This is an ACCEPTED VERSION of the following published document:

Link to published version: https://doi.org/10.1016/j.csda.2016.08.002

General rights:

Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models

Ana López-Cheda[a,*], Ricardo Cao[a], María Amalia Jácome[a], Ingrid Van Keilegom[b]

[a]*Departamento de Matemáticas, Universidade da Coruña, Spain*
[b]*Institut de Statistique, Biostatistique et Sciences Actuarielles, Université catholique de Louvain, Belgium*

Abstract

A completely nonparametric method for the estimation of mixture cure models is proposed. A nonparametric estimator of the incidence is extensively studied and a nonparametric estimator of the latency is presented. These estimators, which are based on the Beran estimator of the conditional survival function, are proved to be the local maximum likelihood estimators. An iid representation is obtained for the nonparametric incidence estimator. As a consequence, an asymptotically optimal bandwidth is found. Moreover, a bootstrap bandwidth selection method for the nonparametric incidence estimator is proposed. The introduced nonparametric estimators are compared with existing semiparametric approaches in a simulation study, in which the performance of the bootstrap bandwidth selector is also assessed. Finally, the method is applied to a database of colorectal cancer from the University Hospital of A Coruña (CHUAC).

Keywords: Survival analysis, censored data, local maximum likelihood, kernel estimation

*Correspondence to: Departamento de Matemáticas, Facultade de Informática, Universidade da Coruña, A Coruña 15071, Spain. Tel.: +34 981 16 70 00 Ext. 1301; fax: +34 981 16 70 00

Email address: `ana.lopez.cheda@udc.es` (Ana López-Cheda)

1. Introduction

Thanks to the effectiveness of current cancer treatments, the proportion of patients who get cured (or who at least survive for a long time) is increasing over time. Therefore, data coming from cancer studies typically have heavy censoring at the end of the follow-up period, and a standard survival model is inappropriate. To accommodate for the cured or insusceptible proportion of subjects, a cure fraction can be explicitly incorporated into survival models and, as a consequence, cure models arise. These models allow to estimate the cured proportion (incidence) and also the probability of survival of the uncured patients up to a given time point (latency). Note that cure models should not be used indiscriminately (Farewell, 1986), there must be good empirical and biological evidence of an insusceptible population.

There are two main classes of cure models: mixture and non-mixture models. The first papers in non-mixture models were due to Haybittle (1959), Haybittle (1965). One category, belonging to this group, is the proportional hazards (PH) cure model, also known as the promotion time cure model, first proposed by Yakovlev and Tsodikov (1996). The unknown terms in this model can be estimated parametrically (Yakovlev et al., 1994; Chen et al., 1999, 2002) or semiparametrically (Tsodikov, 1998, 2003; Zeng et al., 2006). Moreover, Tsodikov (2001) proposed a nonparametric estimator of the incidence, but it cannot handle continuous covariates.

In this paper we consider a model which belongs to the other category of cure models, called two-component mixture cure models. The mixture cure model was proposed by Boag (1949) and it explicitly expresses the survival function in terms of a mixture of the survival of two types of patients: those who are cured and those who are not. An advantage of this model is that it allows the covariates to have different influence on cured and uncured patients. Maller and Zhou (1996) provided a detailed review of this model. In mixture cure models, the incidence is usually assumed to have a logistic form and the latency is usually estimated parametrically (Farewell, 1982, 1986; Cantor and

Shuster, 1992; Ghitany et al., 1994; Denham et al., 1996) or semiparametrically (Kuk and Chen, 1992; Yamaguchi, 1992; Peng et al., 1998; Peng and Dear, 2000; Sy and Taylor, 2000; Li and Taylor, 2002; Zhang and Peng, 2007). Some recent topics covered in the mixture cure models literature include multivariate
35 survival data (Yu and Peng, 2008), clustered survival data (Lai and Yau, 2010) and accelerated models (Zhang et al., 2013).

Due to the fact that the effects of the covariate on the cure rate cannot always be well approximated using parametric or semiparametric methods, a nonparametric approach is needed. In the literature, some nonparametric methods for
40 the estimation of the cure rate have been studied: Maller and Zhou (1992) proposed a consistent nonparametric estimator of the incidence, but it cannot handle covariates. In order to overcome this drawback, Laska and Meisner (1992) proposed another nonparametric estimator of the cure rate, but it only works for discrete covariates. Furthermore, Wang et al. (2012) proposed a cure model with
45 a nonparametric form in the cure probability. To ensure model identifiability, they assumed a nonparametric proportional hazards model for the hazard function. The estimation was carried out by an expectation-maximization algorithm for a penalized likelihood. They defined the smoothing spline function estimates as the minimizers of the penalized likelihood. More recently, Xu and Peng (2014)
50 extended the existing work by proposing a nonparametric incidence estimator which allows for a continuous covariate. Although the above papers have a nonparametric flavor, they fail to consider a completely nonparametric mixture cure model which works for discrete and continuous covariates in both the incidence and the latency.

55 In this paper, we fill this important gap by proposing a two-component mixture model with nonparametric forms for both the cure probability and the survival function of the uncured individuals. Although we consider only one covariate, the method can be directly extended to a case with multiple covariates.

60 Very recently López-Cheda et al. (2016) have carried out a detailed study of the nonparametric kernel latency estimator proposed in this paper. They

3

have proven asymptotic properties for the latency estimator and proposed a bandwidth selector.

The rest of the article is organized as follows. In Section 2 we give a detailed description of our nonparametric mixture cure model, we study the estimator of the incidence proposed in Xu and Peng (2014) and we introduce a nonparametric estimator of the latency. Moreover, we address the model identifiability. We also present a local maximum likelihood result as well as an iid representation and the asymptotic mean squared error for the nonparametric incidence estimator. A bootstrap bandwidth selection method is introduced in Section 3. Section 4 includes a comparison between these nonparametric estimators and the semiparametric one proposed in Peng and Dear (2000) in a simulation study and assesses the practical performance of the bootstrap bandwidth selector. In Section 5 we apply the proposed nonparametric method to real data related to colorectal cancer patients in CHUAC. An appendix contains the proofs.

2. Nonparametric mixture cure model

2.1. Notation

Let ν be a binary variable where $\nu = 0$ indicates if the individual belongs to the susceptible group (the individual will eventually experience the event of interest if followed for long enough) and $\nu = 1$ indicates if the subject is cured (the individual will never experience the event). The proportion of cured patients and the survival function in the group of uncured patients can depend on certain characteristics of the subject, represented by a set of covariates $\mathbf{X}$. Let $p(\mathbf{x}) = P(\nu = 0 | \mathbf{X} = \mathbf{x})$ be the conditional probability of not being cured, and let Y be the time to occurrence of the event. When $\nu = 1$ it is assumed that $Y = \infty$.

The conditional distribution function of Y is $F(t|\mathbf{x}) = P(Y \leq t | \mathbf{X} = \mathbf{x})$. Note that the corresponding survival function, $S(t|\mathbf{x})$, is improper when cured patients exist, since $\lim_{t \to \infty} S(t|\mathbf{x}) = 1 - p(\mathbf{x}) > 0$. The conditional survival

4

function of Y given that the subject is not cured is denoted by

$$S_0(t|\mathbf{x}) = P(Y > t|\mathbf{X} = \mathbf{x}, \nu = 0).$$

Then, the mixture cure model can be written as:

$$S(t|\mathbf{x}) = 1 - p(\mathbf{x}) + p(\mathbf{x})S_0(t|\mathbf{x}), \tag{1}$$

where $1 - p(\mathbf{x})$ is the incidence and $S_0(t|\mathbf{x})$ is the latency. We assume that each individual is subject to random right censoring and that the censoring time C, with distribution function G, is independent of Y given the covariates $\mathbf{X}$. Let $T = \min(Y, C)$ be the observed time with distribution function H and $\delta = I(Y \leq C)$ the uncensoring indicator. Observe that $\delta = 0$ for all the cured patients, and it also happens for uncured patients with censored lifetime. Without loss of generality, let X be a univariate continuous covariate with density function $m(x)$. Therefore, the observations will be $\{(X_i, T_i, \delta_i), i = 1, \ldots, n\}$ independent and identically distributed (iid) copies of the random vector (X, T, δ).

In order to introduce the nonparametric approach in mixture cure models, we consider the generalized Kaplan-Meier estimator by Beran (1981) to estimate the conditional survival function with covariates:

$$\hat{S}_h(t|x) = \prod_{T_{(i)} \leq t} \left(1 - \frac{\delta_{(i)} B_{h(i)}(x)}{\sum_{r=i}^n B_{h(r)}(x)}\right), \tag{2}$$

where

$$B_{h(i)}(x) = \frac{K_h(x - X_{(i)})}{\sum_{j=1}^n K_h(x - X_{(j)})} \tag{3}$$

are the Nadaraya-Watson (NW) weights with $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$ the rescaled kernel with bandwidth $h \to 0$. In the case of fixed design, the Gasser-Müller (GM) weights (Gasser and Muller, 1984) are more common. Here $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(n)}$ are the ordered T_i's, and $\delta_{(i)}$ and $X_{(i)}$ are the corresponding uncensoring indicator and covariate concomitants. We will also denote $\hat{F}_h(t|x) = 1 - \hat{S}_h(t|x)$ for the Beran estimator of $F(t|x)$. The estimator (2) can be extended to the case of multiple covariates $\mathbf{X} = (X_1, \ldots, X_q)$ using, for example, the product kernel (Simonoff, 1996). Discrete covariates can also be included by splitting

5

the sample into subsamples corresponding to the different category combination of the discrete covariates, for each subsample conducting a nonparametric regression on the continuous covariates. Another possibility is smoothing the discrete covariates with certain kernel functions (Li and Racine, 2004).

Departing from the Beran estimator, Xu and Peng (2014) introduced the following kernel type estimator of the incidence:

$$1 - \hat{p}_h(x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{(i)} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right) = \hat{S}_h(T_{\max}^1|x), \qquad (4)$$

where $T_{\max}^1 = \max_{i:\delta_i=1}(T_i)$ is the largest uncensored failure time, and proved its consistency and asymptotic normality.

Using (1), we propose the following nonparametric estimator of the latency:

$$\hat{S}_{0,b}(t|x) = \frac{\hat{S}_b(t|x) - (1 - \hat{p}_b(x))}{\hat{p}_b(x)}, \qquad (5)$$

where $\hat{S}_b(t|x)$ is the Beran estimator of $S(t|x)$ in (2), $1 - \hat{p}_b(x)$ is the estimator by Xu and Peng (2014) in (4) and b is a smoothing parameter not necessarily equal to h in (4).

