

ESTIMATION/IMPUTATION STRATEGIES FOR MISSING DATA IN SURVIVAL ANALYSIS

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

Keywords: Missing at random - hazard rate - penalized contrast estimators -

1. INTRODUCTION

We consider the problem of estimation from right-censored data, when the censoring indicator is possibly 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 $\delta = \mathbf{1}(T \leq C)$ and it is assumed that T and C admit densities respectively denoted by f_T and g . In addition, C is assumed to be independent of T .

When the cause of death is not recorded, the censoring indicator is missing: this is the missing censoring indicator (MCI) model, see Subramanian 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 $i \in \{1, \dots, n\}$:

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

We consider the so-called MAR context, which means that we assume that the indicator is Missing At Random i.e. ξ and δ are independent conditionally to Y (contrary to the MCAR case where it is assumed that the indicator are Missing Completely At Random, i.e. ξ is independent of T and C). In this paper, we mainly concentrate on different strategies for dealing with missing indicators.

This model has been considered by several authors in the last decade. For a complete overview on missing data, we refer to the monography by Little and Rubin Little and Rubin (2002). Most papers are interested in survival function and cumulative hazard rate estimation. In particular, van der Laan and McKeague 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 Subramanian (2004) also proposes an 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 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

⁽¹⁾: 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.guillox@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.

provided by Wang and Ng Wang and Ng (2008). Recently, Wang *et al.* 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 Wang and Shen (2008).

Our aim here is to study nonparametric hazard rate estimation and more specifically, regression and model selection estimation of the function defined, for all $t > 0$ by:

$$(1) \quad \lambda(t) = \frac{f_T(t)}{1 - F_T(t)},$$

where f_T and F_T are respectively the probability density function (p.d.f.) and the cumulative distribution function (c.d.f.) of the survival time T . Two contrasts and estimators are proposed, and both involve the estimation of a specific nuisance functional parameter, π or ζ :

$$(2) \quad \pi(y) = \mathbb{E}(\xi|Y = y) = \mathbb{P}(\xi = 1|Y = y).$$

$$(3) \quad \zeta(y) = \mathbb{E}(\delta|Y = y) = \mathbb{P}(\delta = 1|Y = y),$$

We compare different estimation-imputation strategies for recovering π and ζ . More precisely, we propose either a parametric strategy following a logistic model – standard for π , but involving imputation steps for ζ –, or a pure nonparametric regression strategy. We provide theoretical properties and numerical comparisons for these procedures.

2. MODEL AND STRATEGIES

2.1. Model assumptions. The unknown function λ to be estimated is hazard rate of the random variable T , as defined by (1). We shall denote by G_C the c.d.f. of C . We define the conditional expectations of ξ and δ by (2) and (3).

The global model assumption is denoted **(A0)** and has several parts, specified hereafter.

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

(A0-2) For $i = 1, \dots, n$, we observe $Y_i = T_i \wedge C_i$, $\xi_i \in \{0, 1\}$, 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 .

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

Our estimation strategies are based on penalized contrasts. Namely, we obtain different collections of estimator by minimization of empirical criteria on functional spaces spanned by orthonormal bases defined on an interval $[0, \tau]$. We consider in the following two specific and classical examples of bases and model collections:

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

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

Considering $(\varphi_j)_{0 \leq j \leq m-1}$ yields spaces S_m spanned by them, with odd dimensions m . We denote by \mathcal{S}_n the nesting space of the collection, i.e. the space corresponding to the maximal dimension for S_m and we set $\mathcal{D}_n := \dim(\mathcal{S}_n) \leq \sqrt{n/\log(n)}$.

(2) **Histogram bases.** They are defined by $\varphi_j(x) = \sqrt{2^m/\tau}\mathbf{1}_{[(j-1)\tau/2^m, j\tau/2^m]}(x)$, for $j = 1, \dots, 2^m$ so that $S_m = \text{span}(\varphi_j, j = 1, \dots, 2^m)$ and $D_m = 2^m$. We shall take $m \leq \lceil \log_2(n)/2 \rceil$ where $\lceil z \rceil$ 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 and we set $\mathcal{D}_n := \dim(\mathcal{S}_n) \leq n/\log(n)$.

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.

2.2. Strategy involving knowledge of ζ . The crucial property for the construction of an estimation procedure is the following: for any integrable function h , we have

$$\begin{aligned}\mathbb{E}(\zeta(Y)h(Y)) &= \mathbb{E}[\mathbb{E}(\delta|Y)h(Y)] = \mathbb{E}(\delta h(Y)) \\ &= \mathbb{E}[\mathbf{I}(T \leq C)h(T)] \\ &= \int h(t)(1 - G_C)(t)f_T(t)dt.\end{aligned}$$

This yields the equality

$$(4) \quad \mathbb{E}(\zeta(Y)h(Y)) = \mathbb{E}(\delta h(Y)) = \int h(y)\lambda(y)(1 - L(y))dy$$

with $1 - L(y) = \bar{L}(y) := \mathbb{P}(Y \geq y) = (1 - F_T(y))(1 - G_C(y))$.

Next, we note that, with assumption **(A0-4)** and the definition of ζ , we have

$$(5) \quad \begin{aligned}\mathbb{E}(\delta_i \xi_i + (1 - \xi_i)\zeta(Y_i)|Y_i) &= \mathbb{E}(\delta_i|Y_i)\mathbb{E}(\xi_i|Y_i) + \mathbb{E}[(1 - \xi_i)\mathbb{E}(\delta_i|Y_i)|Y_i] \\ &= \mathbb{E}(\mathbb{E}(\delta_i|Y_i)(\xi_i + (1 - \xi_i))|Y_i) \\ &= \mathbb{E}(\delta_i|Y_i).\end{aligned}$$

If ζ was known, we would consider the contrast:

$$\Gamma_{n,1}^{th}(h) = \frac{1}{n} \sum_{i=1}^n \int_0^1 h^2(y)\mathbf{I}(Y_i \geq y)dy - \frac{2}{n} \sum_{i=1}^n (\xi_i \delta_i + (1 - \xi_i)\zeta(Y_i)) h(Y_i),$$

which is a particular case of the contrast introduced in Brunel *et al.* Brunel *et al.* (2013). Indeed, if we compute the expectation of this theoretical contrast, we obtain, under the MAR assumption and using (4) and (5),

$$\mathbb{E}(\Gamma_{n,1}^{th}(h)) = \|h\|_{\mu_1}^2 - 2 \int h(y)\lambda(y)d\mu_1(y) = \|h - \lambda\|_{\mu_1}^2 - \|\lambda\|_{\mu_1}^2,$$

with $d\mu_1(y) = (1 - L(y))dy$.

Clearly, the above quantity is small if h is near of λ , and the measure denoted by μ_1 plays the role of a reference weighting norm. This explains why minimizing $\Gamma_{n,1}^{th}$ over an appropriate set of functions would be a relevant strategy to estimate λ .

