NONPARAMETRIC ESTIMATION FOR SURVIVAL DATA WITH CENSORING INDICATORS MISSING AT RANDOM

ELODIE BRUNEL[(1)], FABIENNE COMTE[(2)], AGATHE GUILLOUX[(3,4,5)]

Abstract. In this paper, we consider the problem of hazard rate estimation in presence of covariates, for survival data with censoring indicators missing at random. We propose in the context usually denoted by MAR (missing at random, in opposition to MCAR, missing completely at random, which requires an additional independence assumption), nonparametric adaptive strategies based on model selection methods for estimators admitting finite dimensional developments in functional orthonormal bases. Theoretical risks bounds are provided, they prove that the estimators behave well in term of Mean Square Integrated Error (MISE). Simulation experiments illustrate the statistical procedure.

Keywords: Missing at random - conditional hazard rate - penalized contrast estimators - risk bounds.

1. Introduction

We consider the problem of estimation from right-censored data in presence of covariates, when the censoring indicator is missing. Let T be a random variable representing the time to death from the cause of interest. Let C denote a right-censoring random time. Under usual random censorship, the observation is $Y = T \wedge C$ and $\delta = \mathbf{1}(T \leq C)$. Let X denote a real covariate. In what follows, it is assumed that T, C and X admit densities respectively denoted by f_T, g and f_X. In addition, C is assumed to be independent of T conditionally to X, see e.g. Comte *et al.* (2011) for comments on this assumption.

When the cause of death is not recorded, the censoring indicator is missing: this is the missing censoring indicator (MCI) model, see Subramanian (2006), which is defined as follows. Let ξ be the missingness indicator, that is $\xi = 1$ if δ is observed and $\xi = 0$ otherwise. The observed data are then given for individual $i \in \{1, \dots, n\}$:

$$(Y_i, X_i, \delta_i, \xi_i = 1) \quad \text{or} \quad (Y_i, X_i, \xi_i = 0).$$

We shall say that the model is:

- MCAR under the assumption that the indicator are Missing Completely At Random, i.e. ξ is independent of T, C and X.
- MAR under the assumption that the indicator is Missing At Random i.e. ξ and δ are independent conditionally to Y, X.

---

(1): (✉) Corresponding author. I3M, UMR 5149 CNRS, Université Montpellier 2, 34095 Montpellier cedex 5, FRANCE, tel +33 4 67 14 39 57, fax +33 4 67 14 35 58, email: ebrunel@math.univ-montp2.fr

(2): MAP5, UMR 8145 CNRS, Université Paris Descartes, FRANCE, email: fabienne.comte@parisdescartes.fr

(3): LSTA, Université Paris 6, FRANCE, email: agathe.guilloux@upmc.fr

(4) : Centre de Recherche Saint-Antoine (UMR S 938)

(5) :This work is supported by French Agence Nationale de la Recherche (ANR) ANR Grant "Prognostic" ANR-09-JCJC-0101-01.

In this paper, we mainly concentrate on the MAR model. The MCAR model will be consider in Section 2.2.

This model has been considered by several authors in the last decade. Most papers are interested in survival function and cumulative hazard rate estimation. In particular, van der Laan and McKeague (1998) build a sieved nonparametric maximum likelihood estimator of the survival function in the MAR case and prove its efficiency. Their estimator is a generalization of the Kaplan-Meier estimator to this context and is the first proposal reaching the efficiency bound. Subramanian (2004) also proposes an efficient estimator of the survival function in the MAR case; he proves his estimate to be efficient as well.

Kernel methods have also been used to build different estimators in the MAR context. Subramanian (2006) estimates the cumulative hazard rate with a ratio of kernel estimators. He provides an almost sure representation, and a Central Limit Theorem (CLT). He deduces results of the same type for the survival function. A study in a similar context is also provided by Wang and Ng (2008). Recently, Wang *et al.* (2009) proposed density estimator based on kernels and Kaplan Meier-type corrections of censoring. They prove a CLT and suggest a bandwidth selection strategy. Extensions of these works to conditional functions (both cumulative hazard and survival functions) in the presence of covariates is developed in Wang and Shen (2008).

Both our method and our aim are rather different. Indeed, we estimate the conditional hazard rate given a covariate. Moreover, we provide a nonparametric mean square strategy by considering approximations of the target function on finite dimensional linear spaces spanned by convenient and simple orthonormal (functional) bases. A collection of estimators is thus defined, indexed by the dimension of the multidimensional projection space, and a penalization device allows us to select a “good” space among all the proposals.

Our estimator has the advantage of being defined as a contrast minimizer and not a ratio of two estimators, as in standard kernel methodology. As a drawback, it depends on an unknown function, in its definition, which has to be replaced by an estimator, and its mean square risk has consequently the order of the anisotropic rate corresponding to the regularity of the function under estimation, plus the rate of the intermediate plug-in estimator, for which we propose a similar estimation strategy.

The plan of the paper is the following. We first explain in Section 2 how the contrast is built, and how it allows us to compute a collection of estimators. We conclude the section by giving the penalization device that completes the definition of the data driven estimator, up to an estimator to be plugged in the procedure. In Section 3, we state the theoretical results that ensure that the quadratic risk of our estimator behaves well provided that the intermediate estimator has small risk. Then, we show how similar tools can be used to build, compute and control the second estimator. The procedure is tested in a simulation section 4 for both hazard and conditional hazard rates (i.e. with or without covariate) and under different missing scheme. Technical proofs are gathered in Section 5.

2. Definition of the conditional hazard rate estimator

2.1. **Choice of the contrast.** We consider the general MAR case as described in the introduction, the global assumption is denoted **(A0)** and has several parts that we specify below.

(A0-1) The random vectors (X_i, T_i, C_i) are independent copies, for $i = 1, \dots, n$, of (X, Y, C).

(A0-2) For $i = 1, \dots, n$, we observe X_i, $Y_i = T_i \wedge C_i$, ξ_i, and $\delta_i = \mathbf{1}(T_i \leq C_i)$ if $\xi_i = 1$, otherwise $\xi_i = 0$.

(A0-3) C is independent of T given X.

(A0-4) ξ and δ are independent given X, Y.

The unknown function λ to be estimated is the conditional hazard rate of the random variable T given $X = x$ defined, for all $z > 0$ by:

$$\lambda(x,t) = \lambda_{T|X}(x,t) = \frac{f_{T|X}(x,t)}{1 - F_{T|X}(x,t)},$$

where $f_{T|X}$ and $F_{T|X}$ are respectively the conditional probability density function (p.d.f.) and the conditional cumulative distribution function (c.d.f.) of T given X. We shall denote by $G_{C|X}$ the conditional c.d.f. of C given X. We define the conditional expectations of ξ and δ by:

$$\begin{aligned}\pi(x,y) &= \mathbb{E}(\xi|X=x, Y=y) \text{ and}\\ \zeta(x,y) &= \mathbb{E}(\delta|X=x, Y=y).\end{aligned}$$

The crucial property for the construction of an estimation procedure is the following: for any integrable function h, we have

$$\begin{aligned}\mathbb{E}(\zeta(X,Y)h(X,Y)) &= \mathbb{E}[\mathbb{E}(\delta|X,Y)h(X,Y)] = \mathbb{E}(\delta h(X,Y))\\ &= \mathbb{E}[\mathbb{E}(\mathbf{1}(T \le C)h(X,T)|X)]\\ &= \mathbb{E}[(1-G_{C|X})(X,T)h(X,T)] \quad \text{with } \textbf{(A0-3)}\\ &= \iint h(x,t)(1-G_{C|X})(x,t)f_{T|X}(x,t)f_X(x)dxdt.\end{aligned}$$

This yields the equality

$$\mathbb{E}(\zeta(X,Y)h(X,Y)) = \mathbb{E}(\delta h(X,Y)) = \iint h(x,y)\lambda(x,y)d\mu(x,y) \tag{1}$$

with

$$d\mu(x,y) = (1 - L_{Y|X}(y,x))f_X(x)dxdy = f(x,y)dxdy,$$

where $f(x,y) = (1 - L_{Y|X}(y,x))f_X(x)$, and

$$1 - L_{Y|X}(y,x) := \mathbb{P}(Y \ge y|X=x) = (1 - F_{T|X}(x,y))(1 - G_{C|X}(x,y)).$$

If ζ was known, we would consider the contrast:

$$\Gamma_n^{th}(h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 h^2(X_i,y)\mathbf{1}(Y_i \ge y)dy - \frac{2}{n}\sum_{i=1}^{n}\left(\xi_i\delta_i + (1-\xi_i)\zeta(X_i,Y_i)\right)h(X_i,Y_i),$$

which is a natural extension to the MAR case of the contrast introduced in Comte *et al.* (2011). We note that, with assumption **(A0-4)** and the definition of ζ, we have

$$\begin{aligned}\mathbb{E}(\delta_i\xi_i + (1-\xi_i)\zeta(X_i,Y_i)|X_i,Y_i) &= \mathbb{E}(\delta_i|X_i,Y_i)\mathbb{E}(\xi_i|X_i,Y_i) + \mathbb{E}[(1-\xi_i)\mathbb{E}(\delta_i|X_i,Y_i)|X_i,Y_i]\\ &= \mathbb{E}(\mathbb{E}(\delta_i|X_i,Y_i)(\xi_i + (1-\xi_i))|X_i,Y_i),\end{aligned}$$

that is

$$\mathbb{E}(\delta_i\xi_i + (1-\xi_i)\zeta(X_i,Y_i)|X_i,Y_i) = \mathbb{E}(\delta_i|X_i,Y_i). \tag{2}$$

Thus, if we compute the expectation of this theoretical contrast, we obtain, under the MAR assumption and using (1) and (2),

$$\mathbb{E}\big(\Gamma_n^{th}(h)\big) = \|h\|_\mu^2 - 2\iint h(x,y)\lambda(x,y)d\mu(x,y) = \|h-\lambda\|_\mu^2 - \|\lambda\|_\mu^2.$$

Clearly, the above quantity is small if h is near of λ, and the measure denoted by μ plays the role of a reference weighting norm. This explains why minimizing Γ_n^{th} over an appropriate set of functions would be a relevant strategy to estimate λ.

As ζ is unknown, we must replace it by an estimator $\tilde{\zeta}$. Consequently, we consider

$$\Gamma_n(h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 h^2(X_i,y)\mathbf{1}(Y_i \geq y)dy - \frac{2}{n}\sum_{i=1}^{n}\left(\xi_i\delta_i + (1-\xi_i)\tilde{\zeta}(X_i,Y_i)\right)h(X_i,Y_i). \tag{3}$$

An estimator of $\zeta(x,y)$ is constructed in Section 3.3 below. This strategy of estimation of the unknown hazard rate λ, via an estimation of ζ, is also considered in Wang *et al.* (2009).

The empirical reference norm associated with the contrast (3) is defined by

$$\|h\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 h^2(X_i,y)\mathbf{1}(Y_i \geq y)dy$$

and the natural resulting scalar product is denoted by $\langle h, h_2\rangle_n = (1/4)(\|h+h_2\|_n^2 - \|h-h_2\|_n^2)$, where

$$\mathbb{E}(\langle h, h_2\rangle_n) = \langle h, h_2\rangle_\mu.$$

Remark 1. We could consider another strategy for the construction of the contrast, namely

$$\Gamma_n(h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 h^2(X_i,y)\tilde{\pi}(X_i,y)\mathbf{1}(Y_i \geq y)dy - \frac{2}{n}\sum_{i=1}^{n}\delta_i\xi_i h(X_i,Y_i). \tag{4}$$

where $\tilde{\pi}$ is an estimator of π. The second part in Equation (4) is weighted by $\xi_i\delta_i$ which means that fewer observations are used for the estimation. As a consequence, the contrast (3), that we consider, is not only more convenient (from algebraic point of view) but is also expected to be more relevant.

