A General Approach for Cure Models in Survival Analysis

Article *in* The Annals of Statistics · January 2017

DOI: 10.1214/19-AOS1889

Valentin Patilea
ENSAI (Ecole Nationale de la Statistique et de l'Analyse de l'Information), Rennes, Fr…
76 PUBLICATIONS **593** CITATIONS
SEE PROFILE

Ingrid Van Keilegom
KU Leuven
184 PUBLICATIONS **3,166** CITATIONS
SEE PROFILE

A General Approach for Cure Models in Survival Analysis

Valentin Patilea [*] Ingrid Van Keilegom [§]

January 16, 2017

Abstract

In survival analysis it often happens that some subjects under study do not experience the event of interest; they are considered to be 'cured'. The population is thus a mixture of two subpopulations : the one of cured subjects, and the one of 'susceptible' subjects. When covariates are present, a so-called mixture cure model can be used to model the conditional survival function of the population. It depends on two components : the probability of being cured and the conditional survival function of the susceptible subjects.

In this paper we propose a novel approach to estimate a mixture cure model when the data are subject to random right censoring. We work with a parametric model for the cure proportion (like e.g. a logistic model), while the conditional survival function of the uncured subjects is unspecified. The approach is based on an inversion which allows to write the survival function as a function of the distribution of the observable random variables. This leads to a very general class of models, which allows a flexible and rich modeling of the conditional survival function. We show the identifiability of the proposed model, as well as the weak consistency and the asymptotic normality of the model parameters. We also consider in more detail the case where kernel estimators are used for the nonparametric part of the model. The new estimators are compared with the estimators from a Cox mixture cure model via finite sample simulations. Finally, we apply the new model and estimation procedure on two medical data sets.

Key Words: Asymptotic normality; bootstrap; kernel smoothing; logistic regression; mixture cure model; semiparametric model.

MSC2010: 62N01, 62N02, 62E20, 62F12, 62G05.

[*] CREST (Ensai), France. V. Patilea acknowledges support from the research program *New Challenges for New Data* of LCL and Genes. Email address: patilea@ensai.fr.

[§] ORSTAT, Katholieke Universiteit Leuven, Belgium. I. Van Keilegom acknowledges support from the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409), and from IAP Research Network P7/06 of the Belgian State. Email address: ingrid.vankeilegom@kuleuven.be.

1 Introduction

Driven by emerging applications, over the last two decades there has been an increasing interest for time-to-event analysis models allowing the situation where a fraction of the right censored observed lifetimes corresponds to subjects who will never experience the event. In biostatistics such models including covariates are usually called *cure models* and they allow for a positive *cure fraction* that corresponds to the proportion of patients cured of their disease. For a review of these models in survival analysis, see for instance Maller & Zhou (2001) or Peng & Taylor (2014). Economists sometimes call such models *split population models* (see Schmidt & Witte 1989), while the reliability engineers refer to them as *limited-failure population life models* (Meeker 1987).

At first sight, a cure regression model is nothing but a binary outcome, cured versus uncured, regression problem. The difficulty comes from the fact that the cured subjects are unlabeled observations among the censored data. Then one has to use all the observations, censored and uncensored, to complete the missing information and thus to identify, estimate and make inference on the cure fraction regression function. We propose a general approach for this task, a tool that provides a general ground for cure regression models. The idea is to start from the laws of the observed variables and to express the quantities of interest, such as the cure rate and the conditional survival of the uncured subjects, as functionals of these laws. These general expressions, that we call inversion formulae and we derive with no particular constraint on the space of the covariates, are the vehicles that allow for a wide modeling choice, parametric, semiparametric and nonparametric, for both the law of the lifetime of interest and the cure rate. Indeed, the inversion formulae allow to express the likelihood of the binary outcome model as a function of the laws of the observed variables. The likelihood estimator of the parameter vector of the cure fraction function is then simply the maximizer of the likelihood obtained by replacing the laws of the observations by some estimators. With at hand the estimate of the parameter of the cure fraction, the inversion formulae will provide an estimate for the conditional survival of the uncured subjects. For the sake of clarity, we focus on the so-called mixture cure models with a parametric cure fraction function, the type of model that is most popular among practitioners. Meanwhile, the lifetime of interest is left unspecified.

The paper is organized as follows. In Section 2 we provide a general description of the mixture cure models and next we introduce the needed notation and present the inversion formulae on which our approach is built. We finish Section 2 by a discussion of the identification issue and some new insight on the existing approaches in the literature on cure models. Section 3 introduces the general maximum likelihood estimator, while in Section 4 we derive the general asymptotic results. A simple bootstrap procedure for making feasible inference is proposed. Section 4 ends with an illustration of the general approach in the case where the conditional law of the observations is estimated by kernel smoothing. In Sections

5 and 6 we report some empirical results obtained with simulated and two real data sets. Our estimator performs well in simulations and provides similar or more interpretable results in applications compared with a competing logistic/proportional hazards mixture approach. The technical proofs are relegated to the Appendix.

2 The model

2.1 A general class of mixture cure models

Let T denote (a possible monotone transformation of) the lifetime of interest that takes values in $(-\infty, \infty]$. A cured observation corresponds to the event $\{T = \infty\}$, and in the following this event is allowed to have a positive probability. Let X be a covariate vector with support $\mathcal{X}$ belonging to a general covariate space. The covariate vector could include discrete and continuous components. The survival function $F_T((t, \infty] \mid x) = \mathbb{P}(T > t \mid X = x)$ for $t \in \mathbb{R}$ and $x \in \mathcal{X}$ can be written as

$$F_T((t, \infty] \mid x) \;\; = \;\; 1 - \phi(x) + \phi(x) F_{T,0}((t, \infty) \mid x),$$

where $\phi(x) = \mathbb{P}(T < \infty \mid X = x)$ and $F_{T,0}((t, \infty) \mid x) = \mathbb{P}(T > t \mid X = x, T < \infty)$. Depending on which model is used for $\phi(x)$ and $F_{T,0}(\cdot \mid x)$, one obtains a parametric, semiparametric or nonparametric model, called a 'mixture cure model'. In the literature, one often assumes that $\phi(x)$ follows a logistic model, i.e. $\phi(x) = \exp(a + x^\top b)/[1 + \exp(a + x^\top b)]$ for some $(a, b^\top)^\top \in \mathbb{R}^{d+1}$. Recently, semiparametric models (like a single-index model as in Amico *et al.* 2017) or nonparametric models (as in Xu & Peng 2014 or López-Cheda *et al.* 2017) have been proposed. As for the survival function $F_{T,0}(\cdot \mid x)$ of the susceptible subjects, a variety of models have been proposed, including parametric models (see e.g. Boag 1949, Farewell 1982), semiparametric models based on a proportional hazards assumption (see e.g. Kuk & Chen 1992, Sy & Taylor 2000, Fang *et al.* 2005, Lu 2008; see also Othus *et al.* 2009) or nonparametric models (see e.g. Taylor 1995, Xu & Peng 2014).

In this paper we propose to model $\phi(x)$ parametrically, i.e. we assume that $\phi(\cdot)$ belongs to the family of conditional probability functions

$$\{\phi(\cdot, \beta) : \beta \in B\},$$

where $\phi(\cdot, \beta)$ takes values in the interval $(0, 1)$, β is the parameter vector of the model and B is the parameter set. This family could be the logistic family or any other parametric family. For the survival function $F_{T,0}(\cdot \mid x)$ we do not impose any assumptions in order to have a flexible and rich class of models for $F_T(\cdot \mid x)$ to choose from. Later on we will see that for the estimation of $F_{T,0}(\cdot \mid x)$ any estimator that satisfies certain minimal conditions can be used, and hence we allow for a large variety of parametric, semiparametric and nonparametric estimation methods.

As is often the case with time-to-event data, we assume that the lifetime T is subject to random right censoring, i.e. instead of observing T, we only observe the pair (Y, δ), where $Y = T \wedge C$, $\delta = \mathbf{1}\{T \leq C\}$ and C is a non-negative random variable, called the censoring time. Some identification assumptions are required to be able to identify the conditional law of T from the observed variables Y and δ. Let us assume that

$$C \perp T \mid X \qquad \text{and} \qquad \mathbb{P}(C < \infty) = 1. \tag{2.1}$$

The conditional independence between T and C is an usual identification assumption in survival analysis in the presence of covariates. The zero probability at infinity condition for C implies that $\mathbb{P}(C < \infty \mid X) = 1$ almost surely (a.s.). This latter mild condition is required if we admit that the observations Y are finite, which is the case in the common applications. For the sake of simplicity, let us also consider the condition

$$\mathbb{P}(T = C) = 0, \tag{2.2}$$

which is commonly used in survival analysis, and which implies that $\mathbb{P}(T = C \mid X) = 0$ a.s.

2.2 Some notations and preliminaries

We start with some preliminary arguments, which are valid in general without assuming any model on the functions ϕ, F_T and F_C.

The observations are characterized by the conditional sub-probabilities

$$\begin{aligned} H_1((-\infty, t] \mid x) &= \mathbb{P}(Y \leq t, \delta = 1 \mid X = x) \\ H_0((-\infty, t] \mid x) &= \mathbb{P}(Y \leq t, \delta = 0 \mid X = x), \qquad t \in \mathbb{R},\ x \in \mathcal{X}. \end{aligned}$$

Then $H((-\infty, t] \mid x) \overset{def}{=} \mathbb{P}(Y \leq t \mid X = x) = H_0((-\infty, t] \mid x) + H_1((-\infty, t] \mid x)$. Since we assume that Y is finite, we have

$$H((-\infty, \infty) \mid x) = 1, \qquad \forall x \in \mathcal{X}. \tag{2.3}$$

For $j \in \{0, 1\}$ and $x \in \mathcal{X}$, let $\tau_{H_j}(x) = \sup\{t : H_j([t, \infty) \mid x) > 0\}$ denote the right endpoint of the support of the conditional sub-probability H_j. Let us define $\tau_H(x)$ in a similar way and note that $\tau_H(x) = \max\{\tau_{H_0}(x), \tau_{H_1}(x)\}$. Note that $\tau_{H_0}(x)$, $\tau_{H_1}(x)$ and $\tau_H(x)$ can equal infinity, even though Y only takes finite values. For $-\infty < t \leq \infty$, let us define the conditional probabilities

$$F_C((-\infty, t] \mid x) = \mathbb{P}(C \leq t \mid X = x) \quad \text{and} \quad F_T((-\infty, t] \mid x) = \mathbb{P}(T \leq t \mid X = x), \quad x \in \mathcal{X}.$$

Let us show how the probability of being cured could be identified from the observations without any reference to a model for this probability. Under conditions (2.1)-(2.2) we can write

$$H_1(dt \mid x) = F_C([t, \infty) \mid x) F_T(dt \mid x), \qquad H_0(dt \mid x) = F_T([t, \infty] \mid x) F_C(dt \mid x),$$

and $H([t,\infty) \mid x) = F_T([t,\infty] \mid x)F_C([t,\infty) \mid x)$. These equations could be solved and thus they allow to express the functions $F_T(\cdot \mid x)$ and $F_C(\cdot \mid x)$ in an unique way as explicit transformations of the functions $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$. For this purpose, let us consider the conditional cumulative hazard measures

$$\Lambda_T(dt \mid x) = \frac{F_T(dt \mid x)}{F_T([t,\infty] \mid x)} \quad \text{and} \quad \Lambda_C(dt \mid x) = \frac{F_C(dt \mid x)}{F_C([t,\infty) \mid x)}, \qquad x \in \mathcal{X}.$$

The model equations yield

$$\Lambda_T(dt \mid x) = \frac{H_1(dt \mid x)}{H([t,\infty) \mid x)} \quad \text{and} \quad \Lambda_C(dt \mid x) = \frac{H_0(dt \mid x)}{H([t,\infty) \mid x)}. \tag{2.4}$$

Then, we can write the following functionals of $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$:

$$\begin{aligned} F_T((t,\infty] \mid x) &= \prod_{-\infty<s\le t} \{1 - \Lambda_T(ds \mid x)\}, \\ F_C((t,\infty) \mid x) &= \prod_{-\infty<s\le t} \{1 - \Lambda_C(ds \mid x)\}, \qquad t \in \mathbb{R}, \end{aligned} \tag{2.5}$$

where $\prod_{s\in A}$ stands for the product-integral over the set A (see Gill and Johansen 1990).

Moreover, if $\tau_{H_1}(x) < \infty$, then

$$\mathbb{P}(T > \tau_{H_1}(x) \mid x) = \prod_{t\in(-\infty,\tau_{H_1}(x)]} \{1 - \Lambda_T(dt \mid x)\},$$

but there is no way to identify the conditional law of T beyond $\tau_{H_1}(x)$. Therefore, we will impose

$$\mathbb{P}(T > \tau_{H_1}(x) \mid x) = \mathbb{P}(T = \infty \mid x), \tag{2.6}$$

i.e. $\prod_{t\in\mathbb{R}}\{1 - \Lambda_T(dt \mid x)\} = \prod_{-\infty<t\le\tau_{H_1}(x)}\{1 - \Lambda_T(dt \mid x)\}$. Note that if $\tau_{H_1}(x) = \infty$, condition (2.6) is no longer an identification restriction, but just a simple consequence of the definition of $\Lambda_T(\cdot \mid x)$. Finally, the condition that $\mathbb{P}(C < \infty) = 1$ in (2.1) can be re-expressed by saying that we assume that $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$ are such that

$$\mathbb{P}(C = \infty \mid x) = \prod_{t\in\mathbb{R}}\{1 - \Lambda_C(dt \mid x)\} = 0, \qquad \forall x \in \mathcal{X}. \tag{2.7}$$

Let us point out that this condition is satisfied only if $\tau_{H_1}(x) \le \tau_{H_0}(x)$. Indeed, if $\tau_{H_1}(x) > \tau_{H_0}(x)$ then necessarily $\tau_{H_0}(x) < \tau_H(x)$ and so $H([\tau_{H_0}(x),\infty) \mid x) > 0$. Hence, $\Lambda_C(\mathbb{R} \mid x) = \Lambda_C((-\infty,\tau_{H_0}(x)] \mid x) < \infty$, and thus $\mathbb{P}(C = \infty \mid x) > 0$, which contradicts (2.7).

