_{j=0}^{m-1} \hat{a}_j \varphi_j \text{ with } \vec{\hat{a}}^{(m)} = \begin{pmatrix} \hat{a}_0 \\ \vdots \\ \hat{a}_{m-1} \end{pmatrix} = \frac{1}{n} \widehat{\Psi}_{m,Z}^{-1} \, {}^t\widehat{\Phi}_m \vec{\delta}, \tag{4}$$

where $\widehat{\Phi}_m = (\varphi_j(Z_i))_{1\leq i\leq n, 0\leq j\leq m-1}$ and $\vec{\delta} = {}^t(\delta_1, \ldots, \delta_n)$.

Remark 2.1. We can compare with the regression model: $Y_i = b(Z_i) + \varepsilon_i$ where (Z_i, Y_i) are observed, ε_i is a centered unobserved noise and the $(X_i)_i$ and the $(\varepsilon_i)_i$ are i.i.d. independent sequences. To estimate the regression function b, estimators of the m first coefficients of b in the basis are $(1/n)\widehat{\Psi}_m^{-1} \, {}^t\widehat{\Phi}_m \vec{Y}$ where $\vec{Y} = {}^t(Y_1, \ldots, Y_n)$ and $\widehat{\Psi}_m = (1/n) \, {}^t\widehat{\Phi}_m \widehat{\Phi}_m$. Here the fact that the same matrix $\widehat{\Phi}_m$ appears in all terms is very important and convenient. This is what makes an important difference with hazard rate estimation. Here, $\widehat{\Psi}_{m,Z}$ is not directly related to $\widehat{\Phi}_m$.

Formula (4) provides an easy way to compute our projection estimator $\widehat{\lambda}_m$ provided that $\widehat{\Psi}_{m,Z}$ is *a.s.* invertible. So, to guarantee it is always satisfied, we define the trimmed estimator by :

$$\tilde{\lambda}_m = \begin{cases} \widehat{\lambda}_m & \text{if } \|\widehat{\Psi}_{m,Z}^{-1}\|_{\text{op}} \leq \mathfrak{c} \dfrac{n}{\log(n)} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathfrak{c}$ is a constant defined further (see Proposition 2.1).

Convention. We set $\|\widehat{\Psi}_{m,Z}^{-1}\|_{\text{op}} = +\infty$ if $\widehat{\Psi}_{m,Z}$ is not invertible.

2.2. **Bounds for the empirical risk and the integrated risk of one estimator.** We consider a general context where the estimation support A is such that $A \subseteq \mathbb{R}^+$ and

$$\int_A \lambda^2(x) S_Z(x) dx < +\infty. \tag{6}$$

Condition (6) is fulfilled for most classical models. Indeed as $S_Z \leq S$, the condition holds if the distribution of X is such that $\int_A \lambda^2 S < +\infty$. We shall denote $\lambda_A = \lambda \mathbf{1}_A$.

Examples of models satisfying $\int_{\mathbb{R}^+} \lambda^2 S = \int_{\mathbb{R}^+} f^2/S < +\infty$.

(1) Exponential density: f exponential $\mathcal{E}(\theta)$, $\theta > 0$, $S(x) = \exp(-\theta x)\mathbf{1}_{\{x\geq 0\}}$, $\lambda(x) = \theta\,\mathbf{1}_{\{x\geq 0\}}$,
(2) Weibull model, $\lambda(x) = \alpha\theta^\alpha x^{\alpha-1}\mathbf{1}_{\{x\geq 0\}}$, $S(x) = \exp(-(\theta x)^\alpha)\mathbf{1}_{\{x\geq 0\}}$, $\alpha > 1/2$, $\theta > 0$
(3) Gamma model, $f(x) = \theta^\nu x^{\nu-1}e^{-\theta x}/\Gamma(\nu)\mathbf{1}_{\{x\geq 0\}}$, $\nu > 1/2$, $\theta > 0$,
(4) Gompertz-Makeham, $\lambda(x) = \gamma_0 + \gamma_1 e^{\gamma_2 x}$, $S(x) = e^{-\gamma_0 x-(\gamma_2/\gamma_1)(e^{\gamma_2 x}-1)}\mathbf{1}_{\{x\geq 0\}}$, for real numbers $\gamma_0, \gamma_1, \gamma_2 > 0$,
(5) Log-logistic, $\lambda(x) = \theta\nu x^{\nu-1}/(1+\theta x^\nu)\mathbf{1}_{\{x\geq 0\}}$, $\nu > 1/2$, $\theta > 0$, $S(x) = 1/(1+\theta x^\nu)\mathbf{1}_{\{x\geq 0\}}$,
(6) Log-normal $\lambda(x) = (1/(x\sigma))\phi\left((\ln x - \mu)/\sigma\right) / \left[1 - \Phi\left((\ln x - \mu)/\sigma\right)\right]\mathbf{1}_{\{x\geq 0\}}$, where $\phi(x)$ and $\Phi(x)$ are respectively the density and the cumulative distribution function of a standard gaussian, $\mu \in \mathbb{R}$, $\sigma > 0$.

In addition, we assume that the basis $(\varphi_j)_j$ is such that

$$L(m) := \sup_{x\in A} \sum_{j=0}^{m-1} \varphi_j^2(x) < +\infty \tag{7}$$

For most bases, we have $L(m) \leq c_\varphi^2 m$ where c_φ is a constant depending on the bases (see examples in Sections 2.3 and 2.4 below).

Then, we can prove risk bounds with respect to the empirical risk first, and to the integrated S_Z-weighted risk in a second time. In the regression setting, the risk bound obtained for the empirical risk is rather straightforward and relies on projection arguments; it is interesting to see that the problem here also involves linear algebra but still, is more involved.

Proposition 2.1. *Assume that $\Psi_{m,Z}$ is invertible, that condition (7) holds and that*

$$\int_A \lambda^2(z)\sqrt{S_Z(z)}dz < +\infty. \tag{8}$$

Then, for any m such that $L(m) \leq n$ and

$$\|\Psi_{m,Z}^{-1}\|_{\mathrm{op}} \leq \frac{\mathfrak{c}}{2}\frac{n}{\log(n)}, \quad \mathfrak{c} = \frac{3\log(3/2)-1}{10}, \tag{9}$$

we have

$$\mathbb{E}\left[\|\tilde{\lambda}_m - \lambda_A\|_n^2\right] \leq \inf_{t\in S_m} \|t - \lambda_A\|_{S_Z}^2 + 2\frac{\mathrm{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z})}{n} + \frac{\mathcal{C}_1}{n}. \tag{10}$$

where $\mathcal{C}_1$ is a positive constant and

$$\Psi_{m,\lambda S_Z} = \left(\int \varphi_j(x)\varphi_k(x)\lambda(x)S_Z(x)dx\right)_{0\leq j,k\leq m-1}. \tag{11}$$

Note that, as $S_Z(x) \leq 1$, condition (8) implies that $\int_A \lambda^2(z)S_Z(z)dz < +\infty$ and thus condition (6) holds. It is also fulfilled in the examples listed above.

Condition (9) corresponds to what Cohen *et al.* (2013) call *a stability condition* in the classical regression setting described in Remark 2.1. It is expressed in function of similar matrices in Comte and Genon-Catalot (2018), but is is noteworthy that the standard

regression entails a different constraint, namely $m\|\Psi_m^{-1}\|_{\text{op}} \leq (\mathfrak{c}/2)(n/\log(n))$ where $\Psi_m = \mathbb{E}(\widehat{\Psi}_m)$ and $\widehat{\Psi}_m$ is defined in Remark 2.1.

Following ideas developped in Cohen *et al.* (2013), we can obtain a risk bound on the integrated weighted risk, with coefficient in front of the squared bias nearly 1 for large n.

Proposition 2.2. *Assume that $\Psi_{m,Z}$ is invertible, that conditions (6) and (7) hold. Then for any m such that $L(m) \leq n$ and (9) holds, we have*

$$\mathbb{E}\left[\|\tilde{\lambda}_m - \lambda_A\|_{S_Z}^2\right] \leq \left(1 + 8\frac{\mathfrak{c}}{\log(n)}\right) \inf_{t \in S_m} \|t - \lambda_A\|_{S_Z}^2 + 8\frac{\text{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z})}{n} + \frac{\mathcal{C}_2}{n}, \tag{12}$$

where $\mathfrak{c}$ is defined in (9) and $\mathcal{C}_2$ is a positive constant.

Equations (10) and (12) provide empirical and integrated risk bounds involving three terms: a negligible one of order $1/n$, a variance term of order $\text{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z})/n$ and a bias term $\inf_{t\in S_m}\|t-\lambda_A\|_{S_Z}^2$. It is noteworthy that the coefficient in front of the bias term is exactly 1 in the first case and of order 1 for large n in the second one. Clearly, this term is decreasing if the space S_m grows when m increases (with nested collection, $m \leq m' \implies S_m \subset S_{m'}$). On the other hand, the true novelty stands in the variance bound $\text{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z})/n$ obtained in Propositions 2.1 and 2.2, which is more general than in previous works. The result holds without constraint on the support. Moreover, even it is not obvious at first sight, we can prove that $m \mapsto \text{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z})$ is increasing.

Lemma 2.1. *Let the collection S_m be nested, then $m \mapsto \text{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z})$ is increasing.*

Therefore, both bounds in (10) and (12) lead to the same conclusion that a compromise has to be found for the choice of m, making a tradeoff between bias and variance. In the next section, we illustrate that, in the standard case of compact support A, we can obtain a more explicit upper bound on the variance, and recover previous results.

2.3. **Specific cases of compact A.** Let us assume here that A is compact and show in what extent our new results encompass previous ones.

We can consider a trigonometric basis on $A = [0, \mathtt{a}]$: $\varphi_0(x) = (1/\sqrt{\mathtt{a}})\mathbf{1}_{[0,\mathtt{a}]}(x)$, $\varphi_{2j-1}(x) = \sqrt{2/\mathtt{a}}\cos(2\pi jx/\mathtt{a})\mathbf{1}_{[0,\mathtt{a}]}(x)$, $\varphi_{2j}(x) = \sqrt{2/\mathtt{a}}\sin(2\pi jx/\mathtt{a})\mathbf{1}_{[0,\mathtt{a}]}(x)$, $j \geq m$. Clearly, in that case, $L(m) \leq (2/\mathtt{a})m$ and $L(m) = m/\mathtt{a}$ if m is even.

We may also choose the histogram basis on $A = [0, \mathtt{a}]$, we set $\varphi_j(x) = \sqrt{m\mathtt{a}}\mathbb{1}_{[j\mathtt{a}/m,(j+1)\mathtt{a}/m[}$ for $j = 0, \ldots, m-1$. We can consider more general piecewise polynomials with given degree r, by rescaling $Q_0, \ldots, Q_r$ the Legendre basis on each sub-interval $[j\mathtt{a}/m, (j+1)\mathtt{a}/m[$, $j = 0, \ldots, m-1$. In that case, we have $L(m) = \mathtt{a}\,m$ for histograms and $L(m) \leq (r+1)\mathtt{a}\,m$ for piecewise polynomials (see Comte (2017, chap.2)).
Consequently, condition (7) is satisfied for these bases, and $L(m) \leq c_\varphi^2 m$, where c_φ^2 is a known constant depending on the basis and not on m.