2.2. **The MCAR case.** In the MCAR case, the function π is constant, that is $\pi(x,y) = p = \mathbb{E}(\xi)$. The conditional hazard function can thus be estimated via the following contrast function:

$$\gamma_n^{(0)}(h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 h^2(X_i,y)\mathbf{1}(Y_i \geq y)dy - \frac{2}{n}\sum_{i=1}^{n}\frac{\delta_i\xi_i}{\hat{p}_n}h(X_i,Y_i),$$

where

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n}\xi_i.$$

Indeed, it is easy to see that, if p is known and $\hat{p}_n$ can be replaced by p, the expectation of the contrast is equal to

$$\|h\|_\mu^2 - 2\iint h(x,y)\lambda(x,y)\mathbb{E}(\mathbf{1}(Y \geq y)|X=x)f_X(x)dxdy = \|h-\lambda\|_\mu^2 - \|\lambda\|_\mu^2.$$

Also, the following contrast

$$\gamma_n^{(1)}(h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 h^2(X_i,y)\xi_i\mathbf{1}(Y_i \geq y)dy - \frac{2}{n}\sum_{i=1}^{n}\delta_i\xi_i h(X_i,Y_i), \tag{5}$$

would be adequate for conditional hazard rate estimation with reference measure

$$d\mu(x,y) = p(1 - L_{Y|X}(y,x))f_X(x)dxdy.$$

It has the advantage of not involving any estimator and the drawback that there is a ξ_i-factor in all terms, so that only observations with non missing indicator are taken into account.

Remark 2. The contrast $\gamma_n^{(1)}$, defined in (5), can be seen as a rewriting of the contrast (4) taking into account the MCAR assumption. Indeed, in that case, $\tilde{\pi}(X_i, y)$ can be replaced by ξ_i. Notice in addition that, in the simulation study, see Section 4, we used this contrast for the MCAR case because we experimented that it was giving much stabler and better results than the contrast $\gamma_n^{(0)}$.

Lastly, this contrast would be still valid for estimating λ for ξ independent of Y given X, with reference measure:

$$d\mu_2(x, y) = \pi(x)(1 - L_{Y|X}(y, x))f_X(x)dxdy,$$

and

$$\pi(x) = \mathbb{E}(\xi|X = x) = \mathbb{E}(\xi|X = x, Y = y).$$

2.3. **Computing the estimator.** We consider that we estimate the hazard rate on a compact set

$$A = A_1 \times [0, \tau],$$

where A_1 is an interval such that all observations lie in the domain. We take $A_1 = [0, 1]$ for simplicity and without loss of generality. Recall that $f(x, y) = (1-L_{Y|X}(y, x))f_X(x)$ and denote by $f_{(X,Y)}(x, y)$ the joint density of the random pair (X, Y). We set standard assumptions of boundednesses from above and below.

- **(A1)** $\forall (x, y) \in A,\ \ 0 < f_0 \le f(x, y) \le f$ for fixed positive constants f_0 and f.
- **(A2)** $\forall (x, y) \in A,\ \ \lambda(x, y) \le \|\lambda\|_{A,\infty} < +\infty$.
- **(A3)** $\forall (x, y) \in A,\ \ 0 < f_0^* < f_{(X,Y)}(x, y) \le f^* + \infty$.

First, we define an estimator $\hat{\lambda}_m$ on the space S_m by:

$$\hat{\lambda}_m = \operatorname{argmin}_{h\in S_m}\Gamma_n(h) \quad \text{where} \quad S_m = S_{m_1}^{(1)} \otimes S_{m_2}^{(2)}, \quad \text{with}$$
$$S_{m_1}^{(1)} = \operatorname{span}\{\varphi_j, j = 1, \ldots, D_{m_1}^{(1)}\} \quad \text{and} \quad S_{m_2}^{(2)} = \operatorname{span}\{\psi_k, k = 1, \ldots, D_{m_2}^{(2)}\}$$

The φ_j's, as well as the ψ_k's, constitute an $\mathbb{L}^2$-orthonormal basis, and the function h is of the form $h = \sum_{j,k} a_{j,k}\varphi_j \otimes \psi_k$.

We consider in the following two specific and classical examples of bases:

(1) **Trigonometric bases.** They are defined by $\varphi_0(x) = \mathbf{I}_{[0,1]}(x)$,

$$\varphi_{2j+1}(x) = \sqrt{2}\sin(2\pi jx)\mathbf{I}_{[0,1]}(x), \varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)\mathbf{I}_{[0,1]}(x)$$

and $\psi_0(x) = (1/\sqrt{\tau})\mathbf{I}_{[0,\tau]}(x)$,

$$\psi_{2k+1}(x) = \sqrt{2/\tau}\sin(2\pi jx/\tau)\mathbf{I}_{[0,\tau]}(x), \psi_{2k}(x) = \sqrt{2/\tau}\cos(2\pi kx/\tau)\mathbf{I}_{[0,\tau]}(x).$$

Considering $(\varphi_j)_{0\le j\le m_1-1}$ and $(\psi_k)_{0\le k\le m_2-1}$ yields spaces with odd dimensions m_1 and m_2. We denote by $\mathcal{S}_n$ the nesting space of the collection, i.e. the product space corresponding to maximal dimensions for $S_{m_1}^{(1)}$ and $S_{m_2}^{(2)}$.

(2) **Histogram bases.** They are defined by $\varphi_j(x) = \sqrt{2^{m_1}}\mathbf{I}_{[(j-1)/2^{m_1}, j/2^{m_1}[}(x)$, for $j = 1, \ldots, 2^{m_1}$ and $\psi_k(x) = \sqrt{2^{m_2}/\tau}\mathbf{I}_{[(k-1)\tau/2^{m_2}, k\tau/2^{m_2}[}(x)$, for $k = 1, \ldots, 2^{m_2}$, so that $D_{m_1}^{(1)} = 2^{m_1}, D_{m_2}^{(2)} = 2^{m_2}$. We shall take $m \le [\log_2(n)/2]$ and $m_2 \le [\log_2(n)/2]$ where $[z]$ denotes the integer part of z and $\log_2(x) = \log(x)/\log(2)$. We denote by $\mathcal{S}_n$ the nesting space of the collection, that is $\mathcal{S}_n = S_{m_1(n)}^{(1)} \otimes S_{m_2(n)}^{(2)}$, where $2^{m_1(n)}2^{m_2(n)} \le n$.

In both cases, we denote by $\mathcal{D}_n := \dim(\mathcal{S}_n) = \mathcal{D}_n^{(1)}\mathcal{D}_n^{(2)}$.
These bases are representative examples of localized bases for the second one (as piecewise polynomials, wavelets) or bounded non localized bases for the first one.

Now, let us study the contrast minimization. Writing that $\partial\Gamma_n(h)/\partial a_{j_0,k_0}$ equals

$$\frac{2}{n}\sum_{i=1}^n \Big(\int_0^\tau \varphi_{j_0}(X_i)\psi_{k_0}(y)(\sum_{j,k} a_{j,k}\varphi_j(X_i)\psi_k(y))\mathbf{1}(Y_i \geq y)dy \\ - \left(\delta_i\xi_i + (1-\xi_i)\tilde{\zeta}(X_i,Y_i)\right)\varphi_{j_0}(X_i)\psi_{k_0}(Y_i)\Big),$$

we get that the coefficients $\hat{a}_{j,k}$ of the estimate of $\hat{\lambda}_m$ verify

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \int_0^\tau \varphi_{j_0}(X_i)\psi_{k_0}(y)(\sum_{j,k}\hat{a}_{j,k}\varphi_j(X_i)\psi_k(y))\mathbf{1}(Y_i\geq y)dy \\ = \frac{1}{n}\sum_{i=1}^n \left(\delta_i\xi_i + (1-\xi_i)\tilde{\zeta}(X_i,Y_i)\right)\varphi_{j_0}(X_i)\psi_{k_0}(Y_i).\end{aligned}$$

In the histogram case, as $\varphi_j\varphi_{j'} \equiv 0$ for $j \neq j'$ and $\psi_k\psi_{k'} \equiv 0$ for $k \neq k'$, we get

$$\hat{a}_{j_0,k_0} = \frac{\sum_{i=1}^n \left(\delta_i\xi_i + (1-\xi_i)\tilde{\zeta}(X_i,Y_i)\right)\varphi_{j_0}(X_i)\psi_{k_0}(Y_i)}{\sum_{i=1}^n \int_0^\tau \varphi_{j_0}^2(X_i)\psi_{k_0}^2(y)\mathbf{1}(Y_i\geq y)dy}$$

if the denominator is non zero.

More generally, let us define the matrices

$$G_{m_1}^{\varphi}(X_i) = (\varphi_j(X_i)\varphi_{j'}(X_i))_{1\leq j,j'\leq D_{m_1}^{(1)}}, \text{ and } H_{m_2}^{\psi}(y) = (\psi_k(y)\psi_{k'}(y))_{1\leq k,k'\leq D_{m_2}^{(2)}},$$

so that their tensorial Kronecker product $G_{m_1}^{\varphi}(X_i)\otimes H_{m_2}^{\psi}(y)$ is of size $(D_{m_1}^{(1)}+D_{m_2}^{(2)})\times(D_{m_1}^{(1)}+D_{m_2}^{(2)})$. We set

$$\Theta_m := \frac{1}{n}\sum_{i=1}^n \int G_{m_1}^{\varphi}(X_i)\otimes H_{m_2}^{\psi}(y)\mathbf{1}_{\{Y_i\geq y\}}dy.$$

Recall that the vec(.) operator stacks the columns of a matrix and let

$$\begin{aligned}\overrightarrow{\hat{a}}_m &= \mathrm{vec}\left({}^t(\hat{a}_{j,k})_{1\leq j\leq D_{m_1}^{(1)},1\leq k\leq D_{m_2}^{(2)}}\right),\\ \Delta_m &= \mathrm{vec}\left(\frac{1}{n}\sum_{i=1}^n \left(\delta_i\xi_i + (1-\xi_i)\tilde{\zeta}(X_i,Y_i)\right)\varphi_j(X_i)\psi_k(Y_i))_{1\leq j\leq D_{m_1}^{(1)},1\leq k\leq D_{m_2}^{(2)}}\right),\end{aligned}$$

then the coefficients of the estimator must fulfill the matrix constraint:

$$\Theta_m\overrightarrow{\hat{a}}_m = \Delta_m.$$

It follows that the estimator is well defined if Θ_m is invertible. We define $\rho(M)$ as the spectral radius of a matrix M, i.e. the largest eigenvalue in modulus of M. We set

$$\overrightarrow{\hat{a}}_m = \Theta_m^{-1}\Delta_m \text{ if } \rho(\Theta_m) \geq \max(\hat{f}_0/3, n^{-1/2}) \tag{6}$$

and $\overrightarrow{\hat{a}}_m = 0$ otherwise.