As ζ is unknown, we must substitute another quantity in the contrast. The first strategy we could think about, is to use an estimator $\tilde{\zeta}$ of ζ . Consequently, we consider

$$(6) \quad \Gamma_{n,1}^{\text{EST}}(h) = \frac{1}{n} \sum_{i=1}^n \int_0^1 h^2(y)\mathbf{I}(Y_i \geq y)dy - \frac{2}{n} \sum_{i=1}^n \left(\xi_i \delta_i + (1 - \xi_i)\tilde{\zeta}(Y_i) \right) h(Y_i).$$

Estimators of $\zeta(y)$ are constructed below. This strategy of estimation of the unknown hazard rate λ , via an estimation of ζ , is also considered in Wang *et al.* Wang *et al.* (2009). Such a regression strategy was also considered in Brunel and Comte Brunel & Comte (2005) without missing indicator.

2.3. Strategy involving knowledge of π . A similar reasoning can be lead in case π is known. Analogously, we consider the contrast

$$\Gamma_{n,2}^{th}(h) = \frac{1}{n} \sum_{i=1}^n \int_0^1 h^2(y) \pi(y) \mathbf{I}(Y_i \geq y) dy - \frac{2}{n} \sum_{i=1}^n \delta_i \xi_i h(Y_i).$$

Indeed, we can compute the expectation of $\Gamma_{n,2}^{th}(h)$. First, we have

$$\mathbb{E}(\delta_i \xi_i h(Y_i)) = \mathbb{E}(\mathbb{E}(\delta_i \xi_i | Y_i) h(Y_i))$$

and with **(A0-4)**, we get $\mathbb{E}(\delta_i \xi_i h(Y_i)) = \mathbb{E}(\zeta(Y_i) \pi(Y_i) h(Y_i))$. Now, applying (4) to $h\pi$ (instead of h) yields

$$\mathbb{E}(\delta_i \xi_i h(Y_i)) = \int h(y) \pi(y) \lambda(y) (1 - L(y)) dy.$$

Consequently,

$$\mathbb{E}(\Gamma_{n,2}^{th}(h)) = \int h^2(y) d\mu_2(y) - 2 \int h(y) \lambda(y) d\mu_2(y) = \|h - \lambda\|_{\mu_2}^2 - \|\lambda\|_{\mu_2}^2,$$

with $d\mu_2(y) = \pi(y)(1 - L_Y(y)) dy$. As previously, this explains why minimizing $\Gamma_{n,1}^{th}$ should allow us to estimate λ .

As π is unknown, we replace it by an estimator $\tilde{\pi}$ in the contrast. Thus, we consider

$$(7) \quad \Gamma_{n,2}^{\text{EST}}(h) = \frac{1}{n} \sum_{i=1}^n \int_0^1 h^2(y) \tilde{\pi}(y) \mathbf{I}(Y_i \geq y) dy - \frac{2}{n} \sum_{i=1}^n \delta_i \xi_i h(Y_i).$$

2.4. Estimation of ζ or π : logit or nonparametric regression. In both (6) and (7), a step has to be completed: we must make proposals for $\tilde{\zeta}$ and $\tilde{\pi}$. Two strategies can be naturally considered. First, coherently with the nonparametric setting of the first step, we can use a nonparametric regression method.

Nonparametric regression for estimation of $\zeta(y)$ or $\pi(y)$. The definition of ζ and π as conditional expectations given Y (see (3) and (2)), allows us to estimate these functions by nonparametric mean square contrast minimization. More precisely, we obtain a collection of estimators of ζ or π on an interval A by minimizing for T varying over spaces S_m , the contrasts

$$\tilde{\gamma}_{n,1}(T) = \frac{1}{n} \sum_{i=1}^n [\xi_i T^2(Y_i) - 2\xi_i \delta_i T(Y_i)] \text{ or } \tilde{\gamma}_{n,2}(T) = \frac{1}{n} \sum_{i=1}^n [T^2(Y_i) - 2\xi_i T(Y_i)]$$

respectively.

On the other hand, a natural strategy to estimate conditional indicators is the parametric logistic regression.

Logistic regression. If we add a logistic assumption

$$\zeta(y) = \frac{\exp(\alpha_0 + \alpha_1 y)}{1 + \exp(\alpha_0 + \alpha_1 y)}, \quad \pi(y) = \frac{\exp(\beta_0 + \beta_1 y)}{1 + \exp(\beta_0 + \beta_1 y)},$$

then we may use a maximum likelihood (M.L.) procedure. Note that, nevertheless, ζ and π are different from this point of view. Indeed, concerning the estimation of π , ξ and Y are observed and therefore, β_0 and β_1 correspond to standard logit-estimation. But this is not true for ζ : we cannot directly compute the M.L. estimators for (α_0, α_1) since there are censoring indicators δ_i missing. Thus, we follow a procedure proposed by Rubin Rubin (1987). For a binary classification variable, a new logistic regression model is simulated from the posterior

predictive distribution of the parameters and is used to impute the missing values for each variable .

2.5. Computing the hazard estimators. We consider that we estimate the hazard rate on a compact set

$$A = [0, \tau].$$

We set standard assumptions about boundedness from above and below.

(A1.1) $\forall y \in A, \lambda(y) \leq \|\lambda\|_{A, \infty} < +\infty.$

(A1.2) $\forall y \in A, 1 - L_Y(y) \geq \bar{L}_0 = \min_{y \in A} 1 - L_Y(y) > a > 0$ for fixed positive constant $a.$

(A1.3) $\forall y \in A, \pi(y) \geq \pi_0 = \inf_{y \in A} \pi(y) > b > 0$ for fixed positive constant $b.$

(A1.4) $\forall y \in A, 0 < f_0^* \leq f_Y(y) \leq f_1^* < +\infty.$

Assumptions **(A1.1)** and **(A1.2)** are common natural assumptions, and assumption **(A1.3)** is specific to the strategy involving π in the reference measure. Assumption **(A1.4)** is a common technical assumption.

First, we define two types of estimators $\hat{\lambda}_m^{(i)}$ for $i = 1, 2$ on the space S_m by:

$$\hat{\lambda}_m^{(i)} = \arg \min_{h \in S_m} \Gamma_{n,i}^{\text{EST}}(h) \quad \text{where } S_m = \text{span}\{\varphi_j, j = 1, \dots, D_m\}$$

The φ_j 's constitute an \mathbb{L}^2 -orthonormal basis, and the function h is of the form $h = \sum_j a_j \varphi_j.$

The estimators are obtained by minimization of the contrasts (6)-(7) given in Sections 2.2 and 2.3. For the histogram basis, we can give their explicit expression in the MAR setting : $\hat{\lambda}_m^{(i), [\text{H}]}(y) = \sum_{j=1}^{D_m} \hat{a}_j^{(i), [\text{H}]} \varphi_j(y)$ with the coefficient $\hat{a}_j^{(i), [\text{H}]}$ having the following form, provided the denominators are non zero:

$$\hat{a}_j^{(1), [\text{H}]} = \frac{\sum_{i=1}^n [\delta_i \xi_i + (1 - \xi_i) \tilde{\zeta}(Y_i)] \varphi_j(Y_i)}{\sum_{i=1}^n \int_0^1 \varphi_j^2(y) \mathbf{I}_{\{Y_i \geq y\}} dy}, \quad \hat{a}_j^{(2), [\text{H}]} = \frac{\sum_{i=1}^n \delta_i \xi_i \varphi_j(Y_i)}{\sum_{i=1}^n \int_0^1 \varphi_j^2(y) \tilde{\pi}(y) \mathbf{I}_{\{Y_i \geq y\}} dy}.$$

