

Conditional Mean Residual Life Estimation

Elodie Brunel, Fabienne Comte

▶ **To cite this version:**

Elodie Brunel, Fabienne Comte. Conditional Mean Residual Life Estimation. Journal of Nonparametric Statistics, 2011, 23 (2), pp.471-495. 10.1080/10485252.2010.544731 . hal-00413413

HAL Id: hal-00413413
https://hal.science/hal-00413413

Submitted on 4 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CONDITIONAL MEAN RESIDUAL LIFE ESTIMATION

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

ABSTRACT. In this paper, we consider the problem of nonparametric mean residual life (MRL) function estimation in presence of covariates. We propose a contrast that provides estimators of the bivariate conditional MRL function, when minimized over different collections of linear finite dimensional function spaces. Then we describe a model selection device to select the best estimator among the collection, in the MISE sense. A non-asymptotic oracle inequality is proved for the estimator, that both ensures the good finite sample performances of the estimator and allows us to compute asymptotic rates of convergence. Lastly, examples and simulation experiments illustrate the method.

Key words: adaptive methods, covariate, mean residual life, nonparametric estimation, survival analysis.

September 4, 2009

1. INTRODUCTION

In randomized clinical trials, survival times are often measured from randomization or treatment implementations. But studying survival functions or hazard rates may be inadequate to answer a patient asking, during the trial, how much more time he still has or whether the new treatment improves his life expectancy. To correctly address these questions, life expectancy must be studied as a function of time, via the so-called mean residual life (MRL) function:

$$e(y) = \mathbb{E}(Y - y | Y > y) \ , \ y > 0, \tag{1.1}$$

where Y is a lifetime (i.e. a nonnegative random variable) with $\mathbb{E}(Y) < +\infty$. This function − average remaining life of a surviving subject − is of interest in several other application fields, such as reliability or actuarial studies. For a discussion concerning statistical applications of the MRL, we refer to Embrechts *et al.* (1997). If we denote by F the cumulative distribution function (cdf) and by $\bar{F} = 1 - F$ the survival function, we have the following formula for the MRL e:

$$e(y) = \begin{cases} \int_y^{+\infty} \overline{F}(u)du/\overline{F}(y) & \text{if } \overline{F}(y) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

This equality leads to several proposals of nonparametric estimators, built by plug-in of Kaplan-Meier survival estimators, see Hall & Wellner (1981) or Csörgö & Zitikis (1996) and the references therein. Under adequate assumptions, these estimators inherit the parametric rates of the Kaplan-Meier estimator, but unfortunately they are not smooth. To circumvent this drawback, regularized estimators based on kernel smoothing have been proposed by Chaubey & Sen (1999) or Abdous & Berred (2005).

(1) I3M, UMR 5149 CNRS, Montpellier 2 University, FRANCE, email: ebrunel@math.univ-montp2.fr

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

To measure the combined effect of a covariate X on the MRL, we shall rather define and study the conditional MRL:

$$e(y|x) = \mathbb{E}(Y - y|Y > y, X = x) = \begin{cases} \int_y^{+\infty} \overline{F}(u|x)du/\overline{F}(y|x) & \text{if } \overline{F}(y|x) > 0 \\ 0 & \text{otherwise,} \end{cases} \quad (1.2)$$

where $\bar{F}(y|x)$ is the conditional survival function of Y given $X = x$:

$$\bar{F}(y|x) = \mathbb{P}(Y > y|X = x) = \frac{\int_y^{+\infty} f_{(X,Y)}(x,u)du}{f_X(x)} \text{ if } f_X(x) > 0.$$

Here $f_{(X,Y)}$ denotes the joint probability density of (X, Y) and f_X denotes the marginal density of X. In semi-parametric regression analysis, Oakes & Dasu (1990) propose a proportional mean residual life model to study the association with related covariates, when the response is completely observed. This model is studied in Maguluri & Zhang (1994). Then, Chen & Cheng (2005) and Chen *et al.* (2005) have developed strategies in this model for censored response.

In this paper, we propose a minimum contrast estimator of the conditional MRL. To our knowledge, this is the first purely nonparametric approach. To be more precise, we propose a regression-type contrast that we minimize over collections of finite dimensional functional spaces spanned by orthonormal bases, called models. This produces a collection of estimators among which the best one, in a sense to be defined, is chosen by using a penalization device. The resulting estimator is proved to satisfy an oracle type inequality. This type of nonparametric strategy has been introduced by Barron *et al.* (1999) among others, but using it in the present setting is new.

We describe in Section 2 our estimation strategy. First, we present the contrast which is minimized in the following. Then we give the conditions on the spaces over which the contrast is minimized: this corresponds to model collections for which examples are provided. The procedure is completed by a model selection performed via a penalization of the minimal contrast. Then MISE bound is given in Section 3 and illustrated by asymptotic rate over Besov spaces. Illustrations are provided in Section 4 and concluding remarks are stated in Section 5. Most proofs are relegated in Section 6.

2. Estimation strategy

2.1. **Definition of the contrast.** Let Y be a nonnegative random variable and X a one-dimensional covariate. We assume that the joint density $f_{(X,Y)}$ of (X, Y) is such that

$$\bar{F}_1(x,y) = \int_y^{+\infty} f_{(X,Y)}(x,u)du \text{ and } \bar{F}_2(x,y) = \int_y^{+\infty} \bar{F}_1(x,u)du$$

are measurable and finite nonnegative functions. Then, it is interesting to remark that

$$e(y|x) = \frac{\bar{F}_2(x,y)}{\bar{F}_1(x,y)} \text{ if } \bar{F}_1(x,y) > 0.$$

This holds by simplification by $f_X(x)$ in Formula (1.2).

Next, we consider two functions S and T such that $\iint S^2(x,y)\bar{F}_1(x,y)dxdy < +\infty$ and $\iint T^2(x,y)\bar{F}_1(x,y)dxdy < +\infty$. We define the μ-scalar product of S and T by

$$\langle S, T\rangle_\mu = \iint S(x,y)T(x,y)d\mu(x,y) \text{ with } d\mu(x,y) = \bar{F}_1(x,y)dxdy \quad (2.1)$$

and by $\|.\|_\mu$ the associated norm: $\|T\|_\mu^2 = \langle T, T\rangle_\mu$. This is meaningful as $\bar{F}_1(x,y) \geq 0$ and we will work on fields where $\bar{F}_1(x,y) > 0$.

Let $T : (x, y) \mapsto T(x, y)$ be a bivariate measurable compactly supported function, with support denoted by $A = A_1 \times A_2$. We propose to study the following contrast for estimating the conditional MRL $e(y|x)$:

$$\Gamma_n(T) = \frac{1}{n}\sum_{i=1}^{n}\left(\int T^2(X_i, y)\mathbb{1}_{\{Y_i \geq y\}}dy - 2\Psi_T(X_i, Y_i)\right) \tag{2.2}$$

where

$$\Psi_T(x, y) = \int_0^y (y - u)T(x, u)du.$$

This contrast is justified by the following result:

Proposition 2.1. *Assume that T and e are μ-square integrable. Then, under the following assumption:*

(A0) *For all $x \in A_1$, $\lim_{y\to+\infty} y\bar{F}_1(x, y) = 0$,*

we have: $\mathbb{E}(\Gamma_n(T)) = \|T - e\|_\mu^2 - \|e\|_\mu^2$.

Therefore, minimizing the contrast $\Gamma_n(T)$ over a large set of functions should mean minimizing the empirical counterpart of $\|T - e\|_\mu^2$ and lead to find the function T which is the "nearest" of e among a given class of functions.

Proof of Proposition 2.1. Let us compute

$$\mathbb{E}(\Gamma_n(T)) = \mathbb{E}\left(\int T^2(X_1, y)\mathbb{1}_{\{Y_1 \geq y\}}dy - 2\Psi_T(X_1, Y_1)\right). \tag{2.3}$$

First, the Fubini-Tonelli Theorem implies that

$$\begin{aligned}\mathbb{E}\left(\int T^2(X_1, y)\mathbb{1}_{\{Y_1 \geq y\}}dy\right) &= \iint\left(\int T^2(x, y)\mathbb{1}_{\{u \geq y\}}dy\right) f_{(X,Y)}(x, u)dxdu \\ &= \iint T^2(x, y)\bar{F}_1(x, y)dxdy, \qquad (2.4)\end{aligned}$$

and the last term is finite as T is μ-square integrable. Secondly, an integration by part yields

$$\int_u^{+\infty}(y - u)f_{(X,Y)}(x, y)dy = [-(y - u)\bar{F}_1(x, y)]_{y=u}^{y=+\infty} + \int_u^{+\infty}\bar{F}_1(x, y)dy = \bar{F}_2(x, u)$$

under the condition (A0). Therefore, with the Fubini-Tonelli Theorem first and next the Cauchy-Schwarz Inequality, we have

$$\begin{aligned}\iiint_0^y (y - u)|T(x, u)|f_{(X,Y)}(x, y)dudxdy &= \iint |T(x, u)|\bar{F}_2(x, u)dudx \\ &= \iint |T(x, u)|e(u|x)\bar{F}_1(x, u)dudx \\ &\leq \left(\iint T^2(x, u)\bar{F}_1(x, u)dudx \iint e^2(u|x)\bar{F}_1(x, u)dudx\right)^{1/2}.\end{aligned}$$

The last bound is finite since T and e are μ-square integrable. Thus, the Fubini Theorem can be applied to get:

$$\begin{aligned}\mathbb{E}\left(\Psi_T(X_1,Y_1)\right) &= \iint\int_0^y (y-u)T(x,u)du f_{(X,Y)}(x,y)dxdy \\ &= \iint\left(\int \mathbf{I}_{\{u\le y\}}(y-u)f_{(X,Y)}(x,y)dy\right)T(x,u)dxdu \\ &= \iint T(x,u)\bar{F}_2(x,u)dxdu. \qquad (2.5)\end{aligned}$$

Now, gathering (2.3), (2.4) and (2.5) yields

$$\mathbb{E}(\Gamma_n(T)) = \iint [T^2(x,y) - 2T(x,y)e(y|x)]\bar{F}_1(x,y)dxdy = \|T\|_\mu^2 - 2\langle T, e\rangle_\mu$$

To end the proof, we can see that $\mathbb{E}(\Gamma_n(T)) = \|T-e\|_\mu^2 - \|e\|_\mu^2$. □

Remark 2.1. Since $\bar{F}_1(x,y) = \bar{F}(y|x)f_X(x)$, Assumption (A0) is easily satisfied when the cdf $y \mapsto F(y|x)$, for fixed x, belongs to exponential family laws (and has exponential rate of decay w.r.t. y).