For these bases with specifically compact supports, we can assume that:

$$\forall x \in A,\ S_Z(x) \geq S_0 > 0 \quad \text{and} \quad \lambda(x) \leq \|\lambda_A\|_\infty < +\infty. \tag{13}$$

Note that S_Z is lower bounded on A if both S and S_C are; moreover, given the first part, the second part of (13) can be obtained if f is bounded on A, as $\lambda(x) \leq f(x)/S_0$. However,

if condition $\|\lambda\|_\infty < +\infty$ generally holds for compact A, it is not the case for $A = \mathbb{R}^+$, see the Weibull (2) or the Gompertz-Makeham (4) examples.

Lemma 2.2. *Let A be a compact set and consider a basis such that $L(m) \leq c_\varphi^2 m$. Under (13), condition (8) is fulfilled. Moreover,*

(i) $\|\Psi_{m,Z}^{-1}\|_{\text{op}} \leq 1/S_0$,

(ii) $0 \leq \text{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z}) \leq m\|\lambda_A\|_\infty$

(iii) $0 \leq \text{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z}) \leq c_\varphi^2 m/S_0$.

Bound (i) shows that condition (9) is automatically fulfilled for n large enough: this is why this condition does not appear in a compact setting. Moreover, with (ii) and (iii) we recover the variance bounds proposed in Plancade (2011), see equation (11) and Theorem 1 therein, see also Comte *et al.* (2011), Theorem 1 for bound (ii), in presence of covariates.

2.4. **Example of non compact A.** The Laguerre basis on $A = \mathbb{R}^+$ is defined by:

$$P_j(x) = \sum_{k=0}^{j} (-1)^k \binom{j}{k} \frac{x^k}{k!}, \qquad \varphi_j(x) = \sqrt{2} P_j(2x) e^{-x} \mathbb{1}_{x\geq 0}, \quad j \geq 0. \tag{14}$$

The P_j are called Laguerre polynomials (P_j) and the φ_j Laguerre functions. The collection $(\varphi_j)_{j\geq 0}$ is a complete orthonormal system on $\mathbb{L}^2(\mathbb{R}^+)$, such that (see Abramowitz and Stegun (1964)) $\forall j \geq 0, \forall x \in \mathbb{R}^+, |\varphi_j(x)| \leq \sqrt{2}$. Therefore $L(m) \leq 2m$ and condition (7) is satisfied.

However, condition (13) does not hold anymore, and Lemma 2.2 has to be stated neither:

Lemma 2.3. *If $\mu(\mathbb{R}^+ \cap \text{Supp}(S_Z)) > 0$ where μ is the Lebesgue measure and $\text{Supp}(S_Z) = \{x \in \mathbb{R}^+, S(x) > 0\}$ is the support of S_Z, then $\Psi_{m,Z}$ is invertible. Moreover, there exists $c^\star > 0$ such that, for m large enough, $\|\Psi_{m,Z}^{-1}\|_{\text{op}} \geq c^\star \sqrt{m}$.*

Lemma 2.3 shows clearly that in the context of the Laguerre basis, bound (i) of Lemma 2.2 is not true. So, the order of the variance is not obvious.

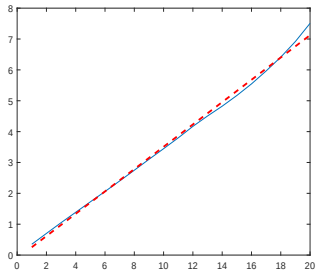
Note that if $X \sim \mathcal{E}(\beta)$ i.e. $f(x) = \beta e^{-\beta x} \mathbf{1}_{\mathbb{R}^+}(x)$ and $S(x) = e^{-\beta x} \mathbf{1}_{\mathbb{R}^+}(x)$, then $\lambda(x) = \beta$. Therefore $\Psi_{m,Z} = (1/\beta)\Psi_{m,\lambda S_Z}$ and

$$\text{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z}) = \beta \text{Tr}(\text{Id}_m) = \beta\, m.$$

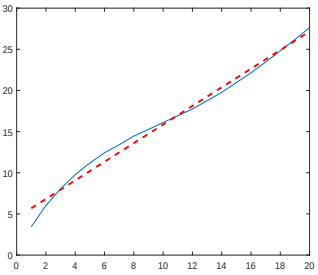
Thus, the variance term can remain of order m/n in the non-compact setting as well. Numerical experiments support the conjecture that that the quantity $\text{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z})$ is generally of order $\mathtt{c}\, m$ with $\mathtt{c}$ a constant that can be evaluated (see Figure 1). However, in the same examples, the term $\|\Psi_{m,Z}^{-1}\|_{\text{op}}$ can grow very fast, so that bounding $\text{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z})$ by $\|\Psi_{m,Z}^{-1/2}\|_{\text{op}}^2 \|\Psi_{m,\lambda S_Z}^{1/2}\|_F^2 = \|\Psi_{m,Z}^{-1}\|_{\text{op}} \text{Tr}(\Psi_{m,\lambda S_Z})$ is not a good strategy in the non-compact setting.

3. Model selection and simulations

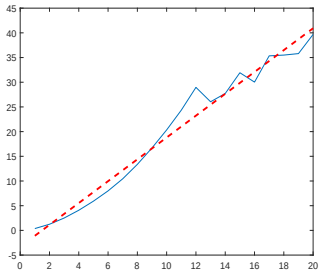
3.1. **Procedure.** In this section, we propose a practical procedure for model selection. A theoretical study has been made for a similar proposal in the case of nonparametric



$f(x) = \beta e^{-\beta x}\mathbf{1}_{x\geq 0}, \beta = 1/3$
$\hat{a} = -0.11, \hat{b} = 0.36$



$f(x) = 3/(1+x)^4\mathbf{1}_{x\geq 0}$
$\hat{a} = 4.56, \hat{b} = 1.31$



$f(x) = xe^{-x^2/2}\mathbf{1}_{x\geq 0}$
$\hat{a} = -3.31, \hat{b} = 2.21$

FIGURE 1. Plots of $m \mapsto \mathrm{Tr}(\widehat{\Psi}_{m,Z}^{-1}\widehat{\Psi}_{m,\lambda S_Z})$ for $m = 1, \dots, 20$, from $n = 10000$ observations with no censoring, in blue. In bold dotted red, the best approximating line $y = \hat{a} + \hat{b}x$, with value of the coefficients in each case.

regression function estimation, see Comte and Genon-Catalot (2020), and we refer the reader to this paper for technicalities, which are numerous. For now, let us describe it.

An important preliminary remark is that, as $\lambda S_Z = fS_C$, the matrix $\Psi_{m,\lambda S_Z}$ can easily be estimated by

$$\widehat{\Psi}_{m,\lambda S_Z} = \left(\frac{1}{n}\sum_{i=1}^{n}\delta_i\varphi_j(Z_i)\varphi_k(Z_i)\right)_{0\leq j,k\leq m-1}. \tag{15}$$

Now, let $\mathcal{M}_n$ be the theoretical collection of models defined by

$$\mathcal{M}_n = \{m \in \{1,\dots,n\}, \|\Psi_{m,Z}^{-1}\|_{\mathrm{op}} \leq \frac{\mathfrak{c}}{2}\frac{n}{\log(n)}\}$$

and its empirical version

$$\widehat{\mathcal{M}}_n = \{m \in \{1,\dots,n\}, \|\widehat{\Psi}_{m,Z}^{-1}\|_{\mathrm{op}} \leq \mathfrak{c}\frac{n}{\log(n)}\}.$$

Then we select

$$\hat{m} = \arg\min_{m\in\widehat{\mathcal{M}}_n}(-\|\hat{\lambda}_m\|_n^2 + \widehat{\mathrm{pen}}(m)), \quad \widehat{\mathrm{pen}}(m) = \kappa\frac{\mathrm{Tr}(\widehat{\Psi}_{m,Z}^{-1}\widehat{\Psi}_{m,\lambda S_Z})}{n}.$$

Indeed, it is easy to check that $\gamma_n(\hat{\lambda}_m) = -\|\hat{\lambda}_m\|_n^2$ and this term is taken as an estimate of the squared bias term. The penalty is the empirical version of the variance order. The criterion is thus an empirical version of the bias variance decomposition. The constant κ is numerical and from a theoretical point of view, it depends neither on λ nor on n; it has to be calibrated once and for all on a set of preliminary simulations.

3.2. **Simulations.** The constant κ is calibrated through preliminary experiments and we take $\kappa = 2$. Then we apply the procedure with only one major change: the set $\widehat{\mathcal{M}}_n$ is too small in most experiments. To be able to consider more models with larger dimension, we replace it by $\widetilde{\mathcal{M}}_n = \{m \in \{1,\dots,n\}, \|\widehat{\Psi}_{m,Z}^{-1}\|_{\mathrm{op}} \leq n^{5/2}\}$. This is much more than expected from the theory, and still a true limitation since $\|\widehat{\Psi}_{m,Z}^{-1}\|_{\mathrm{op}}$ grows really very fast with m.

The matrix $\widehat{\Psi}_{m,\lambda S_Z}$ is straightforward from formula (15), and matrix $\widehat{\Psi}_{m,Z}$ is computed by writing its coefficients $(1/n)\sum_{i=1}^{n}\int_0^{Z_i}\varphi_j(x)\varphi_k(x)dx$ and Riemann discretization of the integrals over $[0, Z_i]$ with 200 steps.

• Comparison with Antoniadis et al. (1999) and others.

First, we consider two cases, which have been studied in previous papers:

(a) The first set of simulations is called in the following the "Gamma case". The X_i's are generated from a Gamma distribution with shape parameter 5 and scale 1 and the independent C_i's from an exponential distribution with mean 6.

(b) The second set is called "the bimodal case". The X_i's have a bimodal density defined by

$$f = 0.8u + 0.2v$$

where u is the density of $\exp(Y/2)$ with $Y \sim \mathcal{N}(0,1)$ and $v = 0.17Y + 2$. The $C_i's$ are generated from an exponential distribution with mean 2.5.

Examples (a) and (b) have been studied by Antoniadis et al. (1999) (wavelet estimator with selection of the coefficients by cross-validation), Reynaud-Bouret (2006) (histogram and Fourier estimators of the Aalen intensity) and Brunel and Comte (2005) (two ratio strategies for projection estimators on compact support). Antoniadis et al. (1999) and Brunel and Comte (2005) estimate both the subdensity fS_C and the hazard rate λ, whereas Reynaud-Bouret (2006) estimates λ only.