The quantity $\hat{f}_0$ is an estimator of $f_0 = \min_{(x,y)\in A} f(x,y)$, where $f(x,y) = (1-L_{Y|X}(y,x))f_X(x)$. It is defined in Comte *et al.* (2011), and proved to satisfy, for n large enough:

(A4) For any integer $k \geq 1$, there exists a constant $C_k^{(f_0)} > 0$ such that

$$\mathbb{P}(|\hat{f}_0 - f_0| > f_0/2) \leq C_k^{(f_0)}/n^k$$

At this stage, we are in a position of defining an estimator of λ on S_m:

$$\hat{\lambda}_m(x,y) = \sum_{j,k} \hat{a}_{j,k}\varphi_j(x)\psi_k(y), \tag{7}$$

where the $\hat{a}_{j,k}$'s are defined in Equation (6).

2.4. **Model selection by penalization.** The model selection device is now based on the following criterion

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} (\Gamma_n(\hat{\lambda}_m) + \mathrm{pen}(m)) \tag{8}$$

where

$$\mathcal{M}_n = \{m = (m_1, m_2) \in \mathbb{N}\times\mathbb{N}, \dim(S_{m_2}^{(2)}) \geq \log(n), \dim(S_{m_1}^{(1)} \otimes S_{m_2}^{(2)}) \leq \mathcal{D}_n\},$$

and the penalty is defined by

$$\mathrm{pen}(m) = \kappa\|\hat{\lambda}\|_{\infty,A}\frac{\dim(S_m)}{n}, \tag{9}$$

where $\hat{\lambda} = \hat{\lambda}_{m_0}$ is an estimator in the collection, on a space S_{m_0} which is specified below. Note that the properties required on $(m_1, m_2) \in \mathcal{M}_n$ mean that all spaces of the collection are included in a nesting space with dimension $\mathcal{D}_n$. Moreover, the dimension $D_{m_2}^{(2)}$ of the space $S_{m_2}^{(2)}$ has to be larger than $\log(n)$, see the proof of Proposition 1. Lastly, we define the theoretical counterpart of the penalty:

$$\mathrm{pen}^{th}(m) = \kappa\|\lambda\|_{\infty,A}\frac{\dim(S_m)}{n}.$$

3. Results

3.1. **Main Theorem.** In order to state our Theorem 1, we have to define the integral norm with respect to $d\varrho(x,y) = f_{(X,Y)}(x,y)dxdy$ where $f_{(X,Y)}$ is the density of the bivariate vector (X,Y), that is

$$\|\psi\|_\varrho^2 = \iint \psi^2(x,y)d\varrho(x,y) = \iint \psi^2(x,y)f_{(X,Y)}(x,y)dxdy \tag{10}$$

and the associated empirical norm:

$$\|\psi\|_{\varrho,n}^2 = \frac{1}{n}\sum_{i=1}^n \psi^2(X_i, Y_i) \tag{11}$$

Theorem 1. *Let $\hat{\lambda}_{\hat{m}}$ be the estimator defined by (6)-(7)-(8)-(9). Under Assumptions* **(A1)-(A4)**, *and if $\mathcal{D}_n^2 \leq n/\log^2(n)$ for basis (1) and $\mathcal{D}_n \leq n/\log^2(n)$ for basis (2), there exists a constant κ such that, for n large enough*

$$\mathbb{E}(\|\lambda\mathbf{I}_A - \hat{\lambda}_{\hat{m}}\|_n^2) \leq C\inf_{m\in\mathcal{M}_n}(\|\lambda\mathbf{I}_A - \lambda_m\|_\mu^2 + \mathrm{pen}^{th}(m)) + C'\mathbb{E}(\|\tilde{\zeta} - \zeta\|_\varrho^2) + \frac{C''}{n}, \tag{12}$$

where C is a numerical constant and C', C'' are constants depending on f_0, f, f_0^*, f^* and $\|\lambda\|_{\infty,A}$.

The result stated in (12) involves a first term: $\inf_{m\in\mathcal{M}_n}(\|\lambda\mathbf{I}_A-\lambda_m\|^2+\text{pen}^{th}(m))$ which is the usual squared-bias ($\|\lambda\mathbf{I}_A-\lambda_m\|^2$)/variance ($\text{pen}^{th}(m)$) compromise, and will lead to an optimal anisotropic rate for a given regularity $\alpha=(\alpha,\alpha_2)$ of λ. The second term in (12) is $\mathbb{E}(\|\tilde{\zeta}-\zeta\|_\varrho^2)$, that is the mean-square risk of the estimator of ζ on A. The last term C''/n is negligible.

3.2. **Consequence on the rate.** The next corollary shows that $\hat{\lambda}_{\hat{m}}$ adapts to the unknown anisotropic smoothness of λ, up to the performance of $\tilde{\zeta}$. Toward that end, assume that λ restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ on A with regularity $\boldsymbol{\alpha}=(\alpha,\alpha_2)$. We mention that anisotropy is almost mandatory in this context, because the regularity in the covariate direction has no reason to be the same as the regularity in the y-direction.

Let us recall the definition of $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$. Let $\{e,e_2\}$ the canonical basis of $\mathbb{R}^2$ and take $A^r_{h,i}:=\{x\in\mathbb{R}^2; x,x+he_i,\dots,x+rhe_i\in A\}$, for $i=1,2$. For $x\in A^r_{h,i}$, let

$$\Delta^r_{h,i}g(x)=\sum_{k=0}^{r}(-1)^{r-k}\binom{r}{k}g(x+khe_i)$$

be the rth difference operator with step h. For $t>0$, the directional moduli of smoothness are given by

$$\omega_{r_i,i}(g,t)=\sup_{|h|\le t}\Big(\int_{A^{r_i}_{h,i}}|\Delta^{r_i}_{h,i}g(x)|^2dx\Big)^{1/2}.$$

We say that g is in the Besov space $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ if $\sup_{t>0}(t^{-\alpha}\omega_{r,1}(g,t)+t^{-\alpha_2}\omega_{r_2,2}(g,t)<\infty$ for r,r integers larger than α,α_2 respectively. More details concerning Besov spaces can be found in Triebel (2006).

Corollary 1. *Assume that λ restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ with regularity $\boldsymbol{\alpha}=(\alpha,\alpha_2)$ such that $\alpha>1/2$ and $\alpha_2>1/2$. Consider the estimator in the histogram basis. Then, under the assumptions of Theorem 1, we have*

$$\mathbb{E}\left(\|\lambda-\hat{\lambda}_{\hat{m}}\|_A^2\right)=O(n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}})+\mathbb{E}(\|\tilde{\zeta}-\zeta\mathbf{I}_A\|_\varrho^2). \tag{13}$$

where $\bar{\alpha}$ is the harmonic mean of α and α_2 (i.e. $2/\bar{\alpha}=1/\alpha+1/\alpha_2$).

The proof follows the lines of Corollary 1 (p.1178) in Comte et al. (2011). At this point, to state our final result for the estimation of λ (stated in Corollary 2), we have to construct and study an estimator of ζ.

3.3. **Estimation of $\zeta(x,y)$.** Here we want to exhibit an estimator of ζ on A for which we can prove a bound for $\mathbb{E}(\|\tilde{\zeta}-\zeta\|_\varrho^2)$. We consider the mean-square regression estimator of ζ defined as the minimizer of

$$\tilde{\gamma}_n(T)=\frac{1}{n}\sum_{i=1}^{n}[\xi_iT^2(X_i,Y_i)-2\xi_i\delta_iT(X_i,Y_i)],$$

for T in $S_m=S^{(1)}_{m_1}\otimes S^{(2)}_{m_2}$, with penalization

$$\widetilde{\text{pen}}(m)=\tilde{\kappa}\frac{\dim(S_m)}{n}.$$

Here the reference norm must be $\|.\|_\varrho$ defined by (10) but the empirical norm associated with the problem is

$$N^2_{\xi,n}(\psi)=\frac{1}{n}\sum_{i=1}^{n}\xi_i\psi^2(X_i,Y_i),\quad \mathbb{E}(N^2_{\xi,n}(\psi))=\iint\psi^2(x,y)\pi(x,y)f_{(X,Y)}(x,y)dxdy:=\|\psi\|^2_\xi.$$

We assume that there exists a constant π_0, such that:

(B1) $\forall (x,y)\in A,\ \ 0<\pi_0\le \pi(x,y)\le 1.$

If one is interested in a control of $\mathbb{E}(N^2_{\xi,n}(\hat\zeta_{\hat m}-\zeta))$, one may consider that only the vector $(\hat\zeta_m(X_i,Y_i))$ has to be correctly defined, and in this case, classical projection arguments can be used to prove that the definition is consistent without any additional tools.

But considering that our aim here is related to the estimation of conditional hazard rate of the previous section, we wish to provide a $\mathbb{L}^2$ control.

Let us consider the same bases as in Section 2.3, and the matrices

$$G^{\psi}_{m_2}(Y_i)=(\psi_k(Y_i)\psi_{k'}(Y_i))_{1\le k,k'\le D^{(2)}_{m_2}}.$$

Let us define

$$\Upsilon_m=\frac{1}{n}\sum_{i=1}^n \xi_i G^{\varphi}_{m_1}(X_i)\otimes G^{\psi}_{m_2}(Y_i).$$

If the estimate of ζ is denoted by $\hat\zeta_m(x,y)=\sum_{j,k}\hat\zeta_{j,k}\varphi_j(x)\psi_k(y)$ and $\hat Z_m=(\mathrm{vec}(\,{}^t(\hat\zeta_{j,k})_{j,k}))$ and

$$\Xi_m=\mathrm{vec}\left((\frac{1}{n}\sum_{i=1}^n \xi_i\delta_i\varphi_j(X_i)\psi_k(Y_i))_{1\le j\le D^{(1)}_{m_1},1\le k\le D^{(2)}_{m_2}}\right).$$

Then we get in the same way as previously that, if Υ_m is invertible, $\hat Z_m=\Upsilon_m^{-1}\Xi_m$ and we set more restrictively

$$\hat Z_m=\Upsilon_m^{-1}\Xi_m \text{ if } \rho(\Upsilon_m)\ge \max(\hat\rho_0/2,n^{-1/2}),$$

and $\hat Z_m=0$ otherwise. Here $\hat\rho_0$ is an estimate of ρ_0, which can be defined as the minimum of a well-chosen estimator of $\pi f_{(X,Y)}$: for instance $\hat\rho_0=\sqrt{\dim(S_{m^*})}\min_{j,k}|\hat a_{j,k}|$ where

$$\hat a_{j,k}=\frac{1}{n}\sum_{i=1}^n \xi_i\varphi_j(X_i)\psi_k(Y_i)$$

and S_{m^*} is associated to a large enough subdivision for histogram bases (φ_j) and (ψ_k). We consider the assumption

(B2) For any integer $k\ge 1$, there exists a constant $C_k^{(\rho_0)}>0$ such that $\mathbb{P}(|\hat\rho_0-\rho_0|>\rho_0/2):=\mathbb{P}(\Omega^c_{\rho_0})\le C_k^{(\rho_0)}/n^k$.

Then we have the following result bounding the $\mathbb{L}^2_\varrho$-risk of the estimator.

Theorem 2. *Under assumptions* **(A1)**, **(B1)-(B2)**, *and if $\mathcal{D}_n^2\le n/\log^2(n)$ for basis (1) and $\mathcal{D}_n\le n/\log^2(n)$ for basis (2), there exists a choice of $\tilde\kappa$ such that,*

$$\mathbb{E}(\|\hat\zeta_{\hat m}-\zeta\mathbf{1}_A\|^2_\varrho)\le C\inf_{m\in M_n}\left(\|\zeta_m-\zeta\mathbf{1}_A\|^2_\varrho+\widetilde{\mathrm{pen}}(m)\right)+\frac{C'}{n}.$$

The next corollary is an immediate consequence of Corollary 1 and Theorem 2.