Remark 2.2. Note that the simpler contrast defined for a function $t(\cdot)$ of one variable by:

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^n \left(\int t^2(y)\mathbb{1}_{\{Y_i\ge y\}}dy - 2\psi_t(Y_i)\right) \text{ where } \psi_t(y) = \int_0^y (y-u)t(u)du \qquad (2.6)$$

would lead to build an estimator of $e(.)$ in the non-conditional setting (in which case $e(.)$ is univariate). But this would not improve the rate of convergence. Indeed nonparametric estimators obtained by substituting Kaplan-Meier estimators (or empirical distribution functions) to the true survival functions achieve the parametric rate $\sqrt{n}$ (see Hall & Wellner (1981) or Csörgö & Zitikis (1996)): therefore, it cannot be improved by any strategy. The contrast proposed in (2.6) would only avoid to consider directly a quotient estimator. Note that this estimator would have interesting properties in a nonasymptotic point of view. But, this is not in the scope of this paper.

Now, we need to specify the set of functions T that are considered here, to check that the minimum Γ_n-contrast estimator can adequately be defined, including possibly model selection, and to prove an oracle risk-bound result.

2.2. **Assumptions and collections of linear spaces.** Let us mention first that we provide an estimator of e on a compact set only. We denote this compact by $A = A_1 \times A_2$ and the collection of spaces are defined with respect to this compact set. We will use norms referring to this compact set, for $T \in (\mathbb{L}^2 \cap \mathbb{L}^\infty)(A)$:

$$\|T\|_A^2 = \iint_A T^2(x,y)dxdy, \quad \|T\|_{\infty,A} = \sup_{(x,y)\in A}|T(x,y)|.$$

Moreover, our assumptions are also related to this compact set:

(A1) There exist $\bar{F}_0, f_1 > 0$ such that $\forall (x,y) \in A_1 \times A_2$, $\bar{F}_1(x,y) \ge \bar{F}_0$ and $f_X(x) \le f_1$.
(A2) $\forall (x,y) \in A_1 \times A_2$, $e(y|x) \le \|e\|_{\infty,A} < +\infty$.

Assumptions (A1) and (A2) are weak because the bounds are required on a compact set only. As $\bar{F}_1(x,y)/f_X(x)$ is a conditional survival function, it is bounded by 1, thus $f_X(x) \leq f_1$ in (A1) implies $\bar{F}_1(x,y) \leq f_1$ for $(x,y) \in A$. Therefore Assumption (A1) implies that $\forall (x,y) \in A,\ \bar{F}_0 \leq \bar{F}_1(x,y) \leq f_1$ i.e. the reference measure of the problem here $d\mu(x,y) = \bar{F}_1(x,y)dxdy$ is equivalent to the Lebesgue measure on A.

Now, we introduce a collection $\{S_m : m \in \mathcal{M}_n\}$ of projection spaces: S_m is called a model and $\mathcal{M}_n$ is a set of multi-indexes (see the examples below). For each m, the space S_m of functions with support in $A = A_1 \times A_2$ is defined by:

$$S_m = F_m \otimes \mathcal{H}_n = \Big\{ h, \quad h(x,z) = \sum_{j \in J_m} \sum_{k \in \mathcal{K}_n} a_{j,k} \varphi_j^m(x) \psi_k(z),\ a_{j,k} \in \mathbb{R} \Big\},$$

where F_m and $\mathcal{H}_n$ are subspaces of $(\mathbb{L}^2 \cap \mathbb{L}^\infty)(\mathbb{R})$ respectively spanned by two orthonormal bases: $(\varphi_j^m)_{j \in J_m}$ with $|J_m| = D_m$, where D_m is varying and $(\psi_k)_{k \in \mathcal{K}_n}$ with $|\mathcal{K}_n| = \mathcal{D}_n^{(2)}$ is fixed. For all j and all k, the supports of φ_j^m and ψ_k are respectively included in A_1 and A_2. Here, indexes j and k are not necessarily integers, they can be pairs of integers, as in the case of a piecewise polynomial space specified below.

Remark 2.3. From a theoretical point of view, we may consider that the covariates X are in $\mathbb{R}^d$ and consider models of the form $S_m = F_{m_1} \otimes \dots F_{m_d} \otimes \mathcal{H}_n$. The convergence rate would be slower because of the curse of dimensionality.

Let us introduce the following set of assumptions on the models $\{S_m : m \in \mathcal{M}_n\}$, which are usual in model selection techniques. The specificity here is that they mainly concern the x-direction and the spaces $\{F_m : m \in \mathcal{M}_n\}$. We denote by $\mathcal{F}_n$ the space in the collection of the F_m's with maximal dimension denoted by $\mathcal{D}_n^{(1)}$ and note that $\dim(\mathcal{H}_n) = \mathcal{D}_n^{(2)}$.

- ($\mathcal{M}1$) $\mathcal{D}_n^{(1)} \leq n^{1/4}/\sqrt{\log n}$, $\mathcal{D}_n^{(2)} \leq n^{1/4}/\sqrt{\log n}$ and $\forall m, F_m \subset \mathcal{F}_n$.
- ($\mathcal{M}2$) There exists a positive real ϕ_1 such that, for all $m \in \mathcal{M}_n$, we have

$$\forall x \in A_1,\ \sum_{j \in J_m} (\varphi_j^m(x))^2 \leq \phi_1 D_m$$

- ($\mathcal{M}3$) Nesting condition:

$$D_m \leq D_{m'} \Rightarrow F_m \subset F_{m'}.$$

Assumptions ($\mathcal{M}1$)–($\mathcal{M}3$) are not too restrictive. Indeed, they are verified for the spaces F_m on $A_1 = [0,1]$ without loss of generality, spanned by the following bases (see Barron *et al.* (1999)):

- $[T]$ Trigonometric basis: $\text{span}(\varphi_0, \dots, \varphi_{m-1})$ with $\varphi_0 = \mathbb{I}_{[0,1]}$, $\varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)\mathbb{I}_{[0,1]}(x)$, $\varphi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx)\mathbb{I}_{[0,1]}(x)$ for $j \geq 1$. For this model $D_m = m$ and $\phi_1 = 2$ hold.
- $[DP]$ Regular piecewise polynomial basis: polynomials of degree $0, \dots, r$ (where r is fixed) on each interval $[(\ell-1)/2^D, \ell/2^D[$ with $\ell = 1, \dots, 2^D$. In this case, we have $m = (D,r)$, $J_m = \{j = (\ell, d) \in \mathbb{N} \times \mathbb{N},\ 1 \leq \ell \leq 2^D, 0 \leq d \leq r\}$, $D_m = (r+1)2^D$ and $\phi_1 = \sqrt{r+1}$.
- $[W]$ Regular wavelet basis on an interval, as described by Cohen *et al.* (1993).

Remark 2.4. If we denote by $[H]$ the histogram basis defined by: for $A_1 = [0,1]$, $\text{span}(\varphi_1, \ldots, \varphi_{2^m})$ with $\varphi_j = 2^{m/2}\mathbb{I}_{[(j-1)/2^m, j/2^m[}$ for $j = 1, \ldots, 2^m$, we have $D_m = 2^m$, $\phi_1 = 1$. We want to emphasize that $[H]$ is a particular case of both $[DP]$ and $[W]$, and that practical computations with $[H]$ are easier than with any other basis.

Remark 2.5. The first assumption can be weakened for localized basis: for histogram basis, piecewise polynomial basis and wavelets, $(\mathcal{M}1)$ reduces to $\mathcal{D}_n^{(i)} \leq \sqrt{n/\log n}$. Assumption $(\mathcal{M}1)$ implies that there exists for the $F_m \otimes \mathcal{H}_n$'s a global nesting space $\mathcal{S}_n := \mathcal{F}_n \otimes \mathcal{H}_n$ with dimension denoted by N_n. By assumption $(\mathcal{M}1)$, we have $N_n \leq \sqrt{n/\log n}$ but for localized basis $N_n \leq n/\log n$ would be sufficient to prove Theorem 3.1 in Section 3. The condition $(\mathcal{M}2)$ implies a useful link between the sum of the squared basis functions and the dimension of the space F_m. The third assumption $(\mathcal{M}3)$ implies in particular that $\forall m, m' \in \mathcal{M}_n$, $S_m + S_{m'} \subset \mathcal{S}_n := \mathcal{F}_n \otimes \mathcal{H}_n$.

2.3. **Definition of the estimator.** The first step would be to define $\hat{e}_m = \arg\min_{T \in S_m} \Gamma_n(T)$. To that end, let $T(x,y) = \sum_{j \in J_m} \sum_{k \in \mathcal{K}_n} a_{j,k} \varphi_j^m(x) \psi_k(y)$ be a function in S_m. To compute $\hat{e}_m$, we have to solve:

$$\forall j_0 \in J_m \forall k_0 \in \mathcal{K}_n, \quad \frac{\partial \Gamma_n(T)}{\partial a_{j_0,k_0}} = 0$$

or equivalently for all $j_0 \in J_m, k_0 \in \mathcal{K}_n$,

$$\sum_{j \in J_m} \sum_{k \in \mathcal{K}_n} a_{j,k} \frac{1}{n} \sum_{i=1}^n \varphi_j^m(X_i) \varphi_{j_0}^m(X_i) \int \psi_k(z) \psi_{k_0}(z) \mathbb{I}_{\{Y_i \geq z\}} dz = \frac{1}{n} \sum_{i=1}^n \varphi_{j_0}^m(X_i) \int_0^{Y_i} (Y_i - u) \psi_{k_0}(u) du.$$

Let $vec(.)$ denote the operator that stacks the columns of a matrix into a vector. The above equation can be summarized by

$$G_m \hat{A}_m = \Upsilon_m,$$

where $\hat{A}_m$ denotes the vector $vec((\hat{a}_{j,k})_{j \in J_m, k \in \mathcal{K}_n})$ of the coefficients of the development of the estimator in the basis,

$$G_m := \Big(\frac{1}{n} \sum_{i=1}^n \varphi_j^m(X_i) \varphi_l^m(X_i) \int \psi_k(z) \psi_p(z) \mathbb{I}_{\{Y_i \geq z\}} dz \Big)_{(j,k),(l,p) \in (J_m \times \mathcal{K}_n)^2}$$

and

$$\Upsilon_m := vec\left(\Big(\frac{1}{n} \sum_{i=1}^n \varphi_j^m(X_i) \int_0^{Y_i} (Y_i - u) \psi_k(u) du \Big)_{j \in J_m, k \in \mathcal{K}_n} \right).$$

Remark 2.6. We want to point several features of G_m. First, it is a square matrix with size $|J_m||\mathcal{K}_n| \times |J_m||\mathcal{K}_n|$. Next, it has nonnegative eigenvalues. Indeed, if $u = vec((u_{j,k})_{j \in J_m, k \in \mathcal{K}_n})$ is a vector, and $u^\top$ denotes its transpose, then

$$u^\top G_m u = \frac{1}{n} \sum_{i=1}^n \int \left(\sum_{j,k} u_{j,k} \varphi_j^m(X_i) \psi_k(z) \right)^2 \mathbb{I}_{\{Y_i \geq z\}} dz \geq 0.$$

Lastly, the matrix G_m can also be written

$$G_m = \frac{1}{n} \sum_{i=1}^n \Phi_m^{(i)} \otimes \Psi_m^{(i)}$$

where $\Phi_m^{(i)} \otimes \Psi_m^{(i)}$ is the tensorial product of two square matrices $\Phi_m^{(i)} := (\varphi_j^m(X_i) \varphi_{j_0}^m(X_i))_{(j,j_0) \in J_m^2}$ and $\Psi_m^{(i)} := (\int \psi_k(z) \psi_{k_0}(z) \mathbb{I}_{\{Y_i \geq z\}} dz)_{(k,k_0) \in \mathcal{K}_n^2}$. For the practical implementation of the estimator, we need to compute the inverse of the matrix G_m.