The identifiability of a cure model is needed to obtain unique estimates of the model functions. In a cure model, all observed uncensored lifetimes ($\delta_i = 1$) correspond necessarily to uncured subjects ($\nu_i = 0$); but it is impossible to distinguish if a subject with a censored time ($\delta_i = 0$) belongs to the susceptible group ($\nu_i = 0$) or to the non-susceptible group ($\nu_i = 1$), because some censored subjects may experience failures beyond the study period. This leads to difficulties in making a distinction between models with high incidence and long tails of the latency distribution, and low incidence and short tails of the latency distribution. To address this problem, we present Lemma 1.

Lemma 1. *Let D be the support of X. Model (1), with $p(x)$ and $S_0(t|x)$ unspecified, is identifiable if $S_0(t|x)$ is a proper survival function for $x \in D$.*

Since the proof is straightforward, it is omitted.

2.2. Theoretical properties

The Beran estimator of the conditional survival function has been deeply studied in the literature. Dabrowska (1989), in Theorem 2.1, shows its asymptotic unbiasedness, considering NW weights. Furthermore, using GM weights, González-Manteiga and Cadarso-Suárez (1994) give an almost sure iid representation for the estimator, and Van Keilegom and Veraverbeke (1997a) prove an asymptotic representation for the bootstrapped estimator and obtain the strong consistency of the bootstrap approximation for the conditional distribution function.

Let $\widehat{\Lambda}_h(t|x)$ be the estimator of the conditional cumulative hazard function:

$$\widehat{\Lambda}_h(t|x) = \sum_{i=1}^n \frac{\delta_{(i)} B_{h(i)}(x)}{\sum_{r=i}^n B_{h(r)}(x)} I(T_{(i)} \leq t) = \int_0^t \frac{d\hat{H}_h^1(v|x)}{1 - \hat{H}_h(v^-|x)},$$

where

$$\hat{H}_h(t|x) = \sum_{i=1}^n B_{hi}(x) I(T_i \leq t)$$

and

$$\hat{H}_h^1(t|x) = \sum_{i=1}^n B_{hi}(x) I(T_i \leq t, \delta_i = 1)$$

are the empirical estimators of

$$H(t|x) = P\left(T \leq t | X = x\right) \quad \text{and} \quad H^1(t|x) = P\left(T \leq t, \delta = 1 | X = x\right),$$

respectively. Let us define: $\tau_H(x) = \sup\{t : H(t|x) < 1\}$, $\tau_{S_0}(x) = \sup\{t : S_0(t|x) > 0\}$ and $\tau_G(x) = \sup\{t : G(t|x) < 1\}$. Since $S(t|x)$ is an improper survival function, then $S(t|x) > 0$ for any $t \in [0, \infty)$, and $1 - H(t|x) = S(t|x) \times \bar{G}(t|x)$ with $\bar{G}(t|x) = 1 - G(t|x)$ the proper conditional survival function of the censoring time C, we have $\tau_H(x) = \tau_G(x)$.

Let $\tau_0 = \sup_{x \in D} \tau_{S_0}(x)$. As in Xu and Peng (2014), we assume

$$\tau_0 < \tau_G(x), \forall x \in D. \tag{6}$$

This condition states that the support of the censoring variable is not contained in the support of Y, which guarantees that censored subjects beyond

the largest observable failure time are cured. Hence, our estimator does not overestimate the true cure rate. A similar assumption was used by Maller and Zhou (1992, 1994) in homogeneous cases. As pointed out in Laska and Meisner (1992), if the censoring variable takes values always below a time $\tau_G < \tau_0$, for example in a clinical trial with a fixed maximum follow-up period, the largest uncensored observation $T^1_{\max}$ may occur at a time not larger than τ_G and therefore always before τ_0. In such a case, for a large sample size, the estimator in (4) is an estimator of $1 - p(x) + p(x)S_0(\tau_G)$ which is strictly larger than $1 - p(x)$. This comment shows the need of considering the length of follow-up in the design of a clinical trial carefully, so that $S_0(\tau_G)$ is sufficiently small to take the estimator (4) of $1 - p(x) + p(x)S_0(\tau_G)$ as a good estimator of $1 - p(x)$ for practical purposes. The simulations in Xu and Peng (2014) show that if the censoring distribution $G(t|x)$ has a heavier tail than $S_0(t|x)$, the estimates from the proposed method will tend to have smaller biases regardless of the value of $\tau_{S_0}(x)$.

Maller and Zhou (1992) dealt with the problem of testing a similar condition to (6) in an unconditional setting. They proposed to test $H_0 : \tau_0 > \tau_G$ versus the alternative $H_1 : \tau_0 \leq \tau_G$. One of the weak points of this approach is to include condition (6) in the alternative hypothesis. Since this is a neutral assumption, it seems more reasonable to keep (6) if there are no strong evidences against it. In that sense, it is more natural to include (6) in the null hypothesis. Apart from that, the ideas by Maller and Zhou (1992) can be extended to a conditional setting as follows. Let us consider $\Pi(t) = E(\delta|T = t)$ and define $\underline{\tau}_G = \inf_{x \in D} \tau_G(x)$. Condition (6) implies that $\exists a < \underline{\tau}_G$ such that $\Pi(t) = 0\ \forall t \geq a$. Consequently, this condition can be checked in practice by the following hypothesis test:

$$\begin{cases} H_0 : \exists a < \underline{\tau}_G\ /\ \Pi(t) = 0,\ \forall t \geq a \\ H_1 : \forall a < \underline{\tau}_G,\ \exists t \geq a\ /\ \Pi(t) > 0 \end{cases}.$$

This can be tested by means of nonparametric regression estimators of $\Pi(t)$ based on the sample $(T_1, \delta_1), \ldots, (T_n, \delta_n)$. To do that, it is necessary to estimate $\underline{\tau}_G$ in a nonparametric way. This can be done by just estimating the support of

8

every $G(t|x)$, via the Beran estimator.

We need to consider the following assumptions, to be used in the asymptotic results for the incidence estimator:

(A1) X, Y and C are absolutely continuous random variables.

(A2) Condition (6) holds.

(A3) (a) Let $I = [x_1, x_2]$ be an interval contained in the support of m, and $I_\delta = [x_1 - \delta, x_2 + \delta]$ for some $\delta > 0$ such that

$$0 < \gamma = \inf[m(x) : x \in I_\delta] < \sup[m(x) : x \in I_\delta] = \Gamma < \infty$$

and $0 < \delta\Gamma < 1$. For all $x \in I_\delta$ the random variables Y and C are conditionally independent given $X = x$.

(b) There exist $a, b \in \mathbb{R}$, with $a < b$ satisfying $1 - H(t|x) \geq \theta > 0$ for $(t, x) \in [a, b] \times I_\delta$.

(A4) The first derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the first derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.

(A5) The second derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the second derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.

(A6) The first derivatives with respect to t of the functions $G(t|x)$, $H(t|x)$, $H^1(t|x)$ and $S_0(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times D$.

(A7) The second derivatives with respect to t of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times D$.

(A8) The second partial derivatives with respect to t and x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded for $(t, x) \in [0, \infty) \times D$.

(A9) Let us define $H_{c,1}(t) = P(T < t|\delta = 1)$. The first and second derivatives of the distribution and subdistribution functions $H(t)$ and $H_{c,1}(t)$ are bounded away from zero in $[a, b]$. Moreover, $H'_{c,1}(\tau_0) > 0$.

(A10) The functions $H(t|x)$, $S_0(t|x)$ and $G(t|x)$ have bounded second-order derivatives with respect to x for any given value of t.

9

(A11) $\int_0^\infty \dfrac{dH^1(t|x)}{(1-H(t|x))^2} < \infty \; \forall x \in I.$

(A12) The kernel function, K, is a symmetric density vanishing outside $(-1,1)$ and the total variation of K is less than some $\lambda < \infty$.

(A13) The density function of T, f_T, is bounded away from 0 in $[0,\infty)$.

Assumptions $(A1)$, $(A3)-(A9)$ and $(A12)-(A13)$ are necessary in Theorem 3 because its proof is strongly based on Theorem 2 of Iglesias-Pérez and González-Manteiga (1999). Assumptions (A2) and (A10) are needed to prove Lemma 4 and, consequently, required so existing results in the literature, stated for a fixed t such that $1 - H(t|x) \geq \theta > 0$ in $(t,x) \in [a,b] \times I_\delta$, can be applied with the random value $t = T_{\max}^1$. Assumption (A11) is necessary to bound the result of an integral in Lemma 7.

In the next theorem we show that both the proposed nonparametric incidence and latency estimators are the local maximum likelihood estimators of $1 - p(x)$ and $S_0(t|x)$.

Theorem 2. *The estimators $1 - \hat{p}_h(x)$ and $\hat{S}_{0,b}(t|x)$, given in (4) and (5) respectively, are the local maximum likelihood estimators of $1 - p(x)$ and $S_0(t|x)$ for the mixture cure model (1), for any $x \in D$ and $t \geq 0$.*

We also obtain an iid representation of the incidence estimator.

Theorem 3. *Under assumptions $(A1) - (A13)$, for any sequence of bandwidths satisfying $nh^5(\ln n)^{-1} = O(1)$ and $\ln n/(nh) \to 0$, then*

$$(1 - \hat{p}_h(x)) - (1 - p(x)) = (1 - p(x)) \sum_{i=1}^n \tilde{B}_{hi}(x)\xi\left(T_i, \delta_i, x\right) + R_n\left(x\right),$$

where

$$\tilde{B}_{hi}(x) = \frac{\frac{1}{nh}K\left(\frac{x-X_i}{h}\right)}{m(x)}, \tag{7}$$

$$\xi\left(T_i, \delta_i, x\right) = \frac{I(\delta_i = 1)}{1 - H(T_i|x)} - \int_0^{T_i} \frac{dH^1(t|x)}{\left(1 - H(t|x)\right)^2} \tag{8}$$

and

$$\sup_{x \in I} |R_n(x)| = O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) \quad a.s.$$

10

Finally, from the representation in Theorem 3 with straightforward calculations, the asymptotic expression of the mean squared error of the incidence estimator,

$$MSE_x(h_x) = E[(\hat{p}_{h_x}(x) - p(x))^2], \tag{9}$$

is given by:

$$AMSE_x(h_x) = \frac{1}{nh_x}(1 - p(x))^2 c_K \sigma^2(x) + \left[h_x^2 \frac{1}{2} d_K (1 - p(x)) \mu(x) \right]^2, \tag{10}$$

where the first term corresponds to the asymptotic variance and the second one to the asymptotic squared bias, with $d_K = \int v^2 K(v) dv$, $c_K = \int K^2(v) dv$ and, following a notation similar to that in Dabrowska (1992):

$$\mu(x) = \frac{2\Phi'(x,x)m'(x) + \Phi''(x,x)m(x)}{m(x)},$$

$$\sigma^2(x) = \frac{1}{m(x)} \int_0^\infty \frac{dH^1(t|x)}{(1 - H(t|x))^2},$$

where

$$\Phi(u,x) = \int_0^\infty \frac{dH^1(t|u)}{1 - H(t|x)} - \int_0^\infty \frac{1 - H(t|u)}{(1 - H(t|x))^2} dH^1(t|x),$$

with $\Phi'(u,x) = \partial/(\partial u)\Phi(u,x)$ and $\Phi''(u,x) = \partial^2/(\partial u^2)\Phi(u,x)$.