Corollary 2. *Under the assumptions of Corollary 1 and assuming that ζ restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol\beta}_{2,\infty}(A)$ with regularity $\boldsymbol\beta=(\beta,\beta_2)$ such that $\beta>1/2$ and $\beta_2>1/2$. We take the estimator in the histogram basis. Then, under the assumptions of Corollary 1, the estimator $\hat\lambda_{\hat m}$ of λ verifies:*

$$\mathbb{E}\left(\|\lambda-\hat\lambda_{\hat m}\|^2_A\right)=O(n^{-\frac{2\bar\alpha}{2\bar\alpha+2}})+O(n^{-\frac{2\bar\beta}{2\bar\beta+2}}). \tag{14}$$

where $\bar\alpha$ (resp. $\bar\beta$) is the harmonic mean of α and α_2 (resp. β and β_2).

4. Simulations

To evaluate the finite sample performances of our different proposals for hazard rate estimation, we made Monte Carlo studies in different settings. We study the (possibly conditional) hazard rate estimators with or without covariate, and under both settings of dependence for the missing of censoring indicators.

4.1. **Hazard estimation.** In this paragraph, we illustrate both settings of missing indicators in the absence of covariate. Two distributions are taken for the survival time T : a Weibull distribution (with parameters a and b) and a Pareto distribution (with parameters k and t_0) associated respectively to the hazard rates :

- Weibull $\lambda(t) = ba^{-b}t^{b-1}$ with $a = 3$ and $b = 4$
- Pareto $\lambda(t) = k/t$ if $t \geq t_0$ with $k = 3$ and $t_0 = 1$.

The censoring time C is generated as an exponential random variable with parameter μ calibrated to give a censoring level around 40 % ($\mu = 3$ for the Pareto distribution and $\mu = 5$ for the Weibull distribution). Then, we set $Y = T \wedge C$ and $\delta = \mathbf{1}_{(T \leq C)}$. We follow the Missing at Random mechanism proposed by Wang *et al.* (2009) where the missing indicator ξ is generated as a binomial random variable with parameter $P(\xi = 1) = (1 + \exp(-\theta - \theta_2 Y))^{-1}$ in the MAR case, and $P(\xi = 1) = p$ a constant probability in the MCAR setting. The parameters $\theta = 0.3$ and $\theta_2 = 0.7$ were calibrated to give a non missing rate of indicators around 70 % in the MAR case, as well as the probability of missing p is taken equal to 0.7 in the MCAR one.

The estimators are obtained by minimization of the contrasts (5) and (3) given in Section 2. For the histogram basis, we can give their explicit expression in both MAR and MCAR settings : $\hat{\lambda}^{[\mathrm{H}]}_{\hat{m}_2}(y) = \sum_{j=1}^{D^{(2)}_{\hat{m}_2}} \hat{a}_k \psi_k(y)$ with the coefficient $\hat{a}_k$ having the form:

$$\hat{a}_k^{\mathrm{MAR}} = \frac{\sum_{i=1}^n [\delta_i \xi_i + (1-\xi_i)\tilde{\zeta}(Y_i)]\psi_k(Y_i)}{\sum_{i=1}^n \int_0^1 \psi_k^2(y)\mathbf{1}_{(Y_i \geq y)}dy} \quad \text{and} \quad \hat{a}_k^{\mathrm{MCAR}} = \frac{\sum_{i=1}^n \delta_i \xi_i \psi_k(Y_i)}{\sum_{i=1}^n \xi_i \int_0^1 \psi_k^2(y)\mathbf{1}_{(Y_i \geq y)}dy}$$

Histogram basis	MAR estimator under MAR setting		MCAR estimator under MCAR setting		MAR estimator under MCAR setting
Missing rate	0%	30 %	0%	30 %	30 %
Pareto					
$n = 250$	0.40	0.41	0.53	0.56	0.42
	(0.37)	(0.37)	(0.46)	(0.48)	(0.39)
$n = 1000$	0.19	0.19	0.23	0.24	0.19
	(0.19)	(0.19)	(0.22)	(0.24)	(0.18)
Weibull					
$n = 250$	0.14	0.14	0.19	0.19	0.14
	(0.12)	(0.12)	(0.15)	(0.15)	(0.12)
$n = 1000$	0.05	0.05	0.07	0.07	0.05
	(0.05)	(0.04)	(0.05)	(0.05)	(0.05)

Table 1. Average and median (in parenthesis) of the MISE over 500 replicated samples for hazard rate estimators in MAR and MCAR settings with histogram basis. Censoring rate $\simeq 40\%$

We can improve the results by using a polynomial basis of degree one. Thus, the penalized estimators are given by $\hat{\lambda}^{[\mathrm{P}]}_{\hat{m}_2}(y) = \sum_{j=1}^{D^{(2)}_{\hat{m}_2}} \hat{a}_{k,1}\psi_{k,1}(y) + \hat{a}_{k,2}\psi_{k,2}(y)$ with pairs of coefficients $(\hat{a}_{k,1}, \hat{a}_{k,2})$ for $k = 1, \cdots, D^{(2)}_{\hat{m}_2}$ solving the Cramer system:

$$\Theta_{k,\hat{m}_2} \begin{pmatrix} \hat{a}_{k,1} \\ \\ \hat{a}_{k,2} \end{pmatrix} = \Delta_{k,\hat{m}_2}$$

where, for instance in the MCAR case, we can evaluate explicitly

$$\Theta_{k,\hat{m}_2} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} \xi_i \int \psi^2_{k,1}(y)\mathbf{1}_{(Y_i \geq y)}dy & \frac{1}{n}\sum_{i=1}^{n} \xi_i \int \psi_{k,1}(y)\psi_{k,2}(y)\mathbf{1}_{(Y_i \geq y)}dy \\ \\ \frac{1}{n}\sum_{i=1}^{n} \xi_i \int \psi_{k,1}(y)\psi_{k,2}(y)\mathbf{1}_{(Y_i \geq y)}dy & \frac{1}{n}\sum_{i=1}^{n} \xi_i \int \psi^2_{k,2}(y)\mathbf{1}_{(Y_i \geq y)}dy \end{pmatrix}$$

and

$$\Delta_{k,\hat{m}_2} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} \delta_i\xi_i\psi_{k,1}(Y_i) \\ \\ \frac{1}{n}\sum_{i=1}^{n} \delta_i\xi_i\psi_{k,2}(Y_i) \end{pmatrix}.$$

Very similar expressions can be shown in the MAR setting. We also apply a curve fitting (with a mean square polynomial of degree 2) to the points $(t_i, \tilde{\lambda}(t_i))$ for smoothing the break points we have with the local basis. This obviously improves the MISE (see Tab. 2).

For $K = 500$ replications over different paths, we compute the (empirical) average MISE of the penalized estimators $\tilde{\lambda}$ over a grid of size 100:

$$MISE = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\tau_k}{100}\sum_{i=1}^{100}\left(\lambda(t_i) - \tilde{\lambda}^{(k)}(t_i)\right)^2\right),$$

where τ_k is the inter-quantile interval length associated with the 10% and 90% empirical quantiles of the Y_i's . The value of the constant κ appearing in the penalty has been calibrated and fixed to 1 for both models. We also give in parenthesis the median value of the MISE evaluated over the 500 samples. The results are summarized in Tab. 1 and 2. The three columns of Tables 1 and 2 give the MISEs of the estimator obtained with the contrast Γ_n in the MAR setting (first column) and in the MCAR setting (last column), the estimator obtained with the contrast $\gamma^{(1)}_n$ in the MCAR case is shown in the second column. As MAR and MCAR are the same when there is no missing ($\xi = 1$), the column is not repeated.

Remark that in the MCAR setting, the MAR estimator always behaves slightly better than the MCAR one, see Tables 1 and 2. At first sight, this may seem surprising but this is related to Remark 2. Indeed, the MAR estimator is obtained via the contrast (3) which is based on imputation and, as a consequence, uses all the data. On the contrary, the MCAR estimator is obtained via the contrast (5) which uses only the non missing data.

Polynomial basis	MAR estimator under MAR setting		MCAR estimator under MCAR setting		MAR estimator under MCAR setting
Missing rate	0%	30 %	0%	30 %	30 %
Weibull					
$n = 250$	0.15	0.12	0.17	0.19	0.13
	(0.08)	(0.08)	(0.09)	(0.09)	(0.08)
fitting	0.05	0.05	0.05	0.06	0.05
$n = 1000$	0.04	0.04	0.04	0.05	0.04
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)
fitting	0.03	0.02	0.02	0.02	0.02

TABLE 2. Average and median in parenthesis of the MISE over 500 replicated samples for hazard rate estimators in MAR and MCAR settings with local polynomial basis and average MISE of the *a posteriori* quadratic fitting. Censoring rate $\simeq 40\%$

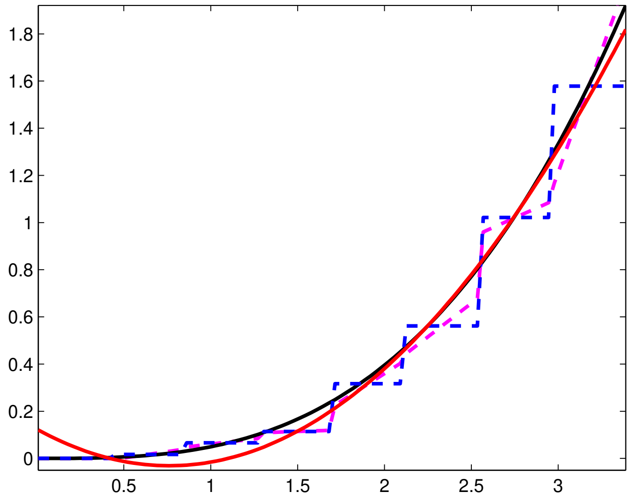

FIGURE 1. Hazard rate estimators in the MAR setting for a sample of size $n = 1000$ with 40% of censoring and 30% of missing indicators for the Weibull distribution : true hazard curve (black plain), polynomial estimator (magenta dotted line), histogram estimator (blue dotted line) and *a posteriori* quadratic fitting (red plain).

4.2. **Conditional hazard estimation.** Now, we illustrate the conditional hazard rate estimation which is the core of the theoretical part of the paper. We choose again several conditional distributions of T given $X = x$, most of them have already been studied in the literature in others contexts:

- Exponential distribution with hazard rate $\lambda(t, x) = 1 + 2x^2$,
- Pareto distribution $\lambda(t, x) = k(x)/t$ if $t \geq t_0$ with $k(x) = 3/(1 + x)$ and $t_0 = 1$.