As G_m may be non invertible, we modify the definition of $\hat{e}_m$ in the following way:

$$\hat{e}_m := \begin{cases} \arg\min_{T\in S_m} \Gamma_n(T) & \text{on } \hat{H}_m \\ 0 & \text{on } \hat{H}_m^c \end{cases}, \tag{2.7}$$

where

$$\hat{H}_m := \left\{ \min \mathrm{Sp}(G_m) \geq \max(\hat{\bar{F}}_0/3, n^{-1/2}) \right\}$$

where $\mathrm{Sp}(G_m)$ denotes the spectrum of G_m i.e. the set of the nonnegative eigenvalues of the matrix G_m. The quantity $\hat{\bar{F}}_0$ is an estimator of the bound $\bar{F}_0$ (the minimum of $\bar{F}_1$ on A, see (A1)). We require that it fulfills the following assumption:

(A3) For any integer $k \geq 1$, $\mathbb{P}(|\hat{\bar{F}}_0 - \bar{F}_0| > \bar{F}_0/2) \leq C_k/n^k$.

An estimator $\hat{\bar{F}}_0$ satisfying (A3) is defined in Comte *et al.* (2008). The definition of $\hat{\bar{F}}_0$ is recalled in Section 3.3. Note that $k = 5$ is enough for the proofs.

The final step is to select the relevant space via the penalized criterion. Here, only one direction requires model selection, namely the x-direction. Indeed, the y-direction keeps the good properties of empirical estimators provided that the y-space is simply chosen as large as possible, as it would be done if no covariate was involved. Therefore, we select a model $\hat{m}$ defined by

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \Big(\Gamma_n(\hat{e}_m) + \mathrm{pen}(m) \Big), \tag{2.8}$$

where $\mathrm{pen}(m)$ is defined in Theorem 3.1 below. Our estimator of e on A is then $\tilde{e} = \hat{e}_{\hat{m}}$.

3. Oracle inequality and rate of convergence

3.1. **Mise bound.** We can prove an oracle-type inequality under the following assumption:

(A4) $Y_1, \cdots, Y_n$ are $\mathbb{R}^+$-supported and $\mathbb{E}(Y_1^k) < +\infty$ for $k = 6$.

Moreover, we denote by e_m the $\mathbb{L}^2$-orthogonal projection on S_m of e rectricted to A. Our main Theorem is the following.

Theorem 3.1. *Assume that* (A0)–(A4) *hold and the model collection satisfies (M1)-(M2)-(M3). Let*

$$\mathrm{pen}(m) = \kappa\phi_1 \frac{\mathbb{E}(Y_1^3) + \ell(A_2)\mathbb{E}(Y_1^2)}{\bar{F}_0} \frac{D_m}{n}, \tag{3.1}$$

where κ is a numerical constant. Then $\tilde{e} = \hat{e}_{\hat{m}}$ with $\hat{m}$ defined by (2.8) with $\mathrm{pen}(m)$ *given by (3.1), satisfies:*

$$\mathbb{E}(\|\tilde{e} - e\|_A^2) \leq C \inf_m \big(\|e - e_m\|_A^2 + \mathrm{pen}(m) \big) + \frac{C'}{n}, \tag{3.2}$$

where C is a constant depending on $\bar{F}_0$ and C' is a constant depending on $\mathbb{E}(Y_1^6)$, $\bar{F}_0$, $\|e\|_{\infty,A}$.

Remark 3.1. Inequality (3.2) shows that the estimator automatically makes the compromise between the square bias $\|e - e_m\|_A^2$ and the variance term which is proportional to the order D_m/n of the penalty.

Remark 3.2. The constant terms in the penalty do not have the same status. The constant κ is numerical and does not depend on any unknown quantity. Roughly speaking, it is universal in the sense that it is not affected by the sampling changes and it can be calibrated over a wide range of models by simulation experiments. The constant ϕ_1 is known when the basis is chosen. On the other hand, the unknown quantities $\mathbb{E}(Y_1^2)$ and $\mathbb{E}(Y_1^3)$ can be estimated by empirical

moments, and $\bar{F}_0$ can be replaced by $\hat{\bar{F}}_0$. For an example of theoretical study of such random penalty, see Comte *et al.* (2008). The results of Theorem 3.1 would be generalized but in an asymptotic setting.

Remark 3.3. We might have worked under the following stronger assumption instead of (A4):

(A'4) $Y_1, \cdots, Y_n$ are bounded random variables and there exists a positive constant $B < +\infty$ such that $0 < Y_1 \leq B$ a.s.

In that case, the proof of the result (3.2) can be made simpler, and we can take the penalty

$$\text{pen}_B(m) = \kappa' \phi_1 \frac{B^3}{\bar{F}_0} \frac{D_m}{n}. \tag{3.3}$$

Here, the bound B is unknown and depends on the observations. Nevertheless, it is possible to estimate it by the simple estimator $\hat{B} = \max_{1 \leq i \leq n} Y_i$.

3.2. **Rates of convergence on Besov spaces.** We can deduce from Theorem 3.1 the order of the risk and the rate of convergence of the estimator. For that purpose, assume that e restricted to A belongs to the anisotropic Besov space $\mathcal{B}^{\boldsymbol{\alpha}}_{2,\infty}(A)$ on A with regularity $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$. Let us recall the definition of $\mathcal{B}^{\boldsymbol{\alpha}}_{2,\infty}(A)$. Let e_1 and e_2 be the canonical basis vectors in $\mathbb{R}^2$ and for $i = 1, 2$, $A^r_{i,h} = \{x \in \mathbb{R}^2; x, x + he_i, \ldots, x + rhe_i \in A\}$. Next, for x in $A^r_{i,h}$, let

$$\Delta^r_{i,h} g(x) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} g(x + khe_i) \tag{3.4}$$

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

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

We say that g is in the Besov space $\mathcal{B}^{\boldsymbol{\alpha}}_{2,\infty}(A)$ if

$$|g|_{B^{\boldsymbol{\alpha}}_{2,\infty}} := \sup_{t>0} \sum_{i=1}^{2} t^{-\alpha_i} \omega_{r_i,i}(g,t) < \infty \tag{3.5}$$

for r_i integers larger than α_i.

The estimation procedure may allow an adaptation of the approximation space to each directional regularity. But, it happens that in the y-direction, the greatest space is directly chosen. Thus, we just have to select a relevant F_m.

Corollary 3.1. *Assume that e restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ with regularity $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ such that $\alpha_1 > 1/2$ and $\alpha_2 > 1$. We consider the spaces [DP] and [W] described in Subsection 2.2 (with the regularity r of the polynomials and the wavelets larger than $\alpha_i - 1$, $i = 1, 2$). Then, for $\mathcal{D}^{(2)}_n = O(\sqrt{n/\log(n)})$, and under the assumptions of Theorem 3.1,*

$$\mathbb{E}(\|e\mathbb{1}_A - \tilde{e}\|^2) = O(n^{-\frac{2\alpha_1}{2\alpha_1+1}}).$$

The proof of Corollary 3.1 is standard and thus omitted (see Brunel *et al.* (2008)).

Thus we obtain a rate of convergence which would be standard for the estimation of a function of one variable with regularity α_1.

3.3. **About Assumption (A3).** We recall here the definition of $\hat{\bar{F}}_0$ which is given in Comte *et al.* (2008), and the assumptions under which it fulfills (A3).

First define

$$\hat{\bar{F}}_m(x,y) = \sum_{j\in J_m, k\in\mathcal{K}_n} \hat{b}_{j,k}\varphi_j^m(x)\psi_k(y), \quad \text{with} \quad \hat{b}_{j,k} = \frac{1}{n}\sum_{i=1}^n \varphi_j^m(X_i)\int \psi_k(y)\mathbb{1}_{\{Y_i\geq y\}}dy.$$

Indeed, it is easy to see that

$$\mathbb{E}(\hat{b}_{j,k}) = \iint \varphi_j^m(x)\psi_k(y)\bar{F}_1(x,y)dxdy = \langle \varphi_j^m\otimes\psi_k, \bar{F}_1\rangle,$$

so that $\hat{\bar{F}}_m(x,y)$ is a natural projection estimator of $\bar{F}_1$. Then take

$$\hat{\bar{F}}_0 = \inf_{(x,y)\in A} \hat{\bar{F}}_{m^*}(x,y)$$

where m^* is chosen such that $\log(n)\leq D_{m^*}\leq n^{1/4}/\sqrt{\log(n)}$, and $\mathcal{D}_n^{(2)} = n^{1/4}/\sqrt{\log(n)}$.
Then Proposition 1 in Comte *et al.* (2008) states that, if $\bar{F}_1(x,y)\in\mathcal{B}^{\beta}_{2,\infty}(A)$ with $\beta=(\beta_1,\beta_2)$ and $\bar{\beta}>1$ and n large enough, then (A3) is fulfilled.
Here $\bar{\beta}$ denotes the harmonic mean: $\bar{\beta}^{-1} = \frac{1}{2}(\beta_1^{-1}+\beta_2^{-1})$.

4. Examples and illustration

We give numerical illustrations for some classical regression models used in lifetime analysis. The description and the parametric inference of these models are detailed in Chapter 6 of Lawless (2003).