It is important to understand that *any* two conditional sub-probabilities $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$ satisfying conditions (2.1) and (2.2) define uniquely $F_T(\cdot \mid x)$ and $F_C(\cdot \mid x)$. Indeed, $F_T(\cdot \mid x)$ is precisely the probability distribution of T given $X = x$ with all the mass beyond $\tau_{H_1}(x)$ concentrated at infinity. In general, $F_T(\cdot \mid x)$ and $F_C(\cdot \mid x)$ are only functionals of $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$.

We will assume conditions (2.1), (2.2) and (2.6) throughout the paper.

2.3 A key point for the new approach: the inversion formulae

Write

$$\begin{aligned} H([t,\infty)\mid x) &= F_T([t,\infty]\mid x)F_C([t,\infty)\mid x)\\ &= F_T([t,\infty)\mid x)F_C([t,\infty)\mid x) + \mathbb{P}(T=\infty\mid x)F_C([t,\infty)\mid x), \end{aligned}$$

and thus

$$F_T([t,\infty)\mid x) = \frac{H([t,\infty)\mid x) - \mathbb{P}(T=\infty\mid x)F_C([t,\infty)\mid x)}{F_C([t,\infty)\mid x)}. \tag{2.8}$$

Consider the conditional cumulative hazard measure for the finite values of the lifetime of interest:

$$\Lambda_{T,0}(dt\mid x) \overset{def}{=} \frac{F_{T,0}(dt\mid x)}{F_{T,0}([t,\infty)\mid x)} = \frac{F_T(dt\mid x)}{F_T([t,\infty)\mid x)}$$

for $t\in\mathbb{R}$. Since $H_1(dt\mid x) = F_C([t,\infty)\mid x)F_T(dt\mid x)$, using relationship (2.8) we obtain

$$\Lambda_{T,0}(dt\mid x) \;=\; \frac{H_1(dt\mid x)}{H([t,\infty)\mid x) - \mathbb{P}(T=\infty\mid x)F_C([t,\infty)\mid x)}. \tag{2.9}$$

Next, using the product-integral we can write

$$F_{T,0}((t,\infty)\mid x) = \prod_{-\infty<s\le t}\{1-\Lambda_{T,0}(ds\mid x)\},\quad t\in\mathbb{R},\ x\in\mathcal{X}. \tag{2.10}$$

Let us recall that $F_C(\cdot\mid x)$ can be written as a transformation of $H_0(\cdot\mid x)$ and $H_1(\cdot\mid x)$, see equations (2.4) and (2.5). This representation is not surprising since we can consider C as a lifetime of interest and hence T plays the role of a censoring variable. Hence, estimating the conditional distribution function $F_C(\cdot\mid x)$ should not be more complicated than in a classical conditional Kaplan-Meier setup, since the fact that T could be equal to infinity with positive conditional probability is irrelevant when estimating $F_C(\cdot\mid x)$.

Finally, the representation of $F_C(\cdot\mid x)$ given in equation (2.5), plugged into equation (2.9), allows to express $\Lambda_{T,0}(\cdot\mid x)$, and thus $F_{T,0}(\cdot\mid x)$, as maps of $\mathbb{P}(T=\infty\mid x)$ and the measures $H_0(\cdot\mid x)$ and $H_1(\cdot\mid x)$. This will be the key element for providing more insight in the existing approaches and the starting point of our new approach.

2.4 Model identification issues

Let us now investigate the identification issue. Recall that our model involves the functions $F_T(\cdot\mid x)$, $F_C(\cdot\mid x)$ and $\phi(\cdot,\beta)$, and the assumptions (2.1), (2.2) and (2.6). For a fixed value of the parameter β, and for $t\in\mathbb{R}$ and $x\in\mathcal{X}$, let

$$\Lambda^{\beta}_{T,0}(dt\mid x) = \frac{H_1(dt\mid x)}{H([t,\infty)\mid x) - [1-\phi(x,\beta)]F_C([t,\infty)\mid x)}, \tag{2.11}$$

and

$$F^{\beta}_{T,0}((t,\infty)\mid x)=\prod_{-\infty<s\leq t}\left\{1-\Lambda^{\beta}_{T,0}(ds\mid x)\right\}. \tag{2.12}$$

Let $F_{Y,\delta}(\cdot,\cdot\mid x)$ denote the conditional law of (Y,δ) given $X=x$. Moreover, let

$$F^{\beta}_{Y,\delta}(dt,1\mid x)=\phi(x,\beta)F_C((t,\infty)\mid x)F^{\beta}_{T,0}(dt\mid x)$$

and

$$F^{\beta}_{Y,\delta}(dt,0\mid x)=[F^{\beta}_{T,0}((t,\infty)\mid x)\phi(x,\beta)+1-\phi(x,\beta)]F_C(dt\mid x).$$

These equations define a conditional law for the observations (Y,δ) based on the model. More precisely, for a choice of $F_T(\cdot\mid x)$, $F_C(\cdot\mid x)$ and β, the model yields a conditional law for (Y,δ) given $X=x$. If the model is correctly specified, there exists a value β_0 such that

$$F_{Y,\delta}(\cdot,\cdot\mid x)=F^{\beta_0}_{Y,\delta}(\cdot,\cdot\mid x),\quad\forall x\in\mathcal{X}. \tag{2.13}$$

The remaining question is whether the true value of the parameter is identifiable. In other words, one should check if, given the conditional subdistributions $H_0(\cdot\mid x)$ and $H_1(\cdot\mid x)$, $x\in\mathcal{X}$, there exists a unique β_0 satisfying condition (2.13). For this purpose we impose the following mild condition:

$$\phi(X,\beta)=\phi(X,\widetilde{\beta}),\ \text{ almost surely }\quad\Rightarrow\beta=\widetilde{\beta}, \tag{2.14}$$

and we show that

$$F^{\beta_0}_{Y,\delta}(\cdot,\cdot\mid x)=F^{\widetilde{\beta}}_{Y,\delta}(\cdot,\cdot\mid x),\ \forall x\in\mathcal{X}\ \ \Rightarrow\ \ \phi(x,\beta_0)=\phi(x,\widetilde{\beta}),\ \forall x\in\mathcal{X}.$$

Indeed, if $F^{\beta_0}_{Y,\delta}(\cdot,\cdot\mid x)=F^{\widetilde{\beta}}_{Y,\delta}(\cdot,\cdot\mid x)$, then for any x,

$$\phi(x,\beta_0)F_C((t,\infty)\mid x)F^{\beta_0}_{T,0}(dt\mid x)=\phi(x,\widetilde{\beta})F_C((t,\infty)\mid x)F^{\widetilde{\beta}}_{T,0}(dt\mid x),$$

for all $t\in(-\infty,\tau_H(x)]\cap\mathbb{R}$. Our condition (2.7) guarantees that $\tau_H(x)=\tau_{H_0}(x)$ so that $F_C((t,\infty)\mid x)$ should be necessarily positive for $t\in(-\infty,\tau_H(x))$, and thus could be simplified in the last display. Deduce that

$$\phi(x,\beta_0)F^{\beta_0}_{T,0}(dt\mid x)=\phi(x,\widetilde{\beta})F^{\widetilde{\beta}}_{T,0}(dt\mid x),$$

for all $t\in(-\infty,\tau_H(x))$. Finally, recall that by construction, in the model we consider, $F^{\beta}_{T,0}((-\infty,\infty)\mid x)=1$ for any β such that $F^{\beta}_{Y,\delta}(\cdot,\cdot\mid x)$ coincides with the conditional law of (Y,δ) given $X=x$, for any x. Thus taking integrals on $(-\infty,\infty)$ on both sides of the last display we obtain $\phi(\cdot,\beta_0)=\phi(\cdot,\widetilde{\beta})$. Let us gather these facts in the following statement.

Theorem 2.1. *Under conditions (2.1), (2.2), (2.6) and (2.14) the model is identifiable.*

2.5 Interpreting the previous modeling approaches

We suppose here that the function $\phi(x)$ follows a logistic model, and comment on several models for $F_{T,0}$ that have been considered in the literature.

2.5.1 Parametric and proportional hazards mixture model

In a parametric modeling, one usually supposes that $\tau_{H_0}(x) = \tau_{H_1}(x) = \infty$ and that $\Lambda_{T,0}(\cdot \mid x)$ belongs to a parametric family of cumulative hazard functions, like for instance the Weibull model; see Farewell (1982).

Several contributions proposed a more flexible semiparametric proportional hazards (PH) approach; see Fang *et al.* (2005), Lu (2008) and the references therein. In such a model one imposes a PH structure for the $\Lambda_{T,0}(\cdot \mid x)$ measure. More precisely, it is supposed that

$$\Lambda_{T,0}(dt \mid x) = \frac{H_1(dt \mid x)}{H([t,\infty) \mid x) - [1-\phi(x,\beta)]F_C([t,\infty) \mid x)} = \exp(x^\top \gamma)\Lambda_0(dt),$$

where γ is some parameter to be estimated and $\Lambda_0(\cdot)$ is an unknown baseline cumulative hazard function. Our inversion formulae reveal that in this approach the parameters γ and Λ_0 depend on the observed conditional measures $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$, but also on the parameter β. The same is true for the parametric models.

2.5.2 Kaplan-Meier mixture cure model

Taylor (1995) suggested to estimate $F_{T,0}$ using a Kaplan-Meier type estimator. With such an approach one implicitly assumes that the law of T given X and given that $T < \infty$ does not depend on X. This is equivalent to supposing that $\Lambda_{T,0}(\cdot \mid x) = \Lambda_{T,0}(\cdot)$. Next, to estimate $\Lambda_{T,0}(\cdot)$ one has to modify the unconditional version of the usual inversion formulae (2.4) to take into account the conditional probability of the event $\{T = \infty\}$. Following Taylor's approach we rewrite (2.9) as

$$\Lambda_{T,0}(dt) = \frac{H_1(dt \mid x)}{H_1([t,\infty)\,|\,x) + \int_{[t,\infty)} \left\{1 - \frac{1-\phi(x,\beta)}{\phi(x,\beta)F_{T,0}([s,\infty)) + 1 - \phi(x,\beta)}\right\} H_0(ds \mid x)}.$$

Next, assume that the last equality remains true if $H_0(dt \mid x)$ and $H_1(dt \mid x)$ are replaced by their unconditional versions, that is assume that

$$\Lambda_{T,0}(dt) = \frac{H_1(dt)}{H_1([t,\infty)) + \int_{[t,\infty)} \left\{1 - \frac{1-\phi(x,\beta)}{\phi(x,\beta)F_{T,0}([s,\infty)) + 1 - \phi(x,\beta)}\right\} H_0(ds)}. \tag{2.15}$$

See equations (2) and (3) in Taylor (1995). The equation above could be solved iteratively by a EM-type procedure: for a given β and an iteration $F_{T,0}^{(m)}(\cdot)$, build $\Lambda_{T,0}^{(m+1)}(dt)$ and the updated estimate $F_{T,0}^{(m+1)}(\cdot)$; see Taylor (1995) for the details. Let us point out that even if

(T, C) is independent of X and thus $H_1(\cdot \mid x)$ does not depend on x, the subdistribution $H_0(\cdot \mid x)$ still depends on x, since

$$\begin{aligned} H_0(dt \mid x) &= F_T((t,\infty] \mid x) F_C(dt \mid x) \\ &= [F_T((t,\infty) \mid x) + \mathbb{P}(T = \infty \mid x)] F_C(dt \mid x). \end{aligned}$$

Hence, a more natural form of equation (2.15) is

$$\Lambda_{T,0}(dt) = \frac{H_1(dt)}{H_1([t,\infty)) + \int_{[t,\infty)} \left\{ 1 - \frac{1-\phi(x,\beta)}{\phi(x,\beta)F_{T,0}([s,\infty)) + (1-\phi(x,\beta))} \right\} H_0(ds \mid x)}.$$

The investigation of a EM-type procedure based on the latter equation will be considered elsewhere.

3 Maximum likelihood estimation

Let (Y_i, δ_i, X_i) $(i = 1, \ldots, n)$ be a sample of n i.i.d. copies of the vector (Y, δ, X).