Histogram basis	MAR estimator under MAR setting		MCAR estimator under MCAR setting		MAR estimator under MCAR setting
Missing rate	0%	30 %	0%	30 %	30 %
Exponential					
$n = 250$	2.14	0.27	0.25	0.72	0.33
	(0.20)	(0.22)	(0.21)	(0.23)	(0.22)
$n = 1000$	0.16	0.16	0.16	0.17	0.16
	(0.15)	(0.15)	(0.15)	(0.16)	(0.15)
Pareto					
$n = 250$	0.58	0.59	0.60	0.66	0.59
	(0.59)	(0.61)	(0.61)	(0.66)	(0.59)
$n = 1000$	0.35	0.35	0.42	0.53	0.35
	(0.34)	(0.34)	(0.41)	(0.55)	(0.34)
Non linear					
$n = 250$	0.60	0.36	0.44	0.37	0.19
	(0.17)	(0.18)	(0.17)	(0.18)	(0.17)
$n = 1000$	0.14	0.14	0.14	0.16	0.14
	(0.15)	(0.15)	(0.15)	(0.17)	(0.16)

TABLE 3. Average and median of the MISE over 500 replicated samples for conditional hazard rate estimators in MAR and MCAR settings with histogram basis. Censoring rate $\simeq 40\%$

- Non-linear regression $T = 2X + 5 + \sigma\varepsilon$ where the error process ε has a $\chi^2(4)$ distribution, $\sigma = 0.25$. The resulting hazard rate is $\lambda(t, x) = \frac{1}{\sigma}\lambda_\varepsilon\left(\frac{t-2x-5}{\sigma}\right)$ and λ_ε is the hazard rate of the error ε.

Note that these models are drawn from Wang and Shen (2008) or Zhou and Sun (2003) for the Exponential model, and Comte et al. (2011) for the non-linear regression, the Pareto distribution can be viewed as a generalization of the unconditional setting of the previous paragraph.

The censoring time C has an exponential distribution given X with mean $\mu(X)$ depending on the model to be censored: $\mu(x) = 2/(1+3x)$ for the exponential model, $\mu(x) = 3.8/(1+0.1x)$ for the Pareto model and $\mu(x) = 2x+15$ for the non-linear regression model, each one corresponding to a censoring level around 40%. The covariate X has uniform distribution on the interval $[0, 1]$.

The MAR mechanism proceeds from the conditional probability function π of the logistic model:

$$\pi(x, y) = P(\xi = 1|X = x, Y = y) = \frac{1}{1 + \exp(cx + c_2 y)}.$$

The parameters c and c_2 were adjusted to produce a missing rate around 30 % with $c = -0.5$ for all the models, $c_2 = -2; -0.5; -0.1$ for the Exponential, Pareto and non-linear regression model respectively.

The constant κ has been roughly calibrated over the three models, we compute the average MISE and medians for several values of κ from 1 to 10 and the value $\kappa = 5$ seems to give a good compromise. For $K = 500$ replications over different paths, we compute the (empirical) average

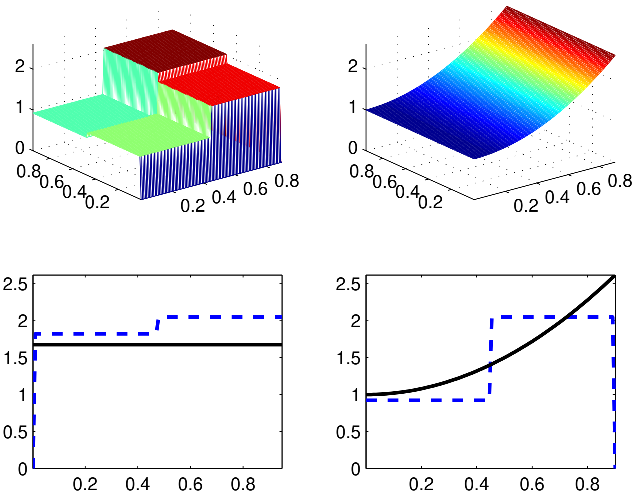

FIGURE 2. Conditional Hazard rate estimators in the MAR setting for a sample of size $n = 1000$ with 40% of censoring and 30% of missing indicators for the Exponential distribution : true hazard curve (right-top), histogram estimator (left-top), $\lambda(t,x)$ (solid black) and $\tilde{\lambda}(t,x)$ (dashed blue) with $x = 0.58$ (left-bottom) and $t = 0.61$ (right-bottom).

MISE of the penalized estimators $\tilde{\lambda}$ over a grid of size 100×100:

$$MISE = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\ell(A_{1,k})\tau_k}{(100)^2}\sum_{i,j=1}^{100}\left(\lambda(t_i,x_j) - \tilde{\lambda}^{(k)}(t_i,x_j)\right)^2\right)$$

where τ_k is defined as in the unconditional setting (see paragraph 4.1) and $\ell(A_{1,k})$ is the length of the range interval of the X_i's. We compute also the median of the empirical error over the $K = 500$ replicated samples. The results are summarized in Tab. 3. For the non linear model, both estimators are unstable for small samples: the mean errors are not of the same order as the median error. Moreover, we can see that the MAR estimator gives systematically better results than the MCAR one, as in the unconditional setting. In conclusion, we recommend the systematic use of the MAR estimator when censoring indicators are missing.

5. Proofs

5.1. **Proof of Theorem 1.** Note that the two bases we use satisfy the following property. For all h in $S_{m_1}^{(1)} \otimes S_{m_2}^{(2)}$,

$$\|h\|_\infty := \sup_{(x,y)\in A\times[0,\tau]} |h(x,y)| \le \sqrt{D_{m_1}^{(1)} D_{m_2}^{(2)}}\|h\|,$$

where $\|h\| = \int_A h^2$. Moreover, we recall that all S_m's are subsets of a nesting space belonging to the collection denoted by $\mathcal{S}_n$ with dimension $\mathcal{D}_n$.

Let λ_A denote the restriction of the unknown function λ to A. For any $h, h_2 \in (L^2 \cap L^\infty)(A)$, we have

$$\Gamma_n(h) - \Gamma_n(h_2) = \|h-\lambda_A\|_n^2 - \|h_2-\lambda_A\|_n^2 - 2\nu_n(h-h_2) - 2R_n(h-h_2) \tag{15}$$

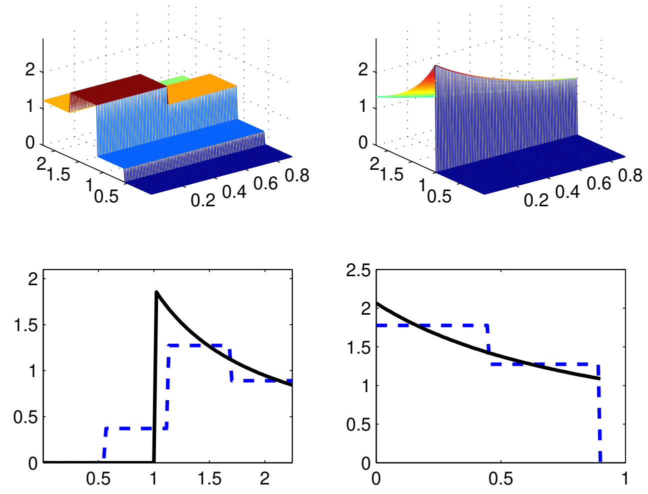

FIGURE 3. Conditional Hazard rate estimators in the MAR setting for a sample of size $n = 1000$ with 40% of censoring and 30% of missing indicators for the Pareto distribution : true hazard curve (right-top), histogram estimator (left-top), $\lambda(t,x)$ (solid black) and $\tilde{\lambda}(t,x)$ (dashed blue) with $x = 0.58$ (left-bottom) and $t = 1.55$ (right-bottom).

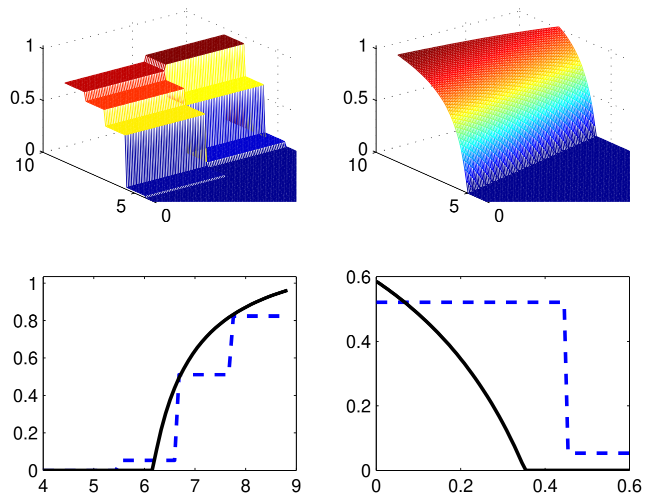

FIGURE 4. Conditional Hazard rate estimators in the MAR setting for a sample of size $n = 1000$ with 40% of censoring and 30% of missing indicators for the Non linear regression model : true hazard curve (right-top), histogram estimator (left-top), $\lambda(t,x)$ (solid black) and $\tilde{\lambda}(t,x)$ (dashed blue) with $x = 0.58$ (left-bottom) and $t = 7.9$ (right-bottom).

where ν_n is the centered empirical process defined by:

$$\nu_n(h) = \frac{1}{n}\sum_{i=1}^{n}\left((\delta_i\xi_i + (1-\xi_i)\zeta(X_i,Y_i))\, h(X_i,Y_i) - \int h(X_i,y)\lambda(X_i,y)\mathbf{I}(Y_i \geq y)dy\right).$$

and the term R_n is defined as:

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n}(1-\xi_i)\left(\tilde{\zeta}(X_i,Y_i) - \zeta(X_i,Y_i)\right) h(X_i,Y_i)$$

Many steps of the proof below will refer to the one given in Comte *et al.* (2011), the main difference lies in the additional term R_n.

We shall use the following sets:

$$\hat{G}_m = \{\min \mathrm{Sp}(\Theta_m) \geq \max(\hat{f}_0/3, n^{-1/2})\},\ \ \hat{G} := \bigcap_{m\in\mathcal{M}_n} \hat{G}_m,$$

$$\Delta_\mu := \left\{\forall h \in \mathcal{S}_n : \left|\frac{\|h\|_n^2}{\|h\|_\mu^2} - 1\right| \leq \frac{1}{2}\right\},\ \ \Delta_\varrho := \left\{\forall h \in \mathcal{S}_n : \left|\frac{\|h\|_{\varrho,n}^2}{\|h\|_\varrho^2} - 1\right| \leq \frac{1}{2}\right\},\ \text{and}$$

$$\Omega_{f_0} := \left\{\left|\frac{\hat{f}_0}{f_0} - 1\right| \leq \frac{1}{2}\right\}. \tag{16}$$

We now define $\Omega = \Delta_\mu \cap \Delta_\varrho \cap \Omega_{f_0}$ and simply write

$$\mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda_A\|_n^2) = \mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda_A\|_n^2\mathbf{I}_\Omega) + \mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda_A\|_n^2\mathbf{I}_{\Omega^\complement}).$$

The second term is bounded by:

$$\mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda_A\|_n^2\mathbf{I}_{\Omega^\complement}) \leq 2\left(\mathbb{E}(\|\hat{\lambda}_{\hat{m}}\|_n^2\mathbf{I}_{\Omega^\complement}) + \mathbb{E}(\|\lambda_A\|_n^2\mathbf{I}_{\Omega^\complement})\right)$$

Now, the following lemma, proved in Section 5.2, holds.

Lemma 1. *Under the assumptions of Theorem 1, we have* $\|\hat{\lambda}_{\hat{m}}\|_n^2 \leq n^3$.

