Bernoulli **23**(4B), 2017, 3437–3468
DOI: 10.3150/16-BEJ852

Efficiency and bootstrap in the promotion time cure model

FRANÇOIS PORTIER[1], ANOUAR EL GHOUCH[2,*] and INGRID VAN KEILEGOM[2,**]

[1]*Télécom ParisTech, Université Paris-Saclay, 46 Rue Barrault, 75013 Paris, France.*
E-mail: francois.portier@gmail.com
[2]*Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Voie du Roman Pays 20, B1348 Louvain-la-Neuve, Belgium.*
*E-mail: *anouar.elghouch@uclouvain.be; **ingrid.vankeilegom@kuleuven.be*

In this paper, we consider a semiparametric promotion time cure model and study the asymptotic properties of its nonparametric maximum likelihood estimator (NPMLE). First, by relying on a profile likelihood approach, we show that the NPMLE may be computed by a single maximization over a set whose dimension equals the dimension of the covariates plus one. Next, using Z-estimation theory for semiparametric models, we derive the asymptotics of both the parametric and nonparametric components of the model and show their efficiency. We also express the asymptotic variance of the estimator of the parametric component. Since the variance is difficult to estimate, we develop a weighted bootstrap procedure that allows for a consistent approximation of the asymptotic law of the estimators. As in the Cox model, it turns out that suitable tools are the martingale theory for counting processes and the infinite dimensional Z-estimation theory. Finally, by means of simulations, we show the accuracy of the bootstrap approximation.

Keywords: asymptotic inference; bootstrap; Cox model; promotion time cure model; semiparametric efficiency

1. Introduction

In traditional survival analysis, it is assumed that all subjects will eventually experience the event of interest. Hence, the survival function of the population will reach zero at infinity. However, there are various situations in practice where this assumption is not met. Consider for example, the situation where one is interested in the time until someone dies or experiences a relapse from a certain disease. Some people will get cured from the disease and so they will never die due to that disease and they will never experience a relapse. The survival function will in that case tend to the proportion of cured individuals. When covariates are present, a number of models have been proposed in the literature that take this special feature into account. They can be broadly classified into two groups: mixture cure models and promotion time cure models. In this paper, we focus on a semiparametric promotion time cure model, which is an extension of the famous Cox model (see [7]) to the presence of cured subjects. The model has been proposed by [34], and has been further studied by [5,13,26–29,35], among many others. We give a formal definition of this model in the next section.

The goal of this paper is to study various aspects of this model in a thorough and mathematically rigorous way. Although there is an ever expanding literature on this type of models, we

respectfully believe that a rigorous theoretical study of this model is still missing in the literature. The paper by [35] is a serious attempt to fill this gap. However, the way they calculate the nonparametric maximum likelihood estimator (NPMLE) of the vector of regression coefficients is unnecessarily complicated. This has an important impact on the computation time of the NPMLE as well as on the development of asymptotic properties of this estimator. Moreover, the estimation of the variance of the estimator lacks clarity, and alternative methods to do inference (based on for example, bootstrap, empirical likelihood or other inferential procedures) are not studied. Finally, efficiency is only shown to hold true for the estimator of the parametric component of the model.

In this paper, we first show that the NPMLE of the vector of regression coefficients can be computed by maximizing a certain criterion function over a set of dimension $d + 1$, where d is the dimension of the covariates. This is an important improvement with respect to the paper by [35], who obtain the same NPMLE after maximizing an objective function over a $(m + d + 1)$-dimensional space, where m is the number of uncensored observations, which increases as the sample size increases, and can be very large in practice.

Next, this new way of expressing the NPMLE of the vector or regression coefficients offers other important advantages. We show the weak consistency and weak convergence of the NPMLE and calculate its asymptotic variance, which has a much simpler expression than in [35], and even more importantly, the formula of the asymptotic variance of the parametric part is completely explicit. Finally, another important contribution of the paper is that we show the efficiency of the NPMLE of both the parametric and the nonparametric components of the model (not just the Euclidean parameter as in [35]), by making use of the results in the books by [3,31] and [14] on semiparametric efficiency theory.

Although the asymptotic variance of the estimator of regression coefficients has an explicit formula, it is difficult to estimate it in practice. We therefore propose a bootstrap approach to estimate the variance or even the whole distribution of the estimator. We propose a general weighted bootstrap procedure, as developed in [23] and [33]. The weighted bootstrap offers important advantages over Efron's classical bootstrap for censored data, since the former leads to less ties than the latter (see, e.g., [6]). A slightly more restrictive bootstrap procedure has been studied in [15] in the context of proportional hazards frailty models.

For showing the weak consistency and weak convergence of the NPMLE and the consistency of the proposed bootstrap procedure, we make use of the theory of empirical processes (see [32]), which is an excellent tool for dealing with the asymptotics of Z-estimators like the NPMLE, and of the theory of martingales and related concepts (see [12]). For similar results in related models, see, for example, [20] and [15] for frailty models, [21] for proportional odds models, [10] for the Cox model with missing covariates, and [17] for mixture cure models.

The paper is organized as follows. In the next section, we introduce the model and explain what type of data we have at hand. We also discuss several aspects related to the model, like the definition of a cure threshold, the likelihood under the model and the identifiability of the model. Section 3 deals with the estimation of the parametric and nonparametric component of the model. In Section 4, the efficient score function and the associated information bound for the vector of regression coefficients are obtained. The asymptotic properties of the proposed estimators are studied in Section 5. In particular, the consistency, weak convergence and efficiency of the estimators of the parametric and nonparametric components of the model are proved. In

Section 6, it is explained how to do inference using a bootstrap approach, and it is shown that the proposed bootstrap procedure is consistent. The finite sample behavior of the proposed estimators is studied in Section 7 through a simulation study. In particular, we study the accuracy of the bootstrap approximation. Section 8 contains the proofs of the main results, whereas some auxiliary results needed for the main asymptotic results are collected in Appendix.

2. The model

Let (T, X) denote a random vector where $T \in \mathbb{R}^+$ is a survival time and $X = (1, Z^T)^T \in \mathbb{R}^d$ contains covariates Z distributed according to some probability density function. The promotion time cure model assumes that the conditional survival function of T given $X = x$ has the form

$$S_0(t|x) = \exp\big(-\eta\big(\beta_0^T x\big)\Lambda_0(t)\big), \tag{1}$$

where $\eta : \mathbb{R} \to \mathbb{R}^+$, $\beta_0 \in \mathbb{R}^d$ is the vector of regression coefficients and Λ_0 is an improper (i.e., bounded) continuous cumulative hazard function with $\Lambda_0(0) = 0$. The parameter β_0 synthesises the effect of the covariates on the response and the non-parametric part Λ_0 models the influence of the time. Clearly the effect of the intercept in β_0 overlaps with the effect of the limiting value of Λ_0. We avoid identifiability issues by fixing arbitrarily that $\Lambda_0(+\infty) := \lim_{t\to+\infty} \Lambda_0(t) = 1$. Moreover, we assume that β_0 lies in the compact set $B \subset \mathbb{R}^d$ and that Λ_0 is absolutely continuous. Therefore, we define the model

$$\mathcal{P} = \big\{(x,t) \mapsto S_{\beta,\Lambda}(t|x) = \exp\big(-\eta\big(\beta^T x\big)\Lambda(t)\big) : (\beta, \Lambda) \in \widetilde{\Theta}\big\},$$

where $\widetilde{\Theta} = B \times (\mathcal{F} \cap \mathcal{C})$, $\mathcal{F}$ is the space of cumulative distribution functions and $\mathcal{C}$ is the space of absolutely continuous functions. Since $S_{\beta,\Lambda}(+\infty|x) > 0$, this model naturally allows to handle situations where the event $T = +\infty$ occurs with positive probability. This arises in survival analysis when some individuals, called cured, never experience the event of interest. See, for example, [5] for a biological interpretation of the promotion time cure model.

2.1. Censoring and cure thresholding

Equation (1) is very similar to the equation of the well-known Cox model. In fact, when $\eta = \exp$, (1) becomes the Cox model except that, for the latter, Λ_0 is a proper cumulative hazard function, that is, $\Lambda_0(+\infty) = +\infty$. Because of identifiability issues, β_0 does not have an intercept in the Cox model whereas in the promotion time cure model the intercept is required for more flexibility. As a consequence, the partial likelihood of model $\mathcal{P}$ is no longer equal to the marginal likelihood (see [27], Section 4). For this reason, our estimation method will not rely on the marginal likelihood.

In this paper, we consider model $\mathcal{P}$ with censored data and more specifically we focus on right censoring. That is, there exists a random variable $C \in \mathbb{R}^+$, called the censoring time, such that we only observe $Y = \min(T, C)$ and $\delta = 1_{\{T \le C\}}$, where the random variable δ informs us whether the failure time is observed or not. From now on, we require that T and C are conditionally

independent given X and that the censoring mechanism is non-informative (see, for example, [25] for details). These are usual assumptions in survival analysis and they are necessary for the identifiability of the model. Unlike [35], we allow the censoring time C to be finite. This is a natural setting since in practice censoring typically results from the non-occurrence of the event of interest before the end of the trial or from loss to follow-up. Unfortunately, without any additional information, this implies that none of the cured subjects is observable. As a consequence, Λ in model $\mathcal{P}$ cannot be estimated nonparametrically. In order for model $\mathcal{P}$ to be identifiable in (β, Λ), following [35], we need to introduce a threshold value τ such that any censored observation beyond τ is treated as cured ($T = +\infty$). This threshold τ should be larger than the largest uncensored observation and is called the cure threshold. Concerning the parameters of the model, it implies that Λ_0 is flat after τ. As it is common practice in the field, the value of the threshold is assumed to be known. In clinical trials for instance, it is often provided by the physician.

2.2. Identifiability

Let P be the probability measure associated with (Y, δ, X) and denote by $\partial_y f$ the partial derivative of a function f with respect to the argument y. According to the values of (Y, δ), there are three types of individuals and each of them has the following contribution into the likelihood: (1) uncensored uncured subject ($\delta = 1, Y = y$) with contribution $\partial_y P(T \leq y, T \leq C | X = x)$, (2) censored uncured subject ($\delta = 0, Y = y \leq \tau$) with contribution $\partial_y P(C \leq y, C < T | X = x)$, and (3) censored cured subject ($\delta = 0, Y = y > \tau$) with contribution $P(Y > \tau | X = x)$. Under the assumptions stated above, the likelihood function $\mathrm{Lik}_{\beta,\Lambda}$ of model $\mathcal{P}$ is given by

$$\mathrm{Lik}_{\beta,\Lambda}(Y, \delta, X) = \left(\eta\left(\beta^T X\right)\Lambda'(Y) S_{\beta,\Lambda}(Y|X)\right)^{\delta} \times S_{\beta,\Lambda}(Y|X)^{(1-\delta)\Delta} \times S_{\beta,\Lambda}(+\infty|X)^{(1-\delta)(1-\Delta)},$$

where $\Delta = 1_{\{Y \leq \tau\}}$. We say that model $\mathcal{P}$ is identifiable if any distribution in $\mathcal{P}$ is uniquely characterized by some (β, Λ). The identifiability is necessary to guarantee the consistency of the MLE. In particular, it implies that the true parameters are the unique maximizers of the expected likelihood (see Lemma 5.35 in [31]). We have the following proposition.

Proposition 1. *Under censoring and under assumptions* (A1)(i), (A2)(i) *and* (A3) *given in Section 5, model $\mathcal{P}$ is identifiable.*

The proof of this result and of all upcoming theorems are provided in Section 8.

3. Estimation

We observe n independent copies of $W = (Y, \delta, X)$ drawn from model $\mathcal{P}$ which is assumed to be identifiable. The elements of the sample are denoted by $W_i = (Y_i, \delta_i, X_i)$, for $i = 1, \ldots, n$, and the support of W is $\mathcal{W}$. For any measurable function $f : \mathcal{W} \to \mathbb{R}$, we denote $\int f(u)\, dP(u)$ by Pf and $\int f(u)\, d\mathbb{P}_n(u)$ by $\mathbb{P}_n f$, where $\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{W_i}$ is the empirical measure associated to the observations of the sample. We define the metric $\rho(f, g) = \sqrt{P(f - g)^2}$ and we denote

by $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ the so-called empirical process. We also introduce the notation $\ell^\infty(\mathcal{W})$ for the space of bounded functions f from $\mathcal{W}$ to $\mathbb{R}$, endowed with the supremum norm $\|f\|_{\mathcal{W}} = \sup_{x\in\mathcal{W}} |f(x)|$. The Euclidean norm is denoted by $|\cdot|_2$ and the total variation norm by $|\cdot|_{\mathrm{tv}}$. In what follows, all the convergences, in probability or in distribution (denoted by $\xrightarrow{P}$ and $\Rightarrow$ respectively), are stated with respect to the outer expectation (see the introduction of [32] for details).