We use a likelihood approach based on formulae (2.9) and (2.5) to build an estimator of $\phi(\cdot, \beta)$ and $F_{T,0}^{\beta}(\cdot \mid x)$. To build the likelihood we use estimates $\widehat{H}_k(\cdot \mid x)$ of the subdistributions $H_k(\cdot \mid x)$, $k \in \{0, 1\}$. These estimates are constructed with the sample of (Y, δ, X), without reference to any model for the conditional probability $\mathbb{P}(T < \infty \mid x)$. At this stage it is not necessary to impose a particular form for $\widehat{H}_k(\cdot \mid x)$. To derive the asymptotic results we will only impose that these estimators satisfy some mild conditions. Let $\widehat{F}_{T,0}^{\beta}(\cdot \mid x)$ be defined as in equations (2.11) and (2.12) with $\widehat{H}_0(\cdot \mid x)$ and $\widehat{H}_1(\cdot \mid x)$ instead of $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$, that is

$$\widehat{F}_{T,0}^{\beta}((t,\infty) \mid x) = \prod_{-\infty < s \leq t} \left\{ 1 - \frac{\widehat{H}_1(ds \mid x)}{\widehat{H}([s,\infty) \mid x) - [1 - \phi(x,\beta)] \widehat{F}_C([s,\infty) \mid x)} \right\},$$

where $\widehat{F}_C(\cdot \mid x)$ is the estimator obtained from equations (2.4) and (2.5) but with $\widehat{H}_0(\cdot \mid x)$ and $\widehat{H}_1(\cdot \mid x)$, i.e.

$$\widehat{F}_C((t,\infty) \mid x) = \prod_{-\infty < s \leq t} \left\{ 1 - \frac{\widehat{H}_0(ds \mid x)}{\widehat{H}([s,\infty) \mid x)} \right\}.$$

Let f_X denote the density of the covariate vector with respect to some dominating measure. The contribution of the observation (Y_i, δ_i, X_i) to the likelihood when $\delta_i = 1$ is then

$$\phi(X_i, \beta) \widehat{F}_{T,0}^{\beta}(\{Y_i\} \mid X_i) f_X(X_i) \widehat{F}_C((Y_i, \infty) \mid X_i),$$

while the contribution when $\delta_i = 0$ is

$$\widehat{F}_C(\{Y_i\} \mid X_i) f_X(X_i) [\phi(X_i, \beta) \widehat{F}_{T,0}^{\beta}((Y_i, \infty) \mid X_i) + 1 - \phi(X_i, \beta)].$$

Since the laws of the censoring variable and of the covariate vector do not carry information of the parameter β, we can drop the factors $\widehat{F}_C((Y_i,\infty) \mid X_i)f_X(X_i)$ and $\widehat{F}_C(\{Y_i\} \mid X_i)f_X(X_i)$. Hence the criterion to be maximized with respect to $\beta \in B$ is $\widehat{L}_n(\beta)$ where

$$\widehat{L}_n(\beta) = \prod_{i=1}^{n} \left\{\phi(X_i,\beta)\widehat{F}^{\beta}_{T,0}(\{Y_i\} \mid X_i)\right\}^{\delta_i} \left\{\phi(X_i,\beta)\widehat{F}^{\beta}_{T,0}((Y_i,\infty) \mid X_i) + 1 - \phi(X_i,\beta)\right\}^{1-\delta_i}.$$

The estimator we propose is

$$\widehat{\beta} = \arg\max_{\beta\in B} \log \widehat{L}_n(\beta). \tag{3.1}$$

Let us review the identification issue in the context of the likelihood estimation approach. If conditions (2.1), (2.2) and (2.6) hold true and the parametric model for the conditional probability of the event $\{T=\infty\}$ is correct and identifiable in the sense of condition (2.14), the true parameter β_0 is the value identified by condition (2.13). This is the conclusion of Theorem 2.1 above. It remains to check that the proposed likelihood approach allows to consistently estimate β_0.

Let

$$\log p(t,\delta,x;\beta) = \delta \log p_1(t,x;\beta) + (1-\delta)\log p_0(t,x;\beta), \tag{3.2}$$

with

$$p_1(t,x;\beta) = \frac{\phi(x,\beta)F^{\beta}_{T,0}(dt \mid x)F_C((t,\infty) \mid x)}{H_1(dt \mid x)}, \tag{3.3}$$

$$p_0(t,x;\beta) = \frac{F_C(dt \mid x)[\phi(x,\beta)F^{\beta}_{T,0}((t,\infty) \mid x) + 1 - \phi(x,\beta)]}{H_0(dt \mid x)}. \tag{3.4}$$

Following a common notational convention, see for instance Gill (1994), here we treat dt not just as the length of a small time interval $[t,t+dt)$ but also as the name of the interval itself. Moreover, we use the convention $0/0=1$. Let us notice that, up to additive terms not containing β, the function $\beta \mapsto \mathbb{E}[\log p(Y,\delta,X;\beta)]$ is expected to be the limit of the random function $\log \widehat{L}_n(\cdot)$. Hence, a minimal condition for guaranteeing the consistency of the likelihood estimation approach is that β_0 is the maximizer of the limit likelihood criterion $\mathbb{E}[\log p(Y,\delta,X;\cdot)]$. This is proved in the following proposition using a Bernoulli sample likelihood inequality. The proof is given in the Appendix.

Proposition 3.1. *Suppose that conditions (2.1), (2.2), (2.6) and (2.14) hold true. If β_0 is the value of the parameter defined by equation (2.13), then for any $\beta \neq \beta_0$,*

$$\mathbb{E}\left[\log p(Y,\delta,X;\beta)\right] < \mathbb{E}\left[\log p(Y,\delta,X;\beta_0)\right].$$

4 General asymptotic results

Little assumptions were needed for our analysis so far. To proceed further with the asymptotic results we need to be more specific with respect to several aspects. In order to prove consistency, we have to control the asymptotic behavior of $\widehat{L}_n(\beta)$ along sequences of values of the parameter β. Such a control requires a control of denominators like

$$\widehat{H}([t,\infty) \mid x) - (1-\phi(x,\beta))\widehat{F}_C([t,\infty) \mid x)$$

on the support of $H_1(\cdot \mid x)$, uniformly with respect to x. A usual way to deal with this technical difficulty is to consider a finite threshold τ beyond which no uncensored lifetime is observed, i.e.

$$\inf_x H_1((-\infty,\tau] \mid x) = 1. \tag{4.1}$$

Moreover, to be able to keep denominators away from zero, we require the condition

$$\inf_{x\in\mathcal{X}} H_1(\{\tau\} \mid x) H_0((\tau,\infty) \mid x) > 0. \tag{4.2}$$

In particular, this condition implies

$$\tau_{H_0}(x) > \tau_{H_1}(x) = \tau, \qquad x \in \mathcal{X}.$$

Moreover, given condition (2.2), necessarily $H_0(\{\tau\} \mid x) = 0$, $\forall x$. This means that $F_C(\{\tau\} \mid x) = 0$, $\forall x$. This constraint on $H_0(\{\tau\} \mid x)$ could be relaxed at the expense of suitable adjustments of the inversion formulae. For simplicity, we keep condition (2.2). Let us also notice that condition (4.2) implies $\inf_x F_C((\tau,\infty) \mid x) > 0$, and $\inf_x F^{\beta}_{T,0}(\{\tau\} \mid x) > 0$, $\forall\beta$.

Conditions like in equations (4.1)-(4.2) are more or less explicitly used in the literature of cure models. Sometimes τ is justified as representing a total follow-up of the study. For instance, Lu (2008) supposes that $Y = \min\{T, \min(C,\tau)\}$ and $\delta = \mathbf{1}\{T \leq \min(C,\tau)\}$, where $T = \eta T^* + (1-\eta)\infty$, with $T^* < \infty$ and $\eta \in \{0,1\}$. The conditional probability of being cured is precisely the conditional probability of the event $\{\eta = 0\}$. Next, Lu (2008) supposes that $\inf_x \mathbb{P}(\tau \leq T \leq C \mid x) > 0$, and $\Lambda_0(\tau) < \infty$, where $\Lambda_0(\cdot)$ is the cumulative hazard function of T^*. All these conditions together clearly imply our conditions (4.1)-(4.2).

Fang *et al.* (2005) implicitly restrict the uncensored lifetimes to some compact interval $[0,\tau]$ and suppose $\mathbb{E}(\delta\mathbf{1}\{Y \geq \tau\}) > 0$. This could be possible only if $H_1(\{\tau\} \mid x) > 0$ for a set of values x with positive probability. In a proportional hazards context with the covariates taking values in a bounded set, as assumed by Fang *et al.* (2005), this is equivalent to $H_1(\{\tau\} \mid x) \geq c > 0$ for almost all x, for some constant c.

The fact that technical conditions similar to our conditions (4.1)-(4.2) could be traced in the cure models literature is not unexpected in view of our Section 2.5. Indeed, the existing approaches could be interpreted through our inversion formulae and thus the technical problems we face in the asymptotic investigation are expected to be also present in the alternative approaches.

4.1 Consistency

Let us sketch the arguments we use in the proof of Theorem 4.1 below for deriving the consistency of $\widehat{\beta}$. On one hand, if the conditional subdistributions $H_k(\cdot \mid x)$ are given, one can build the purely parametric likelihood

$$\mathcal{L}_n(\beta) = \prod_{i=1}^{n} \Big\{ \phi(X_i, \beta) F_{T,0}^{\beta}(\{Y_i\} \mid X_i) \Big\}^{\delta_i} \Big\{ \phi(X_i, \beta) F_{T,0}^{\beta}((Y_i, \infty) \mid X_i) + 1 - \phi(X_i, \beta) \Big\}^{1-\delta_i}. \quad (4.3)$$

By construction, $\mathcal{L}_n(\beta)$ is a functional of $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$, $x \in \mathcal{X}$, while $\widehat{L}_n(\beta)$ is a functional of the estimated versions of $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$. Hence, a prerequisite condition for deriving the consistency of our semiparametric estimator $\widehat{\beta}$ is the consistency of the infeasible maximum likelihood estimator

$$\widetilde{\beta} = \arg\max_{\beta \in B} \log \mathcal{L}_n(\beta).$$

A necessary condition for the consistency of $\widehat{\beta}$ is

$$\sup_{\beta \in B} |\log \widehat{L}_n(\beta) - \log \mathcal{L}_n(\beta)| = o_{\mathbb{P}}(1). \quad (4.4)$$

We then have that

$$\log \mathcal{L}_n(\widehat{\beta}) \geq \log \mathcal{L}_n(\widetilde{\beta}) - o_{\mathbb{P}}(1).$$

From this we will derive the consistency of $\widehat{\beta}$ using Section 5.2 in van der Vaart (1998). See the proof in the Appendix for details.

To prove condition (4.4), we have to guarantee the uniform convergence of $\widehat{H}_k - H_k$ as stated in Assumption (AC1) below. Indeed, this uniform convergence will imply

$$\sup_{x \in \mathcal{X}} \sup_{t \in (-\infty, \tau]} |\widehat{F}_C([t, \infty) \mid x) - F_C([t, \infty) \mid x)| = o_{\mathbb{P}}(1), \quad (4.5)$$

and

$$\sup_{\beta \in B} \sup_{x \in \mathcal{X}} \sup_{t \in (-\infty, \tau]} |\widehat{F}_{T,0}^{\beta}([t, \infty) \mid x) - F_{T,0}^{\beta}([t, \infty) \mid x)| = o_{\mathbb{P}}(1). \quad (4.6)$$

See Lemma 7.1 in the Appendix. The uniform convergence in equation (4.4) then follows.

To prove the consistency of $\widehat{\beta}$, we need the following assumptions :

(AC1) For τ appearing in conditions (4.1)-(4.2),

$$\sup_{x \in \mathcal{X}} \sup_{t \in (-\infty, \tau]} |\widehat{H}_k([t, \infty) \mid x) - H_k([t, \infty) \mid x)| = o_{\mathbb{P}}(1), \qquad k \in \{0, 1\}.$$

(AC2) The parameter set $B \subset \mathbb{R}^p$ is compact.

(AC3) There exist some constants $a > 0$ and $c_1 > 0$ such that

$$|\phi(x,\beta) - \phi(x,\beta')| \leq c_1 \|\beta - \beta'\|^a, \qquad \forall \beta, \beta' \in B, \ \forall x, x' \in \mathcal{X}.$$

(AC4) $\inf_{\beta \in B} \inf_{x \in \mathcal{X}} \phi(x,\beta) > 0$.

Now we can state our consistency result.

Theorem 4.1. *Assume that (AC1)-(AC4) and (2.1), (2.2), (2.6), (2.14), (4.1) and (4.2) hold true. Moreover, assume that there exists a unique value β_0 in the parameter space B such that (2.13) is true. Then, $\widehat{\beta} - \beta_0 = o_{\mathbb{P}}(1)$.*

Let us point out that the consistency result is stated in terms of the subdistributions of the observations and the conditional probability model $\{\phi(x,\cdot) : \beta \in B\}$. If the identification assumptions used in Proposition 3.1 hold true and the model is correctly specified, $\phi(\widehat{\beta}, x)$ consistently estimates the cure probability $\mathbb{P}(T = \infty \mid x)$ for all x in the support of X. Let us also notice that condition (AC3) guarantees the Glivenko-Cantelli property for certain classes of functions. It could be significantly weakened, but in the applications our condition (AC3) will cover the common modeling situations. Condition (AC4) is a weak condition on the model $\phi(x,\beta)$ and is e.g. satisfied for the logistic model if $\mathcal{X}$ and B are compact.

4.2 Asymptotic normality

For the asymptotic normality we will use the approach in Chen *et al.* (2003). For this purpose we use the derivative of $\log \widehat{L}_n(\beta)$ with respect to β.

First note that the vector of partial derivatives of the log-likelihood $\log \widehat{L}_n(\beta)$ with respect to the components of β equals

$$\begin{aligned}\nabla_\beta \log \widehat{L}_n(\beta) &= \frac{1}{n}\sum_{i=1}^n \nabla_\beta \Big\{ \delta_i \log \phi(X_i,\beta) + \delta_i \log[\widehat{H}_1(\{Y_i\} \mid X_i)] \\ &\quad -\delta_i \log\Big[\widehat{H}([Y_i,\infty) \mid X_i) - (1-\phi(X_i,\beta))\widehat{F}_C([Y_i,\infty) \mid X_i)\Big] \\ &\quad +\delta_i \log \widehat{F}^{(1)}_{T,\beta}([Y_i,\infty) \mid X_i) \\ &\quad +(1-\delta_i)\log\Big[\phi(X_i,\beta)\widehat{F}^{(1)}_{T,\beta}([Y_i,\infty) \mid X_i) + 1 - \phi(X_i,\beta)\Big]\Big\}. \\ &= \frac{1}{n}\sum_{i=1}^n \nabla_\beta q_i(\beta, \widehat{H}_0, \widehat{H}_1),\end{aligned}$$

where q_i is defined in the proof of Theorem 4.1.