As moreover, $\|\lambda_A\|_n^2 \leq \|\lambda\|_{\infty,A}^2$ a.s., this yields that there exists a constant C such that

$$\mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda_A\|_n^2\mathbf{I}_{\Omega^\complement}) \leq C(n^3 + \|\lambda\|_{A,\infty})\left(\mathbb{P}(\Delta_\mu^\complement) + \mathbb{P}(\Omega_{f_0}^\complement) + \mathbb{P}(\Delta_\varrho^\complement)\right).$$

Next, Assumption **(A4)** ensures that $\mathbb{P}(\Omega_{f_0}^\complement) \leq C_k^{(f_0)}/n^k$ and Proposition 4 in Comte *et al.* (2011) can be used here to get $\mathbb{P}(\Delta_\mu^\complement) \leq C_k^{(\Delta)}/n^k$, under the condition that $\mathcal{D}_n^2 \leq n/\log^2(n)$ for basis (1) and $\mathcal{D}_n \leq n/\log^2(n)$ for basis (2). And for the third term, we prove below the following bound.

Lemma 2. *Under the assumptions of Theorem 1, we have* $\mathbb{P}(\Delta_\varrho^\complement) \leq C_k^{(\Delta)}/n^k$ *for any* $k \geq 1$, *where* $C_k^{(\Delta)}$ *is a constant depending on* k, *on the basis, and on* f_0^*, f^*.

Gathering these elements yields $\mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda_A\|_n^2\mathbf{I}_{\Omega^\complement}) \leq C/n$ by choosing $k = 4$.

Now, we study $\mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda_A\|_n^2\mathbf{I}_\Omega)$. From Lemma 1 in Comte *et al.* (2011), we know that $\Delta_\mu \cap \Omega_{f_0} \subset \hat{G} \cap \Omega_{f_0}$. Hence on Ω, we can use the definition of $\hat{\lambda}_{\hat{m}}$ given via (6) . We introduce for $m \in \mathcal{M}_n$ the orthogonal projection on S_m of λ restricted to A, denoted by λ_m. We write that, on Ω, $\forall m \in \mathcal{M}_n$,

$$\Gamma_n(\hat{\lambda}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \Gamma_n(\lambda_m) + \mathrm{pen}(m),$$

and we use the decomposition (15). We get on Ω and $\forall m \in \mathcal{M}_n$:

$$\|\hat\lambda_{\hat m} - \lambda_A\|_n^2 \le \|\lambda_m - \lambda_A\|_n^2 + \mathrm{pen}(m) + 2\nu_n(\hat\lambda_{\hat m} - \lambda_m) - \mathrm{pen}(\hat m) + 2R_n(\hat\lambda_{\hat m} - \lambda_m). \tag{17}$$

We write

$$2|\nu_n(\hat\lambda_{\hat m} - \lambda_m)| \le \frac{1}{8}\|\hat\lambda_{\hat m} - \lambda_m\|_\mu^2 + 8 \sup_{h \in B_{m,\hat m}(0,1)} \nu_n^2(h),$$

where $B_{m,m'}(0,1) = \{h \in S_m + S_{m'}, \|h\|_\mu \le 1\}$. It follows that on Ω

$$\begin{aligned} 2|\nu_n(\hat\lambda_{\hat m} - \lambda_m)| &\le \frac{1}{4}\|\hat\lambda_{\hat m} - \lambda_m\|_n^2 + 8\left(\sup_{h \in B_{m,\hat m}(0,1)} \nu_n^2(h) - p(m,\hat m)\right)_+ + 8p(m,\hat m), \\ &\le \frac{1}{2}\|\hat\lambda_{\hat m} - \lambda_A\|_n^2 + \frac{1}{2}\|\lambda_m - \lambda_A\|_n^2 + 8\left(\sup_{h \in B_{m,\hat m}(0,1)} \nu_n^2(h) - p(m,\hat m)\right)_+ \\ &\quad + 8p(m,\hat m), \end{aligned} \tag{18}$$

where $p(m,m')$ will be define later. Let us define $\zeta_A = \zeta \mathbf{I}(A)$. Using the inequality $2xy \le x^2/a + ay^2$, we get on Ω

$$\begin{aligned} 2\left|R_n(\hat\lambda_{\hat m} - \lambda_m)\right| &= \frac{2}{n}\left|\sum_{i=1}^n (1-\xi_i)\left(\tilde\zeta(X_i,Y_i) - \zeta_A(X_i,Y_i)\right)(\hat\lambda_{\hat m} - \lambda_m)(X_i,Y_i)\right| \\ &\le \frac{1}{an}\sum_{i=1}^n (1-\xi_i)^2(\hat\lambda_{\hat m} - \lambda_m)^2(X_i,Y_i) + \frac{a}{n}\sum_{i=1}^n \left(\tilde\zeta(X_i,Y_i) - \zeta_A(X_i,Y_i)\right)^2 \\ &\le \frac{1}{a}\|\hat\lambda_{\hat m} - \lambda_m\|_{\varrho,n}^2 + a\|\tilde\zeta - \zeta_A\|_{\varrho,n}^2 \le \frac{3}{2a}\|\hat\lambda_{\hat m} - \lambda_m\|_\varrho^2 + \frac{3a}{2}\|\tilde\zeta - \zeta_A\|_\varrho^2 \text{ (on } \Delta_\varrho) \\ &\le \frac{3f^*}{2f_0 a}\|\hat\lambda_{\hat m} - \lambda_m\|_\mu^2 + \frac{3a}{2}\|\tilde\zeta - \zeta_A\|_\varrho^2 \le \frac{3f^*}{f_0 a}\|\hat\lambda_{\hat m} - \lambda_m\|_n^2 + \frac{3a}{2}\|\tilde\zeta - \zeta_A\|_\varrho^2 \text{ (with } \mathbf{A1} \text{ and } \mathbf{A3}) \\ &\le \frac{6f^*}{f_0 a}\|\hat\lambda_{\hat m} - \lambda_A\|_n^2 + \frac{6f^*}{f_0 a}\|\lambda_A - \lambda_m\|_n^2 + \frac{3a}{2}\|\tilde\zeta - \zeta_A\|_\varrho^2 \end{aligned} \tag{19}$$

Now, choosing $a = 24f^*/f_0$ and gathering (17)–(19), we get, as $1 - 1/2 - 1/4 = 1/4$ and $1 + 1/2 + 1/4 = 7/4$,

$$\begin{aligned} \frac{1}{4}\mathbb{E}(\|\hat\lambda_{\hat m} - \lambda_A\|_n^2 \mathbf{I}_\Omega) &\le \frac{7}{4}\|\lambda_m - \lambda_A\|_\mu^2 + \mathrm{pen}(m) + \frac{36f^*}{f_0}\mathbb{E}(\|\tilde\zeta - \zeta_A\|_\varrho^2) \\ &\quad + 8\mathbb{E}\left(\left(\sup_{h \in B_{m,\hat m}(0,1)} \nu_n^2(h) - p(m,\hat m)\right)_+\right) + 8\mathbb{E}(p(m,\hat m)) - \mathbb{E}\mathrm{pen}(\hat m). \end{aligned} \tag{20}$$

We can use Talagrand Inequality to prove:

Proposition 1. *Under the Assumptions of Theorem 1, there exists a numerical constant κ such that, for*

$$p(m,m') = (\kappa/8)\|\lambda_A\|_{\infty,A}\frac{D_m + D_{m'}}{n},$$

we have

$$\mathbb{E}\left(\sup_{h \in B_{m,\hat m}(0,1)} (\nu_n^2(h) - p(m,\hat m))_+\right) \le \frac{C}{n}.$$

This inequality, together with (20) yields, as $8p(m,m') \le \mathrm{pen}(m) + \mathrm{pen}(m')$, that

$$\frac{1}{4}\mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda_A\|_n^2 \mathbf{I}_\Omega) \le \frac{7}{4}\|\lambda_m - \lambda_A\|_\mu^2 + 2\mathrm{pen}(m) + \frac{C}{n} + \frac{36 f^*}{f_0}\mathbb{E}(\|\tilde{\zeta} - \zeta_A\|_\varrho^2).\square$$

5.2. **Proof of Lemma 1.** Let us note that $\hat{\lambda}_{\hat{m}}$ is either 0 or $\arg\min_{h\in S_{\hat{m}}} \Gamma_n(h)$. In the second case, $\min \mathrm{Sp}(\Theta_{\hat{m}}) \ge \max(\hat{f}_0, n^{-1/2})$ and thus

$$\begin{aligned}
\|\hat{\lambda}_{\hat{m}}\|^2 &= \sum_{j,k} (\hat{a}_{j,k}^{\hat{m}})^2 = \|\overrightarrow{\hat{a}}_{\hat{m}}\|^2 = \|\Theta_{\hat{m}}^{-1}\Delta_{\hat{m}}\|^2 \le (1/\min \mathrm{Sp}(\Theta_{\hat{m}}))^2 \|\Delta_{\hat{m}}\|^2 \\
&\le \min(1/(\hat{f}_0)^2, n) \sum_{j,k}\left(\frac{1}{n}\sum_{i=1}^n \left(\delta_i\xi_i + (1-\xi_i)\zeta_A(X_i,Y_i)\right)\varphi_j^{\hat{m}}(X_i)\psi_k^{\hat{m}}(Y_i)\right)^2 \\
&\le n\frac{1}{n}\sum_{i=1}^n \sum_j (\varphi_j^{\hat{m}}(X_i))^2 \sum_k (\psi_k^{\hat{m}}(Y_i))^2 \le n\mathcal{D}_n^{(1)}\mathcal{D}_n^{(2)} \le n^2.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\|\hat{\lambda}_{\hat{m}}\|_n^2 &\le \frac{1}{n}\sum_{i=1}^n \int \hat{\lambda}_{\hat{m}}^2(X_i,y)dy = \frac{1}{n}\sum_{i=1}^n \int \left(\sum_{j,k} \hat{a}_{j,k}^{\hat{m}}\varphi_j^{\hat{m}}(X_i)\psi_k^{\hat{m}}(y)dy\right)^2 dy \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{j,k,j',k'} \hat{a}_{j,k}^{\hat{m}}\hat{a}_{j',k'}^{\hat{m}}\varphi_j^{\hat{m}}(X_i)\varphi_{j'}^{\hat{m}}(X_i)\int \psi_k^{\hat{m}}(y)\psi_{k'}^{\hat{m}}(y)dy \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{j,k,j'} \hat{a}_{j,k}^{\hat{m}}\hat{a}_{j',k}^{\hat{m}}\varphi_j^{\hat{m}}(X_i)\varphi_{j'}^{\hat{m}}(X_i) \\
&\le \frac{1}{n}\sum_{i=1}^n \sum_j [\varphi_j^{\hat{m}}(X_i)]^2 \sum_{j,k}(\hat{a}_{j,k}^{\hat{m}})^2 \le \mathcal{D}_n^{(1)}\|\hat{\lambda}_{\hat{m}}\|^2.
\end{aligned}$$

Gathering both bounds yields the result as $\mathcal{D}_n^{(1)} \le n$. $\square$

5.3. **Proof of Proposition 1.** First, we write

$$\mathbb{E}\left(\sup_{h\in B_{m,\hat{m}}(0,1)} (\nu_n^2(h) - p(m,\hat{m}))_+\right) \le \sum_{m'\in\mathcal{M}_n} \mathbb{E}\left(\sup_{h\in B_{m,m'}(0,1)} (\nu_n^2(h) - p(m,m'))_+\right)$$

and we bound

$$\mathbb{E}\left(\sup_{h\in B_{m,m'}(0,1)} (\nu_n^2(h) - p(m,m'))_+\right).$$

Next, we split $\nu_n(h) = \nu_n^{(1)}(h) + \nu_n^{(2)}(h)$ with

$$\nu_n^{(1)}(h) = (1/n)\sum_{i=1}^n (f_h(X_i,Y_i,\delta_i,\xi_i) - \langle h,\lambda\rangle_\mu),\ f_h(x,y,\delta,\xi) = (\delta\xi + (1-\xi)\zeta(x,y))\,h(x,y)$$

and

$$\nu_n^{(2)}(h) = (1/n)\sum_{i=1}^n \left(g_h(X_i,Y_i) - \langle h,\lambda\rangle_\mu\right),\ g_h(x,y) = \int h(x,u)\lambda(x,u)\mathbf{I}_{\{y\ge u\}}du.$$

We have

$$\begin{aligned}\mathbb{E}\left(\sup_{h\in B_{m,m'}(0,1)}(\nu_n^2(h)-p(m,m'))_+\right) &\leq 2\mathbb{E}\left(\sup_{h\in B_{m,m'}(0,1)}([\nu_n^{(1)}(h)]^2-p(m,m')/2)_+\right)\\ &+2\mathbb{E}\left(\sup_{h\in B_{m,m'}(0,1)}([\nu_n^{(2)}(h)]^2-p(m,m')/2)_+\right).\end{aligned}$$

In both cases, we apply Talagrand Inequality.