These authors give the mean squared errors of their estimator computed over $T = 200$ replications of samples of size $n = 200$ and $n = 500$. The error is computed over K regularly spaced points t_k, $k = 1, \ldots, K$, of the interval in which the X_i's fall ($[0, \max X_i]$), as the mean over the replications j of

$$\mathrm{MSE}_j = \frac{1}{K}\sum_{k=1}^{K}(\lambda(t_k) - \hat{\lambda}_j(t_k))^2$$

where $\hat{\lambda}_j$ is the estimate of λ for the sample number j, $j = 1, \ldots, T$.

In order to take into account the sparsity of the observations at the end of the interval, ($\mathbb{P}(X > 6) = 0.25$ in the Gamma case and $\mathbb{P}(X > 2) = 0.16$ in the bimodal case), they also compute an error MSE2 defined by the same kind of mean squared error but with a truncated mean over the t_k's less than 6 in the Gamma case and less than 2 in the bimodal case.

Reynaud-Bouret (2006)'s results, those of Antoniadis et al. (1999) and those of Brunel and Comte (2005) are recalled in Table 1, while ours are given in Table 2.

We remark that the MSE of our new estimator is always substantially smaller than the one of all previous estimators. A contrario, the value of our MSE2 is slightly larger in all cases. This means that locally on this part of the interval, our new estimator is not better, but that, considered on the whole domain, it is globally much more performing. Let us add that the relevant cut for the MSE to compute restricted MSE2 is in general unknown, so that the only reliable result is related to the complete observation interval.

	Estimator of Antoniadis et al.				Estimator of Reynaud-Bouret				Estimator of Brunel-Comte			
Model	Gamma		Bimodal		Gamma		Bimodal		Gamma		Bimodal	
n	200	500	200	500	200	500	200	500	200	500	200	500
10 MSE	1.12	0.995	20.80	19.70	0.55	0.579	12.59	11.22	0.857	0.900	9.02	7.06
10 MSE2	0.025	0.016	0.48	0.32	0.032	0.012	1.50	0.51	0.023	0.013	1.068	0.408

TABLE 1. Results of Antoniadis et al. (1999, Table 2), of the Fourier strategy in Reynaud-Bouret (2006) and of the ratio strategy of Brunel and Comte (2005), for the estimation of λ, $T = 200$ replications

	$\hat{\lambda}_{\hat{m}}$				$\hat{\lambda}_{\hat{m}}$ (larger n)			
Model	Gamma		Bimodal		Gamma		Bimodal	
$n =$	200	500	200	500	1000	2000	1000	2000
10 MSE	0.275	0.084	6.287	4.87	0.032	0.019	3.726	3.069
10 MSE2	0.035	0.014	1.268	0.969	0.0067	0.0035	0.697	0.557

TABLE 2. Results for hazard-regression estimators of λ with Laguerre basis, $T = 200$ replications, $\kappa = 2$.

• Comparison with Bouezmarni *et al.* (2011) and Müller and Wang (1994).
Bouezmarni *et al.* (2011) consider a hazard rate estimator built as a quotient of a Gamma-kernel density estimator divided by a Kaplan-Meier survival function estimator. The bandwidth selection method is not clearly specified. The authors consider two models A and B. Model A corresponds to an exponential distribution with parameter 1 for X and a uniform density on $[0, c]$ for C, where c is chosen to ensure the desired censoring rate. In Model B, X follows a Weibull distribution with scale parameter $b = 2$ and shape parameter $a = 1.2$, and C a Weibull distribution with shape parameter a and scale parameter given by $b = ((1 - p)/p)^{1/a}$. This ensures that the degree of censoring is equal to p.

Table 3 presents the results obtained by Bouezmarni *et al.* (2011) in column G, by Müller and Wang (1994) in column MW, and by our estimator in columns MS. The column MS 99% presents the MSE computed on an interval corresponding to 99% of the observations and MSE 85% on an interval corresponding to 85% of the observations. We can see that the performances of our estimator is in the range of the two others for $n = 125$ and sometimes better for $n = 250$. The performances on the smaller interval are clearly better but of course the comparison is unfair. The sample sizes here are quite small (with possibly 50% of censoring) for nonparametric methods, which makes the resulting performances hardly reliable.

• Comparison with Barbeito and Cao (2018).
Barbeito and Cao (2018) consider hazard rate estimation in a model **without** censoring. Their estimator is a quotient of a standard kernel estimator divided by an integrated version of it, and they concentrate on the bandwidth selection problem, for which they propose two strategies: double one sided cross validation denoted 'DO' and a bootstrap method (the best for those three models) 'Boot2'. We recall in Table 4 their results and compare to our estimator, for three of their models corresponding to nonnegative X: a

		$n = 125$				$n = 250$			
Model	% cens	G	MW	MS 99%	MS 85%	G	MW	MS 99%	MS 85%
	10	6.00	12.46	16.47	3.75	3.20	8.09	8.05	1.85
		(0.36)	(0.67)	(4.16)	(0.12)	(0.08)	(0.26)	(1.13)	(0.02)
A	25	5.90	14.04	15.2	4.07	4.95	13.09	7.68	1.98
		(0.16)	(0.45)	(4.50)	(0.15)	(0.07)	(0.30)	(1.29)	(0.03)
	50	13.20	20.00	14.5	5.50	13.13	19.33	7.85	2.49
		(0.38)	(0.64)	(7.02)	(0.26)	(0.22)	(0.35)	(1.97)	(0.06)
	10	2.32	8.87	7.61	1.30	1.04	5.51	4.26	0.71
		(0.04)	(0.27)	(1.07)	(0.01)	(0.02)	(0.08)	(0.31)	(0.003)
B	25	2.64	10.04	8.26	1.40	2.27	8.49	4.01	0.76
		(0.03)	(0.34)	(1.34)	(0.01)	(0.02)	(0.09)	(0.35)	(0.004)
	50	8.52	11.87	8.85	1.57	8.93	11.99	4.35	0.84
		(0.10)	(0.18)	(1.70)	(0.02)	(0.05)	(0.09)	(0.32)	(0.01)

TABLE 3. 100*MSE (100*variance of ISE), comparison with Bouezmarni *et al.* (2011) and Müller and Wang (1994)

Weibull with scale parameter $\alpha = 2$ and shape parameter $\lambda = \pi$, a $\chi^2(2)$ and a $\chi^2(3)$ density. Here again, the sample size is $n = 100$ and is rather small for nonparametric estimation. However, we can see that our estimator performs analogously to Barbeito and Cao (2018)'s.

Model	Weibull	$\chi^2(2)$	$\chi^2(3)$
DO of B & C	0.017	0.065	0.024
(median)	(0.010)	(0.063)	(0.021)
Boot2 of B & C	0.029	0.056	0.020
(median)	(0.013)	(0.049)	(0.016)
Our estimator	0.028	0.040	0.029
(median)	(0.021)	(0.032)	(0.024)
(std)	0.023	0.028	0.020

TABLE 4. Comparison with Barbeito and Cao (2018), $n = 100$, 500 replications.

4. Real data example

We study a real dataset from the National Longitudinal Survey of Youth of the U.S. Bureau of Labor Statistics (https://www.nlsinfo.org/content/cohorts/nlsy79). In the survey, women, aged 14 to 21 in 1979, have been interviewed yearly from 1983 through 1988. They were asked about any pregnancies and breast feeding. This data set consists of the information from $n = 927$ first-born children to mothers who chose to breast feed their children. The lifetime in the data set is the duration of breast feeding in weeks, followed by

an indicator of whether the breast feeding was completed (i.e. time to weaning of breast-fed newborns). The data was restricted to children born after 1978 and whose gestation was between 20 and 45 weeks and it is available from the **KMsurv** package.

We have 892 events over 927 observed data which correspond to 96% of uncensored lifetimes. The minimum observed duration to weaning is one week and the maximum one is 192 weeks with a median at 12 weeks. See also Section 5.4 in Klein and Moeshberger (2003) for a brief analysis of the dataset. Our estimator with Laguerre basis is applied to observations rescaled on the interval $[0,3]$ in order to deal with the high time values of the duration which may cause numerical error in the digital process. The rescaled observations $(Z_i')_{1\leq i\leq n}$ are obtained by applying the transformation $t \mapsto (t-\min(Z_i))/(\max(Z_i)-\min(Z_i))/b)$, with $b=3$ to the original observations $(Z_i)_{1\leq i\leq n}$. Then, the estimator is plotted in its original scale.

In Figure 2 we present the collection of estimators $\tilde{\lambda}_m$ defined in (5), for $m \in \widetilde{\mathcal{M}}_n = \{m \in \{1,\cdots,D_{max}\}, \|\widehat{\Psi}_{m,Z}^{-1}\|_{\mathrm{op}} \leq n^{5/2}\}$. Setting $\kappa = 2$ as in the simulation studies, and $D_{max}=7$ our selection procedure chooses the model $m_{opt}=5$ with $D_{max}=7$ but only models with $m \leq 6$ are allowed by the numerical constraint required with the set $\widetilde{\mathcal{M}}_n$. As these estimators are not necessarily positive, we take the positive part $\max(\tilde{\lambda}_m(x),0)$. The corresponding estimator is displayed in Figure 3 (Left) along with the Müller and Wang kernel estimators and in Figure 3 (Right) with classical parametric models whose parameters were adjusted by maximum likelihood. The kernel estimators in Figure 3 (Left) are built using the function `muhaz` available in the **muhaz** package with a local optimal bandwidth computed at a grid point obtained by minimizing the local MSE and with the Epanechnikov kernel. We can see that the kernel estimator needs to be corrected at the end of the interval whereas our estimator is not affected by boundary effects. We have fitted an exponential hazard rate $\hat{\lambda}_1(x)=0.059$, a log-logistic hazard rate $\hat{\lambda}_2(x)=\hat{\theta}\hat{\nu}x^{\hat{\nu}-1}/(1+\hat{\theta}x^{\hat{\nu}})$ with $\hat{\nu}=1.44$ and $\hat{\theta}=0.037$ and a log-normal hazard

$$\hat{\lambda}_3(x) = \frac{1}{x\hat{\sigma}}\phi\left(\frac{\ln x-\hat{\mu}}{\hat{\sigma}}\right) \Big/ \left[1-\Phi\left(\frac{\ln x-\hat{\mu}}{\hat{\sigma}}\right)\right]$$

with $\hat{\mu}=2.24$ and $\hat{\sigma}=1.18$ and $\phi(x)$ and $\Phi(x)$ are respectively the density and the cumulative distribution function of a standard gaussian.