To develop the asymptotic normality of our estimator $\widehat{\beta}$, we embed the nuisance functions $H_k([\cdot,\infty) \mid \cdot)$ $(k = 0, 1)$ in a functional space $\mathcal{H}$, which is equipped with a pseudo-norm $\|\cdot\|_{\mathcal{H}}$. Both the space $\mathcal{H}$ and its pseudo-norm $\|\cdot\|_{\mathcal{H}}$ will be chosen depending on the estimators

$\widehat{H}_k([\cdot,\infty) \mid \cdot)$, $k=0,1$, and have to satisfy certain conditions, which we give below. The true vector of nuisance functions is

$$\eta_0(t,x) = (\eta_{01}(t,x), \eta_{02}(t,x)) = (H_0([t,\infty) \mid x), H_1([t,\infty) \mid x)).$$

For each $x \in \mathcal{X}$ and for each $\eta \in \mathcal{H}$, let $\eta(dt,x) = (\eta_1(dt,x), \eta_2(dt,x))$ be the measures associated to the non-increasing functions $\eta(\cdot,x) = (\eta_1(\cdot,x), \eta_2(\cdot,x))$, and define

$$M_n(\beta,\eta) = \frac{1}{n}\sum_{i=1}^{n} m(Y_i, \delta_i, X_i; \beta, \eta) \tag{4.7}$$

and

$$M(\beta,\eta) = \mathbb{E}[m(Y,\delta,X;\beta,\eta)], \tag{4.8}$$

where

$$\begin{aligned} m(t,\delta,x;\beta,\eta) = {} & \frac{\delta\nabla_\beta\phi(x,\beta)}{\phi(x,\beta)} - \frac{\delta\nabla_\beta\phi(x,\beta)T_2(\eta)(t,x)}{T_1(\beta,\eta)(t,x)} + \delta T_4(\beta,\eta)(t,x) \\ & + (1-\delta)\frac{\nabla_\beta\phi(x,\beta)T_3(\beta,\eta)(t,x) + \phi(x,\beta)(T_3T_4)(\beta,\eta)(t,x) - \nabla_\beta\phi(x,\beta)}{\phi(x,\beta)T_3(\beta,\eta)(t,x) + 1 - \phi(x,\beta)}, \end{aligned}$$

where

$$\begin{aligned} T_1(\beta,\eta)(t,x) &= [\eta_1(t,x) + \eta_2(t,x)] - (1-\phi(x,\beta))T_2(\beta,\eta)(t,x) \qquad (4.9) \\ T_2(\beta,\eta)(t,x) &\equiv T_2(\eta)(t,x) = \prod_{-\infty<s<t}\left(1 - \frac{\eta_1(ds,x)}{(\eta_1+\eta_2)(s,x)}\right) \\ T_3(\beta,\eta)(t,x) &= \prod_{-\infty<s<t}\left(1 - \frac{\eta_2(ds,x)}{T_1(\beta,\eta)(s,x)}\right) \\ T_4(\beta,\eta)(t,x) &= \nabla_\beta \log T_3(\beta,\eta)(t,x) \\ &= -\nabla_\beta\phi(x,\beta)\int_{(-\infty,t)} \frac{T_2(\eta)(s,x)\eta_2(ds,x)}{T_1(\beta,\eta)(s,x)[T_1(\beta,\eta)(s,x) - \eta_2(\{s\},x)]}. \end{aligned}$$

Note that for $\eta = \eta_0$, we have

$$T_1(\beta,\eta_0)(t,x) = H([t,\infty) \mid x) - (1-\phi(x,\beta))F_C([t,\infty) \mid x),$$

$T_2(\eta_0)(t,x) = F_C([t,\infty) \mid x)$ and $T_3(\beta,\eta_0)(t,x) = F_{T,0}^{\beta}([t,\infty) \mid x)$. Hence, we have that

$$M(\beta_0,\eta_0) = 0 \quad \text{and} \quad \|M_n(\widehat{\beta},\widehat{\eta})\| = \inf_{\beta\in B}\|M_n(\beta,\widehat{\eta})\|,$$

where $\widehat{\eta}(t,x) = (\widehat{\eta}_1(t,x), \widehat{\eta}_2(t,x)) = (\widehat{H}_0([t,\infty) \mid x), \widehat{H}_1([t,\infty) \mid x))$.

If in addition $\beta = \beta_0$,

$$T_1(\beta_0,\eta_0)(t,x) = F_{T,0}^{\beta_0}([t,\infty) \mid x)F_C([t,\infty) \mid x), \qquad t \in (-\infty,\tau],$$

and thus for any x, the map $t \mapsto T_1(\beta_0, \eta_0)(t, x)$ is decreasing on $(-\infty, \tau]$. Moreover, by condition (4.2),

$$\inf_{x \in \mathcal{X}} T_1(\beta_0, \eta_0)(\tau, x) > 0.$$

We need this lower bound to be valid on a neighborhood around β_0. Hence, let us consider B_0 a neighborhood of β_0 such that

$$\inf_{\beta \in B_0} \inf_{x \in \mathcal{X}} T_1(\beta, \eta_0)(\tau, x) > 0. \tag{4.10}$$

The existence of B_0 is guaranteed by condition (4.2) and the regularity of the function $\phi(\cdot, \cdot)$; see assumption (AN3) below that strengthens assumption (AC3). Finally, let us note that by construction for any $t \in (-\infty, \tau)$, $H(\{t\} \mid x) = H_0(\{t\} \mid x) + H_1(\{t\} \mid x) \geq (1 - \phi(x, \beta)) F_C(\{t\} \mid x) + H_1(\{t\} \mid x)$ and thus

$$\begin{aligned} H([t, \infty) \mid x) - (1 - \phi(x, \beta)) F_C([t, \infty) \mid x) - H_1(\{t\} \mid x) \\ \geq H((t, \infty) \mid x) - (1 - \phi(x, \beta)) F_C((t, \infty) \mid x). \end{aligned}$$

Then, by the arguments guaranteeing the existence of a set B_0 as in equation (4.10),

$$\inf_{\beta \in B_0} \inf_{x \in \mathcal{X}} \inf_{t \in (-\infty, \tau)} [T_1(\beta, \eta_0)(t, x) - H_1(\{t\} \mid x)] > 0. \tag{4.11}$$

Further, define the Gâteaux derivative of $M(\beta, \eta_0)$ in the direction $[\eta - \eta_0]$ by

$$\nabla_\eta M(\beta, \eta_0)[\eta - \eta_0] = \lim_{\tau \to 0} \tau^{-1} \Big[M(\beta, \eta_0 + \tau(\eta - \eta_0)) - M(\beta, \eta_0) \Big],$$

and in a similar way the Gâteaux derivatives $\nabla_\eta T_j(\beta, \eta_0)[\eta - \eta_0]$ are defined.

We need the following assumptions :

(AN1) The matrix $\nabla_\beta M(\beta, \eta)$ exists for β in a neighborhood B_0 of β_0 and is continuous in β for $\beta = \beta_0$. Moreover, $\nabla_\beta M(\beta_0, \eta_0)$ is non-singular.

(AN2) $H_k([\cdot, \infty) \mid \cdot) \in \mathcal{H}$ for $k = 0, 1$.

(AN3) The function $\beta \to \phi(x, \beta)$ is continuously differentiable for all $x \in \mathcal{X}$, and the derivative is bounded uniformly in $x \in \mathcal{X}$ and $\beta \in B_0$. Moreover, B_0 is compact and β_0 belongs to the interior of B_0.

(AN4) For $k = 0, 1$, the estimator $\widehat{H}_k([\cdot, \infty) \mid \cdot)$ satisfies the following :

(i) $\mathbb{P}(\widehat{H}_k([\cdot, \infty) \mid \cdot) \in \mathcal{H}) \to 1$

(ii) $\|(\widehat{H}_k - H_k)([\cdot, \infty) \mid \cdot)\|_{\mathcal{H}} = o_{\mathbb{P}}(n^{-1/4})$

(*iii*) There exist functions Ψ_1 and Ψ_2, such that

$$\sum_{k=0}^{1} \mathbb{E}^* \left[\psi_{1k}(Y,X) \int_{-\infty<u<Y} \psi_{2k}(u,X) d\Big((\widehat{H}_k - H_k)([u,\infty) \mid X) \Big) \right]$$
$$= \frac{1}{n} \sum_{i=1}^{n} \Psi_1(Y_i, \delta_i, X_i) + R_{1n},$$

and

$$\sum_{k,\ell=0}^{1} \mathbb{E}^* \left[\psi_{3k}(Y,X) \int_{-\infty<u<Y} \psi_{4k}(u,X) \psi_{5k}\Big((\widehat{H}_k - H_k)([u,\infty) \mid X), X \Big) dH_\ell(u \mid X) \right]$$
$$= \frac{1}{n} \sum_{i=1}^{n} \Psi_2(Y_i, \delta_i, X_i) + R_{2n},$$

where $\mathbb{E}^*$ denotes conditional expectation given the data (Y_i, δ_i, X_i), $1 \le i \le n$, the functions ψ_{jk} are defined in (7.13) in the Appendix, and where

$$\mathbb{E}[\Psi_\ell(Y,\delta,X)] = 0, \quad R_{\ell n} = o_{\mathbb{P}}(n^{-1/2})$$

$(j = 1, \dots, 5; k = 0, 1; \ell = 1, 2)$. Note that the above expectations are conditionally on the sample and are taken with respect to the generic variables Y, δ, X which have the same law as the sample.

(AN5) The class $\mathcal{H}$ satisfies $\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{H}, \|\cdot\|_{\mathcal{H}})} d\varepsilon < \infty$, where $N(\varepsilon, \mathcal{H}, \|\cdot\|_{\mathcal{H}})$ is the ε-covering number of the space $\mathcal{H}$ with respect to the norm $\|\cdot\|_{\mathcal{H}}$, i.e. the smallest number of balls of $\|\cdot\|_{\mathcal{H}}$-radius ε needed to cover the space $\mathcal{H}$.

Theorem 4.2. *Assume that $\widehat{\beta} - \beta_0 = o_{\mathbb{P}}(1)$ and that (AN1)-(AN5) and (2.1), (2.2), (2.6), (2.14), (4.1) and (4.2) hold true. Then,*

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

where

$$\Omega = \big\{\nabla_\beta M(\beta_0, \eta_0)\big\}^{-1} V \big\{\nabla_\beta M(\beta_0, \eta_0)\big\}^{-1}$$

and $V = Var\big(m(Y,\delta,X;\beta_0,\eta_0) + \Psi_1(Y,\delta,X) + \Psi_2(Y,\delta,X)\big)$.

4.3 Bootstrap consistency

Although in principle one can use Theorem 4.2 above for making inference, the asymptotic variance Ω has a complicated structure, and the estimation of Ω would not only be cumbersome, but its precision for small samples could moreover be rather poor. We continue this section by showing that a bootstrap procedure can be used to estimate the asymptotic variance of $\widehat{\beta}$, to approximate the whole distribution of $\widehat{\beta}$ or to construct confidence intervals or test hypotheses regarding β_0.

Here, we propose to use a naive bootstrap procedure, consisting in drawing triplets $(Y_i^*, \delta_i^*, X_i^*)$, $1 \leq i \leq n$, randomly with replacement from the data (Y_i, δ_i, X_i), $1 \leq i \leq n$. Let $\widehat{H}_k^*$ be the same estimator as $\widehat{H}_k$ $(k = 0, 1)$ but based on the bootstrap data, and for each (β, η) let $M_n^*(\beta, \eta) = n^{-1} \sum_{i=1}^n m(Y_i^*, \delta_i^*, X_i^*; \beta, \eta)$. Define the bootstrap estimator $\widehat{\beta}^*$ to be any sequence that satisfies

$$\|M_n^*(\widehat{\beta}^*, \widehat{\eta}^*) - M_n(\widehat{\beta}, \widehat{\eta})\| = \inf_{\beta \in B} \|M_n^*(\beta, \widehat{\eta}^*) - M_n(\widehat{\beta}, \widehat{\eta})\|,$$

where $\widehat{\eta}^*(t, x) = (\widehat{\eta}_1^*(t, x), \widehat{\eta}_2^*(t, x)) = (\widehat{H}_0^*([t, \infty) \mid x), \widehat{H}_1^*([t, \infty) \mid x))$.

The following result shows that the bootstrap works, in the sense that it allows to recover correctly the distribution of $n^{1/2}(\widehat{\beta} - \beta_0)$.

Theorem 4.3. *Assume that $\widehat{\beta} - \beta_0 = o_{\mathbb{P}}(1)$ and that (AN1)-(AN5) hold true. Moreover, assume that $\nabla_\beta M(\beta, \eta)$ is continuous in η (with respect to $\| \cdot \|_{\mathcal{H}}$) at $(\beta, \eta) = (\beta_0, \eta_0)$, and that (AN4) holds true with $\widehat{H}_k - H_k$ replaced by $\widehat{H}_k^* - \widehat{H}_k$ $(k = 0, 1)$ in $\mathbb{P}^*$-probability. Then,*

$$\sup_{u \in \mathbb{R}^p} \Big| \mathbb{P}^*\big(n^{1/2}(\widehat{\beta}^* - \widehat{\beta}) \leq u\big) - \mathbb{P}\big(n^{1/2}(\widehat{\beta} - \beta_0) \leq u\big) \Big| = o_{\mathbb{P}}(1),$$

where $\mathbb{P}^*$ denotes probability conditionally on the data (Y_i, δ_i, X_i), $i = 1, \ldots, n$, and where the inequality sign means the component-wise inequality for vectors.

4.4 Verification of the assumptions for kernel estimators

We finish this section with an illustration of the verification of the assumptions of our asymptotic results when the conditional subdistributions H_k are estimated by means of kernel smoothing.