More generally, let us define the matrices and vectors: $\vec{\hat{a}}_m^{(i)} = \left(\hat{a}_j^{(i)} \right)_{1 \leq j \leq D_m},$

$$\Theta_m^{(1)} := \left(\frac{1}{n} \sum_{i=1}^n \int \varphi_j(y) \varphi_{j'}(y) \mathbf{I}_{\{Y_i \geq y\}} dy \right)_{1 \leq j, j' \leq D_m},$$

$$\Delta_m^{(1)} = \text{vec} \left(\frac{1}{n} \sum_{i=1}^n \left(\delta_i \xi_i + (1 - \xi_i) \tilde{\zeta}(Y_i) \right) \varphi_j(Y_i) \right)_{1 \leq j \leq D_m},$$

$$\Theta_m^{(2)} := \left(\frac{1}{n} \sum_{i=1}^n \int \varphi_j(y) \varphi_{j'}(y) \tilde{\pi}(y) \mathbf{I}_{\{Y_i \geq y\}} dy \right)_{1 \leq j, j' \leq D_m} \quad \Delta_m^{(2)} = \left(\frac{1}{n} \sum_{i=1}^n \delta_i \xi_i \varphi_j(Y_i) \right)_{1 \leq j \leq D_m}.$$

Then the coefficients of the estimators must fulfill the matrix constraint:

$$\Theta_m^{(i)} \vec{\hat{a}}_m^{(i)} = \Delta_m^{(i)}.$$

It follows that the estimator is well defined if $\Theta_m^{(i)}$ 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

$$(8) \quad \vec{\hat{a}}_m^{(i)} = (\Theta_m^{(i)})^{-1} \Delta_m^{(i)} \text{ if } \rho(\Theta_m^{(i)}) \geq \max(\hat{c}_i, n^{-1/2})$$

and $\vec{\hat{a}}_m^{(i)} = 0$ otherwise, with $\hat{c}_1 = \hat{L}_0/3$ and $\hat{c}_2 = 4\tilde{\pi}_0 \hat{L}_0/9.$

The quantities $\hat{\bar{L}}_0$ and $\tilde{\pi}_0$ are estimators of $\bar{L}_0 = \min_{y \in A} 1 - L_Y(y) > 0$ under **(A1.2)** and $\pi_0 = \min_{y \in A} \pi(y) > 0$ under **(A1.3)**. An estimator $\hat{\bar{L}}_0$ is defined in Comte *et al.* (2009), and proved to have the required properties, namely, to satisfy:

(A2.1) For any integer $k \geq 1$, there exists a constant $C_k^{(L_0)} > 0$ such that $\mathbb{P}(|\hat{\bar{L}}_0 - \bar{L}_0| > \bar{L}_0/2) \leq C_k^{(L_0)}/n^k$.

For $\tilde{\pi}_0$, we can study $\min \tilde{\pi}(y)$ but we state results in term of the conditions $\tilde{\pi}$ must fulfill:

(A2.2) $\mathbb{P}\left(\left\|\frac{\tilde{\pi} - \pi}{\pi}\right\|_{\infty, A} > 1/4\right) := \mathbb{P}(\Omega_\pi^c) \leq \frac{C}{n^k}$ for $k \geq 4$ and C a given constant.

After this stage, the coefficients $\hat{a}_j^{(i)}$, for $i = 1, 2$ provide the development of an estimator

$$(9) \quad \hat{\lambda}_m^{(i)}(y) = \sum_j \hat{a}_j^{(i)} \varphi_j(y).$$

The model selection device is now based on the following criterion: for $i = 1, 2$,

$$(10) \quad \hat{m}^{(i)} = \arg \min_{m \in \mathcal{M}_n} (\Gamma_{n,i}^{\text{EST}}(\hat{\lambda}_m) + \widehat{\text{pen}}^{(i)}(m))$$

where

$$\mathcal{M}_n = \{m \in \mathbb{N}, \dim(S_m) \leq N_n\}.$$

Moreover

$$(11) \quad \widehat{\text{pen}}^{(i)}(m) = \kappa_i \|\hat{\lambda}\|_{\infty, A} \frac{\dim(S_m)}{n},$$

where Φ_0 is such that $\forall t \in S_m, \|t\|_\infty \leq \Phi_0 \|t\|$ and $\hat{\lambda} = \hat{\lambda}_{m_0}$ is an estimator in the collection, on a space S_{m_0} with dimension D_{m_0} such that $\log(n) \leq D_{m_0} \leq n^{1/4}$. Moreover, the dimension D_m of the spaces S_m has to be larger than $\log(n)$.

The computation of $\tilde{\zeta}$ and $\tilde{\pi}$ is similar and the vector of their coefficients in the basis $(\varphi_j)_{1 \leq j \leq d_m}$ is defined by the inverse of matrices

$$\Upsilon_m^{(1)} = \left(\frac{1}{n} \sum_{i=1}^n \xi_i \varphi_j(Y_i) \varphi_{j'}(Y_i) \right)_{1 \leq j, j' \leq D_m}, \quad \Upsilon_m^{(2)} = \left(\frac{1}{n} \sum_{i=1}^n \varphi_j(Y_i) \varphi_{j'}(Y_i) \right)_{1 \leq j, j' \leq D_m},$$

multiplied respectively by vectors

$$\Xi_m^{(1)} = \Delta_m^{(2)}, \quad \Xi_m^{(2)} = \left(\frac{1}{n} \sum_{i=1}^n \xi_i \varphi_j(Y_i) \right)_{1 \leq j \leq D_m}$$

with adequate thresholds.

2.6. Theoretical Results. We consider that we estimate the hazard rate on a compact set $A = [0, \tau]$.

The empirical reference norms are defined by

$$\|h\|_{n,1}^2 = \frac{1}{n} \sum_{i=1}^n \int_0^1 h^2(y) \mathbf{I}(Y_i \geq y) dy, \quad \|h\|_{n,2}^2 = \frac{1}{n} \sum_{i=1}^n \int_0^1 h^2(y) \pi(y) \mathbf{I}(Y_i \geq y) dy$$

associated with the natural scalar product $\langle h_1, h_2 \rangle_{n,i}$ obtained by polarization identity.

Let us also define another integral norm with respect to $d\rho(y) = f_Y(y)dy$ where f_Y is the pdf of Y_1 , that is

$$(12) \quad \|\psi\|_{\rho}^2 = \int \psi^2(y)d\rho(y) = \int \psi^2(y)f_Y(y)dy.$$

The result we can obtain is the following, for estimators $\tilde{\zeta}$ or $\tilde{\pi}$ computed on an independent sample of observations.

Theorem 1. Let $\hat{\lambda}_{\hat{m}^{(i)}}^{(i)}$ be the estimator defined by (8)-(9)-(10)-(11) for $i = 1, 2$, under Assumptions **(A0)**. Assume that the collection of models is nested and the estimators $\tilde{\zeta}$ and $\tilde{\pi}$ are computed on independent samples.