As $S_m+S_{m'}$ is a finite dimension space with dimension less than $\dim(S_m)+\dim(S_{m'})$, we can find by Gram-Schmidt orthonormalisation a basis $\phi_{j,k}(x,y)$ which is orthonormal with respect to the $\mathbb{L}^2(\mu)$-norm, with cardinal equal to the dimension and thus less than $\dim(S_m)+\dim(S_{m'})$. The functions are such that $\iint\phi_{j,k}^2(x,y)d\mu(x,y)=1$. It follows that

$$\begin{aligned}&\mathbb{E}\left(\sup_{h\in B_{m,m'}(0,1)}([\nu_n^{(1)}]^2(h)\right)\\ \leq\ & \sum_{j,k}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n f_{\phi_{j,k}}(X_i,Y_i,\delta_i,\xi_i)\right)=\frac{1}{n}\sum_{j,k}\mathrm{Var}\left(f_{\phi_{j,k}}(X,Y,\delta,\xi)\right)\\ \leq\ & \frac{2}{n}\sum_{j,k}\mathbb{E}\big[\big(\zeta_A(X,Y)\pi_A(X,Y)+(\mathbf{1}(A)-\pi_A(X,Y))\zeta_A(X,Y)\big)\phi_{j,k}^2(X,Y)\big]=\frac{2}{n}\sum_{j,k}\langle\phi_{j,k}^2,\lambda_A\rangle_\mu\\ \leq\ & \frac{2}{n}\sum_{j,k}\|\lambda\|_{A,\infty}\iint\phi_{j,k}^2(x,y)d\mu(x,y)=2\|\lambda\|_{A,\infty}\frac{\dim(S_m+S_{m'})}{n}\\ \leq\ & 2\|\lambda\|_{A,\infty}\frac{\dim(S_m)+\dim(S_{m'})}{n}:=H^2.\end{aligned}$$

For the other empirical process, we first write that

$$\mathbb{E}\left(\sup_{h\in B_{m,m'}(0,1)}([\nu_n^{(2)}]^2(h)\right)\leq\frac{1}{f_0}\mathbb{E}\left(\sup_{\|h\|=1,h\in S_m+S_{m'}}([\nu_n^{(2)}]^2(h)\right)$$

and we consider a basis of $(S_{m_1}+S_{m_1'})\otimes(S_{m_2}+S_{m_2'})=S_{m_1\vee m_1'}\otimes S_{m_2\vee m_2'}$ as both collections are nested. Then we can take basis $(\varphi_j\otimes\psi_k)_{j,k}$ and we write

$$\begin{aligned}&\mathbb{E}\left(\sup_{\|h\|=1,h\in S_m+S_{m'}}([\nu_n^{(2)}]^2(h)\right)\\ \leq\ & \sum_{j,k}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n g_{\varphi_j\otimes\psi_k}(X_i,Y_i)\right)\leq\frac{1}{n}\sum_{j,k}\mathbb{E}\left(\left(\int\varphi_j(X)\psi_k(u)\lambda_A(X,u)\mathbf{1}_{\{Y\geq u\}}du\right)^2\right).\end{aligned}$$

Now we note that

$$\sum_k\left(\int\psi_k(u)\lambda_A(X,u)\mathbf{1}_{\{Y\geq u\}}du\right)^2=\sum_k\langle\psi_k,\lambda_A(X,.)\mathbf{1}_{\{Y\leq .\}}\rangle^2$$

which is equal to the $\mathbb{L}^2-$norm of the projection on $S_{m_2\vee m_2'}$ of $y\mapsto\lambda(X,y)\mathbf{1}_{\{Y\leq y\}}$ and thus less than the $\mathbb{L}^2-$norm of the function. In other words,

$$\sum_k\left(\int\psi_k(u)\lambda_A(X,u)\mathbf{1}_{\{Y\geq u\}}du\right)^2\leq\int\lambda_A^2(X,u)\mathbf{1}_{\{Y\geq u\}}du$$

and

$$\begin{aligned}
&\mathbb{E}\left(\sup_{h\in B_{m,m'}(0,1)}([\nu_n^{(2)}]^2(h)\right)\\
\leq\ & \frac{1}{nf_0}\sum_j \mathbb{E}\left(\varphi_j^2(X)\int\lambda_A^2(X,u)\mathbf{1}_{\{Y\geq u\}}du\right)\\
\leq\ & \frac{D_m\vee D_{m'}}{nf_0}\mathbb{E}\left(\int\lambda_A^2(X,u)\mathbf{1}_{\{Y\geq u\}}du\right)=\frac{(D_m\vee D_{m'})\|\lambda_A\|_\mu^2}{nf_0}\leq H^2
\end{aligned}$$

since

$$(D_m\vee D_{m'})\|\lambda_A\|_\mu^2/f_0\leq\|\lambda\|_{A,\infty}(D_m+D_{m'})\log(n)\leq\|\lambda\|_{A,\infty}(D_{m_1}D_{m_2}+D_{m_1'}D_{m_2'}),$$

for $n\geq n_0:=\exp(\|\lambda_A\|_\mu^2/(f_0\|\lambda\|_{A,\infty}))$ and $\log(n)\leq D_{m_2},\log(n)\leq D_{m_2'}$.

Next we consider that there is a nesting space $\mathcal{S}_n$ for all the S_m's, that is $\forall m\in\mathcal{M}_n$, $S_m\subset\mathcal{S}_n$. Thus for $h\in B_{m,m'}(0,1)$, from property (**M1**), we have $\|h\|_\infty\leq\sqrt{\dim(\mathcal{S}_n)}\|h\|\leq\sqrt{\dim(\mathcal{S}_n)}\|h\|_\mu/f_0=\phi_0\sqrt{\dim(\mathcal{S}_n)}/f_0\leq\sqrt{n}/f_0$. Therefore

$$\sup_{(x,y)\in A}|f_h(x,y,\delta,\xi)|\leq\|h\|_\infty\leq\sqrt{n}/f_0:=M$$

and

$$\sup_{(x,y)\in A}|g_h(x,y)|\leq\|h\|_\infty\leq\tau\|\lambda\|_\infty/f_0:=M'.$$

Lastly

$$\sup_{h\in B_{m,m'}(0,1)}\mathrm{Var}(f_h(X,Y,\delta,\xi))\leq 2\|\lambda\|_{A,\infty}:=v,\quad \sup_{h\in B_{m,m'}(0,1)}\mathrm{Var}(g_h(X,Y))\leq\tau\|\lambda\|_{A,\infty}:=v_2.$$

By applying Talagrand Inequality, we get for $i=1,2$,

$$\mathbb{E}\left(\sup_{h\in B_{m,m'}(0,1)}([\nu_n^{(i)}(h)]^2-p(m,m'))_+\right)\leq\frac{C^{(i)}}{n}\left(e^{-C_2^{(i)}(D_m+D_{m'})}+e^{-C_3^{(i)}\sqrt{D_m+D_{m'}}}\right)$$

and this yields the result. □

5.3.1. *Proof of Lemma 2.* First we observe that:

$$\mathbb{P}(\Delta_\varrho^{\complement})\leq\mathbb{P}\Big(\sup_{h\in B_{\mathcal{S}_n}^\varrho(0,1)}|\vartheta_n(h^2)|>1/2\Big)$$

where $\vartheta_n(\cdot)$ is defined by

$$\vartheta_n(h)=\frac{1}{n}\sum_{i=1}^n[h(X_i,Y_i)-\mathbb{E}(h(X_i,Y_i))],$$

and $B_{\mathcal{S}_n}^\mu(0,1)=\{t\in\mathcal{S}_n,\|h\|_\varrho\leq 1\}$. We denote by $(\chi_\lambda)=(\varphi_j\otimes\psi_k)$ the L^2-orthonormal basis of $\mathcal{S}_n$. If $h(x,y)=\sum_{j,k}a_{j,k}\varphi_j(x)\psi_k(y)=\sum_\lambda a_\lambda\chi_\lambda$, then

$$\vartheta_n(h^2)=\sum_{j,k,j',k'}a_{j,k}a_{j',k'}\vartheta_n((\varphi_j\otimes\psi_k)(\varphi_{j'}\otimes\psi_{k'}))=\sum_{\lambda,\lambda'}a_\lambda a_{\lambda'}\vartheta_n(\chi_\lambda\chi_{\lambda'}).\tag{21}$$

We obtain

$$\sup_{h\in B_{\mathcal{S}_n}^\varrho(0,1)}|\vartheta_n(h^2)|\leq(f_0^*)^{-1}\sup_{\sum a_\lambda^2\leq 1}\Big|\sum_{\lambda,\lambda'}a_\lambda a_{\lambda'}\vartheta_n((\chi_\lambda)(\chi_{\lambda'}))\Big|.\tag{22}$$

Lemma 3. *Baraud et al. (2001) Let* $B_{\lambda,\lambda'} = \|\chi_\lambda \chi_{\lambda'}\|_\infty$ *and* $V_{\lambda,\lambda'} = \|\chi_\lambda \chi_{\lambda'}\|_2$. *Let, for any symmetric matrix* $(A_{\lambda,\lambda'})$

$$\bar\rho(A) := \sup_{\sum b_\lambda^2 \le 1} \sum_{\lambda,\lambda'} |b_\lambda b_{\lambda'}| A_{\lambda,\lambda'}$$

and $L(\chi) := \max\{\bar\rho^2(V), \bar\rho(B)\}$. *Then, if* φ_j *and* ψ_k *are trigonometric bases, we have* $L(\chi) \le 4(\mathcal{D}_n)^2$, *and if* φ_j *and* ψ_k *are histogram bases* $L(\chi) \le \mathcal{D}_n$ *(and more generally if the bases are localized).*

Let us define

$$\begin{aligned} x &:= \frac{(f_0^*)^2}{16 f^* L(\chi)} \text{ and} \\ \Theta &:= \Big\{\forall\lambda\forall\lambda' \quad |\vartheta_n(\chi_\lambda\chi_{\lambda'})| \le 4\Big(B_{\lambda,\lambda'} x + V_{\lambda,\lambda'}\sqrt{2f^* x}\Big)\Big\}. \end{aligned}$$