Consider the case where X is composed of continuous and discrete components, that is $X = (X_c, X_d) \in \mathcal{X}_c \times \mathcal{X}_d \subset \mathbb{R}^{d_c} \times \mathbb{R}^{d_d}$, with $d_c + d_d = d \geq 1$. For simplicity, assume that the support $\mathcal{X}_d$ of the discrete subvector X_d is finite. We also assume that the life time T has not been transformed by a logarithmic or other transformation, so that its support is $[0, \infty]$. The subdistributions $H_k([t, \infty) \mid x)$ could be estimated by means of a kernel estimator :

$$\widehat{H}_k([t, \infty) \mid x) = \sum_{i=1}^n \frac{\widetilde{K}_{h_n}(X_i - x)}{\sum_{j=1}^n \widetilde{K}_{h_n}(X_j - x)} I(Y_i \geq t, \delta_i = k),$$

where for any $(x_c, x_d) \in \mathcal{X}_c \times \mathcal{X}_d$,

$$\widetilde{K}_{h_n}(X_i - x) = K_{h_n}(X_{c,i} - x_c) I(X_{d,i} = x_d),$$

h_n is a bandwidth sequence, $K_h(\cdot) = K(\cdot/h)/h^{d_c}$, $K(u) = k(u_1) \cdot \ldots \cdot k(u_{d_c})$ and k is a probability density function.

Nonparametric smoothing of continuous covariates is possible for dimensions d_c larger than 1. However, the technical arguments necessary to verify the assumptions used for the asymptotic results are tedious. Therefore, in the following we consider $d_c = 1$. The discrete covariates do not contribute to the curse of dimensionality, and therefore d_d could be larger than 1. However, for simplicity, below we do not consider discrete covariates.

To satisfy assumption (AN4), we need to impose the following conditions :

(C1) The sequence h_n satisfies $nh_n^4 \to 0$ and $nh_n^{3+\zeta}(\log n)^{-1} \to \infty$ for some $\zeta > 0$.

(C2) The support $\mathcal{X}$ of X is a compact subset of $\mathbb{R}$.

(C3) The probability density function K has compact support, $\int uK(u)du = 0$ and K is twice continuously differentiable.

Further, let $\mathcal{F}_1$ be the space of functions from $[0, \tau]$ to $[0, 1]$ with variation bounded by M, and let $\mathcal{F}_2$ be the space of continuously differentiable functions f from $\mathcal{X}$ to $[-M, M]$ that satisfy $\sup_{x\in\mathcal{X}} |f'(x)| \le M$ and $\sup_{x_1,x_2\in\mathcal{X}} |f'(x_1) - f'(x_2)|/|x_1 - x_2|^\epsilon \le M$ for some $M < \infty$ and $0 < \epsilon < 1$. Let

$$\begin{aligned}\mathcal{H} \;&=\; \Big\{(t, x) \to \eta(t, x) : \eta(\cdot, x) \in \mathcal{F}_1,\, \frac{\partial}{\partial x}\eta(\cdot, x) \in \mathcal{F}_1 \text{ for all } x \in \mathcal{X}, \\ &\qquad\qquad\qquad \text{and } \eta(t, \cdot) \in \mathcal{F}_2 \text{ for all } 0 \le t \le \tau\Big\}.\end{aligned}$$

We define the following norm associated with the space $\mathcal{H}$: for $\eta \in \mathcal{H}$, let

$$\|\eta\|_{\mathcal{H}} = \sup_{0\le t\le\tau} \sup_{x\in\mathcal{X}} |\eta(t, x)|.$$

Then, it follows from Propositions 1 and 2 in Akritas and Van Keilegom (2001) that $\mathbb{P}(\widehat{H}_k \in \mathcal{H}) \to 1$ provided $nh_n^{3+\zeta}(\log n)^{-1} \to \infty$, with $\zeta > 0$ as in condition (C1). Moreover, $\sup_{x,t} |\widehat{H}_k([t, \infty) \mid x) - H_k([t, \infty) \mid x)| = O_{\mathbb{P}}((nh_n)^{-1/2}(\log n)^{1/2}) = o_{\mathbb{P}}(n^{-1/4})$ (see Proposition 1 in Akritas and Van Keilegom 2001). The class $\mathcal{H}$ satisfies assumption (AN5) thanks to Lemma 6.1 in Lopez (2011). It remains to show the validity of assumption (AN4)(*iii*). We will show the first statement, the second one can be shown in a similar way. Note that the left hand side equals

$$\sum_{k=0}^{1} \mathbb{E}\Big[\psi_{1k}(Y,X)\int_{0<u<Y}\psi_{2k}(u,X)\frac{1}{nh}f_X^{-1}(X)\sum_{i=1}^{n}K_h(X_i-X)\,d\Big(I(Y_i\geq u,\delta_i=k)$$

$$-H_k([u,\infty)\mid X)\Big)\Big|Y_i,\delta_i,X_i\Big]+o_{\mathbb{P}}(n^{-1/2})$$

$$=\sum_{k=0}^{1}\mathbb{E}\Big[\psi_{1k}(Y,X)\frac{1}{nh}f_X^{-1}(X)\sum_{i=1}^{n}K_h(X_i-X)\Big\{-\psi_{2k}(Y_i,X)I(Y_i\leq Y,\delta_i=k)$$

$$-\int_{0<u<Y}\psi_{2k}(u,X)\,dH_k([u,\infty)\mid X)\Big\}\Big|Y_i,\delta_i,X_i\Big]+o_{\mathbb{P}}(n^{-1/2})$$

$$=-\sum_{k=0}^{1}n^{-1}\sum_{i=1}^{n}\mathbb{E}\Big[\psi_{1k}(Y,X_i)\Big\{\psi_{2k}(Y_i,X_i)I(Y_i\leq Y,\delta_i=k)$$

$$-\int_{0<u<Y}\psi_{2k}(u,X_i)\,dH_k((-\infty,u]\mid X_i)\Big\}\Big|Y_i,\delta_i,X_i,X=X_i\Big]+o_{\mathbb{P}}(n^{-1/2}+h^2),$$

which is of the required form.

5 Simulations

In this section we will investigate the small sample performance of our estimation method. We consider the following model. The covariate X is generated from a uniform distribution on $[-1,1]$, and the conditional probability $\phi(x,\beta)$ of not being cured follows a logistic model :

$$\phi(x,\beta)=\frac{\exp(\beta_1+\beta_2 x)}{1+\exp(\beta_1+\beta_2 x)},$$

for any $-1\leq x\leq 1$. We will work with $\beta_0=(\beta_{01},\beta_{02})=(1.75,2)$ and $(1.1,2)$, corresponding to an average cure rate of 20% respectively 30%. The conditional distribution function $F_{T,0}(\cdot|x)$ of the uncured individuals is constructed as follows. For a given X, we draw T from an exponential distribution with mean equal to $\exp[-(\gamma_0+\gamma_1 x+\gamma_2/(1+2x^2))]$, where $\gamma_0=\gamma_1=0.5$ and $\gamma_2\in\{0,1,2\}$. Next, in order to respect condition (4.2), we truncate this distribution at τ, which is the quantile of order 0.97 of an exponential distribution with mean $\mathbb{E}\{\exp[-(\gamma_0+\gamma_1 X+\gamma_2/(1+2X^2))]\}$, i.e.

$$F_{T,0}([0,t]|x)=1-\exp[-\exp(\gamma_0+\gamma_1 x+\gamma_2/(1+2x^2))t]I(0\leq t\leq\tau).$$

Note that this is the distribution function corresponding to a Cox model with baseline hazard equal to $I(0\leq t\leq\tau)$, and exponential factor equal to $\exp(\gamma_0+\gamma_1 x+\gamma_2/(1+2x^2))$.

Next, we generate the censoring variable C independently of X from an exponential distribution with mean equal to 1.65 when $\beta_0=(1.75,2)$, and with mean 1.45 when $\beta_0=(1.1,2)$. In this way we have respectively 40% and 50% of censoring when $\gamma_2=0$.

In what follows we will compare our estimator of β with the estimator proposed by Lu (2008) which assumes a Cox model for the uncured individuals. The exponential factor in the Cox model is assumed to be linear in the covariate X, and hence the Cox model will only be verified when $\gamma_2 = 0$. The estimated β coefficients under the Cox model are obtained using the `R` package `smcure`.

For our estimation procedure we used the kernel estimators given in Section 4.4, and we programmed $\widehat{\beta}$ using the optimization procedure `optim` in `R`. As starting values we used the estimator obtained from a logistic model based on the censoring indicator (as a surrogate for the unobserved cure indicator). However, due to the non-concavity of our likelihood function and due to the inconsistency of this vector of starting values, the procedure `optim` often ends up in a local maximum instead of the global maximum. To circumvent this problem, we added the following intermediate step to the estimation procedure. Based on the initial starting values, we estimate β from a logistic model based on the nonparametric estimator $\widehat{F}_T([0,\infty)|x)$, so we maximize the log-likelihood $\sum_{i=1}^n \{(1 - \widehat{F}_T([0,\infty)|X_i)) \log(\phi(X_i,\beta)) + \widehat{F}_T([0,\infty)|X_i) \log(1 - \phi(X_i,\beta))\}$. Since this log-likelihood is concave it has a unique local and global maximum, expected to be close to the maximizer of our likelihood. We now use this intermediate estimate as starting value for our likelihood maximization.

The results of this two-step maximization procedure are given in Table 1 for the case where $\beta_0 = (1.75, 2)$, and in Table 2 for the case where $\beta_0 = (1.1, 2)$. A total of 500 samples of size $n = 150$ and $n = 300$ are generated, and the tables show the bias and mean squared error (MSE) of the estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ obtained under the Cox model and from our procedure. The kernel function K is taken equal to the Epanechnikov kernel : $K(u) = (3/4)(1-u^2)I(|u| \leq 1)$. The bandwidth h of the kernel estimators $\widehat{H}_k(\cdot|X_i)$ $(k = 0, 1)$ is taken proportional to $n^{-2/7}$ so as to verify regularity condition (C1), i.e. $h = cn^{-2/7}$ for several values of c, namely $c = 2, 3$ and 4. In addition, we also used the cross-validation (CV) procedure proposed by Li, Lin and Racine (2013) for kernel estimators of conditional distribution functions. The CV procedure is implemented in the package `np` in `R`. For each sample in our simulation, we calculated these bandwidths for $\widehat{H}_0$ and $\widehat{H}_1$ and used the average of these two bandwidths.

The tables show that our estimator outperforms the one that is based on the Cox model, even when the Cox model is correct. They also show that our estimator is only mildly sensitive to the bandwidth, which could be explained by the fact that we average out the effect of the bandwidth. We also see that the CV selection of the bandwidth is working rather well, in the sense that the MSE is close to the smallest value among the MSE's corresponding to the three fixed bandwidths.

Next, we look at the estimation of the quartiles of the distribution $F_{T,0}(\cdot|x)$ when $x = 0.25$. We estimate these quartiles by means of our nonparametric estimator $\widehat{F}^{\widehat{\beta}}_{T,0}(\cdot|x)$ and by means of the Cox model studied in Lu (2008). The results given in Tables 3 and 4 show that, as could be expected, when the Cox model is not satisfied (i.e. when $\gamma_2 = 1$ or 2), the

			$c = 2$		$c = 3$		$c = 4$		h_{CV}		Cox	
n	γ_2	Par.	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
150	0	β_1	.092	.230	.012	.183	-.075	.162	.044	.224	.216	.523
		β_2	.002	.540	-.231	.438	-.506	.510	-.158	.536	.291	1.15
	1	β_1	.063	.123	-.021	.101	-.094	.094	.017	.120	.147	.191
		β_2	-.099	.340	-.334	.331	-.605	.505	-.246	.397	.278	.635
	2	β_1	.045	.100	-.029	.086	-.099	.083	-.005	.101	.124	.145
		β_2	-.109	.242	-.356	.277	-.632	.497	-.302	.382	.263	.476
300	0	β_1	.021	.100	-.029	.088	-.081	.081	.004	.099	.099	.139
		β_2	-.081	.266	-.252	.268	-.461	.363	-.148	.288	.135	.365
	1	β_1	-.004	.060	-.048	.055	-.097	.055	-.013	.061	.097	.092
		β_2	-.107	.181	-.278	.208	-.482	.328	-.150	.201	.215	.302
	2	β_1	-.015	.050	-.059	.046	-.107	.049	-.030	.052	.077	.074
		β_2	-.124	.157	-.295	.198	-.498	.329	-.197	.217	.181	.247

Table 1: *Bias and MSE of* $\widehat{\beta}_1$ *and* $\widehat{\beta}_2$ *for two sample sizes, three values of* γ_2*, three bandwidths of the form* $h = cn^{-2/7}$ *and the bandwidth* h_{CV} *obtained from cross-validation. Here,* $P(cured) = 0.2$ *and* $P(censoring) = 0.4$ *for* $\gamma_2 = 0$*. The Cox model is satisfied for* $\gamma_2 = 0$*.*

			$c = 2$		$c = 3$		$c = 4$		h_{CV}		Cox	
n	γ_2	Par.	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
150	0	β_1	.082	.139	.003	.116	-.060	.105	.031	.130	.116	.189
		β_2	-.017	.418	-.244	.365	-.525	.477	-.217	.419	.215	.618
	1	β_1	.057	.076	-.011	.063	-.068	.059	.010	.071	.086	.099
		β_2	-.107	.253	-.328	.277	-.608	.478	-.294	.352	.227	.434
	2	β_1	.049	.065	-.016	.056	-.071	.054	.006	.066	.065	.078
		β_2	-.129	.202	-.361	.260	-.640	.493	-.329	.358	.196	.329
300	0	β_1	.059	.074	-.010	.060	-.058	.056	.031	.073	.060	.083
		β_2	-.068	.216	-.237	.226	-.453	.326	-.153	.241	.103	.257
	1	β_1	-.013	.036	-.043	.034	-.074	.035	-.024	.035	.050	.047
		β_2	-.117	.135	-.282	.175	-.486	.306	-.187	.160	.159	.197
	2	β_1	-.049	.028	-.080	.030	-.037	.030	-.034	.030	.035	.037
		β_2	-.295	.168	-.496	.306	-.244	.194	-.217	.182	.128	.156

