Efficient Estimation from Right-Censored Data when Failure Indicators are Missing at Random

Mark J. van der Laan
University of California, Berkeley

Ian W. McKeague
Florida State University

July 22, 1997

Abstract

The Kaplan–Meier estimator of a survival function is well known to be asymptotically efficient when cause of failure is always observed. It has been an open problem, however, to find an efficient estimator when failure indicators are missing at random. Lo (1991) showed that nonparametric maximum likelihood estimators are inconsistent, and this has led to several proposals of *ad hoc* estimators, none of which are efficient. We now introduce a sieved-nonparametric maximum likelihood estimator, and show that it is efficient. Our approach is related to the estimation of a bivariate survival function from bivariate right-censored data.

1 Introduction

Suppose that we wish to estimate a survival distribution based on right-censored data. When cause of failure is always observed, the method of nonparametric maximum likelihood leads to the well-studied Kaplan–Meier (1958) estimator, which has many desirable properties including asymptotic efficiency (Wellner, 1982). In this paper we address the problem of finding an asymptotically efficient estimator when cause-of-failure information is missing for some individuals.

Cause-of-failure information can be missing for a number of reasons. For example, in epidemiological studies relevant death certificate information can be missing, or autopsy results and hospital case notes can be inconclusive. In such cases it is not possible to determine whether mortality is due to the cause of interest or due to extraneous causes. In a study of the reporting of motorcycle injury fatalities occurring in Connecticut in 1987, Lapidus et al. (1994) found that 40% of death certificates were missing some or

[1] *AMS 1991 subject classifications.* Primary: 62G07; secondary: 62F12

[2] *Key words and Phrases.* Kaplan–Meier estimator, incomplete data, self-consistency, bivariate censorship, influence curve.

all of the required information. A study of mortality patterns among young people in the Netherlands (Bijlsma, 1994) found that 9% of cases had 'ill-defined symptoms, signs and conditions,' and over 90% of those were registered as 'cause-unknown,' mainly due to missing death certificates of people who had died abroad.

Let T be the survival time of interest, let C be a censoring time which is independent of T, and let ξ be a Bernoulli random variable which is allowed to depend on (T, C) in a way to be specified in a moment. Let $X = T \wedge C$ and $\Delta = I(X = T)$. If (X, Δ) is always observed, i.e., we have classical right-censored data on T, then we can estimate the survival function of T using the Kaplan–Meier estimator. In the problem studied here, however, the failure indicator Δ is missing if ξ happens to be 0. That is, we have n i.i.d. observations on $Y = (X, \xi\Delta, \xi)$.

Our goal is to efficiently estimate the survival function S_T of T under the assumption

$$P(\xi = 1 \mid X, \Delta) = P(\xi = 1 \mid X), \tag{1}$$

that ξ and Δ are conditionally independent given X. This assumption places our problem in the framework of 'missing at random,' introduced by Rubin (1976), or, more generally, 'coarsening at random' (CAR), see Heitjan and Rubin (1991), Jacobsen and Keiding (1995) and Gill, van der Laan and Robins (1997). Coarsening is a sampling mechanism in which instead of observing a random quantity of interest one is only able to observe that it takes a value in some possibly randomly determined set of values. CAR isolates those situations in which the coarsening mechanism can be ignored when making inferences.

Our approach is to find the nonparametric maximum likelihood estimator (NPMLE) of S_T based on reduced data produced by a discretization of X. In this way we 'repair' the usual NPMLE, which is inconsistent for estimating S_T. The proposed estimator is found by noticing that our problem can be considered as a special case of nonparametric estimation of a bivariate distribution from bivariate right-censored data. Indeed, the coarsening mechanism acting on (X, Δ) amounts to right-censorship of Δ: observation of Y is equivalent to observation of (X, Δ_ξ), where $\Delta_\xi = \Delta \wedge (2\xi - 1)$. We are then able to use van der Laan's (1996b) efficient sieved-NPMLE of a bivariate distribution function to estimate the distribution F of (X, Δ). This estimator reduces to a simple and explicit form F_n in our case. Finally, using the fact that S_T is a simple functional Φ of F, we construct the proposed estimator $\hat{S}_T = \Phi(F_n)$.

Many authors have studied our problem under the stronger assumption that ξ and Δ are completely independent, i.e., that $P(\xi = 1 \mid X, \Delta)$ does not depend on (X, Δ). The failure indicators are then said to be 'missing completely at random' (MCAR), cf. Little and Rubin (1987). MCAR can be checked from observation of (X, Δ_ξ) given that CAR is in effect, and it allows the use of relatively simple estimators. For example, the survival distribution can be consistently estimated under MCAR by simply ignoring the missing data (cases with $\xi = 0$) and applying the Kaplan–Meier estimator to the complete data. However, this 'complete case estimator' is highly inefficient if there is a significant degree of missingness. The first attempt to improve upon the complete case estimator was made by Dinse (1982) who used the EM algorithm to obtain a NPMLE. Lo

(1991) showed that there are infinitely many NPMLEs and some of them are inconsistent. He constructed two alternative estimators, one of which is consistent and asymptotically normal. Gijbels, Lin and Ying (1993) and McKeague and Subramanian (1996) have proposed further improvements.

Knowing that the coarsening mechanism is MCAR can help in constructing locally efficient estimators at this submodel. However, the function $\pi(x) = P(\xi = 1 \mid X = x)$, which specifies the coarsening mechanism, factors out of the likelihood (i.e., the likelihood factors into a part which only depends on the distribution functions of T and C, and a part which only depends on π) and can be ignored as far as global efficiency is concerned. A globally efficient estimator of F or S_T therefore adapts to arbitrary CAR mechanisms by (implicit) nonparametric estimation of π, see (2). This puts us in the setting of an 'ill-posed inverse problem' (cf. O'Sullivan, 1986) so that some kind of regularization procedure (e.g., kernel smoothing, method of sieves, etc.) is needed; as in density estimation or nonparametric regression, direct NPMLE is not successful.