All these parametric models cannot capture correctly the shape of the hazard rate. We can observe that the shape of the nonparametric estimator makes sense since the risk of stopping breast feeding is high at the very beginning, then the curve is decreasing and achieves a first local minimum between the 12th and 18th week. Then the curve is increasing and after the week 50, only 43 women keep going on with breast feeding. These remaining women stop between week 50 and 100, and finally over the week 100, only 3 women are still breastfeeding, so the curve is increasing with large slope. This is corroborated by the aspect of the cumulative hazard estimators in Figure 4 where the cumulative hazard rate estimators are displayed. Since the cumulative hazard rate $\Lambda(x) = \int_0^x \lambda(u)du$ is the hazard rate primitive, we can obtain the integrated Laguerre estimator by using the following formula giving the primitive of the Laguerre basis $\mathcal{L}_j(x) = \int_0^x \varphi_j(u)du$

$$\mathcal{L}_0(x) = \varphi_0(0) - \varphi_0(x) \quad \text{and } \mathcal{L}_j(x) = -\mathcal{L}_{j-1}(x) - \varphi_j(x) + \varphi_{j-1}(x) \ \text{ for } j \geq 1.$$

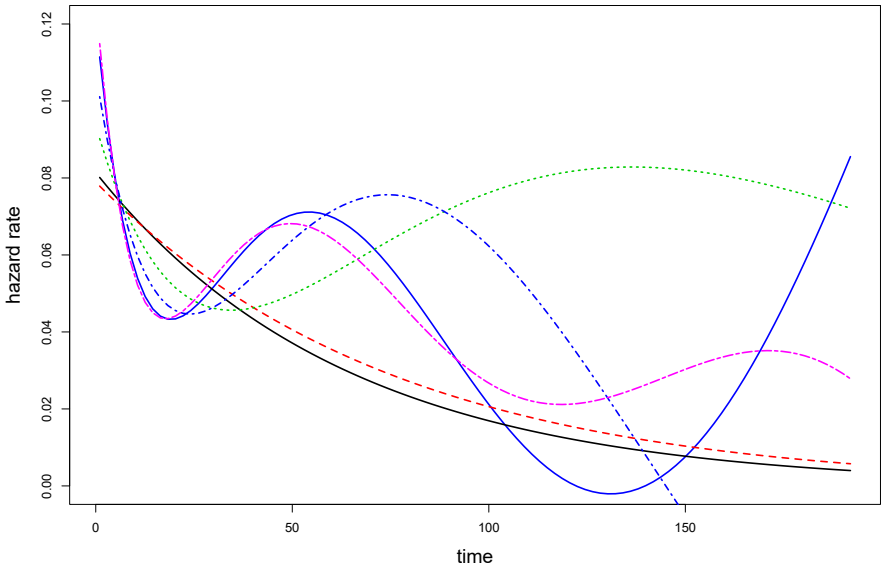


FIGURE 2. Collection of projection estimators with Laguerre basis : $m = 1$ (solid black), $m = 2$ (dashed red), $m = 3$ (dotted green), $m = 4$ (dotdashed blue), $m_{opt} = 5$ (plain blue), $m = 6$ (twodashed magenta)

and thus we obtained the estimator $\tilde{\Lambda}_m(x) = \int_0^x \tilde{\lambda}_m(u)du = \int_0^x \sum_{j=0}^{m-1} \hat{a}_j \varphi_j(u)du = \sum_{j=0}^{m-1} \hat{a}_j \mathcal{L}_j(x)$.

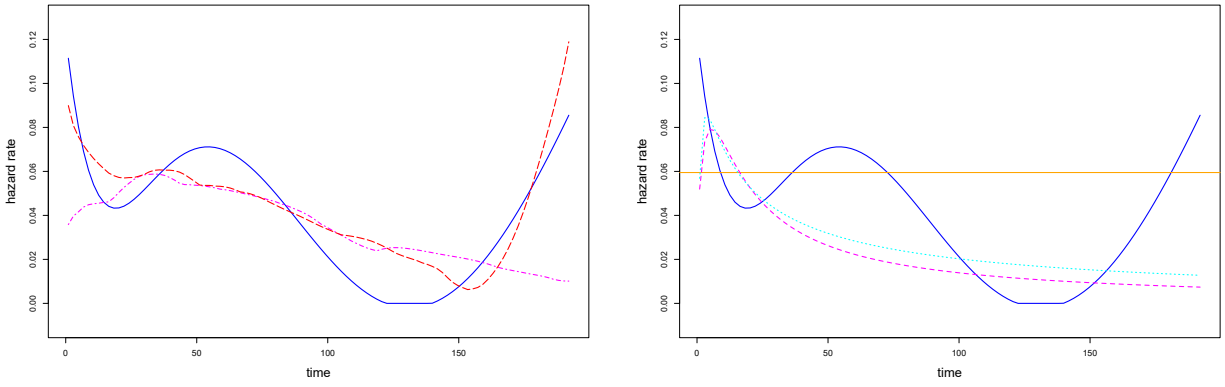


FIGURE 3. Left : Müller and Wang kernel estimator with boundary corrections (dashed red line) and without boundary correction (dotdashed magenta line). Right : exponential (orange plain), log-logistic (cyan dotted) and log-normal (magenta dashed). Both left and right : Our projection estimator with $m_{opt} = 5$ (solid blue line)

In order to check the adequacy of our hazard rate estimator, we compare the integrated estimator $\tilde{\Lambda}_m(x)$ with the nonparametric benchmark estimators : the Nelson Aalen estimator $\hat{\Lambda}_{NA}(x)$ and $-\ln(\hat{S}_{KM}(x))$ where $\hat{S}_{KM}(x)$ is the Kaplan-Meier estimator of the survival function. Obviously, our estimator is in accordance with both estimators, see Figure 4, while the parametric models are not satisfactory. So, our nonparametric estimator appears as a good competitor for the estimation of the hazard rate.

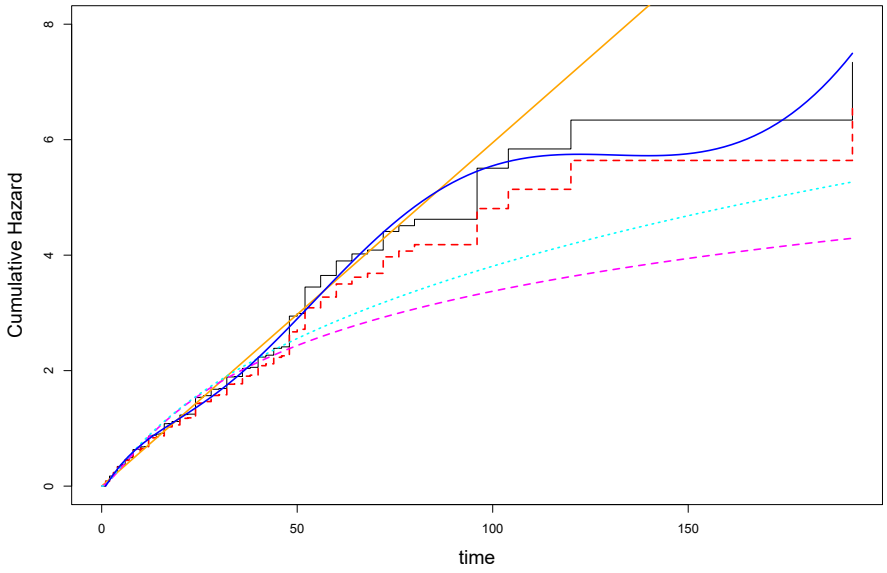


FIGURE 4. Our integrated Laguerre estimator (plain blue), Nelson-Aalen estimator (dashed red), $-\ln(\hat{S}_{KM}(x))$ (plain black) and parametric cumulative hazard curves : exponential (orange plain), log-logistic (cyan dotted) and log-normal (dashed magenta)

5. Concluding remarks

Our study presents a generalization of risk bounds for nonparametric least-squares estimator of the hazard rate, which allows to consider non compactly supported bases. This is very useful to propose developments of the hazard rate in the Laguerre basis, which can be seen as a combination of gamma-type functions. We show that our new result encompasses the ones obtained in Plancade (2011) or Comte *et al.* (2011). We show on simulations that the performance of the new estimator are comparable to or better than previous kernel or wavelet proposals and we also illustrate that it can be used successfully to analyze real data.
Further simulation may certainly be conducted, in particular to improve numerical stability of the computation of Laguerre functions. Indeed in practice, we rescale the real data

to a smaller range to avoid numerical problems with the Laguerre basis. A theoretical study including the choice of the range from the data may be conducted: adding a range parameter in the definition of the Laguerre basis is possible and a selection procedure for this parameter may be developed.
We chose to compare our estimator to previous results, but sample sizes in these examples are sometimes quite small. We believe that such nonparametric method requires a rather large data set, and empirical experiments maybe be conducted to explored this aspect.
Lastly, from theoretical point of view, the model selection procedure has to be studied; it is beyond the scope of the present work, as it would require quite lengthy developments; therefore it is left for further work.

6. Proofs

6.1. **Proof of Proposition 2.1.** Two sets are of interest in the sequel:

$$\Omega_m = \left\{ \forall t \in S_m, \left| \frac{\|t\|_n^2}{\|t\|_{S_Z}^2} - 1 \right| \leq \frac{1}{2} \right\} \tag{16}$$

$$\Lambda_m = \left\{ \|\widehat{\Psi}_{m,Z}^{-1}\|_{\text{op}} \leq \mathfrak{c} \frac{n}{\log(n)} \right\} \tag{17}$$

The following Lemma provides preliminary results which are the main ingredients to bound the empirical risk and the integrated risk of one estimator.

Lemma 6.1. *Under the assumptions of Proposition 2.1,*

$$\mathbb{P}(\Omega_m^c) \leq 2/n^4 \quad \text{and} \quad \mathbb{P}(\Lambda_m^c) \leq 2/n^4$$

(i) Decomposition of the empirical risk.

$$\begin{aligned} \|\tilde{\lambda}_m - \lambda\|_n^2 &= \|\tilde{\lambda}_m - \lambda\|_n^2 \mathbb{1}_{\Omega_m \cap \Lambda_m} + \|\tilde{\lambda}_m - \lambda\|_n^2 \mathbb{1}_{\Omega_m^c \cap \Lambda_m} + \|\lambda\|_n^2 \mathbb{1}_{\Lambda_m^c} \\ &:= \mathbb{T}_1 + \mathbb{T}_2 + \mathbb{T}_3 \end{aligned}$$

We bound successively the expectation of the three terms.