(1) If **(A1.1)**-**(A1.2)**-**(A1.4)** and **(A2.1)** hold, there exists a choice of κ_1 such that, for n large enough

$$(13) \quad \mathbb{E}(\|\lambda\mathbf{I}_A - \hat{\lambda}_{\hat{m}^{(1)}}^{(1)}\|_{n,1}^2) \leq C_1 \inf_{m \in \mathcal{M}_n} (\|\lambda\mathbf{I}_A - \lambda_m\|^2 + \|\lambda\|_{A,\infty} \frac{D_m}{n}) + C'_1 \mathbb{E}(\|\tilde{\zeta} - \zeta\|_{\rho}^2) + \frac{C''_1}{n},$$

where C_1 is a numerical constant and C'_2, C''_2 are constants depending on the constants of the problem.

(2) If **(A1.1)**-**(A1.4)** and **(A2.1)**-**(A2.2)** hold, there exists a choice of κ_2 such that, for n large enough

$$(14) \quad \mathbb{E}(\|\lambda\mathbf{I}_A - \hat{\lambda}_{\hat{m}^{(2)}}^{(2)}\|_{n,2}^2) \leq C_2 \inf_{m \in \mathcal{M}_n} (\|\lambda\mathbf{I}_A - \lambda_m\|^2 + \|\lambda\|_{A,\infty} \frac{D_m}{n}) + C'_2 \mathbb{E}(\|\tilde{\pi} - \pi\|_{\rho}^2) + \frac{C''_2}{n},$$

where C_2 is a numerical constant and C'_2, C''_2 are constants depending on the constants of the problem.

Inequality (13) is proved in Brunel *et al.* (2013) in a more general setting, and a sketch of the proof of (14) is given in appendix.

Note that the larger κ_i , the larger $\text{pen}^{(i)}(m)$ in the upper bound.

The results stated in (13)-(14) involve three terms:

- the first terms are common, $\inf_{m \in \mathcal{M}_n} (\|\lambda\mathbf{I}_A - \lambda_m\|^2 + \|\lambda\|_{A,\infty} D_m/n)$ and correspond to the squared-bias ($\|\lambda\mathbf{I}_A - \lambda_m\|^2$)/variance ($\|\lambda\|_{A,\infty} D_m/n$) compromise,
- the second term is $\mathbb{E}(\|\tilde{\zeta} - \zeta\|_{\rho}^2)$ or $\mathbb{E}(\|\tilde{\pi} - \pi\|_{\rho}^2)$, that is the mean-square risk of the estimator of ζ or π on A . These quantities can have different orders, depending on the second step strategy,
- the last terms C''_1/n or C''_2/n are negligible.

We can obtain adaptive nonparametric rate with the same model selection principle for ζ and π , if we apply the nonparametric strategy (see Theorem 2 for ζ in Brunel *et al.* (2013) Brunel *et al.* (2013)).

On the other hand, with the logit strategy, we have

$$\tilde{\zeta}(y) = \frac{e^{\hat{\alpha}_0 + y\hat{\alpha}_1}}{1 + e^{\hat{\alpha}_0 + y\hat{\alpha}_1}}, \quad \tilde{\pi}(y) = \frac{e^{\hat{\beta}_0 + y\hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + y\hat{\beta}_1}},$$

and we get

$$\begin{aligned} \mathbb{E}(\|\tilde{\zeta} - \zeta\|_{\rho}^2) &\leq 2 \left(\mathbb{E}((\hat{\alpha}_0 - \alpha_0)^2) + \mathbb{E}(Y_1^2)\mathbb{E}((\hat{\alpha}_1 - \alpha_1)^2) \right), \\ \mathbb{E}(\|\tilde{\pi} - \pi\|_{\rho}^2) &\leq 2 \left(\mathbb{E}((\hat{\beta}_0 - \beta_0)^2) + \mathbb{E}(Y_1^2)\mathbb{E}((\hat{\beta}_1 - \beta_1)^2) \right). \end{aligned}$$

All these terms are of order $1/n$ as soon as all parameters are estimated with parametric rate.

The questions that we have to study empirically are:

- (1) If the logit assumptions are fulfilled, is the nonparametric strategy much worse than the parametric one?
- (2) If the logit model is not satisfied, can the logit approximation still correctly estimate π or ζ ?
- (3) How much does the quality of the estimator of these quantities influences that of λ ?

3. IMPUTATION BASED STRATEGY

The second approach, we aim to explore, is inspired from stochastic regression imputation. The idea is no longer to substitute an estimator of the unknown quantity ζ but rather to substitute a random number γ_i to $\zeta(Y_i)$ which has the same expectation.

Let γ_i generating from a Bernoulli distribution with parameter $\zeta(Y_i)$ given $\xi_i = 0$. Here, the quantity $\zeta(Y_i)$ has also to be estimated as in Section 2.4, and we consider again both nonparametric and logit strategies. Then, we are in position to define the following contrast:

$$(15) \quad \Gamma_n^{\text{IMP}}(h) = \frac{1}{n} \sum_{i=1}^n \int_0^1 h^2(y) \mathbf{I}(Y_i \geq y) dy - \frac{2}{n} \sum_{i=1}^n (\xi_i \delta_i + (1 - \xi_i) \gamma_i) h(Y_i).$$

and the imputation strategy gives the following estimator:

$$\hat{\lambda}_m^{(3)} = \arg \min_{h \in S_m} \Gamma_n^{\text{IMP}}(h) \quad \text{where } S_m = \text{span}\{\varphi_j, j = 1, \dots, D_m\}.$$

We do not provide the theoretical study of the estimator defined by this contrast since $\mathbb{E}(\Gamma_n^{\text{IMP}}(h)) = \mathbb{E}(\Gamma_{n,1}^{\text{th}}(h))$. We propose a numerical comparison of this strategy with the other strategies developed in Section 2.

4. NUMERICAL COMPARISON

To illustrate the behaviors of the different estimators, we consider three models to simulate data from the MCI, i.e. n independent replications of

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

In each model, the parameters are set in order to obtain around 30 % of censoring and 85 % (resp. 55 %) of non missing indicators.

Model 1 (Weibull-Logistic). *Each T_i is drawn from a Weibull distribution with scale $a = 10$ and shape $b = 4$, each C_i from an exponential distribution with mean $\mu = 25$. For each individual i , the missing indicator is simulated from a Bernoulli distribution with parameter $\pi(y_i)$ with*

$$\pi(y) = \frac{1}{1 + \exp(-\beta_1 - \beta_2 y)},$$

we set $\beta_1 = 0.1$ and $\beta_2 = 0.25$ (resp. 0.005) to obtain 85 % (resp. 55 %) non missing indicators. In this model:

$$\zeta(y) = \left(1 + \frac{\mu}{ba - by^{b-1}}\right)^{-1}.$$

Model 2 (Monotone). *It is based on an example from Subramanian Subramanian (2009). For each i , Y_i is drawn from an exponential distribution with parameter 1. The censoring and missing indicators are simulated from Bernoulli distributions with parameters resp. $\zeta(y_i)$ and $\pi(y_i)$ with:*

$$\pi(y) = 1 - \exp(-\theta \exp(y)) \quad \text{and} \quad \zeta(y) = \frac{1}{1 + \exp(-\alpha_1 - \alpha_2 y)},$$

with $\theta = 1$ (resp. 0.32) to obtain 85 % (resp. 55 %) non missing indicators, and $\alpha_1 = -2$, $\alpha_2 = 5.2$. Note that, the hazard rate $h(y)$ coincides with the auxiliary function $\zeta(y)$ in this model.