4.1. **A.F.T. model (Accelerated Failure Time model).** Let $\sigma>0$ and $\mu:\mathbb{R}\mapsto\mathbb{R}$ and consider the model:

$$\ln(Y) = \mu(X) + \sigma\varepsilon, \text{ and } \quad \varepsilon \text{ independent of } X.$$

Then we have

$$\begin{aligned}\mathbb{P}(Y>y|X=x) &= \mathbb{P}(\ln(Y)>\ln(y)|X=x) = \mathbb{P}(\varepsilon > \frac{\ln(y)-\mu(x)}{\sigma}|X=x)\\ &= \bar{F}_\varepsilon\left(\frac{\ln(y)-\mu(x)}{\sigma}\right).\end{aligned}$$

where $\bar{F}_\varepsilon$ stands for the survival function of the noise ε. Therefore, we can write

$$e(y|x) = \frac{\int_y^{+\infty}\bar{F}_\varepsilon\left(\frac{\ln(u)-\mu(x)}{\sigma}\right)du}{\bar{F}_\varepsilon\left(\frac{\ln(y)-\mu(x)}{\sigma}\right)}.$$

• *Example 1*: Take $\bar{F}_\varepsilon(x) = \exp(-\exp(x))$ and $\sigma=1$. Then $e_1(y|x) = e^{\mu(x)}$, the conditional expectation does not depend on y.
• *Example 2*: Take $\bar{F}_\varepsilon(x) = \exp(-\exp(x))$ and $\sigma=2$. Then

$$\int_y^{+\infty}\exp(-\sqrt{u}e^{-\frac{\mu(x)}{2}})du = 2e^{\mu(x)/2}\int_{\sqrt{y}e^{-\frac{\mu(x)}{2}}}^{+\infty} e^{-v}vdv = 2e^{\mu(x)}(1+\sqrt{y}e^{-\frac{\mu(x)}{2}})\exp(-\sqrt{y}e^{-\frac{\mu(x)}{2}}).$$

Thus

$$e_2(y|x) = 2e^{\mu(x)}(1+\sqrt{y}e^{-\frac{\mu(x)}{2}}). \tag{4.1}$$

• *Example 3*: Take $\bar{F}_\varepsilon(x) = (1+\exp(x))^{-1}$ and $\sigma = 1/2$. Then

$$\bar{F}(y|x) = \frac{1}{1+y^2e^{-2\mu(x)}}.$$

$$\begin{aligned}\int_y^{+\infty} \frac{du}{1+u^2e^{-2\mu(x)}} &= e^{\mu(x)} \int_{ye^{-\mu(x)}}^{+\infty} \frac{dv}{1+v^2} = e^{\mu(x)}(\arctan(+\infty) - \arctan(ye^{-\mu(x)})) \\ &= e^{\mu(x)}(\frac{\pi}{2} - \arctan(ye^{-\mu(x)})) = e^{\mu(x)} \arctan(y^{-1}e^{\mu(x)}).\end{aligned}$$

This yields

$$e_3(y|x) = (e^{\mu(x)} + y^2e^{-\mu(x)}) \arctan(y^{-1}e^{\mu(x)}). \tag{4.2}$$

We shall take affine functions $\mu(x)$.

4.2. **Generalized Cox model.** The standard Cox model assumes that the conditional hazard rate α can be decomposed in the following multiplicative way: $\alpha(y|x) = \exp(\beta x)\alpha_0(y)$. It was generalized by Castellan & Letué (2000) for nonparametric estimation purpose by the general equation

$$\alpha(y|x) = \exp(\mu(x))\alpha_0(y).$$

As conditional cumulative hazard denoted by $A(y|x)$ is related to conditional survival function $\bar{F}(y|x)$ by: $A(y|x) = \int_0^y \alpha(u|x)du = -\ln(\bar{F}(y|x))$, we have

$$\bar{F}(y|x) = \exp\left(-\int_0^y \alpha(u|x)du\right).$$

Let us denote $A_0(y) = \int_0^y \alpha_0(u)du$. We find

$$e(y|x) = \int_y^{+\infty} \exp(-e^{\mu(x)}(A_0(v) - A_0(y)))dv.$$

It is worth noting that for $A_0(y) = \lambda y$, that is constant hazard $\alpha_0(y) = \lambda$, this model gives $e(y|x) = e^{-\mu(x)}/\lambda$ which is the same model as the first AFT model above.

• *Example 4*: We can consider the case $\alpha_0(y) = \lambda y$ and $A_0(y) = \lambda y^2/2$. Let $\Phi(u) = \int_{-\infty}^u \exp(-v^2/2)dv$. Then we find

$$e_4(y|x) = \frac{1}{\sqrt{\lambda}} \exp\left[\frac{1}{2}(\lambda y^2 e^{\mu(x)} - \mu(x))\right] (1 - \Phi(\sqrt{\lambda} y e^{\frac{\mu(x)}{2}})). \tag{4.3}$$

4.3. **Additive Hazards models.** Additive Hazards models are sometimes useful and are defined, with the same notations as in section 4.2 by:

$$\alpha(y|x) = \alpha_0(y) + \exp(f(x)).$$

Simple calculations give:

$$\bar{F}(y|x) = \exp\left(-\int_0^y \alpha(u|x)du\right) = \exp[-A_0(y) - yf(x)].$$

Then, we find

$$e(y|x) = \int_y^{+\infty} \exp[A_0(y) - A_0(v) + f(x)(y-v)]dv.$$

• *Example 5*: If we take an exponential baseline hazard with parameter λ, $A_0(y) = \lambda y$, we get:

$$e_5(y|x) = \int_y^{+\infty} \exp[(\lambda + f(x))(y-v)]dv = \frac{1}{\lambda + f(x)} \tag{4.4}$$

	$n = 100$	$n = 500$	$n = 1000$
Example 1	0.0372	0.0191	0.0124
	(0.0249)	(0.0093)	(0.0047)
Example 2	0.198	0.067	0.0687
	(0.344)	(0.0645)	(0.0542)
Example 3	0.0274	0.0191	0.0084
	(0.0854)	(0.1184)	(0.0533)
Example 4	0.0045	0.0011	0.0007
	(0.0018)	(0.0003)	0.0001
Example 5	0.3203	0.1291	0.1009
	(0.2691)	(0.0712)	(0.0577)

TABLE 1. Empirical MISE's averaging over 500 sample replications with their standard deviations given in parenthesis for examples (1)–(5) and different sample size $n = 100, 500$ and 1000.

4.4. **Monte-Carlo study.** We study the numerical performances of our penalized estimator by generating samples $(X_i, Y_i)_{i=1}^n$ following the models described in the previous sections:

- *Example 1*: $e_1(y|x) = e^{\mu(x)}$ with $\mu(x) = ax + b$ with $a = 2$, $b = -2$ and $X \sim \mathcal{U}([0, 1])$.
- *Example 2*: $e_2(y|x)$ given by (4.1) with $\mu(x) = ax + b$ with $a = 1$, $b = -2$ and $X \sim \mathcal{U}([0, 1])$.
- *Example 3*: $e_3(y|x)$ given by (4.2) with $\mu(x) = ax + b$ with $a = 0.5$, $b = -2$ and $X \sim \chi^2(8)/16$.
- *Example 4*: $e_4(y|x)$ given by (4.3) with $\lambda = 2$ $\mu(x) = ax$ with $a = 5$ and $X \sim \chi^2(8)/16$.
- *Example 5*: $e_5(y|x)$ given by (4.4) with $f(x) = x^5$, $\lambda = 0.8$ and $X \sim \mathcal{U}([0, 1])$.

The sets $A = A_1 \times A_2$ are fixed intervals, roughly calibrated with respect to each distribution. In practice, we would have chosen the compact sets of estimation with respect to the data (and their extreme values). Here, we fixed them for reproducibility of the experiments in order to have the same set of estimation for all paths. This is of course only possible in a simulation setting. We illustrate the practical implementation of our estimator for histogram bases [H].

The penalty is chosen as follows:

$$C \left(\max_{1 \le i \le n} (Y_i) \right)^3 \frac{D_{m_1}}{n}. \tag{4.5}$$

It corresponds to the empirical version of (3.3) with constant $C = \kappa \phi_1 / \hat{\bar{F}}_0$ calibrated as 10 (here, $\phi_1 = 1$). Even if the generated observations come from $\mathbb{R}^+$-supported probability laws, we can consider in practice that the probability they fall outside a fixed bounded interval is very small. A preliminary and rough study indicates that for all five models considered here this value of κ has the adequate penalization effects. But, a more tedious calibration study maybe investigated in particular to compare both penalization bounds (3.3) and (3.1). The algorithm selects the x-dimension $D_m^{(1)}$ less than $\sqrt{n}$ whereas the y-dimension $\mathcal{D}_n^{(2)}$ is fixed to the maximal value n.

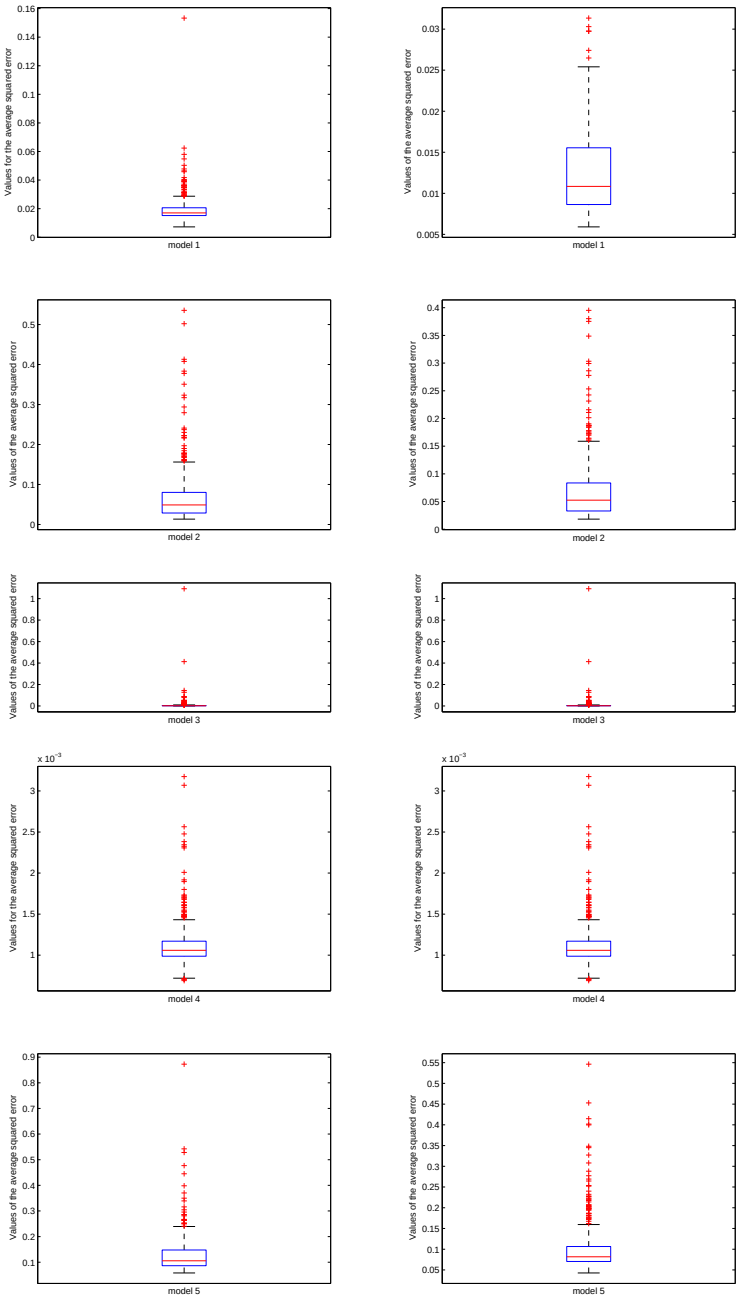


FIGURE 1. Distribution of the MISE's values computed for 500 sample replications for examples (1)–(5) for size $n = 500$ (left) and $n = 1000$ (right).