(ii) Study of $\mathbb{T}_1$, main term in the decomposition. On Λ_m, it holds that $\tilde{\lambda}_m = \widehat{\lambda}_m$, and we get

$$\mathbb{T}_1 = \|\widehat{\lambda}_m - \lambda\|_n^2 \mathbb{1}_{\Omega_m \cap \Lambda_m} = \left(\|\widehat{\lambda}_m - \Pi_m \lambda\|_n^2 + \inf_{t \in S_m} \|t - \lambda\|_n^2 \right) \mathbb{1}_{\Omega_m \cap \Lambda_m}$$

with $\Pi_m \lambda$ the orthogonal projection for norm $\|.\|_n$ of λ on S_m, that is $\Pi_m \lambda = \sum_{j=0}^{m-1} b_j \varphi_j$ is such that $\langle \lambda - \Pi_m \lambda, \varphi_j \rangle_n = 0$, for $j = 1, \cdots, m$. Taking the expectation,

$$\begin{aligned} \mathbb{E}[\mathbb{T}_1] &\leq \mathbb{E}\left(\|\widehat{\lambda}_m - \Pi_m \lambda\|_n^2 \mathbb{1}_{\Omega_m \cap \Lambda_m} \right) + \mathbb{E}\left(\inf_{t \in S_m} \|t - \lambda\|_n^2 \right) \\ &\leq \mathbb{E}\left(\|\widehat{\lambda}_m - \Pi_m \lambda\|_n^2 \mathbb{1}_{\Omega_m \cap \Lambda_m} \right) + \inf_{t \in S_m} \|t - \lambda\|_{S_Z}^2 \end{aligned} \tag{18}$$

and this corresponds to the classical variance/squared bias decomposition.

Let us bound the variance term. Recall that $\widehat{\Phi}_m = (\varphi_j(Z_i))_{1\leq i\leq n, 0\leq j\leq m-1}$, and set,

$$\vec{v} = \frac{1}{n}\,{}^t\widehat{\Phi}_m\vec{\delta} - (\langle\lambda,\varphi_j\rangle_n)_{0\leq j\leq m-1}$$

Note that $\mathbb{E}(\vec{v}) = 0$ and remember that $\langle\lambda - \Pi_m\lambda, \varphi_j\rangle_n = 0$ for $j = 0,\cdots,m-1$ so that :

$$\begin{aligned}\langle\lambda,\varphi_j\rangle_n &= \langle\Pi_m\lambda,\varphi_j\rangle_n \\ &= \sum_{k=0}^{m-1} b_k\langle\varphi_k,\varphi_j\rangle_n \quad \text{with } \Pi_m\lambda = \sum_{j=0}^{m-1} b_j\varphi_j.\end{aligned}$$

Thus, we find that $(\langle\lambda,\varphi_j\rangle_n)_{0\leq j\leq m-1} = \widehat{\Psi}_{m,Z}\vec{b}^{(m)}$ with $\vec{b}^{(m)} = {}^t(b_0,\cdots,b_{m-1})$ and we can write $\widehat{\Psi}_{m,Z}^{-1}\vec{v} = \vec{\hat{a}}^{(m)} - \vec{b}^{(m)}$. Then,

$$\begin{aligned}\|\widehat{\lambda}_m - \Pi_m\lambda\|_n^2 = {}^t(\vec{\hat{a}}^{(m)} - \vec{b}^{(m)})\widehat{\Psi}_{m,Z}(\vec{\hat{a}}^{(m)} - \vec{b}^{(m)}) &= {}^t(\widehat{\Psi}_{m,Z}^{-1}\vec{v})\widehat{\Psi}_{m,Z}\widehat{\Psi}_{m,Z}^{-1}\vec{v} \\ &= {}^t\vec{v}\,\widehat{\Psi}_{m,Z}^{-1}\vec{v} \\ &= {}^t\vec{v}\,\Psi_{m,Z}^{-1/2}\Psi_{m,Z}^{1/2}\widehat{\Psi}_{m,Z}^{-1}\Psi_{m,Z}^{1/2}\Psi_{m,Z}^{-1/2}\vec{v} \\ &\leq \lambda_{max}(\Psi_{m,Z}^{1/2}\widehat{\Psi}_{m,Z}^{-1}\Psi_{m,Z}^{1/2})\,{}^t\vec{v}\Psi_{m,Z}^{-1}\vec{v}\end{aligned}$$

But on Ω_m, we have $\lambda_{max}(\Psi_{m,Z}^{1/2}\widehat{\Psi}_{m,Z}^{-1}\Psi_{m,Z}^{1/2}) \leq 2$ since $\|\Psi_{m,Z}^{-1/2}\widehat{\Psi}_{m,Z}\Psi_{m,Z}^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \leq 1/2$ and we get

$$\mathbb{E}\left(\|\widehat{\lambda}_m - \Pi_m\lambda\|_n^2\mathbb{1}_{\Omega_m\cap\Lambda_m}\right) \leq 2\,\mathbb{E}[{}^t\vec{v}\,\Psi_{m,Z}^{-1}\vec{v}] \tag{19}$$

The study of the variance term will be complete as soon as we have computed $\mathbb{E}[{}^t\vec{v}\,\Psi_{m,Z}^{-1}\vec{v}]$.

$$\mathbb{E}[{}^t\vec{v}\,\Psi_{m,Z}^{-1}\vec{v}] = \mathbb{E}\left[\sum_{j,k} v_j v_k[\Psi_{m,Z}^{-1}]_{j,k}\right]$$

with $v_j = (1/n)\sum_{i=1}^n\left(\delta_i\varphi_j(Z_i) - \int\lambda(x)\varphi_j(x)\mathbb{1}_{\{Z_i>x\}}dx\right)$ the j-th coordinate of $\vec{v}$ and $\mathbb{E}[v_j] = 0$. Note also that

$$\begin{aligned}&\mathbb{E}\left[\left(\delta_1\varphi_j(Z_1) - \int\lambda(x)\varphi_j(x)\mathbb{1}_{\{Z_1>x\}}dx\right)\left(\delta_1\varphi_k(Z_1) - \int\lambda(x)\varphi_k(x)\mathbb{1}_{\{Z_1>x\}}dx\right)\right] \\ =\ &\mathbb{E}(\delta_1\varphi_j(Z_1)\varphi_k(Z_1)) - \int\lambda(x)\varphi_k(x)\mathbb{E}(\delta_1\varphi_j(Z_1)\mathbf{1}_{\{Z_1>x\}})dx \\ &-\int\lambda(x)\varphi_j(x)\mathbb{E}(\delta_1\varphi_k(Z_1)\mathbf{1}_{\{Z_1>x\}})dx + \iint\lambda(x)\lambda(y)S_Z(x\vee y)\varphi_j(x)\varphi_k(y)dxdy \qquad (20) \\ =\ &\mathbb{E}(\delta_1\varphi_j(Z_1)\varphi_k(Z_1)).\end{aligned}$$

Indeed

$$\begin{aligned}\int \lambda(x)\varphi_k(x)\mathbb{E}(\delta_1\varphi_j(Z_1)\mathbf{1}_{\{Z_1>x\}})dx &= \int \lambda(x)\varphi_k(x)\mathbb{E}(S_C(X_1)\varphi_j(X_1)\mathbf{1}_{\{X_1>x\}})dx\\ &= \int \lambda(x)\varphi_k(x)\int \varphi_j(y)S_C(y)f(y)\mathbf{1}_{\{y>x\}}dydx\\ &= \iint \lambda(x)\lambda(y)S_Z(y)\mathbf{1}_{\{y>x\}}\varphi_j(x)\varphi_k(y)dxdy\end{aligned}$$

so that the sum of the two middle terms in (20) cancel with the last one. Then,

$$\begin{aligned}\mathbb{E}[{}^t\vec{v}\,\Psi_{m,Z}^{-1}\vec{v}] &= \sum_{j,k}[\Psi_{m,Z}^{-1}]_{j,k}\mathbb{E}\left[v_jv_k\right] = \sum_{j,k}[\Psi_{m,Z}^{-1}]_{j,k}\left(\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left[\delta_i\varphi_j(Z_i)\varphi_k(Z_i)\right]\right)\\ &= \frac{1}{n}\sum_{j,k}[\Psi_{m,Z}^{-1}]_{j,k}\mathbb{E}\left[\delta_1\varphi_j(Z_1)\varphi_k(Z_1)\right] = \frac{1}{n}\sum_{j,k}[\Psi_{m,Z}^{-1}]_{j,k}\int\varphi_j(x)\varphi_k(x)f(x)S_C(x)dx\\ &= \frac{1}{n}\sum_{j,k}[\Psi_{m,Z}^{-1}]_{j,k}\int\varphi_j(x)\varphi_k(x)\lambda(x)S_Z(x)dx\end{aligned}$$

Finally, with $\Psi_{m,\lambda S_Z}$ defined by (11), we can see that:

$$\mathbb{E}[{}^t\vec{v}\,\Psi_{m,Z}^{-1}\vec{v}] = \frac{1}{n}\sum_{j,k}[\Psi_{m,Z}^{-1}]_{j,k}[\Psi_{m,\lambda S_Z}]_{j,k} = \frac{1}{n}\mathrm{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z}). \tag{21}$$

Thus, plugging this in (19) yields

$$\mathbb{E}\left(\|\widehat{\lambda}_m - \Pi_m\lambda\|_n^2\mathbb{1}_{\Omega_m\cap\Lambda_m}\right) \le \frac{2}{n}\mathrm{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z}) \tag{22}$$

and with (18), we have

$$\mathbb{E}(\mathbf{T}_1) \le \inf_{t\in S_m}\|t-\lambda\|_{S_Z}^2 + \frac{2}{n}\mathrm{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z}).$$

(iii) Residual terms.

$$\mathbb{T}_2 = \|\tilde{\lambda}_m - \lambda\|_n^2\mathbb{1}_{\Omega_m^c\cap\Lambda_m} \le 2\|\widehat{\lambda}_m\|_n^2\mathbb{1}_{\mathbb{1}_{\Omega_m^c\cap\Lambda_m}} + 2\|\lambda\|_n^2\mathbb{1}_{\Omega_m^c}$$

We write

$$\begin{aligned}\|\widehat{\lambda}_m\|_n^2 &= {}^t\vec{a}^{(m)}\widehat{\Psi}_{m,Z}\vec{a}^{(m)} = \frac{1}{n^2}\,{}^t\vec{\delta}\,\widehat{\Phi}_m\widehat{\Psi}_{m,Z}^{-1}\widehat{\Psi}_{m,Z}\widehat{\Psi}_{m,Z}^{-1}\widehat{\Phi}_m\vec{\delta}\\ &\le \frac{1}{n^2}\|\widehat{\Psi}_{m,Z}^{-1}\|_{\mathrm{op}}\;{}^t\vec{\delta}\,\widehat{\Phi}_m^t\widehat{\Phi}_m\vec{\delta} = \frac{1}{n^2}\|\widehat{\Psi}_{m,Z}^{-1}\|_{\mathrm{op}}^2\sum_{j=0}^{m-1}\left(\sum_{i=1}^n\delta_i\varphi_j(Z_i)\right)^2\\ &\le \frac{1}{n^2}\|\widehat{\Psi}_{m,Z}^{-1}\|_{\mathrm{op}}\sum_{j=0}^{m-1}\left(\sum_{i=1}^n\delta_i^2\right)\sum_{i=1}^n\varphi_j^2(Z_i) \quad \text{(Cauchy-Schwarz)}\\ &\le \frac{1}{n^2}\|\widehat{\Psi}_{m,Z}^{-1}\|_{\mathrm{op}}\left(\sum_{j=0}^{m-1}\varphi_j^2(Z_i)\right)\times n^2 \le \|\widehat{\Psi}_{m,Z}^{-1}\|_{\mathrm{op}}L(m) \le \mathfrak{c}L(m)\frac{n}{\log(n)},\end{aligned}$$

since on Λ_m, $\|\widehat{\Psi}_{m,Z}^{-1}\|_{\rm op} \leq \mathfrak{c}n/\log(n)$. Thus, we obtain, by using Lemma 6.1 and $L(m) \leq n$, that

$$\mathbb{E}[\|\widehat{\lambda}_m\|_n^2 \mathbb{1}_{\mathbb{1}_{\Omega_m^c \cap \Lambda_m}}] \leq \mathfrak{c}L(m)\frac{n}{\log(n)}\mathbb{P}(\Omega_m^c) \leq \frac{c}{n}.$$