Table 2: *Bias and MSE of* $\widehat{\beta}_1$ *and* $\widehat{\beta}_2$ *for two sample sizes, three values of* γ_2*, three bandwidths of the form* $h = cn^{-2/7}$ *and the bandwidth* h_{CV} *obtained from cross-validation. Here,* $P(cured) = 0.3$ *and* $P(censoring) = 0.5$ *for* $\gamma_2 = 0$*. The Cox model is satisfied for* $\gamma_2 = 0$*.*

MSE of the quartiles obtained under the Cox model is much higher than the corresponding MSE obtained from our procedure. This shows the importance of having a model that does

			$c = 2$		$c = 3$		$c = 4$		h_{CV}		Cox	
n	γ_2	p	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
150	0	.25	.044	.061	.025	.035	.030	.032	.033	.049	.031	.027
		.50	.022	.049	.003	.031	.003	.025	.011	.043	.028	.023
		.75	.024	.083	.006	.055	.001	.045	.011	.071	.032	.038
	1	.25	.083	.053	.102	.039	.155	.053	.096	.049	.267	.104
		.50	.060	.051	.092	.041	.144	.049	.073	.048	.251	.093
		.75	.072	.089	.114	.075	.154	.077	.091	.085	.254	.117
	2	.25	.126	.060	.218	.082	.325	.139	.202	.085	.592	.401
		.50	.098	.058	.189	.073	.291	.121	.172	.081	.513	.308
		.75	.112	.107	.210	.120	.296	.156	.187	.124	.492	.322
300	0	.25	.007	.032	-.007	.023	-.018	.018	.002	.030	.010	.013
		.50	-.001	.034	-.013	.023	-.026	.018	-.006	.031	.008	.013
		.75	-.003	.053	-.019	.034	-.030	.026	-.009	.046	.010	.023
	1	.25	.027	.026	.049	.021	.081	.021	.033	.026	.252	.081
		.50	.028	.031	.054	.025	.086	.025	.034	.030	.240	.076
		.75	.031	.055	.053	.040	.086	.038	.033	.052	.235	.086
	2	.25	.063	.031	.129	.037	.216	.065	.096	.039	.580	.366
		.50	.055	.033	.121	.040	.200	.062	.086	.038	.498	.274
		.75	.055	.058	.109	.056	.191	.074	.084	.061	.452	.248

Table 3: *Bias and MSE of the conditional quantiles of order* $p = 0.25, 0.50$ *and* 0.75 *at* $x = 0.25$ *for two sample sizes, three bandwidths of the form* $h = cn^{-2/7}$ *and the bandwidth* h_{CV} *obtained from cross-validation. Here,* $P(cured) = 0.2$ *and* $P(censoring) = 0.4$ *for* $\gamma_2 = 0$. *The Cox model is satisfied for* $\gamma_2 = 0$.

not impose any assumptions on the distribution of the uncured individuals and which still provides very accurate estimators for the logistic part of the model.

We also verify how close the distributions of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are to a normal distribution. We know thanks to Theorem 4.2 that the estimators converge to a normal limit when n tends to infinity. Figure 1 shows that for $n = 150$ the distribution is rather close to a normal limit, especially for $\widehat{\beta}_1$. The figure is based on 1000 samples generated from the above model with $P(cured) = 0.2$ and $P(censoring) = 0.4$. The results for $n = 300$ (not shown here for space constraints) are close to a straight line, showing that the results improve when n increases.

Finally, we verify the accuracy of the naive bootstrap proposed in Section 4.3. We consider the above model, but restrict attention to $n = 150$ and to the case where $P(cured) = 0.2$ and $P(censoring) = 0.4$. Figure 2 shows boxplots of the variance of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ obtained from 250 bootstrap resamples for each of 500 samples. The bandwidth is $h = 3n^{-2/7}$. The empirical variance of the 500 estimators of β_1 and β_2 is also added, and shows that the bootstrap variance is well centered around the corresponding empirical variance.

			$c = 2$		$c = 3$		$c = 4$		h_{CV}		Cox	
n	γ_2	p	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
150	0	.25	.074	.095	.020	.049	.021	.044	.046	.071	.039	.038
		.50	.049	.069	.005	.041	-.003	.031	.023	.057	.031	.029
		.75	.042	.098	.007	.063	-.005	.050	.024	.084	.034	.046
	1	.25	.107	.072	.116	.053	.151	.059	.119	.066	.274	.116
		.50	.076	.060	.103	.049	.142	.054	.088	.057	.255	.100
		.75	.099	.113	.118	.083	.154	.086	.109	.102	.257	.126
	2	.25	.159	.085	.232	.097	.322	.144	.233	.109	.586	.402
		.50	.120	.066	.211	.089	.296	.131	.195	.092	.508	.307
		.75	.143	.135	.228	.142	.308	.178	.219	.153	.489	.325
300	0	.25	.032	.042	-.005	.029	-.029	.023	.016	.038	.018	.017
		.50	.018	.039	-.011	.026	-.035	.021	.007	.034	.013	.016
		.75	.015	.060	-.018	.037	-.037	.028	-.001	.054	.016	.022
	1	.25	.026	.033	.046	.025	.071	.023	.034	.031	.252	.083
		.50	.030	.036	.052	.027	.080	.026	.037	.034	.234	.074
		.75	.033	.060	.050	.042	.081	.040	.038	.055	.232	.087
	2	.25	.070	.040	.132	.042	.212	.064	.103	.046	.570	.357
		.50	.060	.038	.125	.044	.201	.064	.089	.043	.492	.269
		.75	.062	.066	.115	.061	.193	.078	.090	.070	.451	.251

Table 4: *Bias and MSE of the conditional quantiles of order* $p = 0.25, 0.50$ *and* 0.75 *at* $x = 0.25$ *for two sample sizes, three bandwidths of the form* $h = cn^{-2/7}$ *and the bandwidth* h_{CV} *obtained from cross-validation. Here,* $P(\mathit{cured}) = 0.3$ *and* $P(\mathit{censoring}) = 0.5$ *for* $\gamma_2 = 0$*. The Cox model is satisfied for* $\gamma_2 = 0$.

6 Data analysis

Let us now apply our estimation procedure on two medical data sets. The first one is about 286 breast cancer patients with lymph-node-negative breast cancer treated between 1980 and 1995 (Wang *et al.* (2005)). The event of interest is distant-metastasis, and the associated survival time is the distant metastasis-free survival time (defined as the time to first distant progression or death, whichever comes first). 107 of the 286 patients experience a relapse from breast cancer. The plot of the Kaplan-Meier estimator of the data is given in Figure 3(a) and shows a large plateau at about 0.60. Furthermore, a large proportion of the censored observations is in the plateau, which suggests that a cure model is appropriate for these data. As a covariate we use the age of the patients, which ranges from 26 to 83 years and the average age is about 54 years.

We estimate β using our estimator and using the estimator based on the Cox model. The bandwidth h is selected using cross-validation, as in the simulation section. The estimated

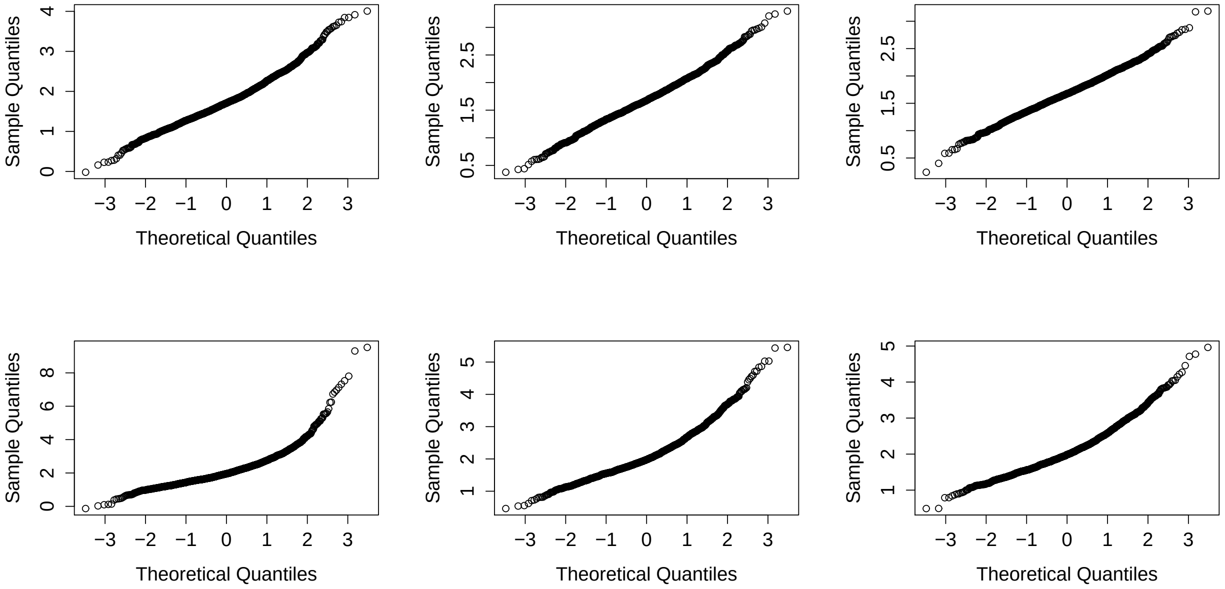

Figure 1: *QQ-plots of $\widehat{\beta}_1$ (first row) and $\widehat{\beta}_2$ (second row) for 1000 samples of size $n = 150$. The first column corresponds to $\gamma_2 = 0$, the second to $\gamma_2 = 1$ and the third to $\gamma_2 = 2$. The bandwidth is $h = 3n^{-2/7}$.*

intercept is -0.224 (with standard deviation equal to 0.447 obtained using a naive bootstrap procedure), and the estimated slope parameter is -0.005 (with standard deviation equal to 0.008). Under the Cox model the estimated intercept and slope are respectively 0.063 and -0.010. A 95% confidence interval is given by $(-1.100, 0.653)$ for the intercept and $(-0.021, 0.011)$ for the slope, where the variance is again based on the naive bootstrap procedure. The graph of the two estimators of the function $\phi(x)$ is given in Figure 3(b). The estimated coefficients and curves are quite close to each other, suggesting that the Cox model might be valid. This is also confirmed by Figure 3(c)-(d), which shows the estimation of the survival function $1 - F_{T,0}(\cdot|x)$ of the uncured patients for $x = 48$ and $x = 60$ based on our estimation procedure and the procedure based on the Cox model. The figure shows that the two estimators are close for both values of x.

Next, we analyse data provided by the Medical Birth Registry of Norway (see http://folk.uio.no/borgan/abg-2008/data/data.html). The data set contains information on births in Norway since 1967, related to a total of 53,558 women. We are interested in the time between the birth of the first and the second child, for those mothers whose first child died within the first year ($n = 262$). The covariate of interest is age (X), which is the age of the mother at the birth of the first child. The age ranges from 16.8 to 29.8 years, with an average of 23.2 years. The cure rate is the fraction of women who gave birth only once.

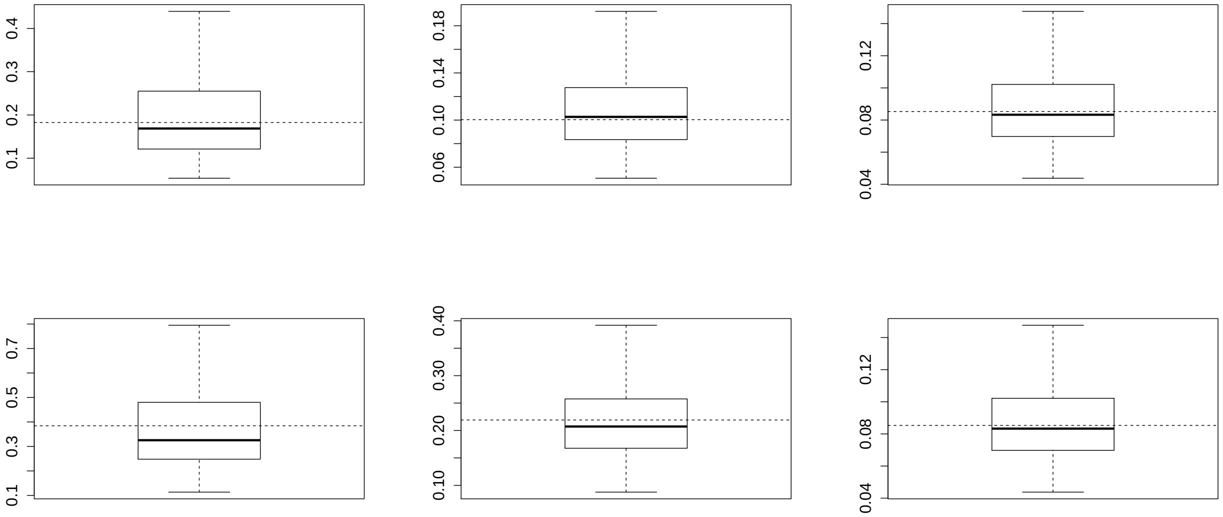

Figure 2: *Boxplots of the variance of* $\widehat{\beta}_1$ *(first row) and* $\widehat{\beta}_2$ *(second row) obtained from 250 bootstrap resamples for each of 500 samples of size* $n = 150$*. The first column corresponds to* $\gamma_2 = 0$*, the second to* $\gamma_2 = 1$ *and the third to* $\gamma_2 = 2$*. The bandwidth is* $h = 3n^{-2/7}$*. The empirical variance of the 500 estimators of* β_1 *and* β_2 *is also added (dashed line).*

Figure 4(a) shows the Kaplan-Meier estimator, and suggests that a cure fraction is present.