We compute the empirical MISE (Mean Integrated Squared Error) over $N = 500$ replications of the samples, by averaging over the paths $j = 1, \dots, N$, the quantities

$$\frac{\ell(A_1)\ell(A_2)}{K^2} \sum_{k,\ell=1}^{K} (\tilde{e}^{(j)}(y_\ell|x_k) - e(y_\ell|x_k))^2, \tag{4.6}$$

where $\ell(A_i)$ is the length of the interval A_i, $i = 1, 2$, $(x_k)_{1\leq k\leq K}$, $(y_k)_{1\leq k\leq K}$ are uniform subdivisions of A_1 and A_2 respectively, and $\tilde{e}^{(j)}$ is the estimator associated to the jth sample path. Note that the computed error given by (4.6) is the empirical version of the $\mathbb{L}^2$- risk

$$\mathbb{E}\left(\iint_{A_1\times A_2}(\tilde{e}(y_\ell|x_k) - e(y_\ell|x_k))^2 dx_k dy_\ell\right)$$

which corresponds to integrated errors in both x- and y-directions instead of the empirical norm in the x-direction for which oracle-inequality is given in Section 3. The values of the average MISE's are not satisfactory in regard to their associated standard deviations for a sample size $n = 100$ which is to small in this context of (bivariate) estimation. We can see in Figure 1 that for larger sample sizes $n = 500$ and $n = 1000$, we obtain better values. However there are still extreme values. This is true for all examples and all the more noticeable for Example 3: the box is so flat (see Figure 1) because of only two extreme values whereas the standard deviations have the smallest values which indicates a good robustness of the estimation mechanism. On the whole, the automatic selection works well. We provide in Figure 2 a view of typical estimates and typical improvement between sizes $n = 500$ and $n = 2000$.

5. Concluding remark: the censored case

If the variable of interest Y is censored, we can generalize the contrast function. Let the observations be $X_i, Z_i = Y_i \wedge C_i, \delta_i = \mathbb{I}_{\{Y_i\leq C_i\}}$ where C is the censoring random variable. Assume that the strong independence assumption holds:

$$C \text{ is independent of } (X, Y).$$

Then the contrast of interest is

$$\Gamma_n^C(T) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i}{\bar{G}_n(Z_i)}\left(\int T^2(X_i, y)\mathbb{I}_{\{Z_i\geq y\}}dy - 2\Psi_T(X_i, Z_i)\right)\right] \tag{5.1}$$

where $\bar{G}_n$ is the modified Kaplan-Meier (1958) estimator for $\bar{G}$, the survival function of the censoring sequence (C_i), $\bar{G}(x) = \mathbb{P}(C \geq x)$, as given in Lo *et al.* (1989). It is defined by

$$\bar{G}_n(x) = \begin{cases} \displaystyle\prod_{i=1,\, Z_{(i)}\leq x}^{n}\left(\frac{n-i+1}{n-i+2}\right)^{1-\delta_{(i)}} & \text{if } x \leq Z_{(n)} \\ \bar{G}_n(Z_{(n)}) & \text{if } x > Z_{(n)}. \end{cases} \tag{5.2}$$

This modification of the Kaplan Meier estimator is proposed because the estimate of $\bar{G}$ appears in a denominator. To justify the proposed contrast, it is easy to check, under the independence assumption, that the expectation of Γ_n^C is the same as the one of Γ_n.

6. Proofs

6.1. **Proof of Theorem 3.1.** The line of the proof follows the line of the one given in Comte *et al.* (2008) and many auxiliary results are borrowed from this work.

The following "empirical" norm is involved by the definition of the contrast. For $T \in S_m$, let

$$\|T\|_n^2 := \frac{1}{n}\sum_{i=1}^{n}\int T^2(X_i, y)\mathbb{I}_{\{Y_i\geq y\}}dy.$$

It is related with the scalar product defined by (2.1) by

$$\mathbb{E}(\|T\|_n^2) = \|T\|_\mu^2.$$

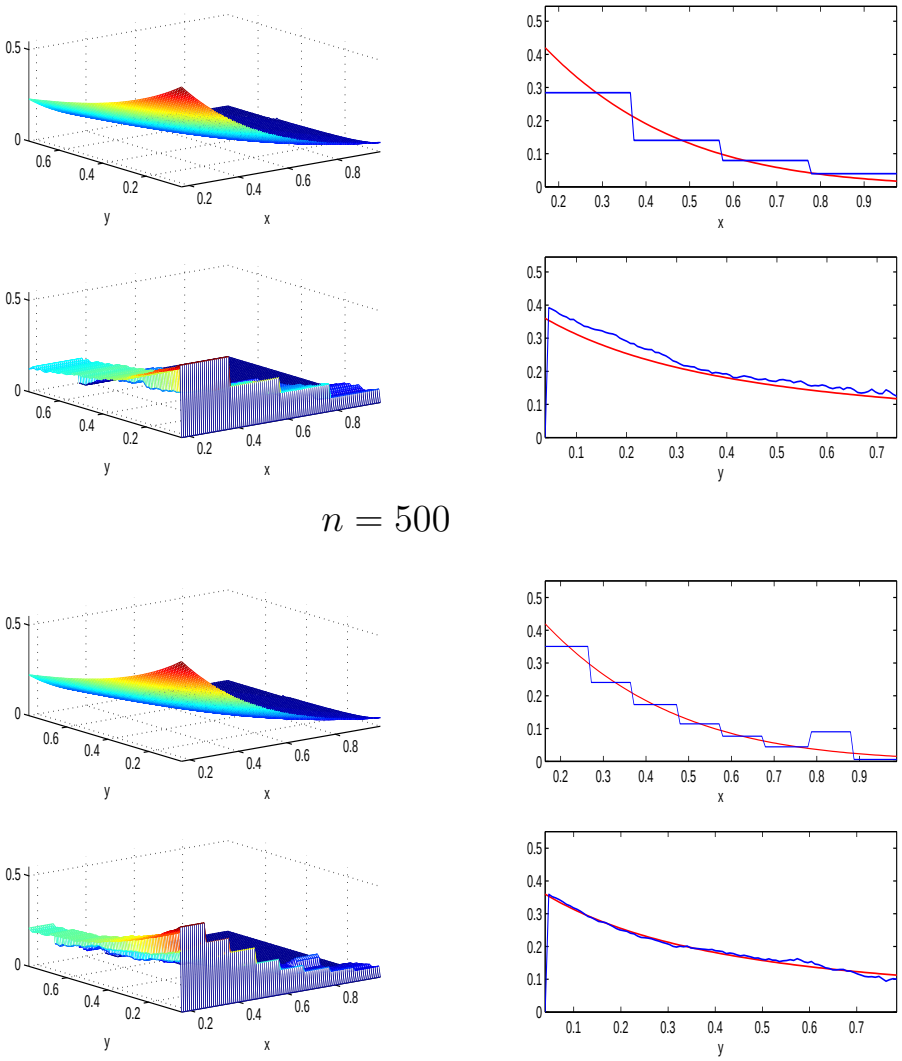


FIGURE 2. Penalized estimator for Example 4 for sample size $n = 500$ and $n = 2000$: top-left: true conditional MRL e, bottom-left: estimator $\tilde{e}$, top-right: $x \mapsto e(y|x)$ and $x \mapsto \tilde{e}(y|x)$ for a fixed value of y, bottom-right: $y \mapsto e(y|x)$ and $y \mapsto \tilde{e}(y|x)$ for a fixed value of x.

Next, we have the following relation between the norm $\|.\|_n$ and the contrast Γ_n:

$$\Gamma_n(T) - \Gamma_n(S) = \|T - e\|_n^2 - \|S - e\|_n^2 - 2\nu_n(T - S), \tag{6.1}$$

where

$$\nu_n(T) = \frac{1}{n}\sum_{i=1}^{n}\left(\Psi_T(X_i, Y_i) - \int T(X_i, y)\mathbf{I}_{\{Y_i \geq y\}} e(y|X_i)dy\right). \tag{6.2}$$

We shall use in the proof the following sets:

$$\hat{H}_m = \{\min \mathrm{Sp}(G_m) \geq \max(\hat{\bar{F}}_0/3, n^{-1/2})\}, \quad \hat{H} := \bigcap_{m \in \mathcal{M}_n} \hat{H}_m,$$

$$\Delta := \Big\{\forall T \in \mathcal{S}_n : \Big|\frac{\|T\|_n^2}{\|T\|_\mu^2} - 1\Big| \leq \frac{1}{2}\Big\}, \text{ and } \Omega := \Big\{\Big|\frac{\hat{\bar{F}}_0}{\bar{F}_0} - 1\Big| \leq \frac{1}{2}\Big\}. \tag{6.3}$$

For $m \in \mathcal{M}_n$, we recall that e_m is the orthogonal projection on $F_m \otimes \mathcal{H}_n$ of e restricted to A. The following bounds hold:

$$\begin{aligned}
\mathbb{E}(\|\hat{e}_{\hat{m}} - e\|_A^2) &\leq 2\|e - e_m\|_A^2 + 2\mathbb{E}(\|\hat{e}_{\hat{m}} - e_m\|_A^2 \mathbb{1}(\Delta \cap \Omega)) \\
&+ 2\mathbb{E}(\|\hat{e}_{\hat{m}} - e_m\|_A^2 \mathbb{1}(\Delta^c \cap \Omega)) + 2\mathbb{E}(\|\hat{e}_{\hat{m}} - e_m\|_A^2 \mathbb{1}(\Omega^c)) \\
&\leq 2\|e - e_m\|_A^2 + 2\mathbb{E}(\|\hat{e}_{\hat{m}} - e_m\|_A^2 \mathbb{1}(\Delta \cap \Omega)) \\
&+ 4\mathbb{E}((\|\hat{e}_{\hat{m}}\|^2 + \|e\|_A^2)\mathbb{1}(\Delta^c \cap \Omega)) + 4\mathbb{E}((\|\hat{e}_{\hat{m}}\|^2 + \|e\|_A^2)\mathbb{1}(\Omega^c)). \qquad (6.4)
\end{aligned}$$

We use the following results, whose proofs can be found in Section 6.2 hereafter for Proposition 6.1 and in Proposition 4 of Comte *et al.* (2008) for Proposition 6.2.

Proposition 6.1. *Provided that* $\mathbb{E}[Y_1^6] < +\infty$, *we have* $\mathbb{E}(\|\hat{e}_{\hat{m}}\|^4) \leq C' n^3$, *where* $C' = \phi_1^2 \mathbb{E}(Y_1^6)/9$ *is a positive constant.*

Proposition 6.2. *If (A1) and* $(\mathcal{M}_1)$ *are fulfilled, we have* $\mathbb{P}(\Delta^c) \leq C_k/n^k$ *for any* $k \geq 1$, *when* n *is large enough, where* C_k *is a constant depending on* $\bar{F}_0$ *and the basis.*

Moreover, (A3) ensures that $\mathbb{P}(\Omega^c) \leq C_k/n^k$ for any integer k. Thus, using Propositions 6.1 and 6.2 and Assumption (A3), we get

$$\begin{aligned}
&\mathbb{E}((\|\hat{e}_{\hat{m}}\|^2 + \|e\|_A^2)\mathbb{1}(\Delta^c \cap \Omega)) + \mathbb{E}((\|\hat{e}_{\hat{m}}\|^2 + \|e\|_A^2)\mathbb{1}(\Omega^c)) \\
\leq\ & \|e\|_A^2(\mathbb{P}(\Omega^c) + \mathbb{P}(\Delta^c)) + \mathbb{E}^{1/2}(\|\hat{e}_{\hat{m}}\|^4)(\mathbb{P}^{1/2}(\Omega^c) + \mathbb{P}^{1/2}(\Delta^c)) \\
\leq\ & C_2/n. \qquad (6.5)
\end{aligned}$$

Thus it remains to study $\mathbb{E}(\|\hat{e}_{\hat{m}} - e_m\|_A^2 \mathbb{1}(\Delta \cap \Omega))$. We state the following Lemma:

Lemma 6.1. *The following embedding holds:*

$$\Delta \cap \Omega \subset \hat{\Gamma} \cap \Omega.$$

As a consequence, for all $m \in \mathcal{M}_n$, *the matrices* G_m *are invertible on* $\Delta \cap \Omega$.