Second,

$$\mathbb{E}[\|\lambda\|_n^2 \mathbb{1}_{\Omega_m^c}] \leq \left(\mathbb{E}[\|\lambda\|_n^4]\right)^{1/2} \sqrt{\mathbb{P}(\Omega_m^c)}$$

with

$$\begin{aligned}
\mathbb{E}[\|\lambda\|_n^4] &= \frac{1}{n^2}\sum_{i,k}\iint \lambda^2(x)\lambda^2(y)\mathbb{E}(\mathbb{1}_{Z_i>x}\mathbb{1}_{Z_k>y})dxdy \\
&\leq \frac{1}{n^2}\sum_{i,k}\iint \lambda^2(x)\lambda^2(y)\mathbb{E}^{1/2}(\mathbb{1}_{Z_i>x})\mathbb{E}^{1/2}(\mathbb{1}_{Z_k>y})dxdy \\
&= \left(\int \lambda^2(x)\sqrt{S_Z(x)}dx\right)^2 < +\infty
\end{aligned}$$

by using assumption (8). Finally,

$$\mathbb{E}[\mathbb{T}_3] = \mathbb{E}[\|\lambda\|_n^2 \mathbb{1}_{\Lambda_m^c}] \leq \left(\mathbb{E}[\|\lambda\|_n^4]\right)^{1/2} \sqrt{\mathbb{P}(\Lambda_m^c)} \leq c/n$$

by using Lemma 6.1. □

6.2. **Proof of Lemma 6.1.** First, note that

$$\mathbb{P}(\Omega_m^c) = \mathbb{P}\left(\sup_{t\in S_m, \|t\|_{S_Z}=1}\left|\frac{1}{n}\sum_{i=1}^n\left(\int t^2(x)\mathbf{1}_{Z_i>x}dx - 1\right)\right| > \frac{1}{2}\right)$$

and then

$$\sup_{t\in S_m, \|t\|_{S_Z}=1}\left|\frac{1}{n}\sum_{i=1}^n\left(\int t^2(x)\mathbf{1}_{Z_i>x}dx - 1\right)\right| = \|\Psi_{m,Z}^{-1/2}(\widehat{\Psi}_{m,Z} - \mathrm{Id}_m)\Psi_{m,Z}^{-1/2}\|_{\rm op}.$$

To apply Matrix Chernoff Inequality given in Theorem 1.1 of Tropp (2012), we denote by

$$K_m(Z_i) = \Psi_{m,Z}^{-1/2}\left(\int \varphi_j(x)\varphi_k(x)\mathbb{1}_{Z_i>x}dx\right)_{j,k}\Psi_{m,Z}^{-1/2}, \quad \text{for } i = 1, \cdots, n.$$

We have $\Psi_{m,Z}^{-1/2}\widehat{\Psi}_{m,Z}\Psi_{m,Z}^{-1/2} = \frac{1}{n}\sum_{i=1}^n K_m(Z_i)$, and

$$\mathbb{E}\left(\sum_{i=1}^n K_m(Z_i)\right) = n\mathrm{Id}_m.$$

This yields that $\mu_{\min} = \mu_{\max} = n$ in Tropp's notations. Moreover

$$\lambda_{max}(K_m(Z_i)) = \sup_{\|\vec{x}\|_{2,m}=1, \vec{x}\in\mathbb{R}^m} {}^t(\Psi_{m,Z}^{-1/2}\vec{x})\left(\int \varphi_j(z)\varphi_k(z)\mathbf{1}_{Z_i>z}dz\right)_{j,k}(\Psi_{m,Z}^{-1/2}\vec{x})$$

and setting $\vec{y} = \Psi_{m,Z}^{-1/2}\vec{x}$,

$${}^t\vec{x}\, K_m(Z_i)\vec{x} = \int\left(\sum_{j=0}^{m-1} y_j\varphi_j(z)\right)^2 \mathbf{1}_{Z_i>z}dz \leq \int\left(\sum_{j=0}^{m-1} y_j\varphi_j(z)\right)^2 dz = \|\vec{y}\|_{2,m}^2.$$

So,

$$\lambda_{max}(K_m(Z_i)) \leq \sup_{\|\vec{x}\|_{2,m}=1, \vec{x}\in\mathbb{R}^m} \|\Psi_{m,Z}^{-1/2}\vec{x}\|_{2,m}^2 = \|\Psi_{m,Z}^{-1}\|_{\text{op}}.$$

Therefore in Tropp (2012)'s notation, we get $R = \|\Psi_{m,Z}^{-1}\|_{\text{op}}$. Now, applying Matrix Chernoff Inequality as stated in Tropp (2012) (Theorem 1.1), we get

$$\mathbb{P}(\Omega_m^c) \leq 2m \exp\left(-c(1/2)\frac{n}{\|\Psi_{m,Z}^{-1}\|_{\text{op}}}\right)$$

provided that $\Psi_{m,Z}$ is invertible and with $c(u) = u + (1-u)\log(1-u)$ for $0 < u < 1$. Under condition (9), as $c(1/2) = (3\log(3/2) - 1)/2$, we obtain

$$\mathbb{P}(\Omega_m^c) \leq 2m \exp(-5\log(n)) \leq \frac{2}{n^4},$$

which is our first statement.

Now, we turn to $\mathbb{P}(\Lambda_m^c)$. Under (9), $\|\Psi_{m,Z}^{-1}\|_{\text{op}} \leq (\mathfrak{c}/2)(n/\log(n))$ and on Λ_m^c, $\|\widehat{\Psi}_{m,Z}^{-1}\|_{\text{op}} > \mathfrak{c}(n/\log(n))$. So,

$$\mathfrak{c}\frac{n}{\log(n)} < \|\widehat{\Psi}_{m,Z}^{-1}\|_{\text{op}} \leq \|\widehat{\Psi}_{m,Z}^{-1} - \Psi_{m,Z}^{-1}\|_{\text{op}} + \|\Psi_{m,Z}^{-1}\|_{\text{op}} \leq \|\widehat{\Psi}_{m,Z}^{-1} - \Psi_{m,Z}^{-1}\|_{\text{op}} + \frac{\mathfrak{c}}{2}\frac{n}{\log(n)}$$

and thus

$$\|\widehat{\Psi}_{m,Z}^{-1} - \Psi_{m,Z}^{-1}\|_{\text{op}} > \frac{\mathfrak{c}}{2}\frac{n}{\log(n)}.$$

At the end

$$\|\widehat{\Psi}_{m,Z}^{-1} - \Psi_{m,Z}^{-1}\|_{\text{op}} > \|\Psi_{m,Z}^{-1}\|_{\text{op}}.$$

Therefore, under (9),

$$\Lambda_m^c \subset \left\{\|\widehat{\Psi}_{m,Z}^{-1} - \Psi_{m,Z}^{-1}\|_{\text{op}} > \|\Psi_{m,Z}^{-1}\|_{\text{op}}\right\}.$$

It follows from Proposition 2.4 (ii) in Comte and Genon-Catalot (2018) that the last set is a subset of Ω_m^c. Thus $\mathbb{P}(\Lambda_m^c) \leq \mathbb{P}(\Omega_m^c) \leq 2/n^4$. □

6.3. **Proof of Proposition 2.2.** We start with a risk decomposition of the same type as in empirical case

$$\begin{aligned}\|\tilde{\lambda}_m - \lambda\|_{S_Z}^2 &= \|\tilde{\lambda}_m - \lambda\|_{S_Z}^2 \mathbb{1}_{\Omega_m\cap\Lambda_m} + \|\tilde{\lambda}_m - \lambda\|_{S_Z}^2 \mathbb{1}_{\Omega_m^c\cap\Lambda_m} + \|\lambda\|_{S_Z}^2 \mathbb{1}_{\Lambda_m^c} \\ &:= \widetilde{\mathbb{T}}_1 + \widetilde{\mathbb{T}}_2 + \widetilde{\mathbb{T}}_3\end{aligned}$$

We bound successively the expectation of the three terms.

Clearly, $\mathbb{E}(\widetilde{\mathbb{T}}_3) = \|\lambda\|_{S_Z}^2 \mathbb{P}(\Lambda_m^c) \leq c/n$.

For $\widetilde{\mathbb{T}}_2$, we write

$$\widetilde{\mathbb{T}}_2 \leq 2(\|\hat{\lambda}_m\|_{S_Z}^2 + \|\lambda\|_{S_Z}^2)\mathbb{1}_{\Omega_m^c\cap\Lambda_m}.$$

Obviously, $\mathbb{E}(\|\lambda\|_{S_Z}^2 \mathbb{1}_{\Omega_m^c \cap \Lambda_m}) \leq \|\lambda\|_{S_Z}^2 \mathbb{P}(\Omega_m^c) \leq c/n$. For the other term, we note that

$$\begin{aligned}\|\Psi_{m,Z}\|_{\text{op}} &= \lambda_{max}(\Psi_{m,Z}) = \sup_{{}^t\vec{x}\,\vec{x}=1} {}^t\vec{x}\,\Psi_{m,Z}\vec{x} = \sup_{{}^t\vec{x}\,\vec{x}=1} \int \left(\sum_{j=0}^{m-1} x_j\varphi_j(u)\right)^2 S_Z(u)du \\ &\leq \sup_{{}^t\vec{x}\vec{x}=1} \int \left(\sum_{j=0}^{m-1} x_j\varphi_j(u)\right)^2 du = \sup_{{}^t\vec{x}\,\vec{x}=1} \sum_{j=0}^{m-1} x_j^2 = 1.\end{aligned}$$

Then we write as for $\|\hat{\lambda}_m\|_n^2$ previously,

$$\begin{aligned}\|\widehat{\lambda}_m\|_{S_Z}^2 &= {}^t(\vec{\widehat{a}}^{(m)})\Psi_{m,Z}\vec{\widehat{a}}^{(m)} \leq \|\Psi_{m,Z}\|_{\text{op}}\, {}^t(\vec{\widehat{a}}^{(m)})\vec{\widehat{a}}^{(m)} \leq \frac{1}{n^2}\, {}^t\vec{\delta}\,\widehat{\Phi}_m\widehat{\Psi}_{m,Z}^{-2}\, {}^t\widehat{\Phi}_m\vec{\delta} \\ &\leq \frac{1}{n^2}\|\widehat{\Psi}_{m,Z}^{-2}\|_{\text{op}}\, {}^t\vec{\delta}\,\widehat{\Phi}_m\, {}^t\widehat{\Phi}_m\vec{\delta} = \frac{1}{n^2}\|\widehat{\Psi}_{m,Z}^{-2}\|_{\text{op}}^2 \sum_{j=0}^{m-1}\left(\sum_{i=1}^{n}\delta_i\varphi_j(Z_i)\right)^2 \\ &\leq \|\widehat{\Psi}_{m,Z}^{-2}\|_{\text{op}} L(m) \leq \mathfrak{c}^2 L(m)\left(\frac{n}{\log(n)}\right)^2,\end{aligned}$$

since on Λ_m, $\|\widehat{\Psi}_{m,Z}^{-1}\|_{\text{op}} \leq \mathfrak{c}n/\log(n)$. Thus, we obtain

$$\mathbb{E}[\|\widehat{\lambda}_m\|_{S_Z}^2 \mathbb{1}_{\mathbb{1}_{\Omega_m^c \cap \Lambda_m}}] \leq \mathfrak{c}^2 L(m)\left(\frac{n}{\log(n)}\right)^2 \mathbb{P}(\Omega_m^c) \leq \frac{c}{n}$$

under assumption (7) and $L(m) \leq n$. As a consequence, $\mathbb{E}(\widetilde{\mathbb{T}}_2) \leq c/n$.