As we did for the first data set, we analyse these data using the approach proposed in this paper, and also using the Cox mixture cure model. The estimated intercept equals 1.952 using our model and 0.034 using the Cox model. The bootstrap confidence interval for the intercept is $(-1.577, 5.481)$ (the estimated standard deviation equals 1.801). The estimated slope equals -0.041 respectively 0.052 using the two models. For our estimation procedure the confidence interval is given by $(-0.193, 0.111)$ (with estimated standard deviation equal to 0.078). Figure 4(b) shows that the two estimators of the function $\phi(x)$ are quite different, and have opposite slopes. Moreover, the survival function $1-F_{T,0}(\cdot|x)$ of the uncured patients is given in Figure 4(c)-(d) for $x = 21$ and $x = 25$. We see that the estimator based on the Cox model is quite different from ours, suggesting that the Cox model might not be valid for these data, although a formal test would need to confirm this. This is however beyond the scope of this paper. Also note that the estimator of the cure proportion $1-\phi(x)$ is increasing under our model and decreasing under the Cox model. It seems however natural to believe that the probability of having no second child (so the cure proportion) is increasing with age, which is again an indication that the Cox model is not valid for these data.

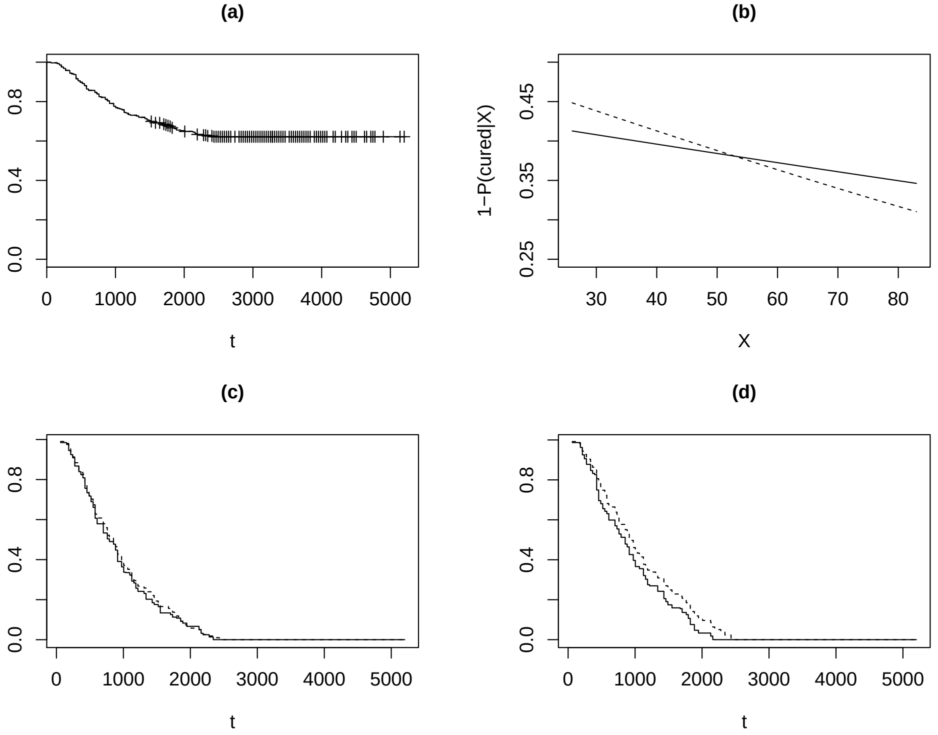

Figure 3: *Analysis of the breast cancer data : (a) Kaplan-Meier estimator; (b) Graph of the proposed estimator of $\phi(x)$ (solid curve) and of the estimator based on the Cox model (dashed curve); (c) Estimation of $1 - F_{T,0}(\cdot|x)$ using the proposed estimator (solid curve) and using the estimator based on the Cox model (dashed curve) when $x = 48$; (d) Idem when $x = 60$.*

7 Appendix : Proofs

Proof of Proposition 3.1. By the properties of the likelihood of a Bernoulli random variable given that $Y \in dt$ and $X = x$, we have

$$\mathbb{E}\left[\delta \log \frac{\phi(x,\beta) F_{T,0}^{\beta}(dt \mid x) F_C((t,\infty) \mid x)}{H_1(dt \mid x)} \right.$$
$$\left. + (1-\delta) \log \frac{F_C(dt \mid x)\left[\phi(x,\beta) F_{T,0}^{\beta}((t,\infty) \mid x) + 1 - \phi(x,\beta)\right]}{H_0(dt \mid x)} \;\middle|\; Y \in dt, X = x \right] \leq 0.$$

Integrate with respect to Y and X and deduce that

$$\mathbb{E}\left[\log p(Y,\delta,X;\beta)\right] \leq \mathbb{E}\left[\log p(Y,\delta,X;\beta_0)\right].$$

If there exists some $\beta \neq \beta_0$ such that the last inequality becomes an equality, then necessarily $p_1(Y, X; \beta) = 1$ almost surely. Then

$$F_C((t,\infty) \mid x)\phi(x,\beta)F^{\beta}_{T,0}(dt \mid x) = F_C((t,\infty) \mid x)\phi(x,\beta_0)F^{\beta_0}_{T,0}(dt \mid x), \ \ \forall -\infty < t \leq \tau_H(x),$$

for almost all $x \in \mathcal{X}$. By Theorem 2.1, we deduce that necessarily $\beta = \beta_0$. □

Lemma 7.1. *Let conditions (4.2), (AC1) and (AC4) hold true. Then,*

$$\sup_{x\in\mathcal{X}} \sup_{t\in(-\infty,\tau]} |\widehat{F}_C([t,\infty) \mid x) - F_C([t,\infty) \mid x)| = o_{\mathbb{P}}(1),$$

$$\sup_{\beta\in B} \sup_{x\in\mathcal{X}} \sup_{t\in(-\infty,\tau]} |T_1(\beta, \widehat{H}_0, \widehat{H}_1)(t,x) - T_1(\beta, H_0, H_1)(t,x)| = o_{\mathbb{P}}(1),$$

where

$$T_1(\beta, H_0, H_1)(t,x) = H([t,\infty) \mid x) - (1-\phi(x,\beta))F_C([t,\infty) \mid x),$$

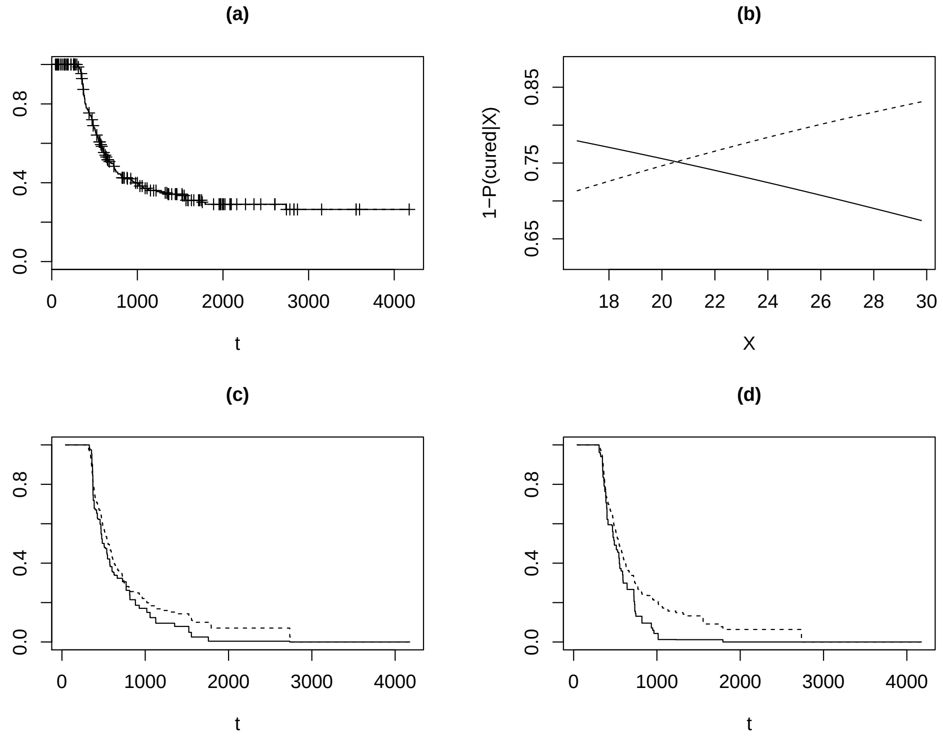

Figure 4: *Analysis of the second birth data : (a) Kaplan-Meier estimator; (b) Graph of the proposed estimator of $\phi(x)$ (solid curve) and of the estimator based on the Cox model (dashed curve); (c) Estimation of $1 - F_{T,0}(\cdot|x)$ using the proposed estimator (solid curve) and using the estimator based on the Cox model (dashed curve) when $x = 21$; (d) Idem when $x = 25$.*

and $T_1(\beta, \widehat{H}_0, \widehat{H}_1)$ *is defined similarly, but with* H_0, H_1 *and* F_C *replaced by* $\widehat{H}_0$, $\widehat{H}_1$ *and* $\widehat{F}_C$, *respectively. Moreover,*

$$\sup_{\beta\in B}\sup_{x\in\mathcal{X}}\sup_{t\in(-\infty,\tau]}|\widehat{F}^{\beta}_{T,0}([t,\infty)\mid x)-F^{\beta}_{T,0}([t,\infty)\mid x)|=o_{\mathbb{P}}(1).$$

Proof of Lemma 7.1. Let us first investigate the uniform convergence of the estimated cumulative hazard measure $\widehat{\Lambda}_C(\cdot\mid x)$. For any $t\in(-\infty,\tau]$ let us write

$$\begin{aligned}\widehat{\Lambda}_C((-\infty,t]\mid x)-\Lambda_C((-\infty,t]\mid x)&=\int_{(-\infty,t]}\frac{\widehat{H}_0(ds\mid x)}{\widehat{H}([s,\infty)\mid x)}-\int_{(-\infty,t]}\frac{H_0(ds\mid x)}{H([s,\infty)\mid x)}\\=\int_{(-\infty,t]}\left[\frac{1}{\widehat{H}([s,\infty)\mid x)}-\frac{1}{H([s,\infty)\mid x)}\right]\widehat{H}_0(ds\mid x)&+\int_{(-\infty,t]}\frac{\widehat{H}_0(ds\mid x)-H_0(ds\mid x)}{H([s,\infty)\mid x)}.\end{aligned}$$

The integrals with respect to $\widehat{H}_0(ds\mid x)$ are well defined, since, with probability tending to 1, for each $x\in\mathcal{X}$, the map $s\mapsto\widehat{H}_0((-\infty,s]\mid x)$, $s\in(-\infty,\tau]$, is a function of bounded variation. The uniform convergence Assumption (AC1) implies that

$$\begin{aligned}&\sup_{x\in\mathcal{X}}\sup_{t\in(-\infty,\tau]}\left|\widehat{\Lambda}_C((-\infty,t]\mid x)-\Lambda_C((-\infty,t]\mid x)\right|\\&\quad\leq c\sup_{x\in\mathcal{X}}\sup_{t\in(-\infty,\tau]}\frac{\left\{\left|\widehat{H}_0([t,\infty)|x)-H_0([t,\infty)|x)\right|+\left|\widehat{H}([t,\infty)\mid x)-H([t,\infty)|x)\right|\right\}}{H([\tau,\infty)\,|x)^2},\end{aligned}$$

for some constant $c>0$. Next, by Duhamel's identity (see Gill and Johansen 1990),

$$\begin{aligned}\widehat{F}_C((t,\infty)\mid x)-F_C((t,\infty)\mid x)&=-F_C((t,\infty)\mid x)\\&\quad\times\int_{(-\infty,t]}\frac{\widehat{F}_C([s,\infty)\mid x)}{F_C((s,\infty)\mid x)}\left(\widehat{\Lambda}_C(ds\mid x)-\Lambda_C(ds\mid x)\right).\end{aligned}$$

Then, the uniform convergence of $\widehat{F}_C(\cdot\mid x)$ follows from the uniform convergence of $\widehat{\Lambda}_C(\cdot\mid x)$ and condition (4.2). The same type of arguments apply for $T_1(\beta,\widehat{H}_0,\widehat{H}_1)$, and hence we omit the details.