Proof of Lemma 6.1. Let $m \in \mathcal{M}_n$ be fixed and let ℓ be an eigenvalue of G_m. There exists $A_m \neq 0$ with coefficients $(a_\lambda)_\lambda$ such that $G_m A_m = \ell A_m$ and thus $A_m^\top G_m A_m = \ell A_m^\top A_m$. Now, take $T := \sum_\lambda a_\lambda \varphi_\lambda \in S_m$. We have $\|T\|_n^2 = A_m^\top G_m A_m$ and $\|T\|_A^2 = A_m^\top A_m$. Thus, on Δ:

$$A_m^\top G_m A_m = \|T\|_n^2 \geq \frac{1}{2}\|T\|_\mu^2 \geq \frac{1}{2}\bar{F}_0 \|T\|_A^2 = \frac{1}{2}\bar{F}_0 A_m^\top A_m.$$

Therefore, on Δ, for all $m \in \mathcal{M}_n$, we have $\min \mathrm{spec}(G_m) \geq \bar{F}_0/2$. Moreover, on Ω, we have $\bar{F}_0 \geq 2\hat{\bar{F}}_0/3$ and $\max(\hat{\bar{F}}_0/3, n^{-1/2}) = \hat{\bar{F}}_0$, for $n \geq 36/\bar{F}_0^2$. □

Now, on $\Delta \cap \Omega$ we have

$$\Gamma_n(\hat{e}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \Gamma_n(e_m) + \mathrm{pen}(m),$$

where $\hat{e}_{\hat{m}} \in F_{\hat{m}} \otimes \mathcal{H}_n$ and $e_m \in F_m \otimes \mathcal{H}_n$. It follows from (6.1) and (6.2) and from the inequality $2xy \leq x^2/\theta^2 + \theta^2 y^2$, with $x, y, \theta \in \mathbb{R}^+$ (here $\theta = 2$), that, on $\Delta \cap \Omega$,

$$\begin{aligned}\|\hat{e}_{\hat{m}} - e_m\|_n^2 &\leq 2\langle \hat{e}_{\hat{m}} - e_m, e - e_m\rangle_n + \mathrm{pen}(m) + 2\nu_n(\hat{e}_{\hat{m}} - e_m) - \mathrm{pen}(\hat{m}) \\ &\leq \frac{1}{4}\|\hat{e}_{\hat{m}} - e_m\|_n^2 + 4\|e - e_m\|_n^2 + \mathrm{pen}(m) \\ &\quad + \frac{1}{4}\|\hat{e}_{\hat{m}} - e_m\|_\mu^2 + 4 \sup_{T \in B^\mu_{m,\hat{m}}(0,1)} \nu_n^2(T) - \mathrm{pen}(\hat{m}),\end{aligned}$$

where $B^\mu_{m,m'}(0,1) := \{T \in (F_m + F_{m'}) \otimes \mathcal{H}_n : \|T\|_\mu \leq 1\}$. This yields

$$\begin{aligned}\frac{3}{4}\|\hat{e}_{\hat{m}} - e_m\|_n^2 &\leq 4\|e - e_m\|_n^2 + \mathrm{pen}(m) + \frac{1}{4}\|\hat{e}_{\hat{m}} - e_m\|_\mu^2 \\ &\quad + 4\Big(\sup_{T \in B^\mu_{m,\hat{m}}(0,1)} \nu_n^2(T) - p(m,\hat{m})\Big)_+ + 4p(m,\hat{m}) - \mathrm{pen}(\hat{m}),\end{aligned}$$

where $p(m,m') \geq 0$ is defined in the following Proposition.

Proposition 6.3. *Let*

$$p(m,m') = \kappa \frac{\mathbb{E}(Y_1^3) + \ell(A_2)\mathbb{E}(Y_1^2)}{4\bar{F}_0} \frac{D_m + D_{m'}}{n}$$

where κ is a numerical constant. Under the assumptions of Theorem 3.1, we have

$$\mathbb{E}\Big(\sup_{T \in B^\mu_{m,\hat{m}}(0,1)} \Big(\nu_n^2(T) - p(m,\hat{m})\Big)_+ \mathbb{1}(\Delta)\Big) \leq \frac{C_1'}{n}.$$

Now, we can see that the penalty is such that

$$\forall m, m',\ 4p(m,m') \leq \mathrm{pen}(m) + \mathrm{pen}(m'), \tag{6.6}$$

and use the definition of Δ. We obtain on $\Delta \cap \Omega$:

$$\begin{aligned}\frac{1}{2}\|\hat{e}_{\hat{m}} - e_m\|_\mu^2 &\leq 4\|e - e_m\|_n^2 + 2\mathrm{pen}(m) \\ &\quad + \frac{1}{4}\|\hat{e}_{\hat{m}} - e_m\|_\mu^2 + 4\Big(\sup_{T \in B^\mu_{m,\hat{m}}(0,1)} \nu_n^2(T) - p(m,\hat{m})\Big)_+\end{aligned}$$

and thus on $\Delta \cap \Omega$:

$$\begin{aligned}\frac{1}{4}\|\hat{e}_{\hat{m}} - e_m\|_\mu^2 &\leq 4\|e - e_m\|_n^2 + 2\mathrm{pen}(m) \\ &\quad + 4\Big(\sup_{T \in B^\mu_{m,\hat{m}}(0,1)} \nu_n^2(T) - p(m,\hat{m})\Big)_+.\end{aligned}$$

Taking the expectation of the last Inequality and using proposition 6.3, we get:

$$\frac{1}{4}\mathbb{E}(\|\hat{e}_{\hat{m}} - e_m\|_\mu^2 \mathbb{1}(\Delta \cap \Omega)) \leq 4\|e - e_m\|_\mu^2 + 2\mathrm{pen}(m) + \frac{C_1}{n}. \tag{6.7}$$

Gathering (6.4), (6.5) and (6.7) leads to

$$\begin{aligned}\mathbb{E}(\|\hat{e}_{\hat{m}} - e\|_A^2) &\leq 2\|e_m - e\|_A^2 + \frac{8}{\bar{F}_0}\Big(4\|e - e_m\|_\mu^2 + 2\mathrm{pen}(m) + \frac{C_1}{n}\Big) + \frac{C_2}{n} \\ &\leq 2\Big(1 + \frac{16}{\bar{F}_0}\Big)\|e_m - e\|_A^2 + \frac{16}{\bar{F}_0}\mathrm{pen}(m) + \frac{C_3}{n} \qquad (6.8)\end{aligned}$$

for any $m \in \mathcal{M}_n$. This concludes the proof of Theorem 3.1. □

6.2. **Proof of Proposition 6.1.** Let us note that $\hat{e}_{\hat{m}}$ is either 0 or $\arg\min_{T\in S_{\hat{m}}}\Gamma_n(T)$. In the second case, $\min \mathrm{Sp}(G_{\hat{m}}) \geq \max(\hat{\bar{F}}_0, n^{-1/2})$ and thus

$$\begin{aligned}
\|\hat{e}_{\hat{m}}\|^2 &= \sum_{j,k}(\hat{a}^{\hat{m}}_{j,k})^2 = \|A_{\hat{m}}\|^2 = \|G^{-1}_{\hat{m}}\Upsilon_{\hat{m}}\|^2 \\
&\leq (1/\min \mathrm{Sp}(G_{\hat{m}}))^2\|\Upsilon_{\hat{m}}\|^2 \\
&\leq \min(1/\hat{\bar{F}}_0^2, n)\sum_{j,k}\left(\frac{1}{n}\sum_{i=1}^n \varphi^{\hat{m}}_j(X_i)\int_{A_2}\mathbb{I}_{(0\leq u\leq Y_i)}(Y_i-u)\psi_k(u)du\right)^2 \\
&\leq n\frac{1}{n}\sum_{i=1}^n\sum_j(\varphi^{\hat{m}}_j(X_i))^2\sum_k\left(\int_{A_2}\mathbb{I}_{(0\leq u\leq Y_i)}(Y_i-u)\psi_k(u)du\right)^2 \\
&\leq \phi_1\mathcal{D}^{(1)}_n\sum_{i=1}^n\int_{A_2}\mathbb{I}_{(0\leq u\leq Y_i)}(Y_i-u)^2du \leq \phi_1\mathcal{D}^{(1)}_n\sum_{i=1}^n\frac{Y_i^3}{3}.
\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}
\mathbb{E}\left(\|\hat{e}_{\hat{m}}\|^4\right) &\leq \phi_1^2(\mathcal{D}^{(1)}_n)^2\mathbb{E}\left[\left(\sum_{i=1}^n\frac{Y_i^3}{3}\right)^2\right] \\
&\leq \phi_1^2(\mathcal{D}^{(1)}_n)^2n^2\mathbb{E}(Y_1^6)/9 \leq \frac{\phi_1^2\mathbb{E}(Y_1^6)}{9}n^3. \ \square
\end{aligned}$$

6.3. **Proof of Proposition 6.3.** We use several times the same very useful inequality based on the property that the squared norm of the orthogonal projection of a function is less than the squared norm of the function itself. We use this property as follows:

$$\sum_k\left(\int_{A_2}h(v)\psi_k(v)dv\right)^2 \leq \|h\|^2 \tag{6.9}$$

for any function $h\in\mathbb{L}^2(A_2)$.