Starting from (22), we have, on Θ:

$$\sup_{h\in B^{\varrho}_{\mathcal{S}_n}(0,1)} |\vartheta_n(h^2)| \le 4(f_0^*)^{-1} \sup_{\sum a_\lambda^2 \le 1} \sum_{\lambda\lambda'} |a_\lambda a_{\lambda'}|)\Big(B_{\lambda,\lambda'} x + V_{\lambda,\lambda'}\sqrt{2f^* x}\Big).$$

Thus, we have on Θ,

$$\begin{aligned} \sup_{h\in B^{\varrho}_{\mathcal{S}_n}(0,1)} |\vartheta_n(h^2)| &\le (f_0^*)^{-1} \sup_{\sum b_\lambda^2 = 1} \sum_{\lambda,\lambda'} |b_\lambda b_{\lambda'}| \Big(B_{\lambda,\lambda'} x + V_{\lambda,\lambda'}\sqrt{2f^* x}\Big) \\ &\le (f_0^*)^{-1}\Big(\bar\rho(B) x + \bar\rho(V)\sqrt{2f^* x}\Big) \\ &\le \frac{1}{2}\Big(\frac{f_0^*}{8f^*}\frac{\bar\rho(B)}{L(\chi)} + \frac{1}{\sqrt2}\Big(\frac{\bar\rho^2(V)}{L(\chi)}\Big)^{1/2}\Big) \\ &\le \frac12\Big(\frac18 + \frac{1}{\sqrt2}\Big) \le \frac12. \end{aligned}$$

Therefore,

$$\mathbb{P}\Big(\sup_{t\in B^{\varrho}_{\mathcal{S}_n}(0,1)} |\vartheta_n(t^2)| > \frac12\Big) \le \mathbb{P}(\Theta^{\complement}).$$

To bound $\mathbb{P}(\vartheta_n(\chi_\lambda\chi_{\lambda'}) \ge B_{\lambda,\lambda'} x + V_{\lambda,\lambda'}\sqrt{2fx})$, we apply the Bernstein inequality given in **?** to the i.i.d. r.v.

$$U_i^{\lambda,\lambda'} = U_i^{(j,k),(j',k')} = \varphi_j(X_i)\varphi_{j'}(X_i)\psi_k(Y_i)\psi_{k'}(Y_i). \tag{23}$$

Thus, the r.v. are bounded

$$|U_i^{\lambda,\lambda'}| \le \|\chi_\lambda\chi_{\lambda'}\|_\infty = B_{\lambda,\lambda'}.$$

Moreover,

$$\mathbb{E}[(U_i^{\lambda,\lambda'})^2] \le \mathbb{E}[(\chi_\lambda(X_i,Y_i)\chi_{\lambda'}(X_i,Y_i))^2] \le f^* V^2_{\lambda,\lambda'}.$$

We get

$$\mathbb{P}(|\vartheta_n(\chi_\lambda\chi_{\lambda'})| \ge B_{\lambda,\lambda'} x + V_{\lambda,\lambda'}\sqrt{2fx}) \le 2e^{-nx}.$$

Given that $\mathbb{P}(\Delta_\rho^{\complement}) \le \mathbb{P}(\Theta^{\complement}) = \sum_{\lambda,\lambda'} \mathbb{P}\Big(|\vartheta_n(\chi_\lambda\chi_{\lambda'})| > B_{\lambda,\lambda'}x + V_{\lambda,\lambda'}\sqrt{2f^*x}\Big)$, we can write:

$$\begin{aligned}\mathbb{P}(\Delta_\rho^{\complement}) &\le 2(\mathcal{D}_n^{(1)}\mathcal{D}_n^{(2)})^2 \exp\Big\{-\frac{n(f_0^*)^2}{16f^*L(\chi)}\Big\}\\ &\le 2n^2\exp\Big\{-\frac{(f_0^*)^2}{16f^*}\frac{n}{L(\chi)}\Big\}.\end{aligned}$$

Following the Lemma of Baraud *et al.* (2001) above, and using

$$L(\chi) \le 4(\mathcal{D}_n^{(2)}\mathcal{D}_n^{(1)})^2 \le \phi n/\log^2(n),$$

we get that, for any k, there exists a constant $C_k^{(\Delta_\varrho)}$ depending on k, f_0^*, f^*, such that

$$\mathbb{P}(\Delta_\varrho^{\complement}) \le 2n^2\exp\Big\{-c\frac{(f_0^*)^2}{160f^*}\log^2(n)\Big\} \le \frac{C_k^{(\Delta_\varrho)}}{n^k}. \tag{24}$$

Now, if the basis is the histogram basis (or localized), the result is better. In this case, $L(\chi) \le \mathcal{D}_n^{(1)}\mathcal{D}_n^{(2)} = \mathcal{D}_n \le n/\log^2(n)$ is enough to get (24) again. This concludes the proof of Lemma 2. □

5.4. **Sketch of proof of Theorem 2.** Here the contrast decomposition is:

$$\tilde{\gamma}_n(T) - \tilde{\gamma}_n(S) = N_{\xi,n}^2(T-\zeta) - N_{\xi,n}^2(S-\zeta) - \frac{2}{n}\sum_{i=1}^n \xi_i(\delta_i - \zeta(X_i,Y_i))(T-S)(X_i,Y_i)$$

and the centered empirical process under study is

$$\tilde{\nu}_n(T) = \frac{1}{n}\sum_{i=1}^n \xi_i(\delta_i - \zeta(X_i,Y_i))T(X_i,Y_i).$$

Let us define $\Omega_\xi = \Delta_\xi \cap \Omega_{\rho_0}$ where Ω_{ρ_0} is defined in **(B2)**,

$$\mathcal{H} = \bigcap_{m\in\mathcal{M}_n} \mathcal{H}_m, \quad \mathcal{H}_m = \{\min sp(\Upsilon_m) \ge \max(\hat{\rho}_0/2, n^{-1/2})\},$$

and

$$\Delta_\xi = \left\{T \in \mathcal{S}_n, \quad \left|\frac{N_{\xi,n}^2(T)}{\|T\|_\xi^2} - 1\right| \le \frac{1}{2}\right\},$$

where we denote by $\|T\|_\xi^2 = \|T\sqrt{\pi}\|_\varrho^2$.

Similarly to the previous study, for n large enough, $\Delta_\rho \cap \Omega_{\rho_0} \subset \mathcal{H}\cap\Omega_{\rho_0}$ and thus, on Ω_ρ, we get

$$\begin{aligned}\|\hat{\zeta}_{\hat{m}} - \zeta_m\|_\xi^2 &\le 2N_{\xi,n}^2(\hat{\zeta}_{\hat{m}} - \zeta_m) \le 4N_{\xi,n}^2(\hat{\zeta}_{\hat{m}} - \zeta) + 4N_{\xi,n}^2(\zeta_m - \zeta)\\ &\le 8N_{\xi,n}^2(\zeta_m-\zeta) + 4\widetilde{\mathrm{pen}}(m) + 4\tilde{\nu}_n(\hat{\zeta}_{\hat{m}} - \zeta_m) - \widetilde{\mathrm{pen}}(\hat{m})\\ &\le 8N_{\xi,n}^2(\zeta_m-\zeta) + 4\widetilde{\mathrm{pen}}(m) + \frac{1}{4}\|\hat{\zeta}_{\hat{m}} - \zeta_m\|_\xi^2\\ &\quad +16\left(\sup_{T\in S_{\hat{m}}+S_m, \|T\|_\xi\le 1}\tilde{\nu}_n^2(T) - \tilde{p}(\hat{m},m)\right)_+ + 16\tilde{p}(\hat{m},m) - \widetilde{\mathrm{pen}}(\hat{m}).\end{aligned}$$

Talagrand's Inequality fixes the value of $\tilde{p}(m,m') = 2(\dim(S_m) + \dim(S_{m'}))/n$ to have

$$\mathbb{E}\left(\sup_{T\in S_{\hat{m}}+S_m, \|T\|_\xi\le 1}\tilde{\nu}_n^2(T) - \tilde{p}(\hat{m},m)\right)_+ \le \frac{C}{n},$$

and we take $\tilde{\kappa}$ large enough so that $\tilde{p}(m, m') \leq \widetilde{\text{pen}}(m) + \widetilde{\text{pen}}(m')$. We get,

$$\frac{1}{2}\mathbb{E}(\|\hat{\zeta}_{\hat{m}} - \zeta_m\|_\xi^2 \mathbf{I}_{\Omega_\xi}) \quad \leq \quad (8 + \frac{1}{2})\|\zeta_m - \zeta\|_\xi^2 + 5\widetilde{\text{pen}}(m) + \frac{C'}{n}.$$

This implies that

$$\mathbb{E}(\|\hat{\zeta}_{\hat{m}} - \zeta\|_\xi^2 \mathbf{I}_{\Omega_\xi}) \leq C(\|\zeta_m - \zeta\|_\xi^2 + \widetilde{\text{pen}}(m)) + \frac{C'}{n}.$$

References

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

Baraud, Y., Comte, F. and Viennet, G. (2001). Adaptive estimation in an autoregression and a geometrical β-mixing regression framework. *The Annals of Statistics* **39**, 839-875.

Birgé, L. and Massart, P. (1998) Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 329-375.

Cohen, A., Daubechies, I. and Vial, P. (1993) Wavelets on the interval and fast wavelet transforms, Applied and *Computational Harmonic Analysis* **1**, 54–81.

Comte, F., Gaïffas, S. and Guilloux, A. (2011). Adaptive estimation of the conditional intensity of marker-dependent counting processes. To appear in *Ann. Inst. Henri Poincaré Probab. Stat.* **47**, 171-1196.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 457-481.

Lawless, J.F. (2003). *Statistical models and methods for lifetime data.* Second edition. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

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**, 461-473.

Nikol'skii, S. M. (1975). *Approximation of functions of several variables and imbedding theorems.* Springer-Verlag, New York. Translated from the Russian by John M. Danskin, Jr., Die Grundlehren der Mathematischen Wissenschaften, Band 205.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126**, 505-563.

van der Laan, M. J. and McKeague, I. W. (1998) Efficient estimation from right-censored data when failure indicators are missing at random. *Ann. Statist.* **26**, 164-182.

Subramanian, S. (2006) Survival analysis for the missing censoring indicator model using kernel density estimation techniques. *Stat. Methodol.* **3**, 125–136.

Subramanian, S. (2004) Asymptotically efficient estimation of a survival function in the missing censoring indicator model. *J. Nonparametr. Stat.* **16**, 797-817.

Triebel, H. (2006). *Theory of function spaces. III.* Monographs in Mathematics, 100. Birkhäuser Verlag, Basel, 2006.

Wang, Q. and Shen, J. (2008) Estimation and confidence bands of a conditional survival function with censoring indicators missing at random. *J. Multivariate Anal.* **99**, 928-948.

Wang, Q. Ng, K. W. (2008) Asymptotically efficient product-limit estimators with censoring indicators missing at random. *Statist. Sinica* **18**, 749-768.

Wang, Q. , Liu, W. and Liu, C. (2009) Probability density estimation for survival data with censoring indicators missing at random. *J. Multivariate Anal.* **100**, 835-850.

Zhou, X. and Sun, L. (2003) Additive hazards regression with missing censoring information. *Statist. Sinica* **13**, 1237–1257.