3. Bandwidth selection

The choice of the bandwidth is a crucial issue in kernel estimation, since it controls the trade-off between bias and variance. Most of the methods for smoothing parameter selection in nonparametric curve estimation look for a small error when approximating the underlying curve by the smooth estimate. The asymptotically optimal local bandwidth to estimate the cure rate, $1 - p(x)$, in the sense of minimizing the asymptotic expression of the MSE_x in (10), is given by:

$$h_{x,AMSE} = \left(\frac{c_K \sigma^2(x)}{d_K^2 \mu^2(x)} \right)^{1/5} n^{-1/5},$$

which is an asymptotic approximation of the bandwidth $h_{x,MSE}$ that minimizes the MSE_x. The optimal bandwidth $h_{x,AMSE}$ depends on unknown functions

11

through $\mu(x)$ and $\sigma^2(x)$. Considering Dabrowska (1989), a plug-in bandwidth selector can be obtained by replacing those unknown functions by consistent nonparametric estimates, giving rise to a never-ending process, which seems even harder than the original problem of incidence estimation. On the other hand, unfortunately, the finite-sample behavior of the cross validation (CV) bandwidth selector in this context turned out to be disappointing. The CV bandwidth was highly variable and tended to undersmooth (results not shown).

3.1. Bootstrap bandwidth selector

Another way to select the bandwidth is to use the bootstrap method. It consists of minimizing a bootstrap estimate of the mean squared error, $MSE_x(h_x)$.

We consider the simple weighted bootstrap, without resampling the covariate X, which is equivalent to the simple weighted bootstrap proposed by Li and Datta (2001). For fixed x and $i = 1, \ldots, n$, we set $X_i^* = X_i$ and generate a pair (T_i^*, δ_i^*) from the weighted empirical distribution $\hat{F}_{g_x}(\cdot, \cdot | X_i^*)$, where

$$\hat{F}_{g_x}(u, v | x) = \sum_{i=1}^{n} B_{g_x i}(x) I(T_i \leq u, \delta_i \leq v)$$

and $B_{g_x i}(x)$ is the NW weight in (3) with pilot bandwidth g_x. The resulting bootstrap resample is $\{(X_1, T_1^*, \delta_1^*), \ldots, (X_n, T_n^*, \delta_n^*)\}$. From now on, we will use the notation E^* and P^* for bootstrap expectation and probability, i.e., conditionally on the original observations.

The bootstrap bandwidth is the minimizer of the bootstrap version of $MSE_x(h_x)$ in (9),

$$MSE_{x,g_x}^*(h_x) = E^*[(\hat{p}_{h_x,g_x}^*(x) - \hat{p}_{g_x}(x))^2], \tag{11}$$

which consists of replacing the original sample by the bootstrap resample, the kernel incidence estimator based on the sample by its bootstrap version and the theoretical incidence function by the estimated incidence based on a pilot bandwidth, g. It can be approximated, using Monte Carlo, by:

$$MSE_{x,g_x}^*(h_x) \simeq \frac{1}{B} \sum_{b=1}^{B} (\hat{p}_{h_x,g_x}^{*b}(x) - \hat{p}_{g_x}(x))^2, \tag{12}$$

12

where $\hat{p}^{*b}_{h_x,g_x}(x)$ is the kernel estimator of $p(x)$ using bandwidth h_x and based on the b-th bootstrap resample generated from $\hat{F}_{g_x}$, and $\hat{p}_{g_x}(x)$ is the kernel estimator of $p(x)$ computed with the original sample and pilot bandwidth g_x.

Considering a bandwidth search grid $\{h_1, \ldots, h_L\}$, the procedure for obtaining the bootstrap bandwidth selector for a fixed covariate value, x, is as follows:

1. Generate B bootstrap resamples of the form:
   $$\left\{ \left(X_1^{(b)}, T_1^{*(b)}, \delta_1^{*(b)}\right), \ldots, \left(X_n^{(b)}, T_n^{*(b)}, \delta_n^{*(b)}\right) \right\}, \, b = 1, \ldots, B.$$

2. For the b-th bootstrap resample ($b = 1, ..., B$), compute the nonparametric estimator $\hat{p}^{*b}_{h_l,g_x}(x)$ with bandwidth h_l, $l = 1, 2, \ldots, L$.

3. With the original sample and the pilot bandwidth g_x, compute $\hat{p}_{g_x}(x)$.

4. For each bandwidth h_l in the grid, compute the Monte Carlo approximation of $MSE^*_{x,g_x}(h_l)$, given by (12).

5. The bootstrap bandwidth, h_x^*, is the minimizer of the Monte Carlo approximation of $MSE^*_{x,g_x}(h_l)$ over the grid of bandwidths $\{h_1, \ldots, h_L\}$.

Based on the results in Van Keilegom and Veraverbeke (1997a,b) for fixed design with GM weights, the optimal pilot bandwidth, g_x, could be chosen so that it minimizes (11) for a given sample. However, simulation results showed that the choice of the pilot bandwidth has a small effect on the final bootstrap bandwidth. Consequently, a simple rule is proposed to select g_x (see equation (13) below).

Remark. The bandwidth sequence $g_x = g_n$ has to be typically asymptotically larger than $h_x = h_n$. This oversmoothing pilot bandwidth is required for the bootstrap bias and variance to be asymptotically efficient estimators for the bias and variance terms. The order $n^{-1/9}$ for this asymptotically optimal pilot bandwidth satisfies the conditions in Theorem 1 of Li and Datta (2001), and it coincides with the order obtained by Cao and González-Manteiga (1993) for the uncensored case.

4. Simulation study

In this section we compare the proposed nonparametric estimators with the semiparametric estimators in Peng and Dear (2000), which are implemented in the *smcure* package in R (Cai et al., 2012). These estimators assume a logistic expression for the incidence and a proportional hazards (PH) model for the latency.

We carry out a simulation study with two purposes. First, we evaluate the finite sample performance of the nonparametric estimators $1 - \hat{p}_{h_x}$ and $\hat{S}_{0,b_x}$, both computed in a grid of bandwidths with the Epanechnikov kernel, and we compare the results with those of the semiparametric estimators. Second, the practical behavior of the bootstrap bandwidth selector is assessed. We consider two different models and for both, the censoring times are generated according to the exponential distribution with mean $1/0.3$ and the covariate X is $U(-20, 20)$.

Model 1. For comparison reasons, this simulated setup is the same as the so-called mixture cure (MC) model considered in Xu and Peng (2014). The data are generated from a logistic-exponential MC model, where the probability of not being cured is

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)},$$

with $\beta_0 = 0.476$ and $\beta_1 = 0.358$, and the survival function of the uncured subjects is:

$$S_0(t|x) = \begin{cases} \dfrac{\exp(-\lambda(x)t) - \exp(-\lambda(x)\tau_0)}{1 - \exp(-\lambda(x)\tau_0)} & \text{if } t \leq \tau_0 \\ 0 & \text{if } t > \tau_0 \end{cases},$$

where $\tau_0 = 4.605$ and $\lambda(x) = \exp((x + 20)/40)$. The percentage of censored data is 54% and of cured data is 47%. In Figure 1 (top) we show the shape of the theoretical incidence and latency functions. Note that in this model the incidence is a logistic function and the latency is a function which is very close to fulfill the proportional hazards model and that has been truncated to guarantee condition (6). Therefore, the semiparametric estimators are expected to give very good results in this model.

14

Model 2. The data are generated from a cubic logistic-exponential mixture model, where the incidence is:

$$1 - p(x) = 1 - \frac{\exp\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3\right)}{1 + \exp\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3\right)},$$

with $\beta_0 = 0.0476$, $\beta_1 = -0.2558$, $\beta_2 = -0.0027$ and $\beta_3 = 0.0020$, and the latency is:

$$S_0(t|x) = \frac{1}{2}\left(\exp(-\alpha(x)t^5) + \exp(-100t^5)\right),$$

with

$$\alpha(x) = \frac{1}{5}\exp((x+20)/40).$$

The percentages of censored and cured data are 62% and 53%, respectively. Figure 1 (bottom) shows the theoretical incidence and latency in this model. The incidence is not a logistic function and the effect of the covariate on the failure time of the uncured patients does not fit a PH model anymore. So, the results will show the gain of using the proposed nonparametric estimators, that do not require any parametric or semiparametric assumptions, with respect to the semiparametric ones.

4.1. Efficiency of the nonparametric estimators

A total of $m = 1000$ samples of size $n = 100$ are drawn to approximate, by Monte Carlo, the mean squared error (MSE) of the incidence estimators, and the mean integrated squared error (MISE) of the latency estimators, for a grid of 100 bandwidths in a logarithm scale, from $h_1 = 1.2$ to $h_{100} = 20$ for the incidence function, and from $b_1 = 10$ to $b_{100} = 40$ for the latency. The results for both models are shown in Figure 2.