To study $\widetilde{\mathbb{T}}_1$, we introduce $\lambda_m^{(S_Z)}$ the orthogonal projection on S_m of λ w.r.t. the scalar product weighted by S_Z and $g := \lambda - \lambda_m^{(S_Z)}$. We write

$$\|\widehat{\lambda}_m - \lambda_A\|_{S_Z}^2 = \|\widehat{\lambda}_m - \Pi_m\lambda + \Pi_m\lambda - \lambda\|_{S_Z}^2$$

and

$$\lambda - \Pi_m\lambda = \lambda - \lambda_m^{(S_Z)} - \Pi_m(\lambda - \lambda_m^{(S_Z)}) = g - \Pi_m g,$$

as $\Pi_m\lambda_m^{(S_Z)} = \lambda_m^{(S_Z)}$. Thus

$$\|\widehat{\lambda}_m - \lambda_A\|_{S_Z}^2 = \|\widehat{\lambda}_m - \Pi_m\lambda + \Pi_m g - g\|_{S_Z}^2 = \|\widehat{\lambda}_m - \Pi_m\lambda + \Pi_m g\|_{S_Z}^2 + \|g\|_{S_Z}^2$$

as g is orthogonal in $\mathbb{L}^2(A, S_Z(x)dx)$ to any function in S_m. Therefore

$$\begin{aligned}\mathbb{E}(\widetilde{\mathbb{T}}_1) &\leq \|g\|_{S_Z}^2 + 2\mathbb{E}(\|\Pi_m g\|_{S_Z}^2 \mathbb{1}_{\Omega_m\cap\Lambda_m}) + 2\mathbb{E}\left(\|\widehat{\lambda}_m - \Pi_m\lambda\|_{S_Z}^2 \mathbb{1}_{\Omega_m\cap\Lambda_m}\right) \\ &\leq \inf_{t\in S_m}\|t-\lambda\|_{S_Z}^2 + 2\mathbb{E}(\|\Pi_m g\|_{S_Z}^2 \mathbb{1}_{\Omega_m\cap\Lambda_m}) + 4\mathbb{E}\left(\|\widehat{\lambda}_m - \Pi_m\lambda\|_n^2 \mathbb{1}_{\Omega_m\cap\Lambda_m}\right),\end{aligned}$$

by using that $\|\widehat{\lambda}_m - \Pi_m\lambda\|_{S_Z}^2 \leq 2\|\widehat{\lambda}_m - \Pi_m\lambda\|_n^2$ on Ω_m (all terms are in S_m). For this last term, we can use the bound obtained w.r.t the empirical norm given by $2\text{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z})/n$, see (19). Now we have the following Lemma, inspired from Cohen *et al.* (2013, 2019):

Lemma 6.2. *Under the assumptions of Proposition 2.2,*

$$\mathbb{E}(\|\Pi_m g\|_{S_Z}^2 \mathbb{1}_{\Omega_m\cap\Lambda_m}) \leq 4\frac{\mathfrak{c}}{\log(n)}\|g\|_{S_Z}^2 = 4\frac{\mathfrak{c}}{\log(n)}\inf_{t\in S_m}\|t-\lambda\|_{S_Z}^2.$$

Thus, we get

$$\mathbb{E}(\widetilde{\mathbb{T}}_1) \leq (1+8\frac{\mathfrak{c}}{\log(n)}) \inf_{t\in S_m} \|t-\lambda\|_{S_Z}^2 + \frac{8}{n}\mathrm{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z}),$$

and we obtain the bound (12). □

Proof of Lemma 6.2. Let $(\bar{\varphi}_j)_{0\leq j\leq m-1}$ be an orthonormal basis w.r.t. S_Z scalar product. If $\bar{\varphi}_j = \sum_{k=0}^{m-1}\alpha_{j,k}\varphi_k$ and $\mathbf{A}_m = (\alpha_{j,k})_{0\leq j,k\leq m-1}$, then $\mathrm{Id}_m = (\int \bar{\varphi}_j\bar{\varphi}_k S_Z)_{j,k} = {}^t\mathbf{A}_m\Psi_{m,Z}\mathbf{A}_m$ so that $\mathbf{A}_m = \Psi_{m,Z}^{-1/2}$. Let $\widehat{\mathbf{G}}_m = (\langle\bar{\varphi}_j,\bar{\varphi}_k\rangle_n)_{j,k} = {}^t\mathbf{A}\widehat{\Psi}_{m,Z}\mathbf{A}_m = \Psi_{m,Z}^{-1/2}\widehat{\Psi}_{m,Z}\Psi_{m,Z}^{-1/2}$. Therefore, on Ω_m, $\|\widehat{\mathbf{G}}_m^{-1}\|_{\mathrm{op}} \leq 2$, as $\|\widehat{\mathbf{G}}_m - \mathrm{Id}_m\|_{\mathrm{op}} \leq 1/2$.

Now if $\Pi_m g = \sum_{j=0}^{m-1}\beta_j\bar{\varphi}_j$, as $\langle g-\Pi_m g,\bar{\varphi}_j\rangle_n = 0$ for $j = 0,1,\ldots,m-1$, we get $\langle g,\bar{\varphi}_j\rangle_n = \langle\Pi_m g,\bar{\varphi}_j\rangle_n = \sum_{k=0}^{m-1}\beta_k\langle\bar{\varphi}_k,\bar{\varphi}_j\rangle_n$ so that

$$\widehat{\mathbf{G}}_m\vec{\beta}_m = (\langle g,\bar{\varphi}_j\rangle_n)_{0\leq j\leq m-1} := \vec{d}_m.$$

Therefore

$$\|\Pi_m g\|_{S_Z}^2 = \|\vec{\beta}_m\|_{2,m}^2 = \|\widehat{\mathbf{G}}_m^{-1}\vec{d}_m\|_{2,m}^2 \leq \|\widehat{\mathbf{G}}_m^{-1}\|_{\mathrm{op}}^2\|\vec{d}_m\|_{2,m}^2 \leq 4\sum_{j=0}^{m-1}\langle g,\bar{\varphi}_j\rangle_n^2. \tag{23}$$

Now, we recall that

$$\langle g,\bar{\varphi}_j\rangle_n^2 = \left(\frac{1}{n}\sum_{i=1}^n\int\bar{\varphi}_j(x)\mathbf{1}_{\{Z_i>x\}}g(x)dx\right)^2$$

and

$$\mathbb{E}\left(\int\bar{\varphi}_j(x)\mathbf{1}_{\{Z_i>x\}}g(x)dx\right) = \int\bar{\varphi}_j(x)g(x)S_Z(x)dx = \langle\bar{\varphi}_j,g\rangle_{S_Z} = 0$$

as $\langle g,\varphi_j\rangle_{S_Z} = 0$. Thus

$$\mathbb{E}[\langle g,\bar{\varphi}_j\rangle_n^2] = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n\int\bar{\varphi}_j(x)\mathbf{1}_{\{Z_i>x\}}g(x)dx\right) = \frac{1}{n}\mathrm{Var}\left(\int\bar{\varphi}_j(x)\mathbf{1}_{\{Z_1>x\}}g(x)dx\right)$$

and

$$\mathbb{E}\left[\sum_{j=0}^{m-1}\langle g,\bar{\varphi}_j\rangle_n^2\right] \leq \frac{1}{n}\sum_{j=0}^{m-1}\mathbb{E}\left[\left(\int\bar{\varphi}_j(x)\mathbf{1}_{\{Z_1>x\}}g(x)dx\right)^2\right] = \mathbb{E}\left[\|\mathbf{A}_m\vec{u}\|_{2,m}^2\right]$$

where $\vec{u} = (\int\varphi_j(x)\mathbf{1}_{\{Z_1>x\}}g(x)dx)_{0\leq j\leq m-1}$. As $\|\mathbf{A}_m\|_{\mathrm{op}}^2 = \|\Psi_{m,Z}^{-1}\|_{\mathrm{op}}$, we get

$$\begin{aligned}\mathbb{E}\left[\sum_{j=0}^{m-1}\langle g,\bar{\varphi}_j\rangle_n^2\right] &\leq \|\Psi_{m,Z}^{-1}\|_{\mathrm{op}}\mathbb{E}(\|\vec{u}\|_{2,m}^2) \leq \|\Psi_{m,Z}^{-1}\|_{\mathrm{op}}\mathbb{E}\left(\sum_{j=0}^{m-1}\left(\int\varphi_j(x)\mathbf{1}_{\{Z_1>x\}}g(x)dx\right)^2\right)\\ &\leq \|\Psi_{m,Z}^{-1}\|_{\mathrm{op}}\mathbb{E}\left(\|\mathrm{Proj}_{S_m}(g\mathbf{1}_{\{Z_1>x\}}\|^2\right) \leq \|\Psi_{m,Z}^{-1}\|_{\mathrm{op}}\mathbb{E}\left(\int g^2(x)\mathbf{1}_{\{Z_1>x\}}dx\right).\end{aligned}$$

We obtain

$$\mathbb{E}\left[\sum_{j=0}^{m-1}\langle g,\bar{\varphi}_j\rangle_n^2\right] \leq \frac{\|\Psi_{m,Z}^{-1}\|_{\mathrm{op}}}{n}\|g\|_{S_Z}^2$$

which, under (9) and reminding (23), implies

$$\|\Pi_m g\|_{S_Z}^2 \leq \frac{4\mathfrak{c}}{\log(n)}\|g\|_{S_Z}^2.$$

This is the announced result. □

6.4. **Proof of Lemma 2.1.** The proof relies on the notations and computations of the proof of Proposition 2.1. Let