Let $W=(X,Y)$ and

$$\xi_T(W) = \Psi_T(X,Y) - \int T(X,v)\mathbb{I}_{\{Y\geq v\}}e(v|X)dv.$$

To study the empirical process, we split $\xi_T(W)$ in 3 parts:

$$\xi_T(W) = \xi_{T,1}(W) + \xi_{T,2}(W) - \xi_{T,3}(W),$$

with

$$\begin{aligned}
\xi_{T,1}(W) &= \Psi_T(X,Y)\mathbb{I}_{\{Y\leq k_n\}} - \mathbb{E}(\Psi_T(X,Y)\mathbb{I}_{\{Y\leq k_n\}}), \\
\xi_{T,2}(W) &= \Psi_T(X,Y)\mathbb{I}_{\{Y> k_n\}} - \mathbb{E}(\Psi_T(X,Y)\mathbb{I}_{\{Y> k_n\}}), \\
\xi_{T,3}(W) &= \int T(X,v)\mathbb{I}_{\{Y\geq v\}}e(v|X)dv - \mathbb{E}\left(\int T(X,v)\mathbb{I}_{\{Y\geq v\}}e(v|X)dv\right),
\end{aligned}$$

where $W=(X,Y)$. Then $\nu_n(T) = (1/n)\sum_{i=1}^n\xi_T(W_i)$ can be split in the same way: $\nu_n(T) = \nu_{n,1}(T) + \nu_{n,2}(T) - \nu_{n,3}(T)$ with $\nu_{n,k}(T) = (1/n)\sum_{i=1}^n\xi_{T,k}(W_i)$ for $k=1,2,3$. We choose

$$k_n = \left(\frac{3n}{\log^4(n)}\right)^{1/3}, \tag{6.10}$$

Now, the main tool of the proof is the checkout of Talagrand Inequality (Talagrand (1996)):

Lemma 6.2. *Let $W_1, \cdots, W_n$ be i.i.d. random variables and $(\xi_T)_{T\in\mathcal{B}}$ a set of bounded functions and $\mathcal{B}$ a unit ball of a finite dimensional subspace of $\mathbb{L}^2(A)$. Let $\nu_n(T) = (1/n)\sum_{i=1}^n \xi_T(W_i)$ where $\mathbb{E}[\xi_T(W_1)] = 0$, and suppose that:*

$$i)\ \sup_{T\in\mathcal{B}} \|\xi_T\|_\infty \le M_1, \quad ii)\ \sup_{T\in\mathcal{B}} \mathrm{Var}[\xi_T(W_1)] \le v \quad iii)\ \mathbb{E}\left(\sup_{T\in\mathcal{B}} |\nu_n(T)|\right)^2 \le H^2.$$

Then, there exists constants $K > 0$, $K_1 > 0$ and $K_2 > 0$ such that:

$$\mathbb{E}\left[\sup_{T\in\mathcal{B}} |\nu_n(T)|^2 - 2H^2\right] \le K\left[\frac{v}{n}e^{-K_1\frac{nH^2}{v}} + \frac{M_1^2}{n^2}e^{-K_2\frac{nH}{M_1}}\right]$$

We apply Talagrand's Inequality given in Lemma 6.2 to the terms involving $\nu_{n,1}$ and $\nu_{n,3}$ in the following inequality:

$$\begin{aligned}\mathbb{E}\Big(\sup_{T\in B^\mu_{m,\hat m}(0,1)}\Big(\nu_n^2(T) - p(m,\hat m)\Big)_+\mathbb{I}(\Delta)\Big) \le\ & 3\mathbb{E}\Big(\sup_{T\in B^\mu_{m,\hat m}(0,1)}\Big(\nu_{n,1}^2(T) - p_1(m,\hat m)/6\Big)_+\Big)\\ &+3\mathbb{E}\Big(\sup_{T\in B^\mu_n(0,1)}\Big(\nu_{n,2}^2(T)\Big)\Big)\\ &+3\mathbb{E}\Big(\sup_{T\in B^\mu_{m,\hat m}(0,1)}\Big(\nu_{n,3}^2(T) - p_3(m,\hat m)/6\Big)_+\Big),\end{aligned}$$

where $p(m,m') = p_1(m,m') + p_3(m,m')$ and with $B_n^\mu(0,1) = \{T \in \mathcal{S}_n : \|T\|_\mu \le 1\}$.

• Study of $\nu_{n,2}$.
Recall that $\mathbb{E}(Y_1^6) < +\infty$. We write:

$$\mathbb{E}\Big(\sup_{T\in B^\mu_n(0,1)}\Big(\nu_{n,2}^2(T)\Big)\Big) \le \frac{1}{\bar F_0}\sum_{j,k}\mathbb{E}\Big(\nu_{n,2}^2(\varphi_j^n\otimes\psi_k)\Big),$$

where $(\varphi_j^n\otimes\psi_k)_{j,k}$ denotes here an orthonormal basis of $\mathcal{S}_n$ w.r.t the norm $\|.\|_A$. This implies, as $\mathbb{E}(\nu_{n,2}^2(T)) = (1/n)\mathrm{Var}(\xi_{T,2}(W))$ and using (6.9), that

$$\mathbb{E}\Big(\sup_{T\in B^\mu_n(0,1)}\Big(\nu_{n,2}^2(T)\Big)\Big) \le \frac{2}{n\bar F_0}\sum_{j,k}\left[\mathbb{E}\left(\int (Y_1 - v)\mathbb{I}_{\{v\le Y_1\}}\varphi_j^n(X_1)\psi_k(v)dv\mathbb{I}_{\{Y_1>k_n\}}\right)^2\right],$$

and with $(\mathcal{M}2)$

$$\begin{aligned}\mathbb{E}\Big(\sup_{T\in B^\mu_n(0,1)}\Big(\nu_{n,2}^2(T)\Big)\Big) &\le \frac{2\phi_1\mathcal{D}_n^{(1)}}{n\bar F_0}\mathbb{E}\left(\int_{A_2}(Y_1 - v)^2\mathbb{I}_{\{v\le Y_1\}}dv\mathbb{I}_{\{Y_1>k_n\}}\right)^2\\ &\le \frac{2\phi_1}{\sqrt{n}\bar F_0}\mathbb{E}(Y_1^3\mathbb{I}_{\{Y_1>k_n\}}) \le \frac{2\phi_1}{\sqrt{n}\bar F_0}\frac{\mathbb{E}(Y_1^{3+p})}{k_n^p}\\ &\le \frac{2\phi_1\mathbb{E}(Y_1^{3+p})}{3^{p/3}\bar F_0}\frac{(\log(n))^{4p/3}}{n^{1/2+p/3}} \le \frac{C}{n}\end{aligned}$$

as soon as we take $p > 3/2$, e.g. $p = 2$.
• Study of $\nu_{n,1}$.
We apply Talagrand's Inequality, and for this purpose, we will have to check $i)$, $ii)$ and $iii)$ and to compute M_1, v and H^2 defined in Lemma 6.2.

i) Search for bound M_1. Under (A1), we have,

$$\sup_{T\in B^{\mu}_{m,m'}(0,1)}\sup_{x\in\mathbb{R},y\in\mathbb{R}}|\xi_{T,1}(x,y)|\leq\frac{1}{\sqrt{\bar{F}_0}}\sup_{T\in B_{m,m'}(0,1)}\sup_{x\in\mathbb{R},y\in\mathbb{R}}|\xi_{T,1}(x,y)|$$

Here $B_{m,m'}(0,1)=\{T\in F_{m\vee m'}\otimes\mathcal{H}_n:\|T\|_A\leq 1\}$ and $T=\sum_{j,k}a_{j,k}\varphi_j\psi_k$ where $(\varphi_j\otimes\psi_k)_{(j,k)}$ is an orthonormal basis of $F_{m\vee m'}\otimes\mathcal{H}_n$ w.r.t. the norm $\|.\|_A$, where $F_{m\vee m'}=F_m+F_{m'}$ and $\dim(F_{m\vee m'})=\max(D_m,D_{m'})$ under $(\mathcal{M}3)$ (nested collection).

$$\begin{aligned}
|\Psi_T(x,y)| &= \left|\sum_{j,k}a_{j,k}\varphi_j(x)\int(y-v)\mathbb{I}_{\{v\leq y\}}\psi_k(v)dv\mathbb{I}_{\{y\leq k_n\}}\right| \\
&\leq \left(\sum_{j,k}a_{j,k}^2\sum_j(\varphi_j(x))^2\sum_k\left(\int(y-v)\mathbb{I}_{\{v\leq y\}}\psi_k(v)dv\mathbb{I}_{\{y\leq k_n\}}\right)^2\right)^{1/2} \\
&\leq \|T\|_A\left(\sum_j(\varphi_j(x))^2\int(y-v)^2\mathbb{I}_{\{v\leq y\}}dv\mathbb{I}_{\{y\leq k_n\}}\right)^{1/2} \quad \text{with Inequality (6.9)} \\
&\leq \sqrt{\phi_1(k_n^3/3)\max(D_m,D_{m'})} \text{ with } (\mathcal{M}1)-(\mathcal{M}2), \\
&\leq \sqrt{\phi_1(D_m+D_{m'})n}/\log^2(n)
\end{aligned}$$

for $k_n=(3n/\log^4(n))^{1/3}$ as given by (6.10). Therefore

$$\sup_{T\in B^{\mu}_{m,m'}(0,1)}\|\xi_T\|_\infty\leq 2\frac{\sqrt{\phi_1(D_m+D_{m'})n}}{\log^2(n)\sqrt{\bar{F}_0}}:=M_1.$$

ii) Search for bound v.
First, let $B=\max(D_m,D_{m'})^{1/5}$ and write

$$\begin{aligned}
\mathrm{Var}[\xi_{T,1}(W_1)] &\leq \mathbb{E}\left[\Psi_T^2(X_1,Y_1)\mathbb{I}_{\{Y_1<k_n\}}\right]\leq\mathbb{E}\left[\Psi_T^2(X_1,Y_1)\right] \\
&\leq \mathbb{E}\left[\Psi_T^2(X_1,Y_1)\mathbb{I}_{\{Y_1\leq B\}}\right]+\mathbb{E}\left[\Psi_T^2(X_1,Y_1)\mathbb{I}_{\{Y_1>B\}}\right].
\end{aligned}$$

Now, we study each term. First, we have for $T\in B^{\mu}_{m,m'}(0,1)$

$$\begin{aligned}
\mathbb{E}\left[\Psi_T^2(X_1,Y_1)\mathbb{I}_{\{Y_1\leq B\}}\right] &\leq \mathbb{E}\left[\left(\int T(X_1,v)(Y_1-v)\mathbb{I}_{\{v\leq Y_1\}}dv\right)^2\mathbb{I}_{\{Y_1\leq B\}}\right] \\
&\leq \ell(A_2)B^2\mathbb{E}\left[\int_{A_2}T^2(X_1,v)\mathbb{I}_{\{v\leq Y_1\}}dv\right] \\
&= \ell(A_2)B^2\iint_A T^2(x,v)\bar{F}_1(x,v)dxdv \\
&= \ell(A_2)B^2\|T\|_\mu^2=O(B^2)=O((\max(D_m,D_{m'})^{2/5}).
\end{aligned}$$

On the other hand, for $T(x,y)=\sum_{j,k}a_{j,k}\varphi_j(x)\psi_k(y)$, and $\|T\|_\mu^2\leq 1$ we have,

$$\begin{aligned}
\mathbb{E}\left[\Psi_T^2(X_1,Y_1)\mathbb{1}_{\{Y_1>B\}}\right] &\leq \sum_{j,k}a_{j,k}^2\sum_{j,k}\mathbb{E}\left[\varphi_j^2(X_1)\left((Y_1-v)\psi_k(v)\mathbb{1}_{\{v\leq Y_1\}}\mathbb{1}_{\{Y_1>B\}}dv\right)^2\right]\\
&\leq \frac{1}{\bar F_0}\mathbb{E}\left[\sum_j\varphi_j^2(X_1)\int_{A_2}(Y_1-v)^2\mathbb{1}_{\{v\leq Y_1\}}\mathbb{1}_{\{Y_1>B\}}dv\right]\\
&\leq \frac{\phi_1\max(D_m,D_{m'})}{\bar F_0}\mathbb{E}\left[\int_{A_2}(Y_1-v)^2\mathbb{1}_{\{v\leq Y_1\}}\mathbb{1}_{\{Y_1>B\}}dv\right]\\
&= \frac{\phi_1\max(D_m,D_{m'})}{\bar F_0}\frac{\mathbb{E}(Y_1^3\mathbb{1}_{\{Y_1>B\}})}{3}\\
&\leq \frac{\phi_1\mathbb{E}(Y_1^6)\max(D_m,D_{m'})}{3\bar F_0B^3}=O((\max(D_m,D_{m'})^{2/5}).
\end{aligned}$$