3.1. Definition of the NPMLE

The maximum likelihood estimator (MLE) of model $\mathcal{P}$ is the maximizer of $\prod_{i=1}^n \mathrm{Lik}_{\beta,\Lambda}(W_i)$ over $(\beta, \Lambda) \in \widetilde{\Theta}$. For any (β, Λ) such that, for some $\delta_i = 1$, $\Lambda'(Y_i) = +\infty$, the maximand function equals $+\infty$. Therefore, the MLE does not exist. Following [22], we can circumvent this problem by, on the one hand, extending the parameter set to discrete cumulative distribution functions, and, on the other hand, modifying slightly the maximand to account for the discreteness of Λ. This leads to the NPMLE, formally defined by

$$(\hat{\beta}, \hat{\Lambda}) = \operatorname*{argmax}_{\beta\in B, \Lambda\in\mathcal{F}} \prod_{i=1}^n L_{\beta,\Lambda}(Y_i, \delta_i, X_i),$$

where

$$L_{\beta,\Lambda}(Y, \delta, X) = \big(\eta\big(\beta^T X\big)\Lambda\{Y\}\exp\big(-\eta\big(\beta^T X\big)\Lambda(Y)\big)\big)^{\delta} \exp\big(-\Delta(1-\delta)\eta\big(\beta^T X\big)\Lambda(Y)\big) \times \exp\big(-(1-\Delta)(1-\delta)\eta\big(\beta^T X\big)\big),$$

with $\Lambda\{y\} = \Lambda(y) - \lim_{t\to y^-}\Lambda(t)$ the size of the jump of Λ at y. Passing to the logarithmic scale, since an uncensored observed time necessarily lies below τ, that is, $\delta\Delta = \delta$, the above optimisation is equivalent to

$$(\hat{\beta}, \hat{\Lambda}) = \operatorname*{argmax}_{\beta\in B, \Lambda\in\mathcal{F}} \mathbb{P}_n l_{\beta,\Lambda}, \tag{2}$$

where $l_{\beta,\Lambda}(Y, \delta, X) = \delta\log(\Lambda\{Y\}) + \delta\log(\eta(\beta^T X)) - \Delta\eta(\beta^T X)\Lambda(Y) - (1-\Delta)\eta(\beta^T X) \times \Lambda(+\infty)$. Let

$$r_{u,\beta}(Y, X) = \eta\big(\beta^T X\big)\big(\Delta 1_{\{Y\geq u\}} + (1-\Delta)\big),$$
$$d_\beta(X) = X\eta'\big(\beta^T X\big)/\eta\big(\beta^T X\big).$$

By differentiating $l_{\beta,\Lambda}$ with respect to β, we define $B_1(\beta, \Lambda)$, the score operator for β, by

$$B_1(\beta, \Lambda)[a] = \big(d_\beta(X)^T a\big)\left(\delta - \int r_{u,\beta}(Y, X)\, d\Lambda(u)\right), \tag{3}$$

for every $a \in \mathbb{R}^d$, where we use $\int$ as a shortcut for $\int_0^{+\infty}$. Following [22], for every $s > 0$ and every bounded function h, we consider the one-dimensional submodel S_{β,Λ_s}, with Λ_s defined

by $d\Lambda_s = (1 + sh)\, d\Lambda$. For every $y \in \mathbb{R}^+$ we have $\partial_s \Lambda_s(y)|_{s=0} = \int_0^y h(u)\, d\Lambda(u)$, and if moreover $\Lambda\{y\} > 0$, then $\partial_s \log \Lambda_s\{y\}|_{s=0} = h(y)$. Hence, by differentiating l_{β,Λ_s} at $s = 0$, we define $B_2(\beta, \Lambda)$, the score operator for Λ, by

$$B_2(\beta, \Lambda)[h] = \delta h(Y) - \int r_{u,\beta}(Y, X) h(u)\, d\Lambda(u), \tag{4}$$

for every bounded function h. Note that our estimator is the same as the one introduced in [27] and studied in [18]. The existence of $(\hat{\beta}, \hat{\Lambda})$ can be shown by noticing the following two facts. First, $\mathbb{P}_n l_{\beta,\Lambda}$ equals minus infinity if Λ does not jump at one of the uncensored observations. Second, an increase of Λ outside the set of finite observed survival times always reduces the value of the preceding or forthcoming jump and so the value of the likelihood. As a consequence, $\hat{\Lambda}$ is a step-function with a finite number $m = \sum_{i=1}^n \delta_i$ of bounded jumps that sum up to one. Thus, $(\hat{\beta}, \hat{\Lambda})$ maximizes a continuous function over the compact set $B \times [0, 1]^{m-1}$.

3.2. Computation of the NPMLE

In this section, we describe a new way to compute $(\hat{\beta}, \hat{\Lambda})$ defined by (2). Unlike the procedures available in the literature, see, for example, [35] or [18], our approach follows from an optimization over the Euclidean set of dimension $d + 1$. It works as follows. We begin by profiling out the "nuisance parameter" Λ. More precisely, for every β, we obtain an explicit formula for $\hat{\Lambda}_\beta = \operatorname{argmax}_{\Lambda \in \mathcal{F}} \mathbb{P}_n l_{\beta,\Lambda}$. Next, by some easy calculations, we get a simple formula for $\hat{\beta} = \operatorname{argmax}_{\beta \in B} \mathbb{P}_n l_{\beta, \hat{\Lambda}_\beta}$. Finally, we plug-in the latter into the formula of Λ_β to get $\hat{\Lambda} = \hat{\Lambda}_{\hat{\beta}}$, and so $(\hat{\beta}, \hat{\Lambda})$ is the maximizer of (2). Such a procedure is not new in survival analysis. In fact, a similar idea was applied for the Breslow estimator in the Cox model, see [4] and [19], and for the proportional hazards frailty model, see [15]. More about profile likelihood can be found in [31].

Since the maximizer $\hat{\Lambda}$ belongs to the space $\mathcal{F}$, the NPMLE must be profiled paying a special attention to the constraint $g(\Lambda) := \lim_{y \to +\infty} \Lambda(y) = 1$. This is done by considering a Lagrange procedure, for which the score equation associated to Λ is given by

$$\mathbb{P}_n B_2(\beta, \Lambda)[h] - \lambda \int h(u)\, d\Lambda(u) = 0,$$

for every bounded function h. The second term above is precisely equal to the Lagrange multiplier $\lambda \in \mathbb{R}$ times the derivative of the constraint g along the submodel associated to $s \mapsto \Lambda_s$. Putting $N_i(y) = \delta_i 1_{\{Y_i \le y\}}$ and $\overline{N}(y) = n^{-1} \sum_{i=1}^n N_i(y)$, the latter equality can be written as

$$\int h(u)\, d\overline{N}(u) = \int (\mathbb{P}_n r_{u,\beta} - \lambda) h(u)\, d\Lambda(u).$$

Taking $h(u) = \frac{1_{\{u \le y\}}}{\mathbb{P}_n r_{u,\beta} - \lambda}$, we know that the solution of the above equation is given by

$$\hat{\Lambda}_\beta(y) = \int_0^y \frac{d\overline{N}(u)}{\hat{R}_\beta(u) - \lambda}, \qquad \text{with } \hat{R}_\beta(u) = \mathbb{P}_n r_{u,\beta} = n^{-1} \sum_{i=1}^n r_{u,\beta}(Y_i, X_i),$$

where λ satisfies $\hat{\Lambda}_\beta(+\infty) = \int \frac{d\overline{N}(u)}{\hat{R}_\beta(u)-\lambda} = 1$, or equivalently,

$$n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\hat{R}_\beta(Y_i)-\lambda} = 1. \tag{5}$$

Let $\hat{R}_\beta = \min_{i:\delta_i=1}\hat{R}_\beta(Y_i)$. Among the m solutions of this equation, only (the smallest) one, say $\hat{\lambda}_\beta$, leads to an (increasing) cumulative distribution. Consequently, we define $\hat{\lambda}_\beta$ as the smallest solution of the previous equation. Note in particular that $\hat{\lambda}_\beta$ belongs to $[\hat{R}_\beta - n^{-1}m, \hat{R}_\beta - n^{-1})$. Injecting $\hat{\Lambda}_\beta$ in the likelihood gives

$$\hat{\beta} = \operatorname*{argmax}_{\beta\in B}\left\{\prod_{i=1}^{n}\left(\frac{\eta(\beta^T X_i)}{\hat{R}_\beta(Y_i)-\hat{\lambda}_\beta}\right)^{\delta_i}\right\} \times \exp\left(-\sum_{i=1}^{n}\eta\big(\beta^T X_i\big)\big(\Delta_i\hat{\Lambda}_\beta(Y_i)+(1-\Delta_i)\big)\right).$$

Using (5), since $\hat{R}_\beta(u) = \mathbb{P}_n r_{u,\beta}$, the sum on the right-hand side of the latter equation equals

$$\begin{aligned}&\sum_{i=1}^{n}\eta\big(\beta^T X_i\big)\big(\Delta_i\hat{\Lambda}_\beta(Y_i)+(1-\Delta_i)\big)\\ &\quad= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_j}{\hat{R}_\beta(Y_j)-\hat{\lambda}_\beta}\big(\eta\big(\beta^T X_i\big)\big(\Delta_i 1_{\{Y_j\le Y_i\}}+(1-\Delta_i)\big)\big)\\ &\quad= \sum_{j=1}^{n}\frac{\delta_j\hat{R}_\beta(Y_j)}{\hat{R}_\beta(Y_j)-\hat{\lambda}_\beta} = n\hat{\lambda}_\beta + \sum_{j=1}^{n}\delta_j.\end{aligned}$$

As a consequence, the NPMLE is given by

$$\hat{\beta} = \operatorname*{argmax}_{\beta\in B}\prod_{i=1}^{n}\left\{\left(\frac{\eta(\beta^T X_i)}{\hat{R}_\beta(Y_i)-\hat{\lambda}_\beta}\right)^{\delta_i}\exp(-\hat{\lambda}_\beta)\right\}, \tag{6}$$

$$\hat{\Lambda}(y) = \int_0^y \frac{d\overline{N}(u)}{\hat{R}_{\hat{\beta}}(u)-\hat{\lambda}_{\hat{\beta}}} = n^{-1}\sum_{i=1}^{n}\frac{\delta_i 1_{\{Y_i\le y\}}}{\hat{R}_{\hat{\beta}}(Y_i)-\hat{\lambda}_{\hat{\beta}}}, \tag{7}$$

recalling that $\hat{\lambda}_\beta$ is the smallest solution of (5).

From these equations, it is interesting to note that moving the threshold beyond the largest uncensored observed time $\tau_n = \max_{i=1,\dots,n}(Y_i\delta_i)$ has no effect on $(\hat{\beta}, \hat{\Lambda})$. In fact, observe that $\delta_i\hat{R}_\beta(Y_i) = n^{-1}\sum_k \eta(\beta^T X_k)1_{\{Y_k\ge Y_i\}}\delta_i - n^{-1}\sum_k \eta(\beta^T X_k)1_{\{Y_i>Y_k>\tau\}}\delta_i$. The last term vanishes if $\tau \ge \tau_n$. As a consequence, for any $i = 1, \dots, n$, $\delta_i\hat{R}_\beta(Y_i)$ does not depend on the threshold and so does the NPMLE of (β, Λ).

4. Efficiency bounds

The score operator for β is given by (3), and the score operator for Λ is given by (4). The efficient score operator associated with the whole model is

$$\widetilde{B}(\beta, \Lambda)[a, h] = \int \big(d_\beta(X)^T a + h(u)\big)\, dM_{\beta,\Lambda}(u),$$

where $M_{\beta,\Lambda} = N - A_{\beta,\Lambda}$, with $N(y) = \delta 1_{\{Y \le y\}}$, and $A_{\beta,\Lambda}(y) = \int_0^y r_{u,\beta}(Y, X)\, d\Lambda(u)$. In the following, for the sake of simplicity, we put $M_0 = M_{\beta_0,\Lambda_0}$, $A_0 = A_{\beta_0,\Lambda_0}$, $r_{u,0} = r_{u,\beta_0}$, $d_0 = d_{\beta_0}$.

It is useful to note that the process M_0 is a martingale with respect to the σ-field $\mathcal{F}_y$ induced by the process $\{(Z, \delta 1_{\{Y \le u\}}, (1-\delta)1_{\{Y \le u\}}) : 0 \le u \le y\}$. Such a result is a generalization of Theorems 1.3.1 and 1.3.2 in [12] when T follows an improper distribution (see also [1]). The proof does not require any new ideas compared to the proof of the aforementioned theorems. Denoting by $\langle M_0 \rangle$ the predictable quadratic variation of M_0, one can use Theorem 2.6.1 in [12] to get that $\langle M_0 \rangle_y = A_0(y)$. Moreover,

$$E\left[\int_0^y h(u)\, dM_0(u) \int_0^y f(u)\, dM_0(u)\right] = E\left[\int_0^y h(u) f(u)\, dA_0(u)\right], \tag{8}$$

for any locally bounded predictable processes f and h.

The efficient score function for β is $\widetilde{B}_1 = B_1(\beta_0, \Lambda_0) - \Pi B_1(\beta_0, \Lambda_0)$, i.e. the residual of the score B_1 after projecting it onto the nuisance tangent space generated by $B_2(\beta_0, \Lambda_0)[h]$, when $h \in \mathcal{G} := \{f \in \ell^\infty(\mathbb{R}^+) : \int f(u)\, d\Lambda_0(u) = 0\}$. This space results from differentiating along submodels Λ_s that lie in $\mathcal{F}$. That is to say that, $P(\widetilde{B}_1 B_2(\beta_0, \Lambda_0)) = 0$. By uniqueness of orthogonal projections on closed subsets of Hilbert spaces, there exists a unique $h_0 \in \mathcal{G}$ such that

$$E\big(\big(B_1(\beta_0, \Lambda_0) - B_2(\beta_0, \Lambda_0)[h_0]\big) B_2(\beta_0, \Lambda_0)[h]\big) = 0, \tag{9}$$

for every $h \in \mathcal{G}$. In general the element h_0 defined above is difficult to compute but it is feasible in our case by following the martingale approach from [24] originally developed for the Cox model. By (8), the left-hand side of equation (9) can be written as

$$\begin{aligned} E\left(\int \big(d_0 - h_0(u)\big)\, dM_0(u) \int h(u)\, dM_0(u)\right) &= E\left(\int \big(d_0 - h_0(u)\big) h(u)\, dA_0(u)\right) \\ &= \int \big(D_0(u) - R_0(u) h_0(u)\big) h(u)\, d\Lambda_0(u), \end{aligned}$$

with $D_0(u) = P d_0 r_{u,0}$ and $R_0(u) = P r_{u,0}$. As a consequence, the efficient score function for β is expressed as $\widetilde{B}_1 = \int (d_0 - h_0(u))\, dM_0(u)$, where

$$h_0(u) = \frac{D_0(u) - c}{R_0(u)}, \qquad \text{with } c = \frac{\int D_0(u) R_0(u)^{-1}\, d\Lambda_0(u)}{\int R_0(u)^{-1}\, d\Lambda_0(u)}.$$

We conclude that the associated information bound is I_0 where, by (8),

$$I_0 = \operatorname{var}(\widetilde{B}_1) = \int P\big\{\big(d_0 - h_0(u)\big)\big(d_0 - h_0(u)\big)^T r_{u,0}\big\}\, d\Lambda_0(u). \tag{10}$$

5. Asymptotic properties

To establish the consistency and find the asymptotic distribution of our estimator $(\hat{\beta}, \hat{\Lambda})$, we will make use of the existing theory about Z-estimation developed in [20,31,33], among others. For that, we first need to introduce the following assumptions. Let $\lambda_\beta \in (-\infty, R_\beta(\tau)]$ be the solution of $E \int \frac{dN(u)}{R_\beta(u)-\lambda} = 1$, where $R_\beta(u) = Pr_{u,\beta}$ (note that $\inf_{u\in\mathbb{R}^+} R_\beta(u) = R_\beta(\tau)$).