Regarding the MSE of the incidence estimators, Figure 2 shows that in Model 1 there is a range of bandwidths, from $h_{50} = 4.83$ to $h_{70} = 8.53$ (light blue lines) for which the nonparametric estimator is quite competitive with respect to the semiparametric estimator in values x of the covariate near the endpoints of the support of X, and it works much better when the value of the covariate is around 0. In Model 2, as expected, the nonparametric estimator outperforms
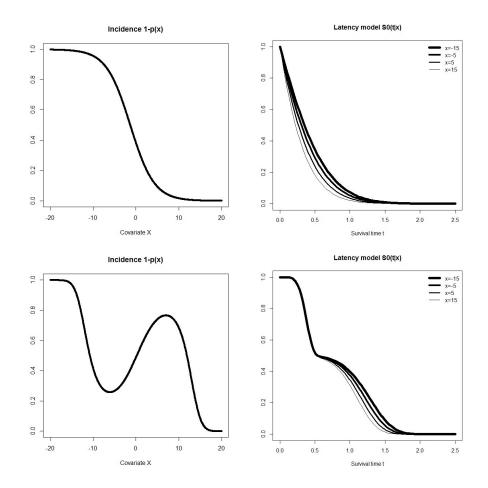
15

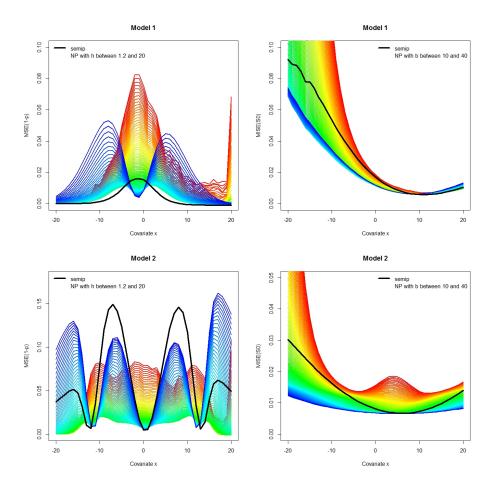Figure 1: Theoretical incidence (left) and latency (right) functions in Model 1 (top) and Model 2 (bottom).

Figure 2: On the left, MSE for the semiparametric (black line) and the nonparametric estimators of $1 - p(x)$ computed with different bandwidths: from $h_1 = 1.2$ (red line) to $h_{100} = 20$ (dark blue line). On the right, MISE for the semiparametric (black line) and the nonparametric estimators of $S_0(t|x)$ computed with different bandwidths: from $b_1 = 10$ (red line) to $b_{100} = 40$ (dark blue line). The data were generated from Model 1 (top) and Model 2 (bottom).

the semiparametric one for a wide range of bandwidths except for 3 singular values of the covariate x, those for which $p''(x) = 0$ and the semiparametric estimate is very close to the true value of $(1 - p(x))$. These 3 values are specific points in which the semiparametric estimation passes through the theoretical function.

Considering the latency estimators, it is noteworthy that in Model 1, for values of the covariate from $x = -20$ to $x = 10$, there is also a very wide range of bandwidths, specifically, between $b_{30} = 15.01$ (light green lines) and $b_{100} = 40$ (dark blue lines), for which the MISE of the nonparametric estimator is smaller than the MISE of the semiparametric estimator, it can be seen in Figure 2 (top, right). In Model 2, the nonparametric estimator of the latency, computed with bandwidths between $b_{20} = 13.05$ (yellow lines) to $b_{100} = 40$ (dark blue lines), outperforms the semiparametric estimator for all the covariate values, except for $x \in [4, 9]$, where the semiparametric estimator is very competitive. In short, the nonparametric estimators are quite comparable to the semiparametric ones in situations where the latter are expected to give better results, as in Model 1, and they outperform the semiparametric estimators when the incidence is not a logistic function and the latency does not fit a PH model (Model 2). The efficiency of the nonparametric estimators depends on the choice of the bandwidth, but although the optimal value of the bandwidth remains unknown, the simulations show that, for quite wide ranges of bandwidths, the proposed nonparametric methods outperform the existing semiparametric estimator by Peng and Dear (2000).

4.2. Efficiency of the bootstrap bandwidth selector for the incidence

In this simulation study, we consider sample sizes of $n = 50, 100$ and 200. For $m = 1000$ trials, we approximate the MSE_x and the optimal bandwidth, $h_{x,MSE}$, of the proposed nonparametric estimator of the incidence. The $MSE_{x,g_x}(h^*)$ and the bootstrap bandwidth h^*_{x,g_x} are also approximated.

Note that minimizing $MSE^*_{x,g_x}(h_x)$ in h_x for each value, x, of the covariate, is a computationally expensive algorithm. For that reason, we carry out a

18

two-step method with a double search in each stage. In the first step, we draw $B = 80$ bootstrap resamples and consider a number of 21 bandwidths equispaced on a logarithmic scale, from $h_1 = 0.2$ to $h_{21} = 50$ in the first search, whereas in the second search the grid is centered around the optimal bandwidth obtained in the first search. Then, we carry out the second step with also a double search in a similar way we did for the first step, but now with two differences: we draw $B = 1000$ bootstrap resamples and we consider a finer smaller grid of 5 bandwidths in both the first and second search.

In view of the fact that the choice of g_x has a low effect on the final bootstrap bandwidth, we propose to use a naive selector, keeping the $n^{-1/9}$ optimal order. Since the distribution of the covariate is uniform, we consider the following global pilot bandwidth, that does not depend on the value x for which the estimation is to be carried out:

$$g = \frac{X_{(n)} - X_{(1)}}{10^{7/9}} n^{-1/9}. \tag{13}$$

Note that, for $X \in U(-20, 20)$, when $n = 100$ the value of the global pilot bandwidth g is $(X_{(n)} - X_{(1)})/10 \simeq 4$. Similarly, $g \simeq 4.32$ ($g \simeq 3.70$) when $n = 50$ ($n = 200$). For a naive pilot bandwidth selector if the distribution for X can not be assumed uniform, see Section 5.

Figure 3 shows the MSE_x evaluated at the median, 25th and 75th percentiles of the proposed bootstrap bandwidth, along the $m = 1000$ simulated samples. The value of the MSE_x for the nonparametric estimator, approximated by Monte Carlo and evaluated at the MSE bandwidth, $h_{x,MSE}$, is also given as reference. We observe that the median, 25th and 75th percentiles of the bootstrap bandwidths have an MSE close to the optimal value. As expected, the similarity increases with the sample size. Moreover, we can also check how $MSE_x(h_{x,MSE})$ and $MSE_x(h^*)$ decrease as n becomes larger.

The performance of the bootstrap bandwidth for Models 1 and 2 is shown in Figure 4. The optimal $h_{x,MSE}$, approximated by Monte Carlo, is displayed together with the median and the 25th and 75th percentiles of the 1000 bootstrap bandwidths, h_x^*. We can appreciate how the bootstrap bandwidth, h_x^*,
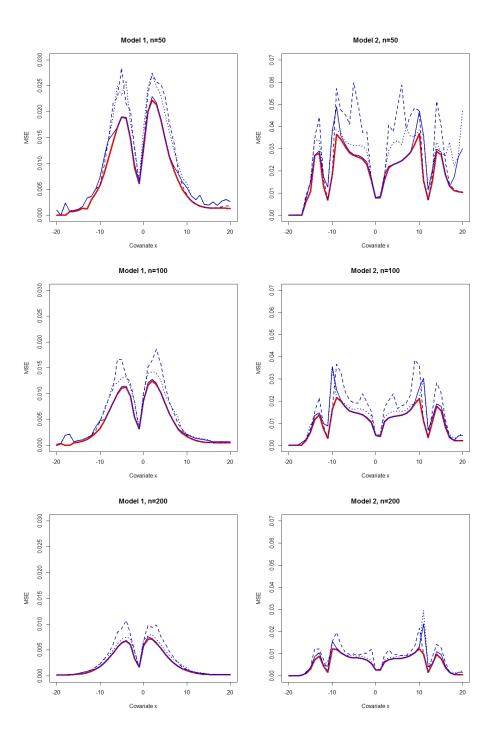
19

Figure 3: MSE_x of the nonparametric estimator of the incidence evaluated at $h_{x,MSE}$ (red line), and MSE_x evaluated at the median (solid blue line), 25th (dotted blue line) and 75th (dashed blue line) percentiles of the bootstrap bandwidth, h_x^*, along $m = 1000$ samples of sizes $n = 50$ (top), $n = 100$ (center) and $n = 200$ (bottom), for Model 1 (left) and Model 2 (right).

approaches $h_{x,MSE}$, adapting properly to the shape of $h_{x,MSE}$ for the three sample sizes. The optimal bandwidth, $h_{x,MSE}$, has got peaks at the values x of the covariate for which $p''(x) = 0$. In other terms, those peaks only occur at points x for which the asymptotically optimal bandwidth is infinitely large because the best choice is to smooth as much as possible, and the best local fit is a global fit. Note that if such large bandwidths are used, those values of x correspond to the values where the MSE_x shows deep valleys, that is, there is a noticeable improvement in the estimation of the incidence.

5. Application to real data

We applied both the semiparametric and the nonparametric estimators to a real dataset of 414 colorectal cancer patients from CHUAC (Complejo Hospitalario Universitario de A Coruña), Spain. We considered two covariates: the stage, from 1 to 4, and the age, from 23 to 103. The variable Y is the follow-up time (months) since the diagnostic until death. About 50% of the observations are censored, with the percentage of censoring varying from 30% to almost 71%, depending on the stage. In Table 1 we show a summary of the data set.

Table 1: Colorectal cancer patients from CHUAC

Stage	Number of patients	Number of censored data	% Censoring
1	62	44	70.97
2	167	92	55.09
3	133	53	39.85
4	52	16	30.77
	414	205	49.52

The incidence is estimated with both the semiparametric and the nonparametric estimators. The age of the patients has been considered as a continuous covariate, and the data have been split into four groups according to the categorical covariate stage.

For the nonparametric estimator of $1 - p$, a naive pilot bandwidth selector has been proposed in (13) if the distribution of X is uniform. The idea is to
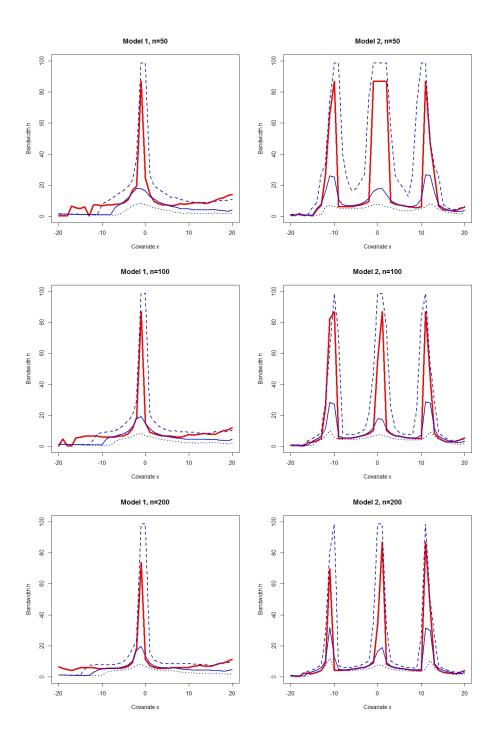
Figure 4: Optimal $h_{x,MSE}$ (red line), median (solid blue line), 25th (dotted blue line) and 75th (dashed blue line) percentiles of the bootstrap bandwidth, h_x^*, along $m = 1000$ samples of sizes $n = 50$ (top), $n = 100$ (center) and $n = 200$ (bottom), for Model 1 (left) and Model 2 (right).

provide a data-driven pilot bandwidth which only depends on both the sample size and on the distribution of the covariate, keeping the $n^{-1/9}$ optimal order. Taking into account that in this case the distribution of the covariate is not uniform (see Figure 5), we propose to use the following local pilot bandwidth:

$$g_x = \frac{d_k^+(x) + d_k^-(x)}{2} 100^{1/9} n^{-1/9},$$

where $d_k^+(x)$ is the distance from x to the k-th nearest neighbor on the right, $d_k^-(x)$ the distance from x to the k-th nearest neighbor on the left, and k a suitable integer depending on the sample size. If there are not at least k neighbors on the right (or left), we use $d_k^+(x) = d_k^-(x)$ (or $d_k^-(x) = d_k^+(x)$) respectively. Our numerical experience shows that a good choice is to consider $k = n/4$. Note that when $n = 100$ the value of the local pilot bandwidth g_x is the mean distance to the 25th nearest neighbor on both the left and right sides.