Model 3 (Non-monotone). Consider the same schema as Model 2, with

$$\zeta(y) = 0.9|\sin(\pi/2 - 1.5y)| + 0.1.$$

| Weibull/ Logistic | Estimation strategy with $\Gamma_{n,1}^{\text{EST}}$ | | Estimation strategy with $\Gamma_{n,2}^{\text{EST}}$ | | Imputation strategy Γ_n^{IMP} | |
|------------------------------|---|------------------|---|------------------|--|------------------|
| Missing rate | 15% | 45 % | 15% | 45 % | 15 % | 45 % |
| <i>n</i> = 200 | | | | | | |
| NP | 0.055 (0.032) | 0.061 (0.043) | 0.120 (0.081) | 0.269 (0.162) | 0.055 (0.032) | 0.067 (0.051) |
| logit | 0.055 (0.029) | 0.053 (0.032) | 0.151 (0.123) | 0.286 (0.160) | 0.053 (0.031) | 0.055 (0.032) |
| <i>n</i> = 1000 | | | | | | |
| NP | 0.017 (0.010) | 0.021 (0.014) | 0.035 (0.018) | 0.086 (0.037) | 0.018 (0.010) | 0.024 (0.015) |
| logit | 0.016 (0.008) | 0.016 (0.009) | 0.062 (0.036) | 0.092 (0.026) | 0.016 (0.008) | 0.019 (0.012) |
| <i>n</i> = 5000 | | | | | | |
| NP | 0.008 (0.007) | 0.009 (0.008) | 0.011 (0.010) | 0.025 (0.023) | 0.008 (0.006) | 0.009 (0.009) |
| logit | 0.007 (0.005) | 0.008 (0.006) | 0.023 (0.022) | 0.024 (0.023) | 0.007 (0.006) | 0.008 (0.006) |

TABLE 1. Average and standard deviation (in parenthesis) of the MISE over 500 replicated samples for hazard rate estimators of Model 1 (Weibull-Logistic), for estimation strategies with $\Gamma_{n,1}^{\text{EST}}$ and $\Gamma_{n,2}^{\text{EST}}$ and imputation strategy Γ_n^{IMP} . Censoring rate $\simeq 30\%$.

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:

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

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 over the three models and fixed to 2 (resp. 3.5) for the contrast $\Gamma_{n,1}^{\text{EST}}$ (resp. $\Gamma_{n,2}^{\text{EST}}$). We also give in parenthesis the standard deviation value of the MISE evaluated over the 500 samples. The results are summarized in Tab. 1–4. For each model, we give the MISEs of our penalized estimators $\hat{\lambda}_{\hat{m}^{(1)}}^{(1)}$ for the estimation strategy involving ζ in column 1-2, $\hat{\lambda}_{\hat{m}^{(2)}}^{(2)}$ for the estimation strategy involving π in column 3-4 and $\hat{\lambda}_{\hat{m}^{(3)}}^{(3)}$ for the imputation strategy in column 5-6.

| Subramanian monotone | Estimation strategy with $\Gamma_{n,1}^{\text{EST}}$ | | Estimation strategy with $\Gamma_{n,2}^{\text{EST}}$ | | Imputation strategy Γ_n^{IMP} | |
|---------------------------------|---|------------------|---|------------------|--|-------------------|
| | 15% | 45 % | 15% | 45 % | 15 % | 45 % |
| $n = 200$ | | | | | | |
| NP | 0.102 (0.051) | 0.125 (0.052) | 0.201 (0.143) | 0.410 (0.445) | 0.102 (0.057) | 0.124 (0.051) |
| logit | 0.105 (0.062) | 0.101 (0.060) | 0.177 (0.083) | 0.182 (0.071) | 0.102 (0.0592) | 0.120 (0.0576) |
| $n = 1000$ | | | | | | |
| NP | 0.037 (0.016) | 0.037 (0.014) | 0.073 (0.043) | 0.181 (0.131) | 0.037 (0.017) | 0.038 (0.014) |
| logit | 0.036 (0.018) | 0.035 (0.018) | 0.063 (0.035) | 0.084 (0.035) | 0.037 (0.017) | 0.039 (0.016) |
| $n = 5000$ | | | | | | |
| NP | 0.012 (0.011) | 0.012 (0.011) | 0.028 (0.025) | 0.080 (0.068) | 0.012 (0.012) | 0.012 (0.011) |
| logit | 0.013 (0.012) | 0.013 (0.011) | 0.019 (0.018) | 0.023 (0.022) | 0.013 (0.012) | 0.012 (0.011) |

TABLE 2. Average and standard deviation (in parenthesis) of the MISE over 500 replicated samples for hazard rate estimators of the Model 2, for estimation strategies with $\Gamma_{n,1}^{\text{EST}}$ and $\Gamma_{n,2}^{\text{EST}}$ and imputation strategy Γ_n^{IMP} . Censoring rate $\simeq 30\%$.

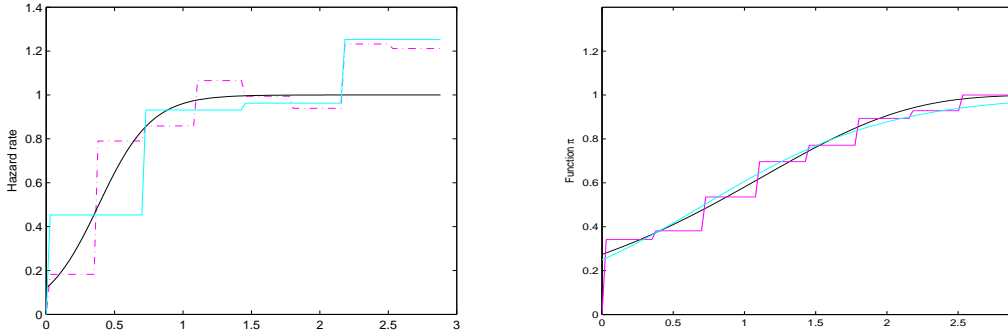


FIGURE 1. Strategy involving estimation of π : Left-side: True hazard rate (black line) for Model 2 and its estimators $\hat{\lambda}_{m(2)}^{(2)}$ with both nonparametric (magenta dotted line) and logit (cyan plain line) estimators of π . Right-Side: True auxiliary function π (black line) and its nonparametric (magenta) and logit (cyan) estimators. $n = 1000$ with $\simeq 30\%$ censoring and 45% missing indicators.