(A1) (i) The matrix $\operatorname{var}(Z)$ has full rank.
(ii) Each component of the variable Z is almost surely bounded by M in absolute value.
(A2) (i) The function η is injective and $\eta(x) > 0$ for every $x \in \mathbb{R}$.
(ii) The function η is two times continuously differentiable.
(iii) The parameter β_0 belongs to the interior of a known compact set $B \subset \mathbb{R}^d$ and $\Lambda_0 \in \mathcal{F} \cap \mathcal{C}$.
(iv) The function λ_β is such that $\inf_{\beta\in B}\{R_\beta(\tau) - \lambda_\beta\} > 0$.
(v) The matrix I_0 has full rank.
(A3) (i) The variables T and C are independent given Z.
(ii) $T > \tau$ implies that $T = +\infty$ and $P(C > \tau|Z) > 0$ a.s.

Assumptions (A1)(i), (A2)(i) and (A3) are required for the identifiability of the model; see Proposition 1. These assumptions are also needed for consistency and weak convergence. In fact, they guarantee that the Kullback–Leibler distance between elements of model $\mathcal{P}$ and the true parameters is uniquely minimized. Assumption (A1)(ii) is not strictly necessary and could be relaxed to a finite moment condition, but it simplifies the proofs. Assumption (A2)(ii), (iii) permit to consider a model that is not "too large." In particular, they are needed to control the metric entropy of the class of functions $\{(y, x) \mapsto r_{u,\beta}(y, x) : \beta \in B, u \in \mathbb{R}^+\}$. Especially, it will imply the weak convergence of the empirical scores. Assumption (A2)(iv) is a restriction on the set B that has to be small enough, and it guarantees that the quantities $\hat{R}_\beta(Y_i) - \hat{\lambda}_\beta$, for every $i = 1, \ldots, n$, remain bounded away from 0 with high probability.

The consistency of $(\hat{\beta}, \hat{\Lambda})$ has already been obtained in [35]. In the following, we present the same result, but with an alternative proof that relies on representations (6) and (7). Our approach to prove this result will be useful to show the consistency of the bootstrap version of the estimator.

Theorem 2. *Under assumptions* (A1)–(A3), *we have*

$$\big|(\hat{\beta}, \hat{\lambda}_{\hat{\beta}}) - (\beta_0, 0)\big|_2 \xrightarrow{P} 0 \quad \textit{and} \quad \|\hat{\Lambda} - \Lambda_0\|_\infty \xrightarrow{P} 0.$$

To study the weak convergence, we start by defining the following score operator:

$$B(\beta, \lambda, \Lambda)[a, b, h] = B_1(\beta, \Lambda)[a] + B_2(\beta, \Lambda)[h] - \lambda \int h(u)\, d\Lambda(u) + b\big(g(\Lambda) - 1\big),$$

for $(\beta, \lambda, \Lambda) \in \Theta = B \times \mathbb{R} \times \mathcal{F}$ and $(a, b, h) \in \mathbb{R}^{d+1} \times \ell^\infty(\mathbb{R}^+)$. This operator plays a crucial role in the derivation of the asymptotic behavior of our estimator. In particular, the choice of its domain is important: it needs to be "small enough" in order to control the metric entropy of the

underlying class but it also needs to be "sufficiently large" so that any zero of the empirical score is indeed the NPMLE.

We introduce the space

$$\mathcal{H} = \big\{(a, b, h) \in \mathbb{R}^{d+1} \times \ell^\infty\big(\mathbb{R}^+\big) : |a|_2 \le 1, |b| \le 1, \|h\|_{\mathrm{tv}} \le 1\big\}.$$

Let the maps $\Psi_n : \Theta \longrightarrow \ell^\infty(\mathcal{H})$ and $\Psi : \Theta \longrightarrow \ell^\infty(\mathcal{H})$ be defined by

$$\Psi_n(\beta, \lambda, \Lambda) = \mathbb{P}_n B(\beta, \lambda, \Lambda) \quad \text{and} \quad \Psi(\beta, \lambda, \Lambda) = P B(\beta, \lambda, \Lambda),$$

and let $\dot{\Psi}_0 : \operatorname{lin}\Theta \longrightarrow \ell^\infty(\mathcal{H})$ be the Fréchet derivative of Ψ at $(\beta_0, 0, \Lambda_0)$, where $\operatorname{lin}\Theta$ denotes the linear span of Θ. As shown in Lemma 9 (see the Appendix), we have (with $d_0 \equiv d_{\beta_0}$ and $r_{u,0} \equiv r_{u,\beta_0}$)

$$\begin{aligned}
&\dot{\Psi}_0[\beta - \beta_0, \lambda, \Lambda - \Lambda_0](a, b, h) \\
&\quad = -\int P\big\{\big(d_0^T a + h(u)\big) d_0^T r_{u,0}\big\}\, d\Lambda_0(u)(\beta - \beta_0) \\
&\qquad - \int P\big\{\big(d_0^T a + h(u)\big) r_{u,0} - b\big\}\, d(\Lambda - \Lambda_0)(u) - \lambda \int h(u)\, d\Lambda_0(u),
\end{aligned}$$

for any $(a, b, h) \in \mathcal{H}$ and $(\beta, \lambda, \Lambda) \in \Theta$. Since our estimator satisfies $\Psi_n(\hat{\beta}, \hat{\lambda}, \hat{\Lambda}) = 0$, we are now in position to apply the classical results from semiparametric Z-estimation theory. This leads to the following statement.

Theorem 3. *Under assumptions* (A1)–(A3), *we have*

$$n^{1/2}\big((\hat{\beta}, \hat{\Lambda}) - (\beta_0, \Lambda_0)\big) \Rightarrow G \qquad \text{on } \mathbb{R}^d \times \ell^\infty\big(\mathbb{R}^+\big),$$

where G is a tight Gaussian process on $\mathbb{R}^d \times \ell^\infty(\mathbb{R}^+)$, whose law is the same as the weak limit of $\dot{\Psi}_0^{-1}\mathbb{G}_n(B(\beta_0, 0, \Lambda_0))$. Moreover, $(\hat{\beta}, \hat{\Lambda})$ is efficient.

Note that the inverse $\dot{\Psi}_0^{-1}$ exists thanks to Lemma 10 in the Appendix. Since $\hat{\beta}$ is efficient we have the following corollary that results from the computation of the efficiency bound done in the previous section. Contrary to the statements in [35], we provide a closed formula for the variance.

Corollary 4. *Under assumptions* (A1)–(A3), *we have*

$$n^{1/2}(\hat{\beta} - \beta_0) \quad \Rightarrow \quad \mathcal{N}\big(0, I_0^{-1}\big),$$

where I_0 is given by (10).

The latter formula for the variance of the parametric component is appealing for its simplicity. It naturally results in new inference procedures concerning the regression coefficients of the promotion time cure model. In the next few lines, we describe a way to estimate I_0 given in (10).

By definition of h_0, $\int P\{(d_0 - h_0(u))h(u)r_{u,0}\}\,d\Lambda_0(u) = 0$ for every bounded real function h such that $\int h(u)\,d\Lambda_0(u) = 0$. Since $\int h_0(u)\,d\Lambda_0(u) = 0$, it follows that

$$\begin{aligned} I_0 &= \int P\big\{\big(d_0 - h_0(u)\big)d_0^T r_{u,0}\big\}\,d\Lambda_0(u) \\ &= \int C_0(u)\,d\Lambda_0(u) - \int h_0(u)D_0(u)^T\,d\Lambda_0(u), \end{aligned}$$

where $C_0(u) = P d_0 d_0^T r_{u,0}$. An estimator $\hat{I}$ is then obtained by replacing the theoretical expectations by empirical means and by replacing β_0 by its estimator $\hat{\beta}$, that is,

$$\hat{I} = \int \hat{C}(u)\,d\hat{\Lambda}(u) - \int \hat{h}(u)\hat{D}(u)^T\,d\hat{\Lambda}(u), \tag{11}$$

where

$$\hat{C}(u) = n^{-1}\sum_{i=1}^{n} d_{\hat{\beta}}(X_i)\,d_{\hat{\beta}}(X_i)^T r_{u,\hat{\beta}}(Y_i, X_i),$$

$$\hat{D}(u) = n^{-1}\sum_{i=1}^{n} d_{\hat{\beta}}(X_i) r_{u,\hat{\beta}}(Y_i, X_i),$$

$$\hat{h}(u) = \frac{\hat{D}(u) - \hat{c}}{\hat{R}_{\hat{\beta}}(u)}, \qquad \text{with } \hat{c} = \frac{\int \hat{D}(u)\hat{R}_{\hat{\beta}}(u)^{-1}\,d\hat{\Lambda}(u)}{\int \hat{R}_{\hat{\beta}}(u)^{-1}\,d\hat{\Lambda}(u)}.$$

The function $\hat{R}_{\hat{\beta}}$ and the quantities $d_{\hat{\beta}}(X_i)$, $r_{u,\hat{\beta}}(Y_i, X_i)$, $i = 1, \ldots, n$, have been introduced in Section 3.

6. Bootstrap inference

The results in the previous section reveal that the asymptotic distributions of the proposed estimators can't be directly used because they depend on many unknown quantities. One solution to this problem is to approximate this distribution by a bootstrap procedure. The classical bootstrap of Efron [11] works by resampling with replacement from the original sample, evaluating the statistic of interest on the bootstrap samples, and then use these statistics to make inference about the population parameter of interest.

In the following, we consider the weighted bootstrap, a more general resampling scheme than Efron's original bootstrap, as developed in [23] and [33]. As mentioned in the introduction, the weighted bootstrap offers important advantages over Efron's classical bootstrap for censored data, since the former leads to less ties than the latter (see, e.g., [6]). The randomness of the bootstrap is produced with the help of sequences of weights $(w_{1,n}, w_{2,n}, \ldots)$ for $n \in \mathbb{N}^*$. These weights are independent from the original sample $(W_1, W_2, \ldots)$ and the underlying probability measure associated to these sequences is denoted by P^*. Additionally we need the following assumptions.

(B1) The sequence $(w_{i,n})_{1\le i\le n}$ is exchangeable, that is, for every permutation $(\pi_1,\dots,\pi_n)$ of $(1,\dots,n)$, $(w_{i,n})$ has the same law as $(w_{\pi_i,n})$.

(B2) Let S_n be the survival function of $w_{1,n}$. Then,

$$\sup_{n\ge 1}\int S_n(u)^{1/2}\,du<+\infty \quad\text{and}\quad \lim_{A\to+\infty}\limsup_{n\to+\infty}\sup_{t\ge A} t^2 S_n(t)=0.$$

(B3) $w_{i,n}\ge 0$ for all i and n, $n^{-1}\sum_{i=1}^n w_{i,n}=1$ and $n^{-1}\sum_{i=1}^n (w_{i,n}-1)^2 \xrightarrow{P^*} 1$ for all n.

Standard examples of weights that verify the previous assumptions are the i.i.d. weighted bootstrap, the double bootstrap and Efron's original bootstrap (see [23] for more examples). Following [33], we define the bootstrap estimator as

$$\big(\hat\beta^*,\hat\Lambda^*\big)=\operatorname*{argmax}_{\beta\in B,\Lambda\in\mathcal{F}}\mathbb{P}_n^* l_{\beta,\Lambda}(W_i), \tag{12}$$

where $\mathbb{P}_n^*$ is the bootstrap empirical measure, i.e. $\mathbb{P}_n^*=n^{-1}\sum_{i=1}^n w_{i,n}\delta_{W_i}$. Using a similar Lagrange optimization procedure as for the original estimator (see Section 2), we maximize (12) for a fixed β with respect to Λ, to obtain

$$\hat\Lambda_\beta^*(y)=n^{-1}\sum_{i=1}^n \frac{w_{i,n}\delta_i 1_{\{Y_i\le y\}}}{\hat R_\beta^*(Y_i)-\hat\lambda_\beta^*},$$

where $\hat R_\beta^*(u)=\mathbb{P}_n^* r_{u,\beta}$ and $\hat\lambda_\beta^*$ is the smallest solution of the equation $n^{-1}\sum_{i=1}^n \frac{w_{i,n}\delta_i}{\hat R_\beta^*(Y_i)-\lambda}=1$. Injecting the previous solution in (12) we obtain

$$\hat\beta^*=\operatorname*{argmax}_{\beta\in B}\left\{\prod_{i=1}^n\left(\frac{\eta(\beta^T X_i)}{\hat R_\beta^*(Y_i)-\hat\lambda_\beta^*}\right)^{\delta_i w_{i,n}}\exp\big(-\hat\lambda_\beta^*\big)\right\},$$

$$\hat\Lambda^*(y)=\hat\Lambda_{\hat\beta^*}^*(y).$$

The following theorem guarantees that the bootstrap works, that is, that it reproduces the asymptotic law of the estimator $\hat\beta$. Formal details about its statement are available in the proof.

Theorem 5. *Under assumptions* (A1)–(A3) *and* (B1)–(B3), *the bootstrap estimator* $\sqrt{n}((\hat\beta^*,\hat\Lambda^*)-(\hat\beta,\hat\Lambda))$ *has the same asymptotic law, conditionally on* $W_1,W_2,\dots$, *as* $\sqrt{n}((\hat\beta,\hat\Lambda)-(\beta_0,\Lambda_0))$.