Next, since by conditions (4.2) and (AC4) we have

$$\inf_{\beta\in B}\inf_{x\in\mathcal{X}}\inf_{t\in(-\infty,\tau]}[H([t,\infty)\mid x)-(1-\phi(x,\beta))F_C([t,\infty)\mid x)]>0,$$

there exists some constant $c>0$ with the property that

$$\mathbb{P}\left(\inf_{\beta\in B}\inf_{x\in\mathcal{X}}\inf_{t\in(-\infty,\tau]}\left[\widehat{H}([t,\infty)\mid x)-(1-\phi(x,\beta))\widehat{F}_C([t,\infty)\mid x)\right]\geq c>0\right)\to 1.$$

Hence, the uniform convergence of $\widehat{F}^{\beta}_{T,0}(\cdot\mid x)$ follows. □

Proof of Theorem 4.1. Let us write

$$F_{T,0}^{\beta}(\{Y_i\} \mid X_i) = \Lambda_{T,0}^{\beta}(\{Y_i\} \mid X_i) F_{T,0}^{\beta}([Y_i, \infty) \mid X_i) = H_1(\{Y_i\} \mid X_i) \frac{F_{T,0}^{\beta}([Y_i, \infty) \mid X_i)}{T_1(\beta, H_0, H_1)(Y_i, X_i)},$$

where

$$T_1(\beta, H_0, H_1)(t, x) = H([t, \infty) \mid x) - (1 - \phi(x, \beta)) F_C([t, \infty) \mid x).$$

Moreover, let

$$q_i(\beta, H_0, H_1) = q(\beta, H_0, H_1)(Y_i, \delta_i, X_i),$$

where, for $t \in \mathbb{R},\ d \in \{0, 1\},\ x \in \mathcal{X}$,

$$\begin{aligned} q(\beta, H_0, H_1)(t, d, x) = d\{&\log \phi(x, \beta) + \log F_{T,0}^{\beta}([t, \infty) \mid x) - \log T_1(\beta, H_0, H_1)(t, x)\} \\ &+ (1 - d) \log\{\phi(x, \beta) F_{T,0}^{\beta}([t, \infty) \mid x) + 1 - \phi(x, \beta)\}. \end{aligned}$$

Let

$$Q_n(\beta, H_0, H_1) = \frac{1}{n} \sum_{i=1}^{n} q_i(\beta, H_0, H_1).$$

Similarly, let us consider $Q_n(\beta, \widehat{H}_0, \widehat{H}_1)$ that is defined as $Q_n(\beta, H_0, H_1)$, but with H_0, H_1, F_C and $F_{T,0}^{\beta}$ replaced by $\widehat{H}_0$, $\widehat{H}_1$, $\widehat{F}_C$ and $\widehat{F}_{T,0}^{\beta}$, respectively. Then the estimator $\widehat{\beta}$ in equation (3.1) becomes

$$\widehat{\beta} = \arg\max_{\beta \in B} Q_n(\beta, \widehat{H}_0, \widehat{H}_1).$$

The first step is to check that

$$\sup_{\beta \in B} \left| Q_n(\beta, \widehat{H}_0, \widehat{H}_1) - Q_n(\beta, H_0, H_1) \right| = o_{\mathbb{P}}(1). \tag{7.12}$$

This follows directly from Lemma 7.1. Next, given our assumptions, it is easy to check that for any $t \in (-\infty, \tau]$, $d \in \{0, 1\}$, $x \in \mathcal{X}$,

$$|q(\beta, H_0, H_1)(t, d, x) - q(\beta', H_0, H_1)(t, d, x)| \leq C \|\beta - \beta'\|^a, \quad \forall \beta, \beta' \in B,$$

with $a > 0$ from Assumption (AC3) and some constant C depending only on c_1 from Assumption (AC3) and the positive values $\inf_{x \in \mathcal{X}} H_1(\{\tau\} \mid x)$ and $\inf_{x \in \mathcal{X}} H_0((\tau, \infty) \mid x)$. It follows that the class $\{(t, d, x) \to q(\beta, H_0, H_1)(t, d, x) : \beta \in B\}$ is Glivenko-Cantelli. Hence,

$$\sup_{\beta \in B} |Q_n(\beta, H_0, H_1) - Q(\beta, H_0, H_1)| = o_{\mathbb{P}}(1),$$

where $Q = \mathbb{E}(Q_n)$. Finally, Proposition 3.1 guarantees that

$$\beta_0 = \arg\max_{\beta \in B} Q(\beta, H_0, H_1).$$

Gathering the facts, we deduce that $\widehat{\beta} - \beta_0 = o_{\mathbb{P}}(1)$. □

Proof of Theorem 4.2. We show the asymptotic normality of our estimator by verifying the high-level conditions in Theorem 2 in Chen *et al.* (2003). First of all, for the consistency we refer to Section 4.1, whereas conditions (2.1) and (2.2) in Chen *et al.* (2003) are satisfied by construction and thanks to assumption (AN1), respectively. Concerning (2.3), first note that the expression inside the expected value in $\nabla_\eta M(\beta,\eta_0)[\eta-\eta_0]$ is linear in $\nabla_\eta T_j(\beta,\eta_0)[\eta-\eta_0]$ ($j=1,2,3,4$). Hence, we will focus attention on the latter Gâteaux derivatives. First,

$$\nabla_\eta T_1(\beta,\eta_0)[\eta-\eta_0](t,x) = (\eta_1-\eta_{01}+\eta_2-\eta_{02})(t,x) - (1-\phi(x,\beta))\nabla_\eta T_2(\eta_0)[\eta-\eta_0](t,x).$$

Using Duhamel's formula (see Gill and Johansen 1990), we can write

$$\begin{aligned}\nabla_\eta T_2(\eta_0)[\eta-\eta_0](t,x) = -T_2(\eta_0)(t,x)\int_{-\infty<u<t} &\frac{1}{(\eta_{01}+\eta_{02})(u,x)-\eta_{01}(\{u\},x)}\\ &\times\Big\{(\eta_1-\eta_{01})(du,x) - \frac{\eta_{01}(du,x)(\eta_1+\eta_2-\eta_{01}-\eta_{02})(u,x)}{(\eta_{01}+\eta_{02})(u,x)}\Big\}.\end{aligned}$$

In a similar way, we find that

$$\begin{aligned}\nabla_\eta T_3(\beta,\eta_0)[\eta-\eta_0](t,x) &= -T_3(\beta,\eta_0)(t,x)\int_{-\infty<u<t}\frac{1}{T_1(\beta,\eta_0)(u,x)-\eta_{02}(\{u\},x)}\\ &\qquad\times\Big\{(\eta_2-\eta_{02})(du,x) - \frac{\eta_{02}(du,x)\big[T_1(\beta,\eta)-T_1(\beta,\eta_0)\big](u,x)}{T_1(\beta,\eta_0)(u,x)}\Big\}.\end{aligned}$$

Finally,

$$\begin{aligned}&\nabla_\eta T_4(\beta,\eta_0)[\eta-\eta_0](t,x)\\ &= -\nabla_\beta\phi(x,\beta)\int_{(-\infty,t)}\Bigg\{\frac{\nabla_\eta T_2(\eta_0)[\eta-\eta_0](s,x)\eta_{02}(ds,x)+T_2(\eta_0)(s,x)(\eta_2-\eta_{02})(ds,x)}{T_1(\beta,\eta_0)(s,x)\big[T_1(\beta,\eta_0)(s,x)-\eta_{02}(\{s\},x)\big]}\\ &\qquad-\frac{T_2(\eta_0)(s,x)\eta_{02}(ds,x)\nabla_\eta T_1(\beta,\eta_0)[\eta-\eta_0](s,x)}{\big[T_1(\beta,\eta_0)(s,x)\big]^2\big[T_1(\beta,\eta_0)(s,x)-\eta_{02}(\{s\},x)\big]}\\ &\qquad-\frac{T_2(\eta_0)(s,x)\eta_{02}(ds,x)\big[\nabla_\eta T_1(\beta,\eta_0)[\eta-\eta_0](s,x)-(\eta_2-\eta_{02})(\{s\},x)\big]}{T_1(\beta,\eta_0)(s,x)\big[T_1(\beta,\eta_0)(s,x)-\eta_{02}(\{s\},x)\big]^2}\Bigg\}.\end{aligned}$$

Note that all denominators in $\nabla_\eta T_j(\beta,\eta_0)[\eta-\eta_0](t,x)$ are bounded away from zero, thanks to (4.10) and (4.11). By tedious but rather elementary arguments, it follows from these formulae that

$$\|\nabla_\eta T_j(\beta,\eta_0)[\eta-\eta_0]-\nabla_\eta T_j(\beta_0,\eta_0)[\eta-\eta_0]\|_{\mathcal{H}} \le C\|\beta-\beta_0\|\|\eta-\eta_0\|_{\mathcal{H}},$$

for some constant C. Hence, it can be easily seen that $\nabla_\eta T_j(\beta,\eta_0)[\eta-\eta_0]$ satisfies the second property in assumption (2.3) in Chen *et al.* (2003), and hence the same holds true for $\nabla_\eta M(\beta,\eta_0)[\eta-\eta_0]$. Similarly, by decomposing $T_j(\beta,\eta)-T_j(\beta,\eta_0)-\nabla_\eta T_j(\beta,\eta_0)[\eta-\eta_0]$

using Taylor-type arguments (in η), the first property in assumption (2.3) is easily seen to hold true.

Next, conditions (2.4) and (2.6) are satisfied thanks to Assumption (AN4) and because it follows from the above calculations of $\nabla_\eta T_j(\beta, \eta_0)[\eta - \eta_0]$ $(j = 1, 2, 3, 4)$ that

$$\begin{aligned}\nabla_\eta M(\beta_0, \eta_0)[\eta - \eta_0] &= \sum_{k \in \{0,1\}} \mathbb{E}\left[\psi_{1k}(Y, X) \int_{-\infty<u<Y} \psi_{2k}(u, X) d\left((\eta_k - \eta_{0k})(u, X)\right)\right] \\ &+ \sum_{k,\ell \in \{0,1\}} \mathbb{E}\left[\psi_{3k}(Y, X) \int_{-\infty<u<Y} \psi_{4k}(u, X)\psi_{5k}\Big((\eta_k - \eta_{0k})(u, X)\Big) dH_\ell(u \mid X)\right] \quad (7.13)\end{aligned}$$

for certain measurable functions ψ_{jk} $(j = 1, \ldots, 5; k = 0, 1)$.

It remains to verify condition (2.5). Note that

$$|m(t, \delta, x; \beta_2, \eta_2) - m(t, \delta, x; \beta_1, \eta_1)| \leq C_1(t, \delta, x)\|\beta_2 - \beta_1\| + C_2(t, \delta, x)\|\eta_2 - \eta_1\|_{\mathcal{H}}$$

for some functions C_j satisfying $\mathbb{E}[C_j^2(Y, \delta, X)] < \infty$ $(j = 1, 2)$, and hence (2.5) follows from assumption (AN5) and Theorem 3 in Chen *et al.* (2003). This finishes the proof. □

Proof of Theorem 4.3. To prove this theorem we will check the conditions of Theorem B in Chen *et al.* (2003), which gives high level conditions under which the naive bootstrap is consistent. The only difference between their setting and our setting is that we are proving bootstrap consistency in $\mathbb{P}$-probability, whereas their result holds true a.s. $[\mathbb{P}]$. As a consequence, in their high level conditions we can replace all a.s. $[\mathbb{P}]$ statements by the corresponding statements in $\mathbb{P}$-probability.

First of all, it follows from assumption (AN1) that condition (2.2) in Chen *et al.* (2003) holds with η_0 replaced by any η in a neighborhood of η_0, and from the proof of Theorem 4.2 it follows that the same holds true for condition (2.3). Next, conditions (2.4B) and (2.6B) in Chen *et al.* follow from the fact that we assume that assumption (AN4) continues to hold true if we replace $\widehat{H}_k - H_k$ by $\widehat{H}_k^* - \widehat{H}_k$ $(k = 0, 1)$. It remains to verify condition (2.5'B) in Chen *et al.* This follows from Theorem 3 in Chen *et al.*, whose conditions have been verified already for our Theorem 4.2. □

References

[1] Akritas, M.G. & Van Keilegom, I. (2001). Nonparametric estimation of the residual distribution. *Scand. J. Statist.* **28**, 549–568.

[2] Amico, M., Legrand, C. & Van Keilegom, I. (2017). The single-index/Cox mixture cure model (submitted).

[3] Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Statist. Soc. - Series B* **11**, 15–53.

[4] Chen, X., Linton, O. & Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71**, 1591–1608.

[5] Fang, H.B., Li, G. & Sun, J. (2005). Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scand. J. Statist.* **32**, 59–75.

[6] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**, 1041–1046.

[7] Gill, R.D. (1994). *Lectures on survival analysis.* Lectures on probability theory: Ecole d'été de probabilités de Saint-Flour XXII. Lecture notes in mathematics 1581. Springer.

[8] Gill, R.D. & Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Ann. Statist.* **18**, 1501–1555.

[9] Kuk, A.Y.C. & Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**, 531–541.

[10] Li, Q., Lin, J. & Racine, J.S. (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *J. Buss. Econ. Statist.* **31**, 57–65.

[11] López-Cheda, A., Cao, R., Jácome M.A. & Van Keilegom, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Comput. Statist. Data Anal.* **105**, 144–165.

[12] Lopez, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications. *Commun. Stat. - Theory Meth.* **40**, 2639–2660.

[13] Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Ann. Inst. Stat. Math.* **60**, 545–574.

[14] Maller, R.A. & Zhou, S. (1996). *Survival Analysis with Long Term Survivors.* Wiley, New York.

[15] Meeker, W.Q. (1987). Limited failure population life tests: Application to integrated circuit reliability. *Technometrics* **29**, 51–65.

[16] Othus, M., Li, Y. & Tiwari, R.C. (2009). A class of semiparametric mixture cure survival models with dependent censoring. *J. Amer. Statist. Assoc.* **104**, 1241–1250.

[17] Peng, Y. & Taylor, J.M.G. (2014). Cure models. In: Klein, J., van Houwelingen, H., Ibrahim, J. G., and Scheike, T. H., editors, *Handbook of Survival Analysis*, Handbooks of Modern Statistical Methods series, chapter 6, pages 113-134. Chapman & Hall, Boca Raton, FL, USA.

[18] Schmidt, P. & Witte, A.D. (1989). Predicting criminal recidivism using split population survival time models. *J. Econometrics* **40**, 141–159.

[19] Sy, J.P. & Taylor, J.M.G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56**, 227–236.

[20] Taylor, J.M.G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics* **51**, 899–907.

[21] van der Vaart, A.D. (1998). *Asymptotic Statistics*. Cambridge University Press.

[22] Wang, Y., Klijn, J.G.M., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijet-van Gelder, M.E.M., Yu, J., Jatkoe, T., Berns, E.M.J.J., Atkins, D. & Foekens, J.A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* **365**, 671–679.

[23] Xu, J. & Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canad. J. Statist.* **42**, 1–17.