For each strategy and each model, we consider various sample sizes $n = 200, 1000$ and 5000 and two levels of missing indicators namely 15% and 45% . The abbreviations "NP" and "logit" stand for the way of estimating the auxiliary functions ζ or π according to the strategy under study (estimation of ζ with $\Gamma_{n,1}^{\text{EST}}$ and Γ_n^{IMP} or estimation of π with $\Gamma_{n,2}^{\text{EST}}$). First, whatever the model we considered, we observe that the strategy involving π gives very poor performances

compared to the other ones. The difference between the strategies involving π or ζ is of great importance for a numerical point of view. In fact, as we can see by examining the sequence of coefficients $\hat{a}_j^{(1),[H]}$ and $\hat{a}_j^{(2),[H]}$, the estimator $\tilde{\pi}$ appear in the denominator (while the estimator $\tilde{\zeta}$ does not) and it results that a small estimation error for π can strongly increase the final error on the hazard estimator. We can see on Figure 1 that despite a quite good estimation of π , the MISE in Table 1–3 is always the worst for the estimation strategy involving π . Consequently, we focus on the other strategies. Both imputation or estimation strategy involving the function ζ have very similar behaviour and perform as well. We also notice that the logit estimator of ζ *versus* the nonparametric one is nearly impossible to beat at least for sample sizes $n = 200, 1000$. Of course, when the logit assumption is violated, for large sample ($n = 5000$) and important missing rate 45%, the results are deteriorating, see Table 3.

| Subramanian non-monotone | Estimation strategy with $\Gamma_{n,1}^{\text{EST}}$ | | Estimation strategy with $\Gamma_{n,2}^{\text{EST}}$ | | Imputation strategy Γ_n^{IMP} | |
|-------------------------------------|---|------------------|---|------------------|--|------------------|
| | 15% | 45 % | 15% | 45 % | 15 % | 45 % |
| <hr/> | | | | | | |
| Missing rate | 15% | 45 % | 15% | 45 % | 15 % | 45 % |
| <hr/> | | | | | | |
| $n = 200$ | | | | | | |
| NP | 0.132 (0.061) | 0.163 (0.060) | 0.218 (0.088) | 0.269 (0.248) | 0.132 (0.061) | 0.163 (0.060) |
| logit | 0.138 (0.060) | 0.180 (0.058) | 0.234 (0.081) | 0.257 (0.073) | 0.137 (0.062) | 0.170 (0.060) |
| <hr/> | | | | | | |
| $n = 1000$ | | | | | | |
| NP | 0.064 (0.019) | 0.076 (0.016) | 0.105 (0.029) | 0.125 (0.055) | 0.065 (0.019) | 0.076 (0.016) |
| logit | 0.066 (0.019) | 0.096 (0.018) | 0.113 (0.026) | 0.128 (0.024) | 0.065 (0.019) | 0.077 (0.016) |
| <hr/> | | | | | | |
| $n = 5000$ | | | | | | |
| NP | 0.038 (0.039) | 0.039 (0.038) | 0.055 (0.054) | 0.067 (0.066) | 0.038 (0.040) | 0.040 (0.040) |
| logit | 0.040 (0.040) | 0.065 (0.064) | 0.061 (0.062) | 0.069 (0.069) | 0.040 (0.040) | 0.065 (0.064) |

TABLE 3. Average and standard deviation (in parenthesis) of the MISE over 500 replicated samples for hazard rate estimators of Model 3, for estimation strategies with $\Gamma_{n,1}^{\text{EST}}$ and $\Gamma_{n,2}^{\text{EST}}$ and imputation strategy Γ_n^{IMP} . Censoring rate $\simeq 30\%$.

5. PROOFS

5.1. Proof of Theorem 1. The first inequality is obtained as a particular case of the bound proved in Brunel *et al.* Brunel *et al.* (2013), see also Comte *et al.* ?. Below, we give steps for obtaining the second inequality.

The reference centered empirical process is

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

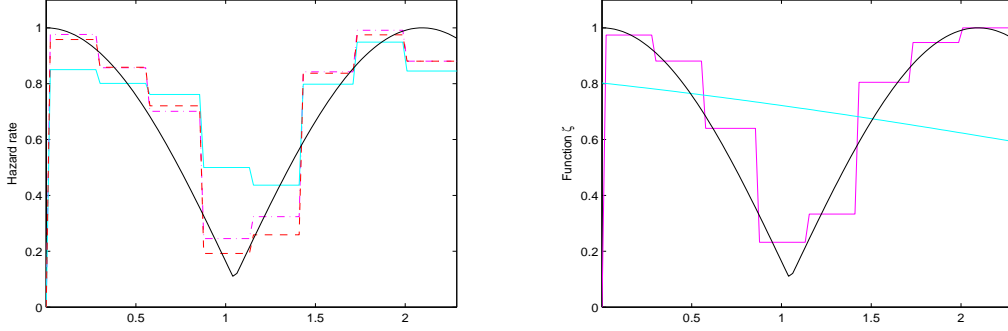


FIGURE 2. Strategies involving estimation of ζ : Left-side: True hazard rate (black line) for Model 3 and its estimators $\hat{\lambda}_{\hat{m}^{(1)}}^{(1)}$ with nonparametric (magenta dotted line) and logit (cyan plain line) estimators of ζ and $\hat{\lambda}_{\hat{m}^{(3)}}^{(3)}$ with nonparametric estimator of ζ (red dotted line). Right-Side: True auxiliary function ζ (black line) and its nonparametric (magenta) and logit (cyan) estimators. $n = 1000$ with $\simeq 30\%$ censoring and 45% missing indicators.

It is constituted of two different processes which have different orders, $\nu_n(h) = \nu_{n,1}(h) - \nu_{n,2}(h)$ with

$$(16) \quad \nu_{n,1}(h) = \frac{1}{n} \sum_{i=1}^n (\delta_i \xi_i h(Y_i) - \langle h, \lambda \rangle_{\mu_2}),$$

$$(17) \quad \nu_{n,2}(h) = \frac{1}{n} \sum_{i=1}^n \left(\int h(y) \lambda(y) \pi(y) \mathbf{1}_{\{Y_i \geq y\}} dy - \langle h, \lambda \rangle_{\mu_2} \right).$$

We define a set $\Omega = \Delta \cap \Omega_{\bar{L}_0} \cap \Omega_\pi$ where Ω_π is defined in **(A2.2)** and

$$\hat{\mathcal{G}} = \bigcap_{m \in \mathcal{M}_n} \hat{\mathcal{G}}_m, \quad \hat{\mathcal{G}}_m = \{ \min Sp(\Theta_m^{(2)}) \geq \max(4\tilde{\pi}_0 \hat{\bar{L}}_0/9, n^{-1/2}) \},$$

$$\Delta = \left\{ \forall h \in \mathcal{S}_n, \left| \frac{\|h\|_{n,2}^2}{\|h\|_{\mu_2}} - 1 \right| \leq \frac{1}{2} \right\} \quad \text{and} \quad \Omega_{\bar{L}_0} = \left\{ \left| \frac{\hat{\bar{L}}_0}{\bar{L}_0} - 1 \right| \leq \frac{1}{2} \right\}.$$

It is easy to see that (see Brunel *et al.* Brunel *et al.* (2013)):

Lemma 1. *Under the assumptions of Theorem 1, for n large enough, the following embedding holds: $\Delta \cap \Omega_{\bar{L}_0} \cap \Omega_\pi \subset \hat{\mathcal{G}} \cap \Omega_{\bar{L}_0} \cap \Omega_\pi$.*

As a consequence, for all $m \in \mathcal{M}_n$, the matrices $\Theta_m^{(2)}$ are invertible on Ω .