7. Simulations

In this section, we study the performance of the estimator given in (6) and the weighted bootstrap procedure described in Section 6. We consider the following model:

$$S(t|Z_1,Z_2)=\exp\big(-\exp(\beta_0+\beta_1 Z_1+\beta_2 Z_2)\Lambda(t)\big),$$

where Z_1 is a uniformly distributed random variable on $[\alpha, \alpha + 1]$ and Z_2 is a Bernoulli random variable that takes a value of 0 or 1 with equal probability. The true parameters are $\beta_0 = 3$, $\beta_1 = -2$ and $\beta_2 = 1$. We choose Λ to be the cumulative distribution function of either an exponential variable with mean 1 (*Case* 1) or a uniform variable on $[0, 1]$ (*Case* 2). The censoring variable is exponential with parameter λ_c. By varying the latter we mainly control the censoring rate, while by varying α we control the cure rate. We use the Newton–Raphson method to maximize the likelihood. In our algorithm, we use as starting values $\beta_0 = 1$ and (β_1, β_2) equal to the estimates obtained from the classical Cox model ignoring the cure proportion. The bootstrap weights are given by $w_{i,n} = e_i / \overline{e}_n$, where e_i, $i = 1, \ldots, n$, are i.i.d. exponential random variables with rate $\lambda = 1$. This is known in the literature as the Bayesian bootstrap, a smooth alternative to the well known nonparametric (multinomial) bootstrap of Efron; see [16]. The main reason for considering the bootstrap approach is the difficulty of using the asymptotic distribution to construct confidence intervals and the fact that, whenever the bootstrap is valid, it typically gives better results, especially for small sample sizes. Our main objective is to investigate the effect of the following factors: the sample size n, the distribution Λ, and the average amount of censoring and cure.

Many types of bootstrap confidence intervals are available in the literature. The most used ones are the basic method, the percentile method and the bias corrected method; see [8]. We studied these three methods and we obtained very similar results in terms of coverage probability and average length. For this reason and for the sake of brevity, in the following, we only report the results obtained with the basic method that was, globally, slightly better than the two others. Given an estimator $\hat{\beta}$ of β and a bootstrap estimate $\hat{\beta}^*$ of it, the basic bootstrap confidence interval of confidence level $1 - \alpha$ is given by $[2\hat{\beta} - q^*_{1-\alpha/2}, 2\hat{\beta} - q^*_{\alpha/2}]$, with q^*_α being the quantile of order α of the bootstrap distribution of $\hat{\beta}^*$. In our simulation, we approximate $q^*_{1-\alpha/2}$ empirically using $B = 1000$ bootstrap replications.

We perform $N = 1000$ repetitions for two sample sizes ($n = 100$ and $n = 200$), three levels of censoring (20%, 40% and 60%) and four levels of cure (10%, 20%, 40% and 60%). The results of our simulation study are partially summarized in Tables 1 and 2. We show the empirical bias (BIAS), the empirical variance (VAR) and the empirical mean squared error (MSE) of $\hat{\beta}_k$, $k = 0, 1, 2$. For the bootstrap estimates $\hat{\beta}^*_k$, we report the average bootstrap bias (BIAS*) and the average bootstrap variance (VAR*) given, respectively, by

$$\mathrm{BIAS}^*\big(\hat{\beta}^*_k\big) = \frac{1}{N}\sum_{i=1}^{N}\big(\overline{\hat{\beta}^{*(i)}_k} - \hat{\beta}^{(i)}_k\big)^2, \quad \text{and} \quad \mathrm{VAR}^*\big(\hat{\beta}^*_k\big) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B}\sum_{j=1}^{B}\big(\hat{\beta}^{*(i)}_{k,j} - \overline{\hat{\beta}^{*(i)}_k}\big)^2,$$

where $\hat{\beta}^{*(i)}_{k,j}$ is the bootstrap estimate of β_k obtained from the jth bootstrap sample of the ith simulated data set and $\overline{\hat{\beta}^{*(i)}_k} = \frac{1}{B}\sum_{j=1}^{B}\hat{\beta}^{*(i)}_{k,j}$. For the basic bootstrap confidence intervals, we report the coverage probability (COV*) and the average length (LEN*). The obtained results for Case 2 (Λ with bounded support) are slightly better than the corresponding results for Case 1 (Λ with unbounded support). For clarity, in the following we will focus on the latter case. With few exceptions, all our comments also apply to Case 2.

Table 1. The bias (BIAS), the variance (VAR) and the mean squared error (MSE) of $\hat{\beta}_k$, $k = 0, 1, 2$, together with the average bootstrap bias (BIAS*), the average bootstrap variance (VAR*), the coverage probability (COV*) and the average length (LEN*) using the Bayesian bootstrap. *Case* 1: $\Lambda(t) = (1 - \exp(-t))I(t \geq 0)$

Cure (%)	Cens. (%)		$n = 100$							$n = 200$						
			COV*	LEN*	BIAS*	VAR*	BIAS	VAR	MSE	COV*	LEN*	BIAS*	VAR*	BIAS	VAR	MSE
10	20	β_0	0.92	2.090	0.032	0.291	0.048	0.351	0.354	0.93	1.482	0.017	0.146	0.020	0.165	0.165
		β_1	0.94	1.650	−0.033	0.181	−0.052	0.199	0.201	0.94	1.157	−0.017	0.089	−0.020	0.097	0.097
		β_2	0.95	0.934	0.016	0.058	0.036	0.060	0.062	0.95	0.654	0.008	0.028	0.004	0.029	0.029
	40	β_0	0.90	2.411	0.029	0.388	−0.160	0.517	0.543	0.88	1.739	0.011	0.201	−0.131	0.280	0.297
		β_1	0.95	1.910	−0.039	0.244	−0.065	0.252	0.257	0.95	1.335	−0.020	0.118	−0.024	0.125	0.126
		β_2	0.97	1.086	0.019	0.078	0.039	0.077	0.078	0.95	0.757	0.009	0.038	0.005	0.039	0.039
20	20	β_0	0.95	2.345	0.041	0.365	0.058	0.391	0.395	0.94	1.656	0.023	0.181	0.043	0.210	0.212
		β_1	0.95	1.633	−0.033	0.177	−0.040	0.185	0.187	0.94	1.151	−0.018	0.088	−0.035	0.100	0.101
		β_2	0.96	0.921	0.015	0.056	0.022	0.056	0.056	0.95	0.649	0.008	0.028	0.008	0.030	0.030
	40	β_0	0.95	2.786	0.038	0.516	0.052	0.589	0.591	0.93	1.974	0.019	0.257	0.038	0.308	0.310
		β_1	0.96	1.897	−0.039	0.240	−0.051	0.246	0.249	0.94	1.333	−0.021	0.117	−0.050	0.132	0.135
		β_2	0.97	1.073	0.018	0.076	0.025	0.076	0.077	0.96	0.755	0.009	0.037	0.008	0.036	0.036
	60	β_0	0.92	3.463	0.036	0.802	−0.216	1.002	1.049	0.90	2.443	0.012	0.396	−0.196	0.537	0.576
		β_1	0.95	2.364	−0.052	0.374	−0.082	0.408	0.415	0.94	1.642	−0.026	0.179	−0.059	0.206	0.209
		β_2	0.97	1.356	0.027	0.122	0.045	0.120	0.122	0.95	0.941	0.013	0.059	0.015	0.062	0.062
40	40	β_0	0.94	3.297	0.049	0.719	0.066	0.740	0.745	0.96	2.318	0.026	0.354	0.059	0.338	0.342
		β_1	0.94	1.884	−0.038	0.236	−0.048	0.243	0.245	0.96	1.325	−0.020	0.116	−0.039	0.110	0.111
		β_2	0.96	1.066	0.017	0.075	0.020	0.074	0.075	0.95	0.754	0.009	0.037	0.006	0.038	0.038
	60	β_0	0.95	4.171	0.048	1.152	0.049	1.248	1.251	0.94	2.930	0.024	0.567	0.073	0.604	0.610
		β_1	0.96	2.343	−0.050	0.367	−0.061	0.384	0.388	0.94	1.641	−0.026	0.179	−0.060	0.187	0.190
		β_2	0.96	1.345	0.026	0.121	0.035	0.120	0.122	0.95	0.940	0.013	0.058	0.004	0.062	0.062
60	60	β_0	0.96	4.891	0.059	1.584	0.074	1.571	1.576	0.94	3.388	0.029	0.756	0.059	0.841	0.844
		β_1	0.97	2.355	−0.049	0.371	−0.063	0.364	0.368	0.94	1.636	−0.024	0.177	−0.036	0.193	0.194
		β_2	0.96	1.363	0.026	0.124	0.036	0.132	0.133	0.95	0.943	0.013	0.059	0.018	0.064	0.064

Table 2. The bias (BIAS), the variance (VAR) and the mean squared error (MSE) of $\hat{\beta}_k$, $k = 0, 1, 2$, together with the average bootstrap bias (BIAS*), the average bootstrap variance (VAR*), the coverage probability (COV*) and the average length (LEN*) using the Bayesian bootstrap. *Case* 2: $\Lambda(t) = tI(0 \le t \le 1) + I(t > 1)$

Cure	Cens.		$n = 100$							$n = 200$						
(%)	(%)		COV*	LEN*	BIAS*	VAR*	BIAS	VAR	MSE	COV*	LEN*	BIAS*	VAR*	BIAS	VAR	MSE
10	20	β_0	0.93	2.082	0.034	0.289	0.056	0.335	0.338	0.95	1.467	0.019	0.142	0.027	0.157	0.157
		β_1	0.94	1.651	−0.033	0.182	−0.052	0.199	0.202	0.94	1.158	−0.018	0.089	−0.021	0.098	0.099
		β_2	0.95	0.934	0.016	0.058	0.037	0.062	0.063	0.95	0.654	0.008	0.028	0.004	0.029	0.029
	40	β_0	0.94	2.453	0.031	0.403	0.024	0.487	0.488	0.92	1.761	0.012	0.206	0.003	0.267	0.267
		β_1	0.95	1.914	−0.039	0.245	−0.072	0.250	0.255	0.95	1.336	−0.020	0.118	−0.023	0.128	0.128
		β_2	0.96	1.091	0.019	0.079	0.038	0.078	0.080	0.94	0.758	0.009	0.038	0.006	0.041	0.041
20		β_0	0.95	2.346	0.041	0.365	0.058	0.391	0.395	0.94	1.655	0.023	0.180	0.043	0.210	0.212
	20	β_1	0.95	1.633	−0.033	0.177	−0.039	0.186	0.187	0.94	1.151	−0.018	0.088	−0.035	0.100	0.101
		β_2	0.96	0.920	0.015	0.056	0.022	0.056	0.056	0.95	0.649	0.008	0.028	0.008	0.029	0.030
	40	β_0	0.95	2.750	0.042	0.501	0.065	0.558	0.562	0.94	1.934	0.023	0.247	0.061	0.290	0.294
		β_1	0.95	1.885	−0.039	0.237	−0.045	0.251	0.253	0.94	1.326	−0.021	0.116	−0.051	0.133	0.136
		β_2	0.96	1.067	0.019	0.076	0.029	0.075	0.076	0.95	0.750	0.009	0.037	0.008	0.037	0.037
	60	β_0	0.94	3.506	0.040	0.822	0.003	0.987	0.988	0.93	2.471	0.014	0.405	0.021	0.508	0.508
		β_1	0.95	2.360	−0.052	0.374	−0.080	0.404	0.410	0.95	1.638	−0.026	0.178	−0.060	0.201	0.204
		β_2	0.97	1.352	0.027	0.122	0.043	0.122	0.123	0.94	0.936	0.013	0.058	0.013	0.063	0.063
40	40	β_0	0.94	3.297	0.049	0.719	0.065	0.741	0.745	0.95	2.317	0.026	0.354	0.059	0.338	0.342
		β_1	0.94	1.883	−0.038	0.236	−0.047	0.243	0.245	0.96	1.325	−0.020	0.116	−0.039	0.110	0.111
		β_2	0.96	1.066	0.017	0.075	0.020	0.074	0.075	0.95	0.754	0.009	0.037	0.006	0.038	0.038
		β_0	0.95	4.144	0.056	1.142	0.086	1.210	1.218	0.95	2.907	0.029	0.557	0.083	0.578	0.585
	60	β_1	0.94	2.341	−0.051	0.367	−0.067	0.385	0.389	0.95	1.641	−0.027	0.179	−0.059	0.185	0.188
		β_2	0.96	1.337	0.026	0.119	0.030	0.120	0.121	0.94	0.938	0.013	0.058	0.005	0.064	0.064
60	60	β_0	0.96	4.892	0.059	1.584	0.072	1.568	1.574	0.94	3.387	0.029	0.755	0.058	0.840	0.843
		β_1	0.97	2.356	−0.049	0.371	−0.063	0.363	0.367	0.94	1.635	−0.024	0.177	−0.036	0.192	0.194
		β_2	0.96	1.363	0.026	0.124	0.036	0.132	0.133	0.95	0.942	0.013	0.059	0.018	0.063	0.064

First, observe that the variance represents almost 100% of the MSE. Second, regarding the sample size, increasing the latter from 100 to 200, decreases the MSE by the (multiplicative) factor of 2. The same remark applies to the variance. The bias also decreases with the sample size except for one case (β_0 with 60% of censoring and 40% of cure). The decrease in the bias varies by a factor of 1 to 9 and is more important for $\hat{\beta}_3$. Except for β_0 with small cure proportion (10%) and large censoring (40% or more), the coverage probability remains stable around the nominal confidence level of 95%. In all studied cases, the average length decreases by a factor of about 1.5. Globally, the estimation of β_0 is more difficult, in the sense that $\hat{\beta}_0$ has more (finite sample) bias and more variance than $\hat{\beta}_1$ and $\hat{\beta}_2$. The cure proportion has no or only a very small effect on the MSE of $\hat{\beta}_1$ and $\hat{\beta}_2$, but it does affect the behavior of $\hat{\beta}_0$. In fact, when the percentage of cure increases, the bias of $\hat{\beta}_0$ decreases but its variance (and so its MSE) increases. For data with small percentage of cure (20% or less) and large percentage of censoring (40% or more), the bias of $\hat{\beta}_0$ can be quite large even for large sample size ($n = 200$). This is due to the fact that the observed percentage of cure, that is, the proportion of censored individuals with observed survival time larger than the last uncensored survival time, can be extremely small. In such case, the resulting point estimate and confidence interval for β_0 should be interpreted with care. Regarding the effect of censoring, one can see that the MSE of the $\hat{\beta}_k$'s increases as the percentage of censoring increases. For $\hat{\beta}_1$ and $\hat{\beta}_2$, this is mainly due to the increase of the variance component. But for $\hat{\beta}_0$, both the bias and the variance may increase significantly with censoring. On average, the bootstrap estimate of the bias and the variance are very accurate ($\max|\mathrm{BIAS}^* - \mathrm{BIAS}| < 5\%$ and $\max|\mathrm{VAR}^* - \mathrm{VAR}| < 10\%$) except for β_0 with high censoring rate and small cure rate.