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i t(X_i) - \int \lambda(x)t(x)\mathbf{1}_{Z_i>x}dx\right).$$

We remark that, for $t = \sum_{j=0}^{m-1} a_j\varphi_j$, then $\nu_n(t) = \sum_{j=0}^{m-1} a_j v_j$, where $v_j = \nu_n(\varphi_j)$. Moreover,

$$\begin{aligned}
\sup_{t\in S_m, \|t\|_{S_Z}=1}[\nu_n(t)]^2 &= \sup_{{}^t\vec{a}\Psi_{m,Z}\vec{a}=1}\left(\sum_{j=0}^{m-1} a_j v_j\right)^2 \\
&= \sup_{\|\vec{b}\|_{2,m}^2=1, \vec{a}=\Psi_{m,Z}^{-1/2}\vec{b}}\left[\sum_{j=0}^{m-1}\left(\sum_{k=0}^{m-1}[\Psi_{m,Z}^{-1/2}]_{j,k}b_k\right)v_j\right]^2 \\
&= \sup_{\|\vec{b}\|_{2,m}^2=1, \vec{a}=\Psi_{m,Z}^{-1/2}\vec{b}}\left[\sum_{k=0}^{m-1} b_k\left(\sum_{j=0}^{m-1}[\Psi_{m,Z}^{-1/2}]_{j,k}v_j\right)\right]^2 \\
&= \sum_{k=0}^{m-1}\left(\sum_{j=0}^{m-1}[\Psi_{m,Z}^{-1/2}]_{j,k}v_j\right)^2 = \sum_{j,\ell=0}^{m-1}\left(\sum_{k=0}^{m-1}[\Psi_{m,Z}^{-1/2}]_{j,k}[\Psi_{m,Z}^{-1/2}]_{\ell,k}\right)v_j v_\ell \\
&= {}^t\vec{v}\Psi_{m,Z}^{-1}\vec{v}.
\end{aligned}$$

Therefore, it follows from (21) in the proof of Proposition 2.2, that

$$\mathbb{E}\left(\sup_{t\in S_m, \|t\|_{S_Z}=1}[\nu_n(t)]^2\right) = \mathbb{E}({}^t\vec{v}\Psi_{m,Z}^{-1}\vec{v}) = \frac{\mathrm{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z})}{n}. \tag{24}$$

This implies that for fixed m, the trace term is increasing with m if the S_m's are nested (and this increasing with m in the inclusion sense). □

6.5. **Proof of Lemma 2.2.** For `(i)`, it follows from the remark: for $\vec{x}\in\mathbb{R}^m$ such that $\|\vec{x}\|_{2,m} = 1$,

$$\vec{x}'\Psi_{m,Z}\vec{x} = \int_A (\sum_{j=0}^{m-1} x_j\varphi_j(u))^2 S_Z(u)du \geq S_0\int_A (\sum_{j=0}^{m-1} x_j\varphi_j(u))^2 du = S_0.$$

Now the trace is nonnegative since $\mathrm{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z}) = \mathrm{Tr}(\Psi_{m,Z}^{-1/2}\Psi_{m,\lambda S_Z}\Psi_{m,Z}^{-1/2})$ where $\Psi_{m,Z}^{-1/2}$ is a symmetric square root of $\Psi_{m,Z}^{-1}$. As the matrix $\Psi_{m,Z}^{-1/2}\Psi_{m,\lambda S_Z}\Psi_{m,Z}^{-1/2}$ is nonnegative ($x'\Psi_{m,Z}^{-1/2}\Psi_{m,\lambda S_Z}\Psi_{m,Z}^{-1/2}x \geq 0$ for all m-dimensional vector x), we get the result.

To prove (ii), let $\varepsilon_0, \ldots, \varepsilon_{m-1}$ be independent centered random variables with unit variance and write:

$$\mathrm{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z}) = \mathrm{Tr}(\Psi_{m,Z}^{-1/2}\Psi_{m,\lambda S_Z}\Psi_{m,Z}^{-1/2}) = \mathbb{E}(\vec{\varepsilon}'\Psi_{m,Z}^{-1/2}\Psi_{m,\lambda S_Z}\Psi_{m,Z}^{-1/2}\vec{\varepsilon}).$$

Then for $\vec{x} = \Psi_{m,Z}^{-1/2}\vec{\varepsilon}$, we have

$$\begin{aligned}\vec{x}'\Psi_{m,\lambda S_Z}\vec{x} &= \int \left(\sum_{j=0}^{m-1} x_j\varphi_j(u)\right)^2 \lambda(u)S_Z(u)du \\ &\leq \|\lambda_A\|_\infty \int \left(\sum_{j=0}^{m-1} x_j\varphi_j(u)\right)^2 S_Z(u)du = \|\lambda_A\|_\infty \vec{x}'\Psi_{m,Z}\vec{x}.\end{aligned}$$

This implies

$$\mathrm{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z}) \leq \|\lambda_A\|_\infty \mathrm{Tr}(\mathrm{Id}_m) = m\|\lambda_A\|_\infty.$$

Inequality (iii) can be obtained from (24), as $\|t\|_{S_Z}^2 = 1 \geq S_0\|t\|^2$, so that

$$\mathbb{E}\left(\sup_{t\in S_m, \|t\|_{S_Z}=1} [\nu_n(t)]^2\right) \leq \frac{1}{S_0}\mathbb{E}({}^t\vec{v}\vec{v}) = \frac{1}{nS_0}\sum_{j=0}^{m-1}\mathbb{E}(\delta_1\varphi_j^2(X_1))$$

by using previous computations of $\mathbb{E}(v_j^2)$. We get that $\mathrm{Tr}(\Psi_{m,Z}^{-1}\Psi_{m,\lambda S_Z}) \leq L(m)/S_0 \leq c_\varphi^2 m/S_0$. □

6.6. **Proof of Lemma 2.3.** Let $\mathbf{u} = (u_0, u_1, \ldots, u_{m-1})'$ be a vector such that $\Psi_{m,Z}\mathbf{u} = 0$. Then $\mathbf{u}'\Psi_{mZ}\mathbf{u} = 0 = \|t\|_{S_Z}^2$ for $t = \sum_{j=0}^{m-1} u_j\varphi_j$. This implies, under $\lambda(\mathbb{R}^+ \cap \mathrm{Supp}(S_Z)) > 0$, that the function t is null on a set with positive Lebesgue measure. Therefore, $x \mapsto P(x) = t(x)e^x$ also; as P is a polynomials of degree $m-1$ with an infinity of zeros, it is null and thus $u_j = 0$, for $j = 0, \ldots, m-1$.

No we turn to the lower bound on $\|\Psi_{m,Z}^{-1}$. First, following the line of the proof of Lemma 8.2 in Comte and Genon-Catalot (2018), we get that, if $\mathbb{E}(Z) < 0$, there exists a constant $c_0 > 0$ such that $\int_0^{+\infty} \varphi_j^2(x)dx \leq c_0/\sqrt{j}$. This is due to the fact that s_Z is bounded, $\int_0^{+\infty} u^{-1/2}S_Z(u)du = \mathbb{E}(\sqrt{Z})$ and $\int_0^{+\infty} S_Z(u)du = \mathbb{E}(Z)$. Then, the conclusion follows as in the proof of Proposition 8 in Comte and Genon-Catalot (2018). □.

References

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

[2] Barbeito, I. and Cao, R. (2019). Smoothed bootstrap bandwidth selection for nonparametric hazard rate estimation. *J. Stat. Comput. Simul.* **89**, 15-37.

[3] Bouezmarni, T., El Ghouch, A. and Mesfioui, M. (2011) Gamma kernel estimators for density and hazard rate of right-censored data. *J. Probab. Stat.* Art. ID 937574, 16 pp.

[4] Brunel, E. and Comte, F. (2005) Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā* **67**, 441-475.

[5] Cohen, A., Davenport, M.A. and Leviatan, D. (2013). On the stability and acuracy of least squares approximations. *Found. Comput. math.* **13**, 819-834.

[6] A. Cohen, M.A. Davenport and D. Leviatan. Correction to: On the stability and accuracy of least squares approximations. *Found. Comput. math.* **19**, 239-239, 2019.

[7] Comte, F. (2017) *Nonparametric Estimation*. Paris: Spartacus-Idh, second edition.
[8] Comte, F., Gaïffas, S. and Guilloux, A. (2011) Adaptive estimation of the conditional intensity of marker-dependent counting processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **47**, 1171-1196.
[9] Diehl, S. and Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring. *J. Multivariate Anal.* **25**, 299-310.
[10] Döhler, S. and Rüschendorf, L. (2002). Adaptive estimation of hazard functions. *Probab. Math. Statist.* **22**, no. 2, 355–379.
[11] Gàmiz, M. L., Mammen, E., Martínez Miranda, M. D., Nielsen, J. P (2016) Double one-sided cross-validation of local linear hazards. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78**, 755-779.
[12] Huber, C. and MacGibbon, B. (2004). Lower bounds for estimating a hazard. *Advances in survival analysis,* 209–226, Handbook of Statist., 23 Elsevier, Amsterdam.-
[13] Klein, J. P., Moeschberger, M. L. (2003). Survival Analysis : Techniques for Censored and Truncated Data, Springer.
[14] Kooperberg, C., Stone, C.J. and Truong, Y.K. (1995). The L_2 rate of convergence for hazard regression. *Scand. J. Statist.* **22**, 143-157.
[15] Lo, S.H., Mack, Y.P. and Wang, J.L. (1989). Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. *Probab. Theory Related Fields* **80**, no. 3, 461-473.
[16] Müller, H. G. and Wang, J. L. (1994) Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* **50**, 61-76.
[17] Patil, P. N. (1997). Nonparametric hazard rate estimation by orthogonal wavelet method. *J. Statist. Plann. Inference* **60**, 153-168.
[18] Patil, P. N., Wells, M. T. and Marron, J. S. (1994). Kernel based estimation of ratio functions. *J. Nonparametric Statist.* **4**, 203-209.
[19] Plancade, S. (2011) Model selection for hazard rate estimation in presence of censoring. *Metrika* **74**, 313-347.
[20] Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11**, 453-466.
[21] Reynaud-Bouret, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli* **12**, 633-661
[22] Tanner, M. A. and Wong, W.H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.* **11**, 989-993.
[23] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434.
[24] Uzunoḡullari, U. and Wang, J.L. (1992). A comparison of hazard rate estimators for left truncated and right censored data. *Biometrika* **79**, 297-310.
[25] Wu, S.S. and Wells, M.T. (2003). Nonparametric estimation of hazard functions by wavelet methods. *Nonparametric Statistics* **15**, 187-203.
[26] Yandell, B. S. (1983). Nonparametric inference for rates with censored survival data. *Ann. Statist.* **11**, 1119-1135.