Therefore,

$$\sup_{T\in B^\mu_{m,m'}(0,1)}\mathrm{Var}[\xi_{T,1}(W_1)]\leq C(D_m+D_{m'})^{2/5}:=v,$$

where C is a constant depending on $\ell(A_2)$, ϕ_1, $\bar F_0$ and $\mathbb{E}(Y_1^6)$.

iii) Search for bound H^2 : Let us write here $T=\sum_{j,k}a_{j,k}\varphi_j\psi_k$ where $(\varphi_j\otimes\psi_k)_{(j,k)}$ is an orthonormal basis of $(F_m+F_{m'})\otimes\mathcal{H}_n$ w.r.t. the norm $\|.\|_A$.

$$\begin{aligned}
\mathbb{E}\left(\sup_{T\in B^\mu_{m,m'}(0,1)}|\nu_{n,1}^2(T)|\right) &\leq \frac{1}{\bar F_0}\mathbb{E}\left(\sup_{T\in B_{m,m'}(0,1)}|\nu_{n,1}^2(T)|\right)\leq\frac{1}{\bar F_0}\sum_{j,k}\mathbb{E}\left(\nu_{n,1}^2(\varphi_j\otimes\psi_k)\right)\\
&\leq \frac{1}{n\bar F_0}\sum_{j,k}\mathbb{E}\left(\int(Y_1-v)\mathbb{1}_{\{v\leq Y_1\}}\varphi_j(X_1)\psi_k(v)dv\right)^2\\
&\leq \frac{1}{n\bar F_0}\sum_j\mathbb{E}\left[\varphi_j^2(X_1)\int(Y_1-v)^2\mathbb{1}_{\{v\leq Y_1\}}dv\right]\ \text{with (6.9)}\\
&\leq \frac{1}{n\bar F_0}\sum_j\iint\varphi_j^2(x)\int(y-v)^2\mathbb{1}_{\{v\leq y\}}dvf_{(X,Y)}(x,y)dxdy\\
&\leq \frac{1}{n\bar F_0}\iint(\sum_j\varphi_j^2(x))\frac{y^3}{3}f_{(X,Y)}(x,y)dxdy\\
&\leq \frac{\phi_1\max(D_m,D_{m'})\mathbb{E}(Y_1^3)}{3n\bar F_0}.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left(\sup_{T\in B^\mu_{m,m'}(0,1)}|\nu_{n,1}^2(T)|\right)\leq\frac{\phi_1(D_m+D_{m'})\mathbb{E}(Y_1^3)}{3n\bar F_0}:=H^2.$$

Applying Lemma 6.2 yields that

$$\begin{aligned}\mathbb{E}\left(\sup_{T\in B^{\mu}_{m,m'}(0,1)}\nu^2_{n,1}(T)-2H^2\right) &\leq K'\left(\frac{(D_m+D_{m'})^{2/5}}{n}e^{-K'_1(D_m+D_{m'})^{3/5}}+\frac{D_m+D_{m'}}{n\log^4(n)}e^{-K'_2\log^2(n)}\right)\\ &\leq \frac{K"}{n}\left(D_m^{2/5}e^{-K'_3D_m^{3/5}}e^{-K'_3D_{m'}^{3/5}}+D_{m'}^{2/5}e^{-K'_3D_{m'}^{3/5}}\right)\\ &\quad+\frac{2K"}{n^{1/2}\log^4(n)}e^{-K'_2\log^2(n)}\end{aligned}$$

using that $(x+y)^a\geq(x^a+y^a)/2$ for $a=2/5$ or $a=3/5$. As $k^{2/5}\exp(-Ck^{3/5})$ is bounded and summable for $k\in\mathbb{N}$ and $K"|\mathcal{M}_n|n^{-1/2}\log^{-4}(n)e^{-K'_2\log^2(n)}$ is $O(1/n)$, it follows that

$$\mathbb{E}\left(\sup_{T\in B^{\mu}_{m,\hat{m}}(0,1)}\nu^2_{n,1}(T)-2H^2\right)\leq\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\sup_{T\in B^{\mu}_{m,m'}(0,1)}\nu^2_{n,1}(T)-2H^2\right)\leq\frac{C}{n}.$$

• Study of $\nu_{n,3}$.

(i) Search for M_1. First, let $T(x,y)=\sum_{j,k}a_{j,k}\varphi_j(x)\psi_k(y)\in B^{\mu}_{m,m'}(0,1)$, we note that

$$\int_{A_2}T^2(x,v)dv=\sum_{j,j'}(\sum_k a_{j,k}a_{j',k})\varphi_j(x)\varphi_{j'}(x)$$

For $b_j=(\sum_k a^2_{j,k})^{1/2}$, we have $\sum_j b_j^2\leq\sum_{j,k}a^2_{j,k}\leq 1/\bar{F}_0$, and

$$\begin{aligned}\int_{A_2}T^2(x,v)dv &\leq \sum_{j,j'}b_jb_{j'}|\varphi_j(x)\varphi_{j'}(x)|=(\sum_j b_j|\varphi_j(x)|)^2\\ &\leq \sum_j b_j^2\sum_j\varphi_j^2(x)\leq(1/\bar{F}_0)\|\sum_j\varphi_j^2\|_\infty\\ &\leq (1/\bar{F}_0)\phi_1\max(D_m,D_{m'}).\end{aligned}$$

This yields

$$\begin{aligned}|\int T(x,v)\mathbb{I}_{\{y\geq v\}}e(v|x)dv| &\leq \left(\int_{A_2}T^2(x,v)dv\int_{A_2}e^2(v|x)\mathbb{I}_{A_1}(x)dv\right)^{1/2}\\ &\leq [(1/\bar{F}_0)\phi_1\max(D_m,D_{m'})]^{1/2}\sup_{(x,v)\in A}|e(v|x)|\sqrt{\ell(A_2)}\\ &= \left(\frac{\ell(A_2)\|e\|_{\infty,A}\phi_1}{\bar{F}_0}\right)^{1/2}\sqrt{D_m+D_{m'}}:=M_1\end{aligned}$$

(ii) Search for v.

$$\begin{aligned}\mathbb{E}\left(\int T(X_1,v)\mathbb{I}_{\{Y_1\geq v\}}e(v|X_1)dv\right)^2 &\leq \mathbb{E}\left[\int_{A_2}T^2(X_1,v)\mathbb{I}_{\{v\leq Y_1\}}e^2(v|X_1)dv\right]\\ &\leq \iint_A T^2(x,v)e^2(v|x)\int\mathbb{I}_{\{v\leq y\}}f_{(X,Y)}(x,y)dydvdx\\ &\leq \iint_A T^2(x,v)e^2(v|x)\bar{F}_1(x,v)dvdx\\ &\leq \|e\|^2_{\infty,A}\|T\|^2_\mu=\|e\|^2_{\infty,A}:=v.\end{aligned}$$

(*iii*) Search for H^2. We also have with the same argument,

$$\sum\nolimits_{j,k}\mathbb{E}\left(\int\varphi_j(X_1)\psi_k(v)\mathbb{1}_{\{Y_1\geq v\}}e(v|X_1)dv\right)^2=\sum_j\mathbb{E}\left[\varphi_j^2(X_1)\int_{A_2}\mathbb{1}_{\{Y_1\geq v\}}e^2(v|X_1)dv\right]$$

$$\begin{aligned}&\leq \phi_1\max(D_m,D_{m'})\int_{A_2}\mathbb{E}(e^2(v|X_1))dv\leq\phi_1(D_m+D_{m'})\int_{A_2}\mathbb{E}[\mathbb{E}((Y_1-v)^2|Y_1>v,X_1)]dv\\&\leq \phi_1(D_m+D_{m'})\ell(A_2)\mathbb{E}(Y_1^2).\end{aligned}$$

Therefore,

$$\mathbb{E}\left(\sup_{T\in B^{\mu}_{m,m'}(0,1)}|\nu^2_{n,3}(T)|\right)\leq\frac{\phi_1(D_m+D_{m'})\mathbb{E}(Y_1^2)\ell(A_2)}{n\bar{F}_0}:=H^2.$$

Applying Lemma 6.2 yields that

$$\mathbb{E}\left(\sup_{T\in B^{\mu}_{m,m'}(0,1)}\nu^2_{n,3}(T)-2H^2\right)\leq K'\left(\frac{1}{n}e^{-K_1'(D_m+D_{m'})}+\frac{1}{n}e^{-K_2'\sqrt{n}}\right),$$

and this gives the result in the same manner as in the previous cases. □

References

Abdous, B. and Berred, A. (2005). Mean residual life estimation. *J. of Statist. Planning and Inference* **132**, 3-19.

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

Brunel, E., Comte, F. and Lacour, C. (2008). Minimax estimation of the conditional cumulative distribution function under random censorship. Preprint MAP5 2008-6.

Castellan, G. and Letué, F. (2000). Estimation of the Cox regression function via model selection. In F. Letué's PhD Thesis. http://ljk.imag.fr/membres/Frederique.Letue/

Chaubey, Y.P. and Sen, P.K. (1999). On smooth estimation of mean residual life. *J. Statist. Plann. Inference* **75**, 223-236.

Chen, Y. Q. and Cheng (2005), S. Semiparametric regression analysis of mean residual life with censored survival data. *Biometrika* **92**, 19-29.

Chen, Y.Q., Jewell, N. P., Lei, X., Cheng, S.C. (2005). Semiparametric estimation of proportional mean residual life model in presence of censoring. *Biometrics* **61**, 170-178.

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. (2008). Adaptive estimation of the conditional intensity of marker-dependent counting processes. Preprint MAP5 2008-18.

Csörgö, M. and Zitikis, R. (1996). Mean Residual Life Processes. *Ann. Statist.* **24**, 1717-1739.

Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modelling Extremal Events.* Karatzas, I, Yor, M. (Eds.), Springer, Berlin.

Hall, W. J. and Wellner, J. (1981) Mean residual life. Statistics and related topics (Ottawa, Ont., 1980), 169–184, North-Holland, Amsterdam-New York.

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.

Maguluri, G. and Zhang, C.-H. (1994). Estimation in the mean residual life regression model. *J. Roy. Statist. Soc. Ser. B* **56**, 477-489.

Na, M. H. and Kim, J. J. (1999) On inference for mean residual life. *Comm. Statist. Theory Methods* **28**, 2917-2933.

Oakes, D., Dasu, T. (1996). A note on residual life. *Biometrika* **77**, 409-410.

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