To get an idea about the distribution of $\hat{\beta}_k$, we provide in Figure 1 the Q–Q plot of $\hat{\beta}_0 - \beta_0$ (somewhat similar plots were obtained for $\hat{\beta}_1$ and $\hat{\beta}_2$). This plot clearly shows that the normal approximation becomes better as the sample size becomes larger. Figure 2 shows the plot of the kernel density estimator of $\hat{\beta}_0 - \beta_0$ and its corresponding bootstrap estimate. From this plot we can see that (at least for the sample under study) the approximation obtained by the proposed bootstrap method becomes more precise as the sample size grows.

Finally, we ran the entire simulation study and recalculated the confidence intervals for the β's based on Efron's basic bootstrap, that is, we generated $\{w_{i,n}\}_{1\le i\le n}$ from a n-multinomial distribution with parameters $(n, (1/n, \ldots, 1/n))$, and based directly on asymptotic normality using the plug-in variance estimator given by (11). Table 3 gives a short overview of the results. Globally, compared to the Bayesian bootstrap, the multinomial weights lead systematically to slightly conservative and wider confidence intervals. Also, through the bootstrap iterations, we noticed that the Newton–Raphson algorithm sometimes has convergence difficulties. This leads, from time to time, to aberrant bootstrap estimates, especially in the case of bootstrap samples with high percentage of censoring. Globally, the asymptotic normal approximation gives satisfactory results, that is, the obtained confidence intervals are somewhat similar to those obtained via bootstrap. This demonstrates the validity of the proposed estimator of the asymptotic variance-covariance matrix I_0^{-1}. However, as can be seen by comparing Table 3 and Table 1, the weighted bootstrap method typically outperforms the empirical asymptotic variance method especially when the percentage of censoring is high and the sample size is "relatively small."

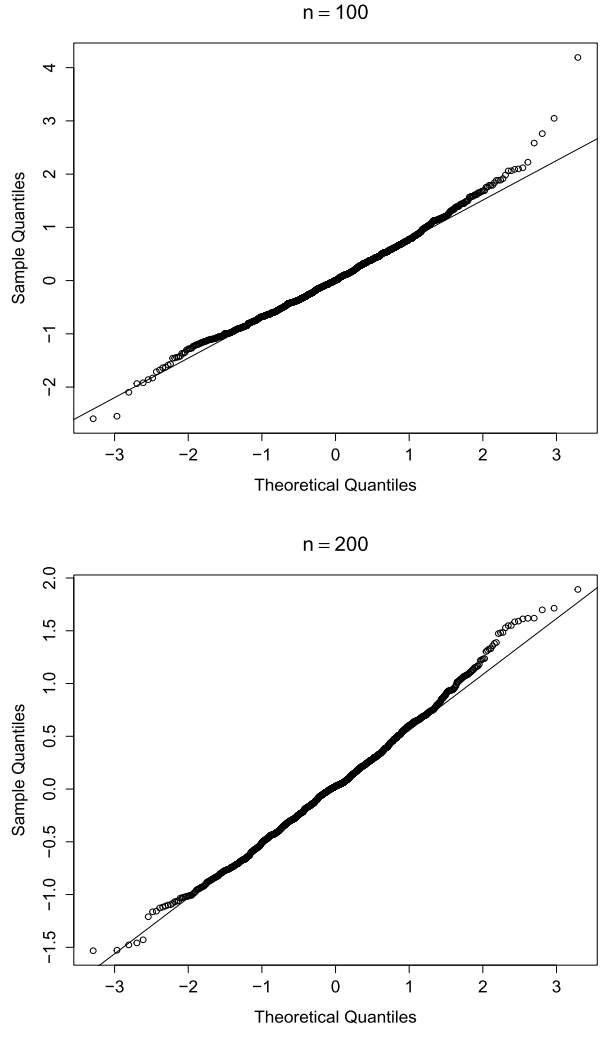

Figure 1. Q–Q plot of $\hat{\beta}_0 - \beta_0$ for Case 1 when the cure rate is 20% and the censoring rate is 40%.

8. Proofs

In the proofs, we use the norm

$$\|\Lambda\|_{\mathrm{BV}} = \sup_{h \in \mathcal{H}} \left| \int h(u)\, d\Lambda(u) \right|,$$

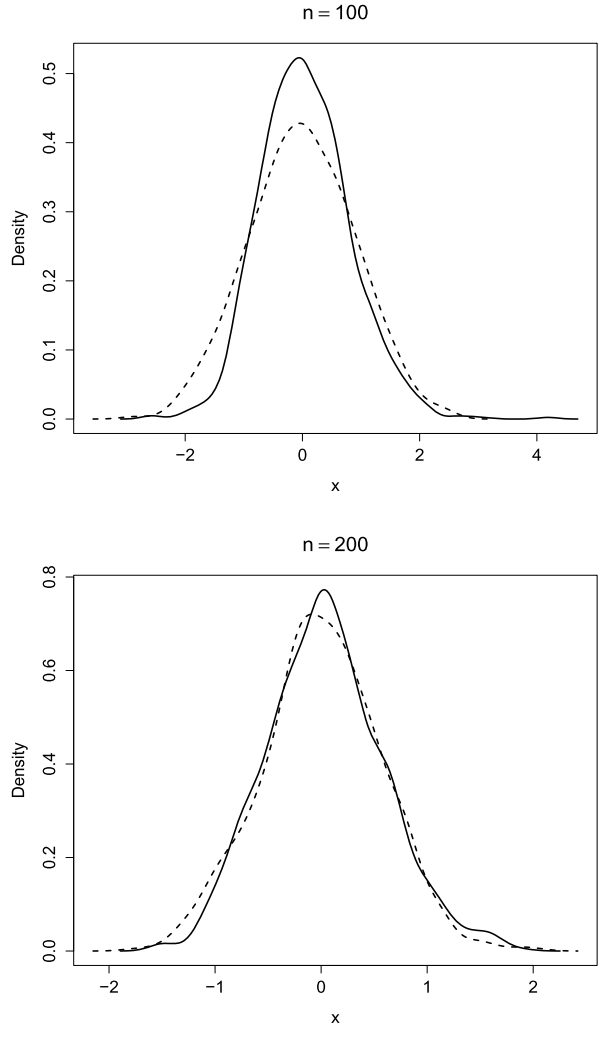

Figure 2. Plot of the density of $\hat{\beta}_0 - \beta_0$ (solid curve) and of $\hat{\beta}_0^* - \hat{\beta}_0$ (dashed curve) for Case 1 when the cure rate is 20% and the censoring rate is 40%. The density estimator is based on an Epanechnikov kernel and the bandwidth equals 0.16.

and we shall keep in mind that both norms $\|\cdot\|_{\mathrm{BV}}$ and $\|\cdot\|_\infty$ are equivalent on the space of real valued functions of bounded variation (see [9]).

8.1. Proof of Proposition 1

Suppose that (β, Λ) and $(\widetilde{\beta}, \widetilde{\Lambda})$ both lie in $\widetilde{\Theta}$ and result in the same distribution for (Y_1, δ_1) given X_1 (this distribution is expressed through the likelihood given in Section 3). Showing

Table 3. The coverage probability and average length based on Efron's bootstrap (COV*, LEN*) and based on the asymptotic normality using (11). *Case* 1: $\Lambda(t) = (1 - \exp(-t))I(t \geq 0)$

Cure (%)	Cens. (%)		$n = 100$				$n = 200$			
			COV*	LEN*	COV	LEN	COV*	LEN*	COV	LEN
20	20	β_0	0.97	2.521	0.98	2.629	0.95	1.718	0.95	1.826
		β_1	0.96	1.750	0.95	1.677	0.95	1.195	0.94	1.166
		β_2	0.97	0.972	0.96	0.950	0.96	0.667	0.94	0.663
	40	β_0	0.96	3.040	0.96	3.010	0.94	2.076	0.94	2.077
		β_1	0.97	2.053	0.94	1.921	0.95	1.389	0.95	1.325
		β_2	0.98	1.147	0.96	1.080	0.96	0.779	0.94	0.754
	60	β_0	0.94	3.847	0.90	3.559	0.91	2.609	0.90	2.469
		β_1	0.97	2.599	0.93	2.316	0.95	1.722	0.93	1.604
		β_2	0.98	1.476	0.94	1.314	0.96	0.980	0.93	0.914

identifiability is showing that $(\beta, \Lambda) = (\widetilde{\beta}, \widetilde{\Lambda})$. Taking $y > \tau$, we have that

$$\eta\big(\beta^T x\big) = \eta\big(\widetilde{\beta}^T x\big),$$

for almost every x that lies in the support of the conditional law of X. Hence, because η is injective by (A2)(i), we obtain that $\beta^T x = \widetilde{\beta}^T x$ for almost every x. Now, since β has an intercept, we write $\beta = (\alpha, \gamma)$ with $\alpha \in \mathbb{R}$, $\gamma \in \mathbb{R}^{d-1}$ and $\widetilde{\beta} = (\widetilde{\alpha}, \widetilde{\gamma})$ with $\widetilde{\alpha} \in \mathbb{R}$, $\widetilde{\gamma} \in \mathbb{R}^{d-1}$. We know that

$$(\gamma - \widetilde{\gamma})^T z = \widetilde{\alpha} - \alpha,$$

where the last equality happens for almost every z in the support of the law of Z. As a consequence, $\operatorname{var}((\gamma - \widetilde{\gamma})^T Z) = 0$, and we obtain by assumption (A1)(i) that $\gamma = \widetilde{\gamma}$ and $\alpha = \widetilde{\alpha}$. Finally, taking $\delta = 1$, we have that

$$\Lambda'(y)\eta\big(\beta^T x\big)\exp\big(-\eta\big(\beta^T x\big)\Lambda(y)\big) = \widetilde{\Lambda}'(y)\eta\big(\beta^T x\big)\exp\big(-\eta\big(\beta^T x\big)\widetilde{\Lambda}(y)\big),$$

for almost every y in the support of the conditional distribution of T when $T \leq C$ given $X = x$, which is, by (A3)(i), equal to the support of $t \mapsto -\partial_t S(t)P(C > t|X = x)$, if S is the survival function associated to Λ. By (A3)(ii), this is true a.e. $(d\Lambda)$ and a.e. $(d\widetilde{\Lambda})$. Integrating from 0 to $y \in \mathbb{R}^+$, we get that $\Lambda = \widetilde{\Lambda}$.

8.2. Proof of Theorem 2

We first show the consistency of $\hat{\beta}$. The consistency of $\hat{\lambda}$ and $\hat{\Lambda}$ will follow. By (6), we have

$$\hat{\beta} = \operatorname*{argmax}_{\beta \in B} M_n(\beta),$$

$$\text{with } M_n(\beta) = n^{-1}\sum_{i=1}^{n}\big\{\delta_i \log\big(\eta\big(\beta^T X_i\big)\big) - \delta_i \log\big(\hat{R}_\beta(Y_i) - \hat{\lambda}_\beta\big) - \hat{\lambda}_\beta\big\},$$

with $\hat{\lambda}_\beta$ defined in (5). Similarly as for obtaining (6), we can show that the map Λ given by $t \mapsto E\int_0^t \frac{dN(u)}{R_\beta(u)-\lambda_\beta}$ maximizes $Pl_{\beta,\Lambda}$, for every $\beta \in B$. By plugging it in the likelihood, we obtain

$$\beta_0 = \operatorname*{argmax}_{\beta\in B} M(\beta),$$
$$\text{with } M(\beta) = E\delta\log\big(\eta\big(\beta^T X\big)\big) - E\delta\log\big(R_\beta(Y)-\lambda_\beta\big) - \lambda_\beta.$$

As a consequence, we can use Theorem 5.7 in [31]. This is done by checking that

$$\sup_{\beta\in B}\big|M_n(\beta) - M(\beta)\big| \xrightarrow{P} 0, \tag{13}$$

$$\sup_{|\beta-\beta_0|_2\geq\varepsilon}\big|M(\beta)\big| < M(\beta_0) \qquad \text{for all } \varepsilon > 0. \tag{14}$$

The second condition is a direct consequence of the identifiability of the model (see Theorem 5.35 in [31]) and the continuity of the map M on B which is compact. The first one can be obtained in the following way. The difference $M_n(\beta) - M(\beta)$ results naturally in three terms. We focus on

$$\int \log\big(\hat{R}_\beta(u)-\hat{\lambda}_\beta\big)\,d\overline{N}(u) - E\int \log\big(R_\beta(u)-\lambda_\beta\big)\,d\overline{N}(u),$$

which is the most difficult term to handle. It equals

$$\int \log\big(\hat{R}_\beta(u)-\hat{\lambda}_\beta\big)\big(d\overline{N}(u) - E\,d\overline{N}(u)\big) + \int \log\left(\frac{\hat{R}_\beta(u)-\hat{\lambda}_\beta}{R_\beta(u)-\lambda_\beta}\right) E\,d\overline{N}(u). \tag{15}$$

By (A2)(iv) and Lemma 7, we have that $\inf_{\beta\in B}(\hat{R}_\beta(\tau) - \hat{\lambda}_\beta) > 0$ with probability going to one. Hence, the first term in (15) goes to 0 in probability provided that $\sup_h |\int h(u)(d\overline{N}(u) - E\,d\overline{N}(u))|$ does, where the supremum is taken over the set of bounded increasing functions. By [9], this is equivalent to the uniform convergence of $\overline{N}(u)$ to $E\overline{N}(u)$ which is indeed true by the Glivenko–Cantelli theorem. The second term goes to 0 as a direct consequence of Lemma 7.