Although the CAR assumption itself is fairly strong, it is the minimal condition on the coarsening mechanism under which the survival distribution is identifiable from observations on (X, Δ_ξ). Indeed, the independence between the survival and censoring mechanisms ensures that the distribution of T is identifiable (see (3) and (4)) from the distribution of (X, Δ), which can be expressed as

$$F(dx, \delta) \equiv P(X \in dx, \Delta = \delta) = \frac{P(\xi = 1, X \in dx, \Delta = \delta)}{P(\xi = 1 \mid X \in dx, \Delta = \delta)}, \tag{2}$$

$\delta \in \{0, 1\}$. In general, the numerator in F is identifiable but the denominator is not because Δ is unobserved unless $\xi = 1$. The CAR assumption is precisely what is needed to make the denominator identifiable, and implies that F will be identifiable if $\pi(x)$ is bounded away from zero.

The CAR assumption can of course be violated in practice, e.g., in the motorcycle injury fatalities example, the relevant death certificate information is more likely to be missing when death is due to motorcycle injuries than when it is due to other (less specific) causes. However, CAR cannot be checked from data on (X, Δ_ξ) alone, cf. the assumption of independence between T and C in the classical right-censored data model. To judge whether CAR is in effect it would be necessary to have data on the coarsening mechanism itself, that is, data on (X, Δ) when $\xi = 0$. In the motorcycle example, such data is available through police accident reports (Lapidus et al., 1994), and the survival distribution could be identified through estimation of F. In the present paper we shall restrict attention to the situation where only (X, Δ_ξ) is observed and CAR is in effect.

The paper is organized as follows. The proposed estimator $\hat{S}_T$ is constructed in Section 2, and shown to be asymptotically efficient in Section 3. An alternative approach to the problem, based on some general results of Robins and Rotnitzky (1992), is discussed in Section 4. Some numerical results assessing the performance of $\hat{S}_T$ are presented in Section 5.

2 The proposed estimator

2.1 Special case of bivariate right-censored data

By the well-known product integral representation of the survival function S_T on which the Kaplan-Meier estimator is based we have

$$S_T(t) = \prod_{(0,t]}(1 - \Lambda_T(dx)), \tag{3}$$

where

$$\Lambda_T(dx) = \frac{F(dx,1)}{S_X(x-)}. \tag{4}$$

Let $D[0,\tau]$ be the cadlag function space of real valued functions on $[0,\tau]$ endowed with the supremum norm. The equations (3) and (4) define S_T as a mapping $\Phi : (D[0,\tau])^2 \to D[0,\tau]$ from $(F(x,1), F(x,0))$ to S_T:

$$S_T(t) = \Phi(F)(t). \tag{5}$$

Gill and Johansen (1990) (the product integral mapping) and Gill (1989) proved that Φ is compactly differentiable in the sense required by the functional delta-method (see Gill, 1989). Hence if we construct an efficient estimator of the bivariate distribution $F(x,\delta)$, then plugging this estimator in (5) provides us with an efficient estimator of S_T. Here we use the result that a compactly differentiable functional of an efficient estimator is efficient (van der Vaart, 1991).

Another well-known fact concerning the univariate right-censored data model is that for any bivariate distribution $F(dx,\delta)$ there exist independent random variables T and C such that $(X = T \wedge C, \Delta = I(T < C)) \sim F$ (see e.g. Bickel, Klaassen, Ritov, Wellner, 1993). In other words, F is completely unspecified. Hence the problem is to estimate F nonparametrically using the i.i.d. data on (X, Δ_ξ).

If we define $C_1 = 2$ if $\xi = 1$ and $C_1 = -1$ if $\xi = 0$, then

$$(X, \Delta_\xi, \xi) = (X, \Delta \wedge C_1, I(\Delta \wedge C_1 = \Delta)).$$

In other words, Δ is right-censored by the discrete random variable C_1. This shows that indeed estimating S_T comes down to estimating a bivariate distribution of (X, Δ), $\Delta \in \{0,1\}$, where X is always uncensored, but Δ is right-censored.

Estimation of a bivariate survival function based on bivariate right-censored data is an extensively studied topic. The NPMLE for this problem is inconsistent due to the fact that the lines induced by the singly-censored observations do not contain any uncensored observations for continuous data. This lack of interaction with the uncensored observations implies that the self-consistency equation (Efron, 1967) for the NPMLE has a wide class of solutions. We refer to Pruitt (1991) and van der Laan (1996a,b) for discussion on inconsistency of NPMLE in missing data models where the induced regions contain no uncensored observations. The inconsistency of the NPMLE has led to many proposals of ad hoc estimators, but these are not useful to us since they invariably require independence of X and C_1 (i.e., MCAR would be needed). Here we do not give a description of the literature, but refer to the later developments in Bickel, Klaassen, Ritov and Wellner (1993) or van der Laan (1996a,b).

In van der Laan (1996b) an asymptotically efficient estimator of the bivariate survival function is proposed which is an NPMLE based on reduced data in the sense that the uncensored components of the singly censored observations are interval censored by a given grid partition. It was shown that for a fixed grid (that does not depend on n) this estimator is asymptotically efficient for the reduced data, and if the width of the grid converges to zero slowly enough with n, then the estimator is efficient. Moreover, this estimator is suited