Let us denote by $\lambda_A = \lambda \mathbf{I}_A$ and $\pi_A = \pi \mathbf{I}_A$. We start with the decomposition

$$\begin{aligned} \Gamma_{n,2}^{\text{EST}}(h) - \Gamma_{n,2}^{\text{EST}}(\ell) &= \|h - \lambda_A\|_{n,2}^2 - \|\ell - \lambda_A\|_{n,2}^2 - 2\nu_n(h - \ell) \\ &+ \frac{1}{n} \sum_{i=1}^n \int (h - \lambda_A)^2(y) \mathbf{I}_{\{Y_i \geq y\}} (\tilde{\pi} - \pi_A)(y) dy \\ &- \frac{1}{n} \sum_{i=1}^n \int (\ell - \lambda_A)^2(y) \mathbf{I}_{\{Y_i \geq y\}} (\tilde{\pi} - \pi_A)(y) dy \\ &+ \frac{2}{n} \sum_{i=1}^n \int (h - \ell)(y) \lambda(y) \mathbf{I}_{\{Y_i \geq y\}} (\tilde{\pi} - \pi_A)(y) dy. \end{aligned}$$

Then, let us denote by λ_m the orthogonal projection of λ_A on S_m and use the definition of $\hat{\lambda}_{\hat{m}}$ on Ω (see Lemma 1). We assume first that $\|\lambda\|_{A,\infty}$ is known and denote by $\text{pen}(m) = \kappa \|\lambda\|_{A,\infty} D_m/n$, and write $\Gamma_{n,2}^{\text{EST}}(\hat{\lambda}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \Gamma_{n,2}^{\text{EST}}(\lambda_m) + \text{pen}(m)$.

Then we get, on Ω ,

$$\begin{aligned} \|\hat{\lambda}_{\hat{m}} - \lambda_A\|_{n,2}^2 &\leq \|\lambda_m - \lambda_A\|_{n,2}^2 + \text{pen}(m) + 2\nu_n(\hat{\lambda}_{\hat{m}} - \lambda_m) - \text{pen}(\hat{m}) \\ &\quad - \frac{1}{n} \sum_{i=1}^n \int (\hat{\lambda}_{\hat{m}} - \lambda_A)^2(y) \mathbf{I}_{\{Y_i \geq y\}} (\tilde{\pi} - \pi_A)(y) dy \\ &\quad + \frac{1}{n} \sum_{i=1}^n \int (\lambda_m - \lambda_A)^2(y) \mathbf{I}_{\{Y_i \geq y\}} (\tilde{\pi} - \pi_A)(y) dy \\ (18) \quad &\quad - \frac{2}{n} \sum_{i=1}^n \int (\hat{\lambda}_{\hat{m}} - \lambda_m)(y) \lambda(y) \mathbf{I}_{\{Y_i \geq y\}} (\tilde{\pi} - \pi_A)(y) dy. \end{aligned}$$

First, we write

$$2|\nu_n(\hat{\lambda}_{\hat{m}} - \lambda_m)| \leq \frac{1}{8} \|\hat{\lambda}_{\hat{m}} - \lambda_m\|_{\mu_2}^2 + 8 \sup_{h \in B_{m,\hat{m}}(0,1)} \nu_n^2(h),$$

where $B_{m,m'}(0,1) = \{h \in S_m + S_{m'}, \|h\|_{\mu_2} \leq 1\}$. It follows that

$$\begin{aligned} 2|\nu_n(\hat{\lambda}_{\hat{m}} - \lambda_m)| \mathbf{I}_\Omega &\leq \frac{1}{4} \|\hat{\lambda}_{\hat{m}} - \lambda_m\|_{n,2}^2 \mathbf{I}_\Omega + 16 \left(\sup_{h \in B_{m,\hat{m}}(0,1)} \nu_{n,1}^2(h) - p(m, \hat{m}) \right)_+ \\ &\quad + 16p(m, \hat{m}) + 16 \sup_{h \in B_{m,\hat{m}}(0,1)} \nu_{n,2}^2(h) \\ (19) \quad &\leq \frac{1}{2} \|\hat{\lambda}_{\hat{m}} - \lambda\|_{n,2}^2 \mathbf{I}_\Omega + \frac{1}{2} \|\lambda_m - \lambda\|_{n,2}^2 + 16 \left(\sup_{h \in B_{m,\hat{m}}(0,1)} \nu_n^2(h) - p(m, \hat{m}) \right)_+ \\ &\quad + 16p(m, \hat{m}) + 16 \sup_{h \in B_{m,\hat{m}}(0,1)} \nu_{n,2}^2(h), \end{aligned}$$

where $p(m, m')$ is defined below.

Next, using the definition of Ω_π and assumption **(A2.2)**, we have

$$(20) \quad \begin{aligned} & \frac{1}{n} \left| \sum_{i=1}^n \int (\lambda_m - \lambda_A)^2(y) \mathbf{I}_{\{Y_i \geq y\}} (\tilde{\pi} - \pi_A)(y) dy \right| \mathbf{I}_{\Omega_\pi} \\ & \leq \frac{1}{4n} \sum_{i=1}^n \int (\lambda_m - \lambda_A)^2(y) \mathbf{I}_{\{Y_i \geq y\}} \pi_A(y) dy \leq \frac{1}{4} \|\lambda_m - \lambda_A\|_{n,2}^2. \end{aligned}$$

In the same way,

$$(21) \quad \frac{1}{n} \left| \sum_{i=1}^n \int (\hat{\lambda}_{\hat{m}} - \lambda)^2(y) \mathbf{I}_{\{Y_i \geq y\}} (\tilde{\pi} - \pi)(y) dy \right| \mathbf{I}_{\Omega_\pi} \leq \frac{1}{4} \|\hat{\lambda}_{\hat{m}} - \lambda\|_{n,2}^2.$$

Moreover, using first the Schwarz Inequality and then $2xy \leq x^2/a + ay^2$ for x, y and $a > 0$ (here $a = 16$), we get

$$\begin{aligned} & \frac{2}{n} \left| \sum_{i=1}^n \int (\hat{\lambda}_{\hat{m}} - \lambda_m)(y) \lambda(y) \mathbf{I}_{\{Y_i \geq y\}} (\tilde{\pi} - \pi_A)(y) dy \right| \leq \frac{1}{16n} \sum_{i=1}^n \int (\hat{\lambda}_{\hat{m}} - \lambda_m)^2(y) \mathbf{I}_{\{Y_i \geq y\}} \pi(y) dy \\ & \quad + \frac{16}{n} \sum_{i=1}^n \int [(\tilde{\pi} - \pi_A)^2(y) / \pi(y)] \lambda^2(y) \mathbf{I}_A(y) dy \\ & \leq \frac{1}{16} \|\hat{\lambda}_{\hat{m}} - \lambda_m\|_{n,2}^2 + \frac{16 \|\lambda\|_{\infty, A}}{n \pi_0} \sum_{i=1}^n \int (\tilde{\pi} - \pi_A)^2(y) \lambda(y) \mathbf{I}_A(y) dy \end{aligned}$$