Now it remains to treat $\hat{\lambda}_{\hat{\beta}}$ and $\hat{\Lambda}$. For $\hat{\lambda}_{\hat{\beta}}$, since $\hat{\lambda}_{\hat{\beta}} = \lambda_{\hat{\beta}} + \hat{\lambda}_{\hat{\beta}} - \lambda_{\hat{\beta}} = \lambda_{\hat{\beta}} + o_P(1)$ by Lemma 7, we get the stated result by the continuity of the map $\beta \mapsto \lambda_\beta$. For $\hat{\Lambda}$, we write

$$\hat{\Lambda}(y) - \Lambda_0(y) = \int_0^y \frac{(d\overline{N}(u) - E\,d\overline{N}(u))}{\hat{R}_{\hat{\beta}}(u) - \hat{\lambda}_{\hat{\beta}}} + \int_0^y \big\{\big(R_{\hat{\beta}}(u) - \hat{\lambda}_{\hat{\beta}}\big)^{-1} - R_{\beta_0}(u)^{-1}\big\} E\,d\overline{N}(u),$$

and both terms are treated similarly as the two terms of (15).

8.3. Proof of Theorem 3

We first apply the well-known weak convergence theorem for Z-estimators [30], quoted in the Appendix as Theorem 11. It gives us the weak convergence of the NPMLE. Second, we obtain

the efficiency by using a suitable characterization of the influence function. The proof is divided in several important steps related to Theorem 11 and to the efficiency for the last step.

For any $(a, b, h) \in \mathbb{R}^{d+1} \times \ell^\infty(\mathbb{R}^+)$, we define the norm $\|(a, b, h)\| = |a|_2 + |b| + |h|_{\mathrm{tv}}$, on the parameter space.

First step: The class of scores is Donsker.

Let V_{β_0}, V_{λ_0} and V_{Λ_0} be neighborhoods of β_0, λ_0 and Λ_0, respectively, and denote $w = (y, \delta, x)$. We need to show that the class

$$\big\{w \mapsto B(\beta, \lambda, \Lambda)[a, b, h](w) : (a, b, h) \in \mathcal{H}, \beta \in V_{\beta_0}, \lambda \in V_{\lambda_0}, \Lambda \in V_{\Lambda_0}\big\}$$

is Donsker. This class is included in $\mathcal{B}_1 + \mathcal{B}_2 + \mathcal{B}_3 + \mathcal{B}_4$ with

$$\mathcal{B}_1 = \big\{w \mapsto \delta d_\beta(x)^T a + \delta h(y) : (a, b, h) \in \mathcal{H}\big\},$$

$$\mathcal{B}_2 = \left\{w \mapsto -\int d_\beta(x)^T a r_{u,\beta}(y, x)\, d\Lambda(u) : (a, b, h) \in \mathcal{H}, \beta \in V_{\beta_0}, \Lambda \in V_{\Lambda_0}\right\},$$

$$\mathcal{B}_3 = \left\{w \mapsto -\int h(u) r_{u,\beta}(y, x)\, d\Lambda(u) : (a, b, h) \in \mathcal{H}, \beta \in V_{\beta_0}, \Lambda \in V_{\Lambda_0}\right\},$$

$$\mathcal{B}_4 = \left\{w \mapsto -\lambda \int h(u)\, d\Lambda(u) + b\big(g(\Lambda) - 1\big) : (a, b, h) \in \mathcal{H}, \lambda \in V_{\lambda_0}, \Lambda \in V_{\Lambda_0}\right\}.$$

The class $\mathcal{B}_1$ is Donsker because the class of functions of bounded variation is Donsker [9] and because the class $\{w \mapsto \delta d_\beta(x) : \beta \in V_{\beta_0}\}$ is Donsker by Lemma 6 in the Appendix. Now it is easy to see that the class $\mathcal{B}_2$ will be Donsker provided that $\mathcal{B}_3$ is Donsker. For the class $\mathcal{B}_3$, we use the continuous mapping theorem (see Theorem 1.3.6 in [33]), by considering the mapping $T : \ell^\infty(\mathbb{R}^+ \times V_{\beta_0}) \to \ell^\infty(\mathcal{H} \times V_{\beta_0} \times V_{\Lambda_0})$ given by

$$T\widetilde{r} : (h, \beta, \Lambda) \mapsto \int h(u) \widetilde{r}_{u,\beta}\, d\Lambda(u),$$

for any $\widetilde{r} \in \ell^\infty(\mathbb{R}^+ \times V_{\beta_0})$. We have the identity $\mathbb{G}_n(\mathcal{B}_3) = T\mathbb{G}_n(\widetilde{\mathcal{B}}_3)$ with $\widetilde{\mathcal{B}}_3 = \{w \mapsto r_{u,\beta}(y, x) : \beta \in V_{\beta_0}, u \in \mathbb{R}^+\}$. Now, since $\widetilde{\mathcal{B}}_3$ is Donsker by Lemma 6 and since T is continuous ($|Tr - T\widetilde{r}| \le \|r - \widetilde{r}\|_\infty \|\Lambda\|_{\mathrm{BV}}$), $\mathbb{G}_n(\mathcal{B}_3)$ converges to a tight element in $\ell^\infty(\mathcal{H} \times V_{\beta_0} \times V_{\Lambda_0})$. Equivalently, B_3 is Donsker. The class B_4 is Donsker since it only contains constant and uniformly bounded functions.

Second step: The scores are ρ-continuous.

More precisely, we will show that

$$\sup_{(a,b,h)\in\mathcal{H}} \big|P\big(B(\beta, \lambda, \Lambda) - B(\beta_0, 0, \Lambda_0)\big)^2\big| \longrightarrow 0$$

as $\|(\beta - \beta_0, \lambda, \Lambda - \Lambda_0)\| \to 0$. We write $B(\beta, \lambda, \Lambda) - B(\beta_0, 0, \Lambda_0) = [B(\beta, \lambda, \Lambda) - B(\beta, 0, \Lambda)] + [B(\beta, 0, \Lambda) - B(\beta, 0, \Lambda_0)] + [B(\beta, 0, \Lambda_0) - B(\beta_0, 0, \Lambda_0)]$ and below we give upper

bounds for each term of this decomposition. First, we have

$$\big|B(\beta,\lambda,\Lambda)-B(\beta,0,\Lambda)\big|=\left|(\lambda-0)\int h(u)\,d\Lambda(u)\right|\le|\lambda-0|\|\Lambda\|_{\mathrm{BV}}.$$

Since $|g(\Lambda)-g(\Lambda_0)|\le\|\Lambda-\Lambda_0\|_{\mathrm{BV}}$, one has

$$\begin{aligned}&\big|B(\beta,0,\Lambda)-B(\beta,0,\Lambda_0)\big|\\&\quad=\left|-\int\big(d_\beta(X)^Ta+h(u)\big)r_{u,\beta}(Y,X)\,d(\Lambda-\Lambda_0)(u)+b\big(g(\Lambda)-g(\Lambda_0)\big)\right|\\&\quad\le\|\Lambda-\Lambda_0\|_{\mathrm{BV}}\big|\eta'\big(\beta^TX\big)\big(X^Ta\big)+\eta\big(\beta^TX\big)+b\big|,\end{aligned}$$

and finally,

$$\begin{aligned}&\big|B(\beta,0,\Lambda_0)-B(\beta_0,0,\Lambda_0)\big|\\&\quad=\left|\int\big(d_\beta(X)^Ta+h(u)\big)\,d(M_{\beta,\Lambda_0}-M_0)(u)+\int\big(d_\beta(X)-d_0(X)\big)^Ta\,dM_0(u)\right|\\&\quad=\left|\int\big(d_\beta(X)^Ta+h(u)\big)\big(r_{u,\beta}(Y,X)-r_{u,0}(Y,X)\big)\,d\Lambda_0(u)\right.\\&\qquad\left.+\int\big(d_\beta(X)-d_0(X)\big)^Ta\,dM_0(u)\right|\\&\quad\le\big|\eta\big(\beta^TX\big)-\eta\big(\beta_0^TX\big)\big|\|\Lambda_0\|_{\mathrm{BV}}\big|d_\beta(X)^Ta+1\big|\\&\qquad+\big|\big(d_\beta(X)-d_0(X)\big)^Ta\big|\Big|\lim_{y\to+\infty}M_0(y)\Big|.\end{aligned}$$

The conclusion follows from Lebesgue's dominated convergence theorem.

Third step: $\dot\Psi_0$ is continuously invertible.

By Lemma 9, Ψ is Fréchet differentiable at $(\beta_0,0,\Lambda_0)$ with derivative given in the aforementioned Lemma. Moreover, by Lemma 10, it is continuously invertible.

So far, since $(\hat\beta,\hat\lambda,\hat\Lambda)\to(\beta_0,0,\Lambda_0)$ in probability, all the conditions of Theorem 11 quoted in the Appendix are satisfied. Therefore, we have the following decomposition:

$$\dot\Psi_0n^{1/2}[\hat\beta-\beta_0,\hat\lambda,\hat\Lambda-\Lambda_0]=-\mathbb{G}_nB(\beta_0,0,\Lambda_0)+o_P(1).$$

This implies that

$$n^{1/2}(\hat\beta-\beta_0,\hat\lambda,\hat\Lambda-\Lambda_0)=-\dot\Psi_0^{-1}\big[\mathbb{G}_nB(\beta_0,0,\Lambda_0)\big]+o_P(1).$$

Fourth step: Efficiency.

To show the efficiency, we follow some ideas of the proof of Corollary 3.2 in [14]. From the previous decomposition, we deduce that the influence function φ associated with the random

sequence $n^{1/2}(\hat{\beta}-\beta_0, \hat{\Lambda}-\Lambda_0)$ can be expressed as

$$\varphi \equiv \big[\varphi_1, \varphi_2(\cdot)\big] = -\dot{\Psi}_0^{-1}\big(B(\beta_0, 0, \Lambda_0)\big)\begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \text{with } \varphi_1 \in \mathbb{R}^d.$$

By Theorem 25.23 in [31], we only need to show that φ is the efficient influence function of model $\mathcal{P}$. Equivalently, characterizing the influence function by Riesz theorem (see page 363 in [31] or Proposition 18.2 in [14]), we have that

$$P\big(\varphi \widetilde{B}(\beta_0, 0, \Lambda_0)[a, h]\big) = \left[a, y \mapsto \int_0^y h(u)\, d\Lambda_0(u)\right],$$

for every $a \in \mathbb{R}^d$ and every $h \in \mathcal{G}$, where $\mathcal{G} = \{f \in \ell^\infty(\mathbb{R}^+) : \int f(u)\, d\Lambda_0(u) = 0\}$ is introduced in Section 4. The operator $\widetilde{B}$ is defined in Section 4. It is the efficient score function associated with model $\mathcal{P}$. The previous equation holds whenever

$$P\big(-\dot{\Psi}_0^{-1}\big[B(\beta_0, 0, \Lambda_0)\big]\widetilde{B}(\beta_0, 0, \Lambda_0)[a, h]\big) = \left[a, 0, y \mapsto \int_0^y h(u)\, d\Lambda_0(u)\right], \tag{16}$$

for every $a \in \mathbb{R}^d$ and every $h \in \mathcal{G}$. By using the linearity of $\dot{\Psi}_0^{-1}$ and then the fact that $P\widetilde{B}(\beta_0, \Lambda_0)[a, h] = 0$ for every $(a, h) \in \mathbb{R}^d \times \mathcal{G}$, we get

$$\begin{aligned} P\big(-\dot{\Psi}_0^{-1}\big[B(\beta_0, 0, \Lambda_0)\big]\widetilde{B}(\beta_0, 0, \Lambda_0)[a, h]\big) &= -\dot{\Psi}_0^{-1}\big[P\big\{B(\beta_0, 0, \Lambda_0)\widetilde{B}(\beta_0, 0, \Lambda_0)[a, h]\big\}\big] \\ &= -\dot{\Psi}_0^{-1}\big[P\big\{\widetilde{B}(\beta_0, 0, \Lambda_0)\widetilde{B}(\beta_0, 0, \Lambda_0)[a, h]\big\}\big]. \end{aligned}$$

By (8), the right-hand side of the above equation equals

$$-\dot{\Psi}_0^{-1}\left[(\widetilde{a}, \widetilde{h}) \mapsto \int P\big\{\big(\widetilde{a}^T d_0 + \widetilde{h}(u)\big)\big(a^T d_0 + h(u)\big) r_{u,0}\big\}\, d\Lambda_0(u)\right].$$

By the definition of $\dot{\Psi}_0$, for any $(a, h) \in \mathbb{R}^d \times \mathcal{G}$, the map $(\widetilde{a}, \widetilde{h}) \mapsto \int P\{(\widetilde{a}^T d_0 + \widetilde{h}(u))(a^T d_0 + h(u)) r_{u,0}\}\, d\Lambda_0(u)$ is the image of $[a, 0, y \mapsto \int_0^y h(u)\, d\Lambda_0(u)]$ by the operator $-\dot{\Psi}_0$. This implies that the quantity in the previous equation is equal to

$$\dot{\Psi}_0^{-1}\dot{\Psi}_0\left[a, 0, y \mapsto \int_0^y h(u)\, d\Lambda_0(u)\right] = \left[a, 0, y \mapsto \int_0^y h(u)\, d\Lambda_0(u)\right].$$

Hence, we have shown (16).