For the nonparametric estimator, alongside the bootstrap bandwidth, we have also used a smoothed bootstrap bandwidth. We followed Cao et al. (2001), who applied a method for smoothing local bandwidths in another context. The bootstrap bandwidths have been computed in the equispaced grid $x_0 < x_1 < \cdots < x_m$ of the interval $[X_{(1)}, X_{(n)}]$ given by $x_i = X_{(1)} + \Delta i, i = 0, 1, 2, \ldots, m$ where $\Delta = (X_{(n)} - X_{(1)})/m$. The smoothed bootstrap bandwidth in point x_l is computed as follows:

$$h_{x_l}^{* \, smooth} = \begin{cases} \frac{\sum_{j=0}^{l+5} h_{x_j}^*}{l+6}, & l = 0, 1, 2, 3, 4 \\ \frac{\sum_{j=l-5}^{l+5} h_{x_j}^*}{11}, & l = 5, 6, 7, \ldots, m-5 \\ \frac{\sum_{j=l-5}^{m} h_{x_j}^*}{m-l+6}, & l = m-4, m-3, m-2, m-1, m \end{cases}.$$

Figure 5 shows the estimations of the probability of being cured for the different stages with respect to the age of the patients. We can see that the effect of the age on the incidence changes with the stage. For example, using the nonparametric incidence estimation, in Stage 1, patients have a probability of survival between 25% and 65%, depending on the age; whereas in Stage 3, for patients above 60, in a 10 years gap that probability decreases considerably from 40% to almost 0%. It is important to highlight the difference between the nonpara-

metric and the semiparametric curves, that seems to indicate that the logistic model is not valid for the data. The results in Stage 4 deserve some comments. A total of 11 in the 12 greatest lifetimes in Stage 4, including the largest lifetime, are uncensored and, consequently, uncured. This causes that the nonparametric estimation of the probability of being cured is equal to 0. Although it should not be stated that it is impossible for a patient with Stage 4 colorectal cancer to survive, this estimation reinforces the assertion that long-term survival in patients with Stage 4 colorectal cancer is uncommon (Miyamoto et al., 2015). This fact, far from being a weakness of the nonparametric method, is an important advantage, since it allows to detect situations in which introducing the possibility of cure does not contribute to improve the model.

Note that in order to obtain the optimal bootstrap bandwidth, $B = 1000$ bootstrap resamples are used. In a similar way as we did in Section 4.2, we carry out a one-step procedure with a double search. We consider a number of 21 bandwidths equispaced on a logarithmic scale in both searches. The first search is performed between 0.2 and the empirical range of X. The second one is carried out using another grid centered around the optimal bandwidth obtained in the first search. We show the resulting bootstrap bandwidths, with the corresponding local pilot bandwidths, for the different values of the covariate age in Figure 6.

In Figures 7 and 8 we show the latency estimation for Stages $1, 2, 3$ and 4 for two different ages, 45 and 76. The nonparametric estimator $\hat{S}_{0,b_x}$ is computed with five different constant bandwidths: $b = 10, 15, 20, 25$ and 30. It is noteworthy that in Stages 1 and 2 for 45 years, the bandwidth selection influences considerably latency estimation. This is due to the low density of the covariate around this age, as we can see in Figure 5. Suggested by the results in Section 4.1, it is reasonable to choose a large bandwidth. Nevertheless, the choice of the bandwidth for the latency estimation is out of the scope of this paper. This issue has been very recently studied by López-Cheda et al. (2016).
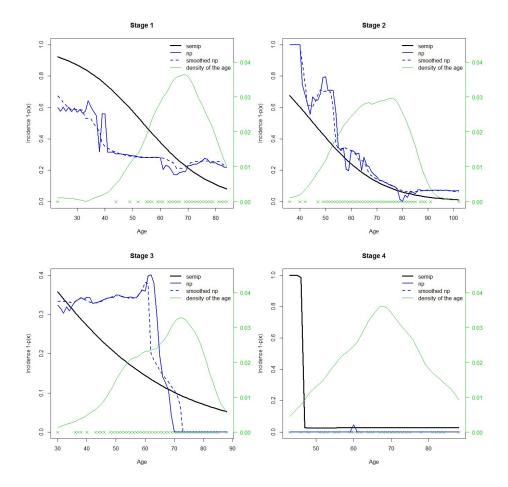
Figure 5: Semiparametric (black line) and nonparametric estimations of the incidence in Stages 1-4 depending on the age computed with the bootstrap bandwidth h_x^* (solid blue line) and with the smoothed bootstrap bandwidth $h_x^{*\,smoothed}$ (dashed blue line). The green line represents the Parzen-Rosenblatt density estimations of the covariate age, using Sheather and Jones' plug-in bandwidth.
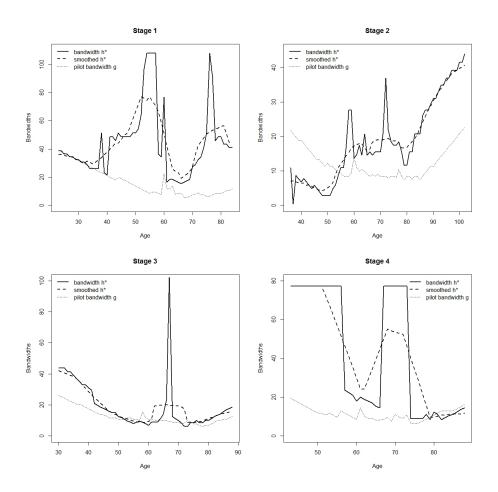
Figure 6: Bootstrap bandwidth h_x^* (solid line), smoothed bootstrap bandwidth $h_x^{*\ smoothed}$ (dashed line) and local pilot bandwidth g_x (dotted line) used for the nonparametric incidence estimator for patients in Stages 1-4.
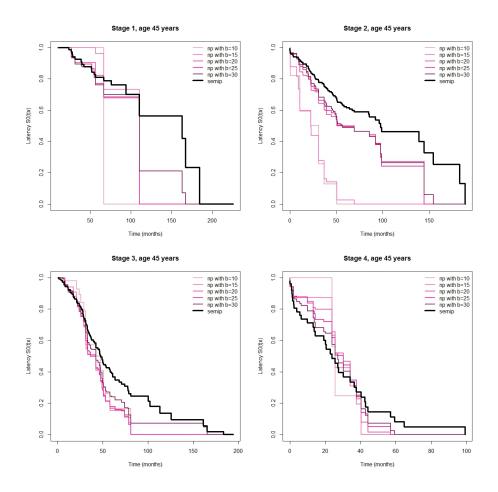
26

Figure 7: Estimated latency for patients of age 45 in Stages 1-4, using the semiparametric (black line) and nonparametric estimators with 5 equispaced bandwidths ranging from $b_0 = 10$ (light pink line) to $b_4 = 30$ (dark pink line).
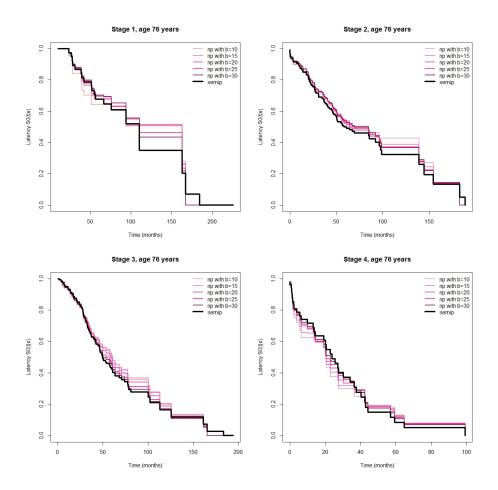
Figure 8: Estimated latency for patients of age 76 in Stages 1-4, using the semiparametric (black line) and nonparametric estimators with 5 equispaced bandwidths ranging from $b_0 = 10$ (light pink line) to $b_4 = 30$ (dark pink line).

Appendix

Proof of Theorem 2. The idea is to estimate $p(x)$ locally, maximizing the observed local likelihood function around x. It can be proved that the maximum likelihood estimator of the survival function $S_0(t|x) = 1 - F_0(t|x)$ has jumps only at the observations (X_i, T_i, δ_i), $i = 1, \ldots, n$ with jumps

$$q_i(x) = S_0(T_{(i)}^-|x) - S_0(T_{(i)}|x).$$

The local likelihood of the model is

$$L(p(x), S_0(\cdot|x)) = \prod_{i=1}^{n} \left\{ [p(x) q_i(x)]^{B_{h(i)}(x)\delta_{(i)}} [1 - p(x) + p(x) \right.$$
$$\left. \times \left(1 - \sum_{j=1}^{i-1} q_j(x)\right)\right]^{\left(1 - \delta_{(i)}\right) B_{h(i)}(x)} \right\}.$$

Let $D_i(x) = B_{h(i)}(x) \delta_{(i)}$ and $P_i(x) = p(x) q_i(x)$, then

$$L(p(x), S_0(\cdot|x)) = \prod_{i=1}^{n} \left\{ P_i(x)^{D_i(x)} \left(1 - \sum_{j=1}^{i-1} P_j(x)\right)^{B_{h(i)}(x) - D_i(x)} \right\}.$$

Consider now the functions $\lambda_i(x) = P_i(x) / \left(1 - \sum_{j=1}^{i-1} P_j(x)\right)$ satisfying

$$1 - \sum_{j=1}^{k} P_j(x) = \prod_{j=1}^{k} (1 - \lambda_j(x)). \tag{A.1}$$

Straightforward calculations yield

$$L(\lambda_1(x), \ldots, \lambda_n(x)) = \prod_{i=1}^{n} \lambda_i(x)^{D_i(x)} (1 - \lambda_i(x))^{\sum_{r=i+1}^{n} B_{h(r)}(x)}.$$

Maximizing the likelihood of the observations for the cure model is equivalent to maximizing

$$\max_{\lambda_i \geq 0; i=1,\ldots,n} \Psi(\lambda_1, \ldots, \lambda_n),$$

where Ψ is the local loglikelihood:

$$\Psi(\lambda_1(x), \ldots, \lambda_n(x)) = \sum_{i=1}^{n} \left[D_i(x) \log \lambda_i(x) + \left(\sum_{r=i+1}^{n} B_{h(r)}(x)\right) \log(1 - \lambda_i(x)) \right],$$

subject to

$$\prod_{i=1}^{n}(1 - \lambda_i(x)) = 1 - \sum_{j=1}^{n} P_j(x) = 1 - p(x). \qquad (A.2)$$

Using standard maximization techniques, we obtain

$$\widehat{\lambda}_i(x) = \frac{D_i(x)}{\sum\limits_{r=i+1}^{n} B_{h(r)}(x) + D_i(x)} = \frac{\delta_{(i)} B_{h(i)}(x)}{\sum\limits_{r=i+1}^{n} B_{h(r)}(x) + \delta_{(i)} B_{h(i)}(x)}.$$

Replacing λ_i in (A.2) by $\widehat{\lambda}_i(x)$, we obtain the estimator of $1 - p(x)$ given in (4).