Therefore, as $\tilde{\pi}$ is computed on an independent sample of observations, we get

$$(22) \quad \begin{aligned} & \frac{2}{n} \mathbb{E} \left(\left| \sum_{i=1}^n \int (\hat{\lambda}_{\hat{m}} - \lambda_m)(y) \lambda(y) \mathbf{I}_{\{Y_i \geq y\}} (\tilde{\pi} - \pi)(y) dy \right| \mathbf{I}_\Omega \right) \\ & \leq \frac{1}{8} \mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda\|_{n,2}^2 \mathbf{I}_\Omega) + \frac{1}{8} \|\lambda - \lambda_m\|_{\mu_2}^2 + \frac{16 \|\lambda\|_{\infty, A}}{\pi_0} \mathbb{E} \left(\int_A (\tilde{\pi} - \pi)^2(y) \lambda(y) dy \right) \\ & \leq \frac{1}{8} \mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda\|_{n,2}^2 \mathbf{I}_\Omega) + \frac{1}{8} \|\lambda - \lambda_m\|_{\mu_2}^2 + \frac{16 \|\lambda\|_{\infty, A}}{\bar{L}_0 \pi_0} \mathbb{E} \left(\int_A (\tilde{\pi} - \pi)^2(y) f_Y(y) dy \right) \\ & \leq \frac{1}{8} \mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda\|_{n,2}^2 \mathbf{I}_\Omega) + \frac{1}{8} \|\lambda - \lambda_m\|_{\mu_2}^2 + \frac{16 \|\lambda\|_{\infty, A}}{\bar{L}_0 \pi_0} \mathbb{E}(\|\tilde{\pi} - \pi\|_\rho^2). \end{aligned}$$

Now, gathering (18)–(22), we get, as $1 - 1/2 - 1/4 - 1/8 = 1/8$ and $1 + 1/2 + 1/4 + 1/8 = 15/8$,

$$(23) \quad \begin{aligned} \frac{1}{8} \mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda_A\|_{n,2}^2 \mathbf{I}_\Omega) & \leq \frac{15}{8} \|\lambda_m - \lambda_A\|_{\mu_2}^2 + \text{pen}(m) + \frac{16 \|\lambda\|_{\infty, A}}{\bar{L}_0 \pi_0} \mathbb{E}(\|\tilde{\pi} - \pi\|_\rho^2) \\ & \quad + 16 \mathbb{E} \left(\sup_{h \in B_{m, \hat{m}}(0,1)} \nu_{n,2}^2(h) \right) \\ & \quad + 16 \mathbb{E} \left(\left(\sup_{h \in B_{m, \hat{m}}(0,1)} \nu_{n,1}^2(h) - p(m, \hat{m}) \right)_+ \right) + 16 \mathbb{E}(p(m, \hat{m}) - \text{pen}(\hat{m})). \end{aligned}$$

Now write $\nu_{n,2}(h) = \int_A h(y)\lambda(y)(\hat{L}(y) - \mathbb{E}(\hat{L}(y)))\pi(y)dy$ with $\hat{L}(y) = (1/n) \sum_{i=1}^n \mathbf{1}_{Y_i \geq y}$. Clearly

$$\begin{aligned} \sup_{h \in B_{m, \hat{m}}(0,1)} \nu_{n,2}^2(h) &\leq \frac{1}{\bar{L}_0^2} \sup_{h \in B_{m, \hat{m}}(0,1)} \int_A h^2(y) d\mu_2(y) \int_A \lambda^2(y) (\hat{L}(y) - \mathbb{E}(\hat{L}(y)))^2 d\mu_2(y) \\ &= \frac{1}{\bar{L}_0^2} \int_A \lambda^2(y) (\hat{L}(y) - \mathbb{E}(\hat{L}(y)))^2 d\mu_2(y). \end{aligned}$$

Therefore, since $\mathbb{E}[(\hat{L}(y) - \mathbb{E}(\hat{L}(y)))^2] = \bar{L}(y)(1 - \bar{L}(y))/n \leq 1/(4n)$ and $\pi(y) \leq 1$, we get

$$(24) \quad \mathbb{E} \left(\sup_{h \in B_{m, \hat{m}}(0,1)} \nu_{n,2}^2(h) \right) \leq \frac{\|\lambda\|_{A, \infty}^2}{4n\bar{L}_0^2} \int \bar{L}(y) dy = \frac{\|\lambda\|_{A, \infty}^2 \mathbb{E}(Y_1)}{4n\bar{L}_0^2}.$$

We can use Talagrand Inequality (see Talagrand Talagrand (1996)) to prove the proposition below, and we refer to Brunel *et al.* Brunel *et al.* (2013) for details.

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

$$p(m, m') = (\kappa/16) \left(\|\lambda\|_{A, \infty}^2 \frac{D_m + D_{m'}}{n} \right),$$

we have

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

and this, together with (23) and (24) yields, as $16p(m, m') \leq \text{pen}(m) + \text{pen}(m')$, that

$$\frac{1}{8} \mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda_A\|_{n,2}^2 \mathbf{1}_\Omega) \leq \frac{15}{8} \|\lambda_m - \lambda_A\|_{\mu_2}^2 + 2\text{pen}(m) + \frac{C}{n}.$$

Next, we study $\mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda\|_{n,2}^2 \mathbf{1}_{\Omega^c})$. We have first (see Brunel *et al.* Brunel *et al.* (2013)):

Lemma 2. *Under the assumptions of Theorem 1, we have $\|\hat{\lambda}_{\hat{m}}\|_{n,2}^2 \leq Cn^3$ for a constant C depending on the basis.*

Moreover **(A2.2)** ensures that $\mathbb{P}(\Omega_\pi^c) \leq C/n^k$, **(A2.1)** implies that $\mathbb{P}(\Omega_{L_0}^c) \leq C_k^{(L_0)}/n^k$ and Proposition 4 in Comte *et al.* (2008) can be adapted here to get $\mathbb{P}(\Delta^c) \leq C_k^{(\Delta)}/n^k$, under the condition that $\mathcal{D}_n^2 \leq n/\log^2(n)$ if basis (1) is used and $\mathcal{D}_n \leq n/\log^2(n)$ if basis (2) is used. Gathering these elements yields $\mathbb{E}(\|\hat{\lambda}_{\hat{m}} - \lambda\|_{n,2}^2 \mathbf{1}_{\Omega^c}) \leq C/n$ by choosing $k = 4$.

For the method allowing to replace $\|\lambda\|_{A, \infty}$ by an estimator, we refer to Comte *et al.* (2008). \square

6. BIBLIOGRAPHY

REFERENCES

- Brunel E., Comte F., (2005) Penalized contrast estimation of density and hazard rate with censored data, *Sankhya*, 67, Part 3, 441-475.
- Brunel, E., Comte, F. and Guilloux, A. (2013). Nonparametric estimation for survival data with censoring indicators missing at random. Preprint HAL <http://hal.archives-ouvertes.fr/hal-00679799>.
- 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.

- Little, R.J.A Rubin, D.B. (2002). *Statistical analysis with missing data*. Second edition. Wiley Series in Probability and Statistics.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley and Sons, Inc., New York.
- 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. (2004) Asymptotically efficient estimation of a survival function in the missing censoring indicator model. *J. Nonparametr. Stat.* **16**, 797-817.
- Subramanian, S. (2006) Survival analysis for the missing censoring indicator model using kernel density estimation techniques. *Stat. Methodol.* **3**, 125-136.
- Subramanian, S. (2009). The multiple imputations based Kaplan-Meier estimator. *Statist. Probab. Lett.* **79**, 1906-1914.
- Subramanian, S. (2011). Multiple imputations and the missing censoring indicator model. *J. Multivariate Anal.* **102**, 105-117
- 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.