8.4. Proof of Theorem 5

The bootstrap estimator is influenced by two different sources of randomness: the original sample $(W_i)_{i=1,\ldots,n}$, and the weights $(w_{i,n})_{i=1,\ldots,n}$. In the following, we say that $\Delta_n = o_{P^*}(1)$ in P-probability, if $P(P^*(|\Delta_n| > \eta) > \varepsilon) \to 0$, for any $\varepsilon > 0$ and $\eta > 0$. The formal statement of

Theorem 5 is

$$n^{1/2}\big((\hat{\beta}^*, \hat{\lambda}^*) - (\hat{\beta}, \hat{\lambda})\big) = G + o_{P^*}(1),$$

in P-probability, with G introduced in Theorem 3. To show this, we apply Theorem 3.1 in [33] in which the consistency (in P-probability) of the bootstrap is required in the first place. For us, this means showing that

$$\big|(\hat{\beta}^*, \hat{\lambda}^*_{\hat{\beta}^*}) - (\beta_0, 0)\big|_2 = o_{P^*}(1) \quad \text{and} \quad \big\|\hat{\Lambda}^* - \Lambda_0\big\|_\infty = o_{P^*}(1),$$

both in P-probability. We follow the proof of Theorem 2. We know that

$$\hat{\beta}^* = \underset{\beta \in B}{\operatorname{argmax}}\, M_n^*(\beta)$$

$$\text{with } M_n^*(\beta) = n^{-1} \sum_{i=1}^{n} \left\{ \delta_i w_{i,n} \log\left(\frac{\eta(\beta^T X_i)}{\hat{R}^*_\beta(Y_i) - \hat{\lambda}^*_\beta} \right) - \hat{\lambda}^*_\beta \right\}.$$

Hence, it suffices to check both conditions (13) and (14), but with M_n replaced by M_n^*. The latter one is already verified. For the former, we introduce $\overline{N}^*(y) = n^{-1} \sum_{i=1}^n \delta_i w_{i,n} 1_{\{Y_i \le y\}}$, and write $M_n^*(\beta) - M_n(\beta)$ as

$$\int \log\big(\hat{R}^*_\beta(u) - \hat{\lambda}^*_\beta\big)\, d\big(\overline{N^*} - \overline{N}\big)(u) + \int \log\left(\frac{\hat{R}^*_\beta(u) - \hat{\lambda}^*_\beta}{\hat{R}_\beta(u) - \hat{\lambda}_\beta} \right) d\overline{N}(u), \tag{17}$$

and then, relying on Lemma 8, we can follow what has been done in the proof of Theorem 2 to show that each term of (17) is $o_{P^*}(1)$, in P-probability.

Because all other conditions in Theorem 3.1 in [33] have been verified when showing Theorem 3, it only remains to show that

$$\lim_{u \to +\infty} \liminf_{n \to +\infty} \sup_{t \ge u} P^*\big(D_n(W) > t\big) = 0$$

$$\text{with } D_n(W) = \sup_{\|\theta - \theta_0\| \le \delta_n} \frac{\|B(\theta) - B(\theta_0)\|_{\mathcal{H}}}{1 + \sqrt{n}\|\theta - \theta_0\|},$$

with $\theta = (\beta, \lambda, \Lambda)$ and $\theta_0 = (\beta_0, \lambda_0, \Lambda_0)$, for every $\delta_n \to 0$. Following Example 1 in [33], for n sufficiently large, we have that

$$D_n(W) \le 2 \sup_{\|\theta - \theta_0\| \le 1} \big\|B(\theta)\big\|_{\mathcal{H}}.$$

Then, since it is fairly straightforward to show that the above quantity is bounded, the conclusion follows.

Appendix: Auxiliary results

For sake of clarity and brevity in the presentation, when possible, we omit from now on the integration variables in the proofs.

A.1. Some lemmas

Lemma 6. *Under* (A1)(ii) *and* (A2)(ii), *the function classes* $\{x \mapsto d_\beta(x) : \beta \in B\}$ *and* $\{(y,x) \mapsto r_{u,\beta}(y,x) : \beta \in B, u \in \mathbb{R}^+\}$ *are P-Donsker.*

Proof. Since B is compact, we can embed B in a ball of finite radius. Let β, $\widetilde{\beta}$ be elements of B. By assumption (A2)(ii), η'/η is Lipschitz on compact sets, and therefore we have for $k = 1, \ldots, d$,

$$P(d_{\beta,k} - d_{\widetilde{\beta},k})^2 \le C_1 E\big(X_k^2\big(\beta^T X - \widetilde{\beta}^T X\big)^2\big) \le C_1 |\beta - \widetilde{\beta}|_2^2 E\big(X_k^2 |X|_2^2\big) \le C_1 M^4 d |\beta - \widetilde{\beta}|_2^2,$$

for some $0 < C_1 < +\infty$, where the last bound is derived using (A1)(ii). It is well known that an ε-covering of B (with respect to the norm $|\cdot|_2$) may have a cardinality of order ε^{-d}. Hence, the ε-covering number of $\{x \mapsto d_\beta(x) : \beta \in B\}$ with respect to the metric ρ has order ε^{-d}, whose square root logarithm is integrable. This implies the first statement.

For the second class, remark that $\{(y,\Delta) \mapsto 1_{\{y \ge u\}}\Delta + (1-\Delta) : u \in \mathbb{R}^+\}$ is bounded and Donsker with covering number of order ε^{-1} (see, for instance, Example 2.5.7 in [33]). Moreover, we can act similarly as before to show that the class $\{x \mapsto \eta(\beta^T x) : \beta \in B\}$ is bounded and Donsker. Finally, the result follows by using that the product of two bounded Donsker classes is again Donsker (see, for instance, Example 2.10.8 in [33]). □

Lemma 7. *Under* (A1)(ii) *and* (A2)(ii), (iv), $\sup_{\beta \in B} \|\hat{R}_\beta - R_\beta\|_\infty \xrightarrow{P} 0$ *and* $\sup_{\beta \in B} |\hat{\lambda}_\beta - \lambda_\beta| \xrightarrow{P} 0$.

Proof. The first convergence follows from Lemma 6, because $(u,\beta) \mapsto R_\beta(u)$ equals the empirical process $(u,\beta) \mapsto \mathbb{P}_n r_{u,\beta}$, and because Donsker classes are also Glivenko–Cantelli. The second convergence is equivalent to the uniform consistency (in β) of a certain class of Z-estimators, indexed by β. For $A > 0$, define $\lambda_\beta(A)$ as the solution (if it exists) on $(-\infty, R_\beta(\tau)]$ of

$$E \int \frac{dN}{R_\beta - \lambda} = A.$$

Then, clearly $\lambda_\beta(1) = \lambda_\beta$. Since $\lambda \mapsto E\int \frac{dN}{R_\beta - \lambda}$ is continuously increasing on the set $(-\infty, R_\beta(\tau)]$, by (A2)(iv), we know that $\inf_{\beta \in B} R_\beta(\tau) - \lambda_\beta(A) > 0$ for A sufficiently close to 1 (we should further assume that this is the case). Moreover, the function $A \mapsto \lambda_\beta(A)$ is uniformly (in $\beta \in B$) continuous at the point $A = 1$, that is,

$$\lim_{A \to 1} \sup_{\beta \in B} \big|\lambda_\beta(A) - \lambda_\beta\big| = 0. \tag{18}$$

Then showing that

$$P\big(\lambda_\beta(1-\eta) \le \hat{\lambda}_\beta \le \lambda_\beta(1+\eta) \text{ for all } \beta \in B\big) \longrightarrow 1, \tag{19}$$

for every $\eta > 0$, will conclude the proof. Indeed since this is true for any $\eta > 0$, one can find a sequence $\eta_n \to 0$, such that (19) holds replacing η by η_n. It will remain to invoke (18) in order to obtain the stated result. Hence, we finish the proof by showing that (19) holds true. Let $\eta > 0$. Since $\lambda \mapsto \Phi_{\beta,n}(\lambda) = n^{-1}\sum_{i=1}^n \frac{\delta_i}{\hat{R}_\beta(Y_i)-\lambda}$ is an increasing function on $(-\infty, R_\beta(\tau)]$ for every $\beta \in B$, we have that the event in (19) is equivalent to

$$\big\{\Phi_{\beta,n}\big(\lambda_\beta(1-\eta)\big) \le 1 \le \Phi_{\beta,n}\big(\lambda_\beta(1+\eta)\big) \text{ for all } \beta \in B\big\}$$

which happens with probability going to 1 as soon as

$$\sup_{\beta\in B}\big|\Phi_{\beta,n}\big(\lambda_\beta(A)\big) - A\big| \xrightarrow{P} 0,$$

for any $A > 0$. We have

$$\Phi_{\beta,n}\big(\lambda_\beta(A)\big) - A = \int \frac{d\overline{N} - E\,\overline{dN}}{\hat{R}_\beta - \lambda_\beta(A)} + \int \big\{\big(\hat{R}_\beta - \lambda_\beta(A)\big)^{-1} - \big(R_\beta - \lambda_\beta(A)\big)^{-1}\big\} E\,\overline{dN}.$$

As a consequence, we can follow the arguments used to obtain the convergence of both terms in (15) in the proof of Theorem 2. □

Lemma 8. *Under* (A1)(ii), (A2)(ii), (iv) *and* (B1)–(B3), *we have that* $\sup_{\beta\in B}\|\hat{R}^*_\beta - \hat{R}_\beta\|_\infty = o_{P^*}(1)$ *in P-probability and* $\sup_{\beta\in B}|\hat{\lambda}^*_\beta - \hat{\lambda}_\beta| = o_{P^*}(1)$ *in P-probability.*

Proof. The first convergence is a consequence of Lemma 6 and Theorem 2.1 in [23]. The technique to obtain the second convergence is the same as in the proof of Lemma 7, showing that, for any $\eta > 0$, we have that $\hat{\lambda}_\beta(1-\eta) \le \hat{\lambda}^*_\beta \le \hat{\lambda}_\beta(1+\eta)$ with high probability. We omit the details, since they are similar to those of the proof of Lemma 7, replacing $\Phi_{\beta,n}$ by $\lambda \mapsto n^{-1}\sum_{i=1}^n \frac{w_{i,n}\delta_i}{\hat{R}^*_\beta(Y_i)-\lambda}$. □

Lemma 9. *Under* (A1)(ii) *and* (A2)(ii), *the map* Ψ *is Fréchet differentiable at* $(\beta_0, 0, \Lambda_0)$, *and the derivative* $\dot{\Psi}_0 : \lin\Theta \to \ell^\infty(\mathcal{H})$ *is given by the map* (*with* $d_0 \equiv d_{\beta_0}$ *and* $r_{u,0} \equiv r_{u,\beta_0}$)

$$\begin{aligned}
&\dot{\Psi}_0[\beta-\beta_0, \lambda, \Lambda-\Lambda_0](a,b,h) \\
&\quad = -\int P\big\{\big(d_0^T a + h(u)\big) d_0^T r_{u,0}\big\}\, d\Lambda_0(u)(\beta-\beta_0) \\
&\qquad - \int \big(P\big\{\big(d_0^T a + h(u)\big) r_{u,0}\big\} - b\big)\, d(\Lambda-\Lambda_0)(u) - \lambda\int h(u)\, d\Lambda_0(u),
\end{aligned}$$

for any $(a,b,h) \in \mathcal{H}$ *and* $(\beta,\lambda,\Lambda) \in \Theta$.

Proof. By definition of the Fréchet differentiability, we need to show that

$$\begin{aligned}&\big\|\Psi(\beta,\lambda,\Lambda)-\Psi(\beta_0,0,\Lambda_0)-\dot{\Psi}_0[\beta-\beta_0,\lambda,\Lambda-\Lambda_0]\big\|_{\mathcal{H}}\\&\quad=o\big(\big\|(\beta-\beta_0,\lambda,\Lambda-\Lambda_0)\big\|\big),\end{aligned}\tag{20}$$

as $\|(\beta-\beta_0,\lambda,\Lambda-\Lambda_0)\|\to 0$. First, we will show that $\dot{\Psi}_0$ is the Gâteau derivative of the map Ψ at the point $(\beta_0,\lambda_0,\Lambda_0)$. Second, we will show that (20) holds. Let $(\beta,\lambda,\Lambda)\in\Theta$. On the one hand, since

$$d(M_{\beta,\Lambda}-M_0)(u)=(r_{u,0}-r_{u,\beta})\,d\Lambda_0(u)+r_{u,0}\,d(\Lambda_0-\Lambda)(u)+(r_{u,\beta}-r_{u,0})\,d(\Lambda_0-\Lambda)(u),$$

and because $Ef(X)\,dM_0=0$ for any bounded f, we have

$$\begin{aligned}&E\int\big(a^Td_\beta(X)+h\big)\,dM_{\beta,\Lambda}\\&\quad=E\int\big(a^Td_\beta(X)+h\big)\,d(M_{\beta,\Lambda}-M_0)\\&\quad=P\int\big(a^Td_\beta+h\big)(r_{u,0}-r_{u,\beta})\,d\Lambda_0(u)\\&\qquad+P\int\big(a^Td_\beta+h(u)\big)r_{u,0}\,d(\Lambda_0-\Lambda)(u)+r_1\\&\quad=P\int\big(a^Td_0+h(u)\big)(r_{u,0}-r_{u,\beta})\,d\Lambda_0(u)\\&\qquad+P\int\big(a^Td_0+h(u)\big)r_{u,0}\,d(\Lambda_0-\Lambda)(u)+r_1+r_2+r_3,\end{aligned}\tag{21}$$

with $r_1=P\int(a^Td_\beta+h(u))(r_{u,\beta}-r_{u,0})\,d(\Lambda_0-\Lambda)(u)$, $r_2=P\int a^T(d_\beta-d_0)(r_{u,0}-r_{u,\beta})\,d\Lambda_0(u)$ and $r_3=P\int a^T(d_\beta-d_0)r_{u,0}\,d(\Lambda_0-\Lambda)(u)$. On the other hand, we have

$$\begin{aligned}-\lambda\int h\,d\Lambda+b\big(g(\Lambda)-1\big)&=-\lambda\int h\,d\Lambda_0+b\big(g(\Lambda)-g(\Lambda_0)\big)+r_4\\&=-\lambda\int h\,d\Lambda_0+b\int d(\Lambda-\Lambda_0)+r_4,\end{aligned}\tag{22}$$

with $r_4=\lambda\int h\,d(\Lambda_0-\Lambda)$. Combining (21) and (22), we get that

$$\begin{aligned}&\Psi(\beta,\lambda,\Lambda)[a,b,h]\\&\quad=-\int P\big(a^Td_0+h(u)\big)(r_{u,\beta}-r_{u,0})\,d\Lambda_0(u)\\&\qquad-\int\big(P\big\{\big(a^Td_0+h(u)\big)r_{u,0}\big\}-b\big)\,d(\Lambda-\Lambda_0)(u)-\lambda\int h\,d\Lambda_0+\sum_{k=1}^{4}r_k.\end{aligned}$$