With respect to the distribution of the uncured subjects, note that

$$F_0\left(T_{(i)}|x\right) = \sum_{j=1}^{i} q_j(x).$$

Since the jumps satisfy $P_i(x) = p(x) q_i(x)$ and using (A.1), we find that the local maximum likelihood estimator is given by

$$\hat{F}_0\left(T_{(i)}|x\right) = \frac{1}{\hat{p}_h(x)}\left[1 - \prod_{j=1}^{i}\left(1 - \hat{\lambda}_j(x)\right)\right] = \frac{\hat{F}_h\left(T_{(i)}|x\right)}{\hat{p}_h(x)},$$

with $\hat{F}_h\left(T_{(i)}|x\right)$ the Beran estimator of $F = 1 - S$ computed at time $T_{(i)}$. $\quad\square$

The following auxiliary results are necessary to prove Theorem 3.

Lemma 4 (Xu and Peng (2014)). *Under assumption (A10),*

$$T_{\max}^1 = \max_{i:\delta_i=1}(T_i) \to \tau_0 \text{ in probability as } n \to \infty.$$

Lemma 5. *Under assumption (A9), we have that*

$$n^{\alpha}(\tau_0 - T_{\max}^1) \to 0 \quad a.s.$$

for any $\alpha \in (0,1)$. In particular, for a sequence of bandwidths satisfying $nh^5(\ln n)^{-1} = O(1)$, we have

$$\tau_0 - T_{\max}^1 = o\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) \quad a.s. \qquad (A.3)$$

Proof of Lemma 5. Using the Borel-Cantelli lemma, it is sufficient to prove that

$$\sum_{n=1}^{\infty} P\left(|a_n(\tau_0 - T_{\max}^1)| > \epsilon\right) < \infty, \text{ for all } \epsilon > 0, \tag{A.4}$$

where $a_n = n^{\alpha}$. Let us fix $\epsilon > 0$ and consider:

$$P(|a_n(\tau_0 - T_{\max}^1)| > \epsilon)$$

$$= P\left(T_{\max}^1 < \tau_0 - \frac{\epsilon}{a_n}\right)$$

$$= P\left(T_i < \tau_0 - \frac{\epsilon}{a_n}, \text{ for all } i = 1, 2, \ldots n \text{ where } \delta_i = 1\right)$$

$$= E\left[P\left(T_i < \tau_0 - \frac{\epsilon}{a_n}, \text{ for all } i = 1, 2, \ldots n \text{ where } \delta_i = 1 | \delta_1, \delta_2, \ldots, \delta_n\right)\right]$$

$$= E\left[\prod_{i=1}^{n} P\left(T_i < \tau_0 - \frac{\epsilon}{a_n} | \delta_i = 1\right)^{\delta_i}\right] = E\left[P\left(T_1 < \tau_0 - \frac{\epsilon}{a_n} | \delta_1 = 1\right)^{\sum_{i=1}^{n} \delta_i}\right]$$

$$= E\left[\left(H_{c,1}\left(\tau_0 - \frac{\epsilon}{a_n}\right)\right)^{\sum_{i=1}^{n} \delta_i}\right],$$

where

$$H_{c,1}(t) = P\left(T < t | \delta = 1\right) = \frac{P(T < t, \delta = 1)}{P(\delta = 1)} = \frac{H_1(t)}{\rho},$$

with $\rho = P(\delta = 1) = E(\delta)$ and $H_1(t) = P(T < t, \delta = 1)$. Consequently, since $\sum_{i=1}^{n} \delta_i \stackrel{\mathrm{d}}{=} B(n, \rho)$, we get:

$$P(|a_n(\tau_0 - T_{\max}^1)| > \epsilon) \tag{A.5}$$

$$= E\left[H_{c,1}\left(\tau_0 - \frac{\epsilon}{a_n}\right)^{\sum_{i=1}^{n} \delta_i}\right] = \sum_{j=0}^{n} \binom{n}{j} \rho^j (1-\rho)^{n-j} H_{c,1}\left(\tau_0 - \frac{\epsilon}{a_n}\right)^j$$

$$= \sum_{j=0}^{n} \binom{n}{j} \left[\rho H_{c,1}\left(\tau_0 - \frac{\epsilon}{a_n}\right)\right]^j (1-\rho)^{nj} = \left[\rho H_{c,1}\left(\tau_0 - \frac{\epsilon}{a_n}\right) + 1 - \rho\right]^n$$

$$= \left[\rho\left(H_{c,1}(\tau_0) - \frac{\epsilon}{a_n} H_{c,1}'(\tau_0) + \frac{\epsilon^2}{2a_n^2} H_{c,1}''(\xi_n)\right) + 1 - \rho\right]^n$$

$$= \left[\rho - \rho\frac{\epsilon}{a_n} H_{c,1}'(\tau_0) + \rho\frac{\epsilon^2}{2a_n^2} H_{c,1}''(\xi_n) + 1 - \rho\right]^n$$

$$= \left(1 - \rho\frac{\epsilon}{a_n} H_{c,1}'(\tau_0) + \rho\frac{\epsilon^2}{2a_n^2} H_{c,1}''(\xi_n)\right)^n,$$

for some $\xi_n \in \left[\tau_0 - \frac{\epsilon}{a_n}, \tau_0\right]$, since $H_{c,1}(\tau_0) = 1$.

Using assumption (A9), $\sup_{t\geq 0} |H_{c,1}''(t)| = C < \infty$. As a consequence, since $\epsilon/a_n \to 0$ as $n \to \infty$, then there exists some $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$:

$$\left| \rho \frac{\epsilon^2}{2a_n^2} H_{c,1}''(\xi_n) \right| \leq \frac{\rho\epsilon^2}{2a_n^2} C \leq \rho \frac{\epsilon}{2a_n} H_{c,1}'(\tau_0). \tag{A.6}$$

From (A.5) and (A.6), we have that:

$$P(|a_n(\tau_0 - T_{\max}^1)| > \epsilon) \leq \left(1 - \rho \frac{\epsilon}{2a_n} H_{c,1}'(\tau_0) \right)^n = \left(1 - \frac{\epsilon}{2a_n} H_1'(\tau_0) \right)^n = b_n^{n/a_n}, \tag{A.7}$$

where

$$b_n = \left(1 - \frac{\epsilon}{2a_n} H_1'(\tau_0) \right)^{a_n} \xrightarrow[n\to\infty]{} r, \tag{A.8}$$

with $r = \exp\left(-\frac{\epsilon H_1'(\tau_0)}{2} \right) < 1$.

Using (A.7) and (A.8), to prove (A.4) it suffices to show that $\sum_{n=1}^{\infty} r^{n/a_n} < \infty$. For that purpose, we will prove that

$$r^{n/a_n} < n^{-2}, \text{ for } n \text{ large enough} \tag{A.9}$$

and, since the hyperharmonic series $\sum_{n=1}^{\infty} n^{-2}$ is convergent, the series $\sum_{n=1}^{\infty} r^{n/a_n}$ will also be convergent.

Note that inequality (A.9) can be written as

$$2\log_R n < \frac{n}{a_n}, \tag{A.10}$$

with $R = r^{-1} \in (1, \infty)$. Recall that $a_n = n^\alpha$ for some $\alpha \in (0,1)$. Now condition (A.10) becomes

$$2\log_R n < n^{1-\alpha},$$

which is true for n large enough, since $n^{-(1-\alpha)} 2\log_R n \to 0$. As a consequence, $n^\alpha(\tau_0 - T_{\max}^1) \to 0$ a.s. for any $\alpha \in (0,1)$. On the other hand, note that:

$$\frac{n^{-\alpha}}{\left(\frac{\ln n}{nh} \right)^{3/4}} = \left[\frac{nh^5}{\ln n} \frac{n^{4-20\alpha/3}}{(\ln n)^4} \right]^{3/20} \xrightarrow[n\to\infty]{} 0,$$

for $\alpha \geq 3/5$ and a sequence of bandwidths satisfying $(\ln n)^{-1} nh^5 = O(1)$. Therefore, the result in (A.3) holds. This completes the proof. $\qquad\square$

In the next three lemmas, we use existing results in the literature for a fixed t such that $1 - H(t|x) \geq \theta > 0$ in $(t,x) \in [a,b] \times I_\delta$, and apply them to the random value $t = T_{\max}^1$. Note that if $\tau_0 < \tau_G(x) = \tau_H(x)$ for all $x \in I_\delta$, then from Lemma 4, under assumption (A10), we have that:

$$T_{\max}^1 = \max_{i:\delta_i=1}(T_i) \to \tau_0 < \tau_H(x) \text{ in probability as } n \to \infty.$$

Therefore, for n large enough, $T_{\max}^1 \leq \tau_0 < \tau_H(x)$ for all $x \in I_\delta$ and taking $b = \tau_0$ we can apply the results considering $t = T_{\max}^1$.

Lemma 6. *Under assumptions (A1)-(A5), (A10) and (A12), and if $nh^5/\ln n = O(1)$ and $\ln n/(nh) \to 0$, then the incidence estimator satisfies:*

$$1 - \hat{p}_h(x) = \exp\left(-\widehat{\Lambda}_h(T_{\max}^1|x)\right) + R_n(x), \text{ for all } x \in I,$$

with

$$\sup_{x \in I} |R_n(x)| = O\left((nh)^{-1}\right) a.s.$$

Proof of Lemma 6. The incidence estimator is equal to:

$$1 - \hat{p}_h(x) = 1 - \hat{F}_h(T_{\max}^1|x),$$

where $\hat{F}_h(t|x) = 1 - \hat{S}_h(t|x)$ is the Beran estimator in (2). The result derives directly for $\hat{F}_h(t|x)$ from the so-called property 3) in the proof of part c) of Theorem 2 in Iglesias-Pérez and González-Manteiga (1999), when the data are subject to random left truncation and right censorship, for which assumptions (A1),(A3)-(A5) and (A12) are required. Assumptions (A2) and (A10) allow to use the aforementioned property when $t = T_{\max}^1$. González-Manteiga and Cadarso-Suárez (1994) proved a similar result under right random censoring with fixed design on the covariate. $\square$

Lemma 7. *Under assumptions (A1)-(A11) and (A13) for $x \in I$ and if $nh^5/\ln n = O(1)$, $\ln n/(nh) \to 0$, then*

$$\widehat{\Lambda}_h(T_{\max}^1|x) - \Lambda(T_{\max}^1|x) = \sum_{i=1}^{n} \tilde{B}_{hi}(x)\xi(T_i,\delta_i,x) + \tilde{R}_n(x),$$

33

with $\tilde{B}_{hi}$ in (7), ξ in (8) and

$$\sup_{x \in I} \left| \tilde{R}_n(x) \right| = O\left(\left(\frac{\ln n}{nh} \right)^{3/4} \right) \quad a.s.$$

Proof of Lemma 7. Under assumptions (A1)-(A8), (A10) and (A13), we apply Theorem 2(b) of Iglesias-Pérez and González-Manteiga (1999) (similarly Theorem 2.2 of González-Manteiga and Cadarso-Suárez (1994) with fixed design using GM weights) to $t = T^1_{\max}$:

$$\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x) = \sum_{i=1}^{n} \tilde{B}_{hi}(x)\xi(T_i, \delta_i, x) + \sum_{i=1}^{n} (B_{hi}(x) - \tilde{B}_{hi}(x))\xi(T_i, \delta_i, x)$$
$$+ \sum_{i=1}^{n} B_{hi}(x) \left(\tilde{\xi}(T_i, \delta_i, x) - \xi(T_i, \delta_i, x) \right) + \tilde{\tilde{R}}_n(x),$$

$$(A.11)$$

with ξ in (8),

$$\tilde{\xi}(T_i, \delta_i, x) = \frac{I(\delta_i = 1)}{1 - H(T_i|x)} - \int_0^{T^1_{\max}} \frac{I(t < T_i)}{(1 - H(t|x))^2} dH^1(t|x)$$

and

$$\sup_{x \in I} \left| \tilde{\tilde{R}}_n(x) \right| = O\left(\left(\frac{\ln n}{nh} \right)^{3/4} \right) \quad a.s.$$

Note that

$$\left| \tilde{\xi}(T_i, \delta_i, x) - \xi(T_i, \delta_i, x) \right| \leq \int_{T^1_{\max}}^{\tau_0} \frac{dH^1(t|x)}{(1 - H(t^-|x))^2} \quad \text{for all } i = 1, \dots, n.$$

Then, under assumption (A9) we apply Lemma 5, and assuming (A11), it is easy to prove that for a sequence of bandwidths satisfying $nh^5(\ln n)^{-1} = O(1)$, the third term in (A.11) is,

$$\sup_{x \in I} \left| \sum_{i=1}^{n} B_{hi}(x) \left(\tilde{\xi}(T_i, \delta_i, x) - \xi(T_i, \delta_i, x) \right) \right| = o\left(\left(\frac{\ln n}{nh} \right)^{3/4} \right) \quad a.s.$$

For the second term in (A.11), it is important to note that:

$$\sum_{i=1}^{n} (B_{hi}(x) - \tilde{B}_{hi}(x))\xi(T_i, \delta_i, x) \quad = \quad \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right) \xi(T_i, \delta_i, x) \frac{m(x) - \hat{m}_h(x)}{\hat{m}_h(x)m(x)},$$

34

with $\hat{m}_h(x)$ the Parzen-Rosenblatt estimator of $m(x)$. Using Theorem 3.3 of Arcones (1997), standard bias and variance calculations and Taylor expansions lead to

$$\sup_{x \in I} \left| \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \xi(T_i, \delta_i, x) \right| = O\left(h^2 + \sqrt{\frac{\ln \ln n}{nh}}\right) \text{ a.s.}$$

Using again Theorem 3.3 of Arcones (1997), it is easy to prove that:

$$\sup_{x \in I} \left| \frac{m(x) - \hat{m}_h(x)}{\hat{m}_h(x)m(x)} \right| = O\left(h^2 + \sqrt{\frac{\ln \ln n}{nh}}\right) \text{ a.s.}$$

Therefore,

$$\sup_{x \in I} \left| \sum_{i=1}^{n} (B_{hi}(x) - \tilde{B}_{hi}(x))\xi(T_i, \delta_i, x) \right| = O\left(\left(h^2 + \sqrt{\frac{\ln \ln n}{nh}}\right)^2\right) \text{ a.s.}$$

For a sequence of bandwidths satisfying $nh^5(\ln n)^{-1} = O(1)$, it is immediate to prove that

$$\sup_{x \in I} \left| \sum_{i=1}^{n} (B_{hi}(x) - \tilde{B}_{hi}(x))\xi(T_i, \delta_i, x) \right| = O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) \text{ a.s.}$$

This completes the proof. $\qquad\square$

Lemma 8. *Under assumptions (A1)-(A8), (A10), (A12) and (A13) and if $nh^5/\ln n = O(1)$, $\ln n/(nh) \to 0$, then*

$$\sup_{x \in I} \left| \widehat{\Lambda}_h\left(T_{\max}^1|x\right) - \Lambda\left(T_{\max}^1|x\right) \right| = O\left(\left(\frac{\ln n}{nh}\right)^{1/2}\right) \text{ a.s.}$$

Proof of Lemma 8. The equivalent result for a fixed $t \in [a, b]$ is within property 2) in the proof of part c) of Theorem 2 in Iglesias-Pérez and González-Manteiga (1999), for which assumptions (A1), (A3)-(A8), (A12) and (A13) are required. Assumptions (A2) and (A10) are needed to apply that result to $t = T_{\max}^1$. For the uniform strong consistency of the Beran estimator $\hat{F}_h(t|x)$, see also Dabrowska (1989). $\qquad\square$

Proof of Theorem 3. The incidence estimator can be split into the following

35

terms:

$$(1 - \hat{p}_h(x)) - (1 - p(x))$$

$$= \hat{S}_h(T^1_{\max}|x) - (1 - p(x))$$

$$= \exp\left[-\widehat{\Lambda}_h(T^1_{\max}|x)\right] - \exp\left[-\Lambda(T^1_{\max}|x)\right] + R_2(x) + R_3(x), \qquad \text{(A.12)}$$

with

$$R_2(x) = \hat{S}_h(T^1_{\max}|x) - \exp\left[-\widehat{\Lambda}_h(T^1_{\max}|x)\right]$$

and

$$R_3(x) = S(T^1_{\max}|x) - (1 - p(x)).$$

To the first term of (A.12) we apply a Taylor expansion of the function $\exp(y)$ around the value $y = -\Lambda(T^1_{\max}|x)$:

$$\exp\left[-\widehat{\Lambda}_h(T^1_{\max}|x)\right] - \exp\left[-\Lambda(T^1_{\max}|x)\right]$$

$$= -\exp\left[-\Lambda(T^1_{\max}|x)\right]\left(\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x)\right) + R_1(x),$$

with

$$R_1(x) = \frac{1}{2}\exp\left[-\Lambda^*(T^1_{\max}|x)\right]\left(\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x)\right)^2$$

and $\Lambda^*(T^1_{\max}|x) = \eta_n(x)$ a value between $\widehat{\Lambda}_h(T^1_{\max}|x)$ and $\Lambda(T^1_{\max}|x)$. Now, adding and substracting $1 - p(x)$, and bearing in mind that $S(T^1_{\max}|x) = \exp[-\Lambda(T^1_{\max})|x]$,

$$\exp\left[-\widehat{\Lambda}_h(T^1_{\max}x)\right] - \exp\left[-\Lambda(T^1_{\max}|x)\right]$$

$$= (1 - p(x))\left(\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x)\right) + R_1(x) + R_4(x), \qquad \text{(A.13)}$$

where

$$R_4(x) = \left[S\left(T^1_{\max}|x\right) - (1 - p(x))\right]\left(\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x)\right).$$

Now, inserting (A.13) in (A.12), we have:

$$(1 - \hat{p}_h(x)) - (1 - p(x))$$

$$= (1 - p(x))\left(\widehat{\Lambda}_h(T^1_{\max}|x) - \Lambda(T^1_{\max}|x)\right) + R_1(x) + R_2(x) + R_3(x) + R_4(x).$$

$$\text{(A.14)}$$

The iid representation of $1 - \hat{p}_h(x)$ now follows, assuming (A1)-(A11) and (A13), from Lemma 7.

Let us study the remainder terms in (A.14), starting with $R_1(x)$. Taking into account that $\exp\left[-\Lambda^*(T_{\max}^1|x)\right]$ is bounded for all $x \in I$, and applying Lemma 8, under the assumptions (A1)-(A8), (A10), (A12) and (A13), we have

$$\sup_{x \in I} |R_1(x)| = O\left(\frac{\ln n}{nh}\right) \quad \text{a.s.}$$

Regarding $R_2(x)$, under the assumptions (A1),(A3)-(A5), (A10) and (A12), directly from Lemma 6 and using $\ln n/(nh) \to 0$ we obtain:

$$\sup_{x \in I} |R_2(x)| = O\left((nh)^{-1}\right) = o\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) \quad \text{a.s.}$$

Focusing on $R_3(x)$, note that it can be bounded as follows:

$$\begin{aligned}
\sup_{x \in I} |R_3(x)| &= \sup_{x \in I} \left| S(T_{\max}^1|x) - (1 - p(x)) \right| \\
&= \sup_{x \in I} \left| \left[(1 - p(x)) + p(x)S_0(T_{\max}^1|x) \right] - (1 - p(x)) \right| \\
&= \sup_{x \in I} \left| p(x)S_0(T_{\max}^1|x) \right| \leq \sup_{x \in I} \left| S_0(T_{\max}^1|x) \right| \\
&= \sup_{x \in I} \left| S_0(T_{\max}^1|x) - S_0(\tau_0|x) \right| \\
&\leq \sup_{x \in I} \left| (T_{\max}^1 - \tau_0)S_0'(\tau_n|x) \right|,
\end{aligned} \qquad (A.15)$$

with $\tau_n \in [T_{\max}^1, \tau_0]$. From condition (A6), that implies that there exists some $\lambda > 0$ such that $\sup_{(t,x) \in [a,b] \times I} |S_0'(t|x)| \leq \lambda$, and using (A.3) and (A.15) for a sequence of bandwidths satisfying $nh^5(\ln n)^{-1} = O(1)$, we have that:

$$\sup_{x \in I} |R_3(x)| = o\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) \text{a.s.}$$

Finally, from Lemma 8, the term R_4 is negligible with respect to R_3, and therefore:

$$\sup_{x \in I} |R_4(x)| = o\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) \text{a.s.}$$

This completes the proof. $\qquad\qquad\square$

Acknowledgements

References

Arcones, M.A., 1997. The law of the iterated logarithm for a triangular array of empirical processes. Electron. J. Probab. 2, 1–39. doi:`10.1214/EJP.v2-19`.