Now, using a Taylor expansion, we obtain that

$$\left|\eta\left(\beta^T X\right) - \eta\left(\beta_0^T X\right) - (\beta-\beta_0)^T X \eta'\left(\beta_0^T X\right)\right| \leq \frac{1}{2}|\beta-\beta_0|_2^2 |X|_2^2 \sup_{u\in K}\left|\eta''(u)\right|,$$

where $K = \{\beta^T x : x \in [-M, M]^d, \beta \in B\}$ is compact. As a consequence, we have

$$\begin{aligned}
&\Psi(\beta,\lambda,\Lambda)[a,b,h] \\
&\quad = -\int P\left\{\left(a^T d_0 + h\right)(\beta-\beta_0)^T d_0 r_{u,0}\right\} d\Lambda_0(u) \\
&\qquad - \int \left(P\left\{\left(a^T d_0 + h(u)\right) r_{u,0}\right\} - b\right) d(\Lambda-\Lambda_0)(u) \\
&\qquad - \lambda \int h\, d\Lambda_0 + \sum_{k=1}^{5} r_k,
\end{aligned}$$

with $r_5 = -\int P(a^T d_0 + h(u))(r_{u,\beta} - r_{u,0} - (\beta-\beta_0)^T d_0 r_{u,0})\, d\Lambda_0(u) = -\int P(a^T d_0 + h(u))(\eta(\beta^T X) - \eta(\beta_0^T X) - (\beta-\beta_0)^T X \eta'(\beta_0^T X))(\Delta 1_{\{Y\geq u\}} + (1-\Delta))\, d\Lambda_0(u)$. Equation (20) holds if and only if

$$\left\|\sum_{k=1}^{5} r_k\right\|_{\mathcal{H}} = o\left(\left\|(\beta-\beta_0,\lambda,\Lambda-\Lambda_0)\right\|\right).$$

It is fairly straightforward to show this using the regularity conditions on η and the boundedness of the support of X. For instance, for r_5 we have

$$|r_5| \leq \frac{1}{2}|\beta-\beta_0|_2^2 \sup_{u\in K}\left|\eta''(u)\right| \int E\left(|X|_2^2 \left|a^T d_0(X) + h\right|\right) d\Lambda_0 = O\left(|\beta-\beta_0|_2^2\right). \qquad \square$$

Lemma 10. *Under* (A1)(i), (ii), (A2)(ii) *and* (A3)(i), (ii), *the operator* $\dot{\Psi}_0 : \operatorname{lin}\Theta \to \ell^\infty(\mathcal{H})$ *is continuously invertible*.

Proof. Since $\dot{\Psi}_0$ (given in Lemma 9) is a linear operator between two Banach spaces we can apply Lemma 6.17 in [14] to $\dot{\Psi}_0 = T + K$, where T and K are given by

$$T[\beta-\beta_0,\lambda,\Lambda-\Lambda_0](a,b,h) = -a^T I_0(\beta-\beta_0) - \int (hR_0 - b)\, d(\Lambda-\Lambda_0) - \lambda \int h\, d\Lambda_0,$$

$$K[\beta-\beta_0,\Lambda-\Lambda_0](a,b,h) = -\int \left(a^T h_0 + h\right) D_0^T\, d\Lambda_0(\beta-\beta_0) - \int a^T D_0\, d(\Lambda-\Lambda_0),$$

and we recall that $I_0 = \int P\{(d_0 - h_0(u))(d_0 - h_0(u))^T r_{u,0}\}\, d\Lambda_0(u)$, $R_0(u) = P r_{u,0}$ and $D_0(u) = P d_0 r_{u,0}$. First, we show that T is continuously invertible by proving that it is bounded and that

$$\left\|T[\beta-\beta_0,\lambda,\Lambda-\Lambda_0]\right\|_{\mathcal{H}} \geq \varepsilon \left\|(\beta-\beta_0,\lambda,\Lambda-\Lambda_0)\right\| \tag{A.23}$$

for some $\varepsilon > 0$. The boundedness of T follows from the boundedness of R_0. Since the right-hand side of (A.23) equals

$$\sup_{|a|_2\le\varepsilon,|b|\le\varepsilon,\|h\|_{\mathrm{tv}}\le\varepsilon}\left\{a^T(\beta-\beta_0)+b\lambda+\int h\,d(\Lambda-\Lambda_0)\right\},$$

it suffices to show that $\{(a,b,h)\in\mathbb{R}^{d+1}\times\ell^\infty(\mathbb{R}^+) : |a|_2\le\varepsilon, |b|\le\varepsilon, \|h\|_{\mathrm{tv}}\le\varepsilon\}$ is included in

$$\left\{\left(a^T I_0, \int h\,d\Lambda_0, hR_0-b\right) : |a|_2\le 1, |b|\le 1, \|h\|_{\mathrm{tv}}\le 1\right\}.$$

This is straightforward, because I^* is invertible by (A1)(i) and the fact that R_0 is bounded away from 0 with bounded variations (by (A2)(ii) and (A3)(i), (ii)).

Second, we demonstrate that K is a compact operator, that is, for every sequence in $K(\{(\alpha,H) : \|(\alpha,H)\|\le 1\})$ there exists a subsequence that converges. Let (α_n, H_n) be such a sequence. By Helly's selection theorem (see [2]) and the compactness of $\{\alpha\in\mathbb{R}^d : |\alpha|_2\le 1\}$, there exists a subsequence $(\alpha_{k(n)}, H_{k(n)})$ that converges pointwise to $(\alpha_\infty, H_\infty)$. According to Gini's theorem, the pointwise convergence can be extended to uniform convergence (see, for instance, the proof of Theorem 19.1 in [31]). Consequently, we have that

$$\begin{aligned}
&\big\|K[\alpha_{k(n)},H_{k(n)}]-K[\alpha_\infty,H_\infty]\big\|_{\mathcal{H}}\\
&\quad=\sup_{(\alpha,h)\in\mathcal{H}}\left|\int\big(\alpha^T h_0+h\big)D_0^T\,d\Lambda_0(\alpha_{k(n)}-\alpha_\infty)+\alpha^T\int D_0\,d(H_{k(n)}-H_\infty)\right|\\
&\quad\le C_2\big(|\alpha_{k(n)}-\alpha_\infty|_2+\|H_{k(n)}-H_\infty\|_{\mathrm{BV}}\big),
\end{aligned}$$

for some $0<C_2<+\infty$, and this tends to 0.

Third, we show that $\ker(\dot\Psi_0)=\{0\}$. Let $(\alpha,\gamma,H)\in\operatorname{lin}\Theta$ such that $\dot\Psi_0[\alpha,\gamma,H]=0$, with $\dot\Psi_0$ given in Lemma 9. By taking $a=h=0$, we get that $\int dH=0$. Then, by taking $h=-a^T h_0$, we obtain

$$a^T\int P\big(d_0-h_0(u)\big)\,d_0^T r_{u,0}\,d\Lambda_0(u)\alpha=0,$$

for every a, since it is easily verified that $\int P(d_0^T a-h_0^T a)r_{u,0}\,dH=0$ and $\int h_0\,d\Lambda_0=0$. This is equivalent to $I_0\alpha=0$, which implies that $\alpha=0$ according to (A1)(i). By taking $h=(1+a^T h_0)$ and for the same reason as previously, we get that $\gamma=0$. We conclude with $\int h\,dH=0$, which implies that $H\equiv 0$. □

A.2. A weak convergence theorem for Z-estimators

Infinite dimensional Z-estimators have been studied by many authors: see Theorem 25.90 in [31], Theorem 3.3.1 in [33], Theorem 2.11 in [14], among others. The following statement is not exactly the original one, due to Van der Vaart [30], since we use Lemma 3.3.5 of [32] to

replace one of the conditions of the original statement. The new conditions (namely (c) and (d) in the following theorem) are not strictly necessary but are easy to check in many cases. Let $(\Theta, \|\cdot\|)$ be a subset of a Banach space and let $\mathcal{H}$ be a given set. Let $B(\theta) : \mathcal{H} \to L_2(P)$ where $L_2(P)$ denotes the set of functions with second moment bounded with respect to P, and define the operator $\dot{\Psi}_0 : \operatorname{lin}\Theta \to \ell^\infty(\mathcal{H})$ as the Fréchet derivative of the map PB at some point $\theta_0 \in \Theta$.

Theorem 11 (van der Vaart [30]). *Let $\theta_0 \in \Theta$ and let V_{θ_0} be some neighborhood of θ_0. Suppose that*

(a) $\|\mathbb{P}_n B(\hat{\theta})\|_{\mathcal{H}} = o_P(n^{-1/2})$ *and* $\|PB(\theta_0)\|_{\mathcal{H}} = 0$.
(b) $\hat{\theta} \xrightarrow{P} \theta_0$.
(c) *The class* $\{B(\theta)[h] : h \in \mathcal{H}, \theta \in V_{\theta_0}\}$ *is Donsker.*
(d) $\|P(B(\theta) - B(\theta_0))^2\|_{\mathcal{H}} \longrightarrow 0$ *whenever* $\theta \to \theta_0$.
(e) *The operator* $\dot{\Psi}_0$ *is continuously invertible.*

Then,

$$\dot{\Psi}_0 n^{1/2}(\hat{\theta} - \theta_0) = -\mathbb{P}_n B(\theta_0) + o_P(1).$$

Acknowledgements

This work was supported by IAP Research Network P7/06 of the Belgian State (Belgian Science Policy), and by the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16-039 of the "Communauté française de Belgique" (granted by the "Académie universitaire Louvain").

F. Portier was supported by Fonds de la Recherche Scientifique (FNRS) A4/5 FC 2779/2014-2017 No. 22342320.

I. Van Keilegom was supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) ERC Grant agreement No. 203650.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. *Springer Series in Statistics*. New York: Springer. MR1198884

[2] Ash, R.B. (1972). *Real Analysis and Probability*. New York: Academic Press. Probability and Mathematical Statistics, No. 11. MR0435320

[3] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. *Johns Hopkins Series in the Mathematical Sciences*. Baltimore, MD: Johns Hopkins Univ. Press. MR1245941

[4] Breslow, N. (1972). Contribution to the discussion of the paper by D.R. Cox entitled: "Regression models and life-tables." *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

[5] Chen, M.-H., Ibrahim, J.G. and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *J. Amer. Statist. Assoc.* **94** 909–919. MR1723307

[6] Cheng, G. and Huang, J.Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *Ann. Statist.* **38** 2884–2915. MR2722459

[7] Cox, D.R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758
[8] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*. *Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge: Cambridge Univ. Press. MR1478673
[9] Dudley, R.M. (1992). Fréchet differentiability, p-variation and uniform Donsker classes. *Ann. Probab.* **20** 1968–1982. MR1188050
[10] Dupuy, J.-F., Grama, I. and Mesbah, M. (2006). Asymptotic theory for the Cox model with missing time-dependent covariate. *Ann. Statist.* **34** 903–924. MR2283397
[11] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681
[12] Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. New York: Wiley. MR1100924
[13] Ibrahim, J.G., Chen, M.-H. and Sinha, D. (2001). Bayesian semiparametric models for survival data with a cure fraction. *Biometrics* **57** 383–388. MR1854229
[14] Kosorok, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. *Springer Series in Statistics*. New York: Springer. MR2724368
[15] Kosorok, M.R., Lee, B.L. and Fine, J.P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Ann. Statist.* **32** 1448–1491. MR2089130
[16] Lo, A.Y. (1993). A Bayesian bootstrap for censored data. *Ann. Statist.* **21** 100–123. MR1212168
[17] Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Ann. Inst. Statist. Math.* **60** 545–574. MR2434411
[18] Ma, Y. and Yin, G. (2008). Cure rate model with mismeasured covariates under transformation. *J. Amer. Statist. Assoc.* **103** 743–756. MR2524007
[19] Murphy, S.A. (1994). Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.* **22** 712–731. MR1292537
[20] Murphy, S.A. (1995). Asymptotic theory for the frailty model. *Ann. Statist.* **23** 182–198. MR1331663
[21] Murphy, S.A., Rossini, A.J. and van der Vaart, A.W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Amer. Statist. Assoc.* **92** 968–976. MR1482127
[22] Murphy, S.A. and van der Vaart, A.W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168
[23] Præstgaard, J. and Wellner, J.A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086. MR1245301
[24] Ritov, Y. and Wellner, J.A. (1988). Censoring, martingales, and the Cox model. In *Statistical Inference from Stochastic Processes* (*Ithaca*, *NY*, 1987). *Contemp. Math.* **80** 191–219. Providence, RI: Amer. Math. Soc. MR0999013
[25] Sasieni, P. (1992). Information bounds for the conditional hazard ratio in a nested family of regression models. *J. Roy. Statist. Soc. Ser. B* **54** 617–635. MR1160487
[26] Tsodikov, A. (1998). Asymptotic efficiency of a proportional hazards model with cure. *Statist. Probab. Lett.* **39** 237–244. MR1646259
[27] Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics* **54** 1508–1516.
[28] Tsodikov, A. (2001). Estimation of survival based on proportional hazards when cure is a possibility. *Math. Comput. Modelling* **33** 1227–1236. MR1837407
[29] Tsodikov, A.D., Ibrahim, J.G. and Yakovlev, A.Y. (2003). Estimating cure rates from survival data: An alternative to two-component mixture models. *J. Amer. Statist. Assoc.* **98** 1063–1078. MR2055496
[30] van der Vaart, A.W. (1995). Efficiency of infinite-dimensional M-estimators. *Stat. Neerl.* **49** 9–30. MR1333176

[31] van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247

[32] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671

[33] Wellner, J.A. and Zhan, Y. (1996). Bootstrapping Z-estimators. Technical Report 308. Univ. Washington, Dept. Statistics.

[34] Yakovlev, A. and Tsodikov, A. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. Singapore: World Scientific.

[35] Zeng, D., Yin, G. and Ibrahim, J.G. (2006). Semiparametric transformation models for survival data with a cure fraction. *J. Amer. Statist. Assoc.* **101** 670–684. MR2256182

Received March 2015 and revised January 2016