Beran, R., 1981. Nonparametric regression with randomly censored survival data. Technical Report. University of California, Berkeley.

Boag, J.W., 1949. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J. Roy. Stat. Soc. B Met. 11, 15–53. doi:`10.2307/2983694`.

Cai, C., Zou, Y., Peng, Y., Zhang, J., 2012. smcure: Fit Semiparametric Mixture Cure Models, R package version 2.0. http://CRAN.R-project.org/package=smcure.

38

Cantor, A.B., Shuster, J.J., 1992. Parametric versus non-parametric methods for estimating cure rates based on censored survival data. Stat. Med. 11, 931–937. doi:`10.1002/sim.4780110710`.

Cao, R., González-Manteiga, W., 1993. Bootstrap methods in regression smoothing. J. Nonparametr. Stat. 2, 379–388. doi:`10.1080/10485259308832566`.

Cao, R., Janssen, P., Veraverbeke, N., 2001. Relative density estimation and local bandwidth selection for censored data. Comput. Stat. Data. An. 36, 497–510. doi:`10.1016/S0167-9473(00)00055-4`.

Chen, K., Jin, Z., Ying, Z., 2002. Semiparametric analysis of transformation models with censored data. Biometrika 89, 659–668. doi:`10.1093/biomet/89.3.659`.

Chen, M.H., Ibrahim, J.G., Sinha, D., 1999. A new bayesian model for survival data with a surviving fraction. J. Am. Stat. Assoc. 94, 909–919. doi:`10.1080/01621459.1999.10474196`.

Dabrowska, D., 1989. Uniform consistency of the kernel conditional Kaplan-Meier estimate. Ann. Stat. 17, 1157–1167. doi:`10.1214/aos/1176347261`.

Dabrowska, D., 1992. Variable bandwidth conditional Kaplan-Meier estimate. Scand. J. Stat. 19, 351–361. doi:`10.2307/4616252`.

Denham, J.W., Denham, E., Dear, K.B., Hudson, G.V., 1996. The follicular non-Hodgkin's lymphomas - i. the possibility of cure. Eur. J. Cancer 32, 470–479. doi:`10.1016/0959-8049(95)00607-9`.

Farewell, V.T., 1982. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 38, 1041–1046. doi:`10.2307/2529885`.

Farewell, V.T., 1986. Mixture models in survival analysis: Are they worth the risk? Can. J. Stat. 14, 257–262. doi:`10.2307/3314804`.

Gasser, T., Muller, H.G., 1984. Estimating regression functions and their derivatives by the kernel method. Scand. J. Stat. 11, 171–185. doi:`10.2307/4615954`.

Ghitany, M.E., Maller, R.A., Zhou, S., 1994. Exponential mixture models with long-term survivors and covariates. J. Multivariate Anal. 49, 218–241. doi:`10.1006/jmva.1994.1023`.

González-Manteiga, W., Cadarso-Suárez, C., 1994. Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. J. Nonparametr. Stat. 4, 65–78. doi:`10.1080/10485259408832601`.

Haybittle, J.L., 1959. The estimation of the proportion of patients cured after treatment for cancer of the breast. Brit. J. Radiol. 32, 725–733. doi:`10.1259/0007-1285-32-383-725`.

Haybittle, J.L., 1965. A two-parameter model for the survival curve of treated cancer patients. J. Am. Stat. Assoc. 60, 16–26. doi:`10.1080/01621459.1965.10480772`.

Iglesias-Pérez, M.C., González-Manteiga, W., 1999. Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. J. Nonparametr. Stat. 10, 213–244. doi:`10.1080/10485259908832761`.

Kuk, A.Y.C., Chen, C.H., 1992. A mixture model combining logistic regression with proportional hazards regression. Biometrika 79, 531–541. doi:`10.1093/biomet/79.3.531`.

Lai, X., Yau, K.K.W., 2010. Extending the long-term survivor mixture model with random effects for clustered survival data. CSDA 54, 2103–2112. doi:`10.1016/j.csda.2010.03.017`.

Laska, E.M., Meisner, M.J., 1992. Nonparametric estimation and testing in a cure model. Biometrics 48, 1223–1234. doi:`10.2307/2532714`.

Li, C., Taylor, J.M.G., 2002. A semi-parametric accelerated failure time cure model. Stat. Med. 21, 3235–3247. doi:`10.1002/sim.1260`.

Li, G., Datta, S., 2001. A bootstrap approach to nonparametric regression for right censored data. Ann. I. Stat. Math. 53, 708–729. doi:`10.1023/A:1014644700806`.

Li, Q., Racine, J., 2004. Cross-validated local linear nonparametric regression. Stat. Sinica 14, 485–512. doi:`10.1023/A:1014644700806`.

López-Cheda, A., Jácome, M., Cao, R., 2016. Nonparametric latency estimation for mixture cure models. Submitted paper.

Maller, R.A., Zhou, S., 1992. Estimating the proportion of immunes in a censored sample. Biometrika 79, 731–739. doi:`10.1093/biomet/79.4.731`.

Maller, R.A., Zhou, S., 1994. Testing for sufficient follow-up and outliers in survival data. J. Am. Stat. Assoc. 89, 1499–1506. doi:`10.1080/01621459.1994.10476889`.

Maller, R.A., Zhou, S., 1996. Survival Analysis with Long-term Survivors. Wiley, Chichester, U. K. doi:`10.1002/cbm.318`.

Miyamoto, Y., Hayashi, N., Sakamoto, Y., Ohuchi, M., Tokunagam, R., Kurashige, J., Hiyoshi, Y., Baba, Y., Iwagami, S., Yoshida, N., Yoshida, M., Baba, H., 2015. Predictors of long-term survival in patients with stage IV colorectal cancer with multi-organ metastases: a single-center retrospective analysis. Int. J. Clin. Oncol. 20, 1140–1146. doi:`10.1007/s10147-015-0835-2`.

Peng, Y., Dear, K.B., 2000. A nonparametric mixture model for cure rate estimation. Biometrics 56, 237–243. doi:`10.1111/j.0006-341X.2000.00237.x`.

Peng, Y., Dear, K.B., Denham, J.W., 1998. A generalized F mixture model for cure rate estimation. Stat. Med. 17, 813–830. doi:`10.1002/(SICI)1097-0258(19980430)17`.

41

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York. doi:10.1007/978-1-4612-4026-6.

Sy, J.P., Taylor, J.M.G., 2000. Estimation in a Cox Proportional Hazards cure model. Biometrics 56, 227–236. doi:10.1111/j.0006-341X.2000.00227.x.

Tsodikov, A., 1998. A Proportional Hazards model taking account of long-term survivors. Biometrics 54, 1508–1516. doi:10.2307/2533675.

Tsodikov, A., 2001. Estimation of survival based on proportional hazards when cure is a possibility. Math. Comput. Model. 33, 1227–1236. doi:10.1016/S0895-7177(00)00311-3.

Tsodikov, A., 2003. Semiparametric models: A generalized self-consistency approach. J. Roy. Stat. Soc. B 65, 759–774. doi:10.1111/1467-9868.00414.

Van Keilegom, I., Veraverbeke, N., 1997a. Estimation and bootstrap with censored data in fixed design nonparametric regression. Ann. I. Stat. Math. 49, 467–491. doi:10.1023/A:1003166728321.

Van Keilegom, I., Veraverbeke, N., 1997b. Weak convergence of the bootstrapped conditional Kaplan-Meier process and its quantile process. Commun. Stat. Theor. M. 26, 853–869. doi:10.1080/03610929708831954.

Wang, L., Du, P., Lian, H., 2012. Two-component mixture cure rate model with spline estimated nonparametric components. Biometrics 68, 726–735. doi:10.1111/j.1541-0420.2011.01715.x.

Xu, J., Peng, Y., 2014. Nonparametric cure rate estimation with covariates. Can. J. Stat. 42, 1–17. doi:10.1002/cjs.11197.

Yakovlev, A.Y., Cantor, A.B., Shuster, J.J., 1994. Parametric versus nonparametric methods for estimating cure rates based on censored survival data. Stat. Med. 13, 983–986. doi:10.1002/sim.4780110710.

Yakovlev, A.Y., Tsodikov, A., 1996. Stochastic Models of Tumor Latency and their Biostatistical Applications. World Scientific Pub Co Inc.

Yamaguchi, K., 1992. Accelerated failure-time regression model with a regression model of surviving fraction: an analysis of permanent employment

₅₈₀ in Japan. J. Am. Stat. Assoc. 87, 284–292. doi:`10.1080/01621459.1992.10475207`.

Yu, B., Peng, Y., 2008. Mixture cure models for multivariate survival data. CSDA 52, 1524–1532. doi:`10.1016/j.csda.2007.04.018`.

Zeng, D., Yin, G., Ibrahim, J., 2006. Semiparametric transformation models

₅₈₅ for survival data with a cure fraction. J. Am. Stat. Assoc. 101, 670–684. doi:`10.1198/016214505000001122`.

Zhang, J., Peng, Y., 2007. A new estimation method for the semiparametric accelerated failure time mixture cure model. Stat. Med. 26, 3157–3171. doi:`10.1002/sim.2748`.

₅₉₀ Zhang, J., Peng, Y., Li, H., 2013. A new semiparametric estimation method for accelerated hazards mixture cure model. CSDA 59, 95–102. doi:`10.1111/j.1541-0420.2011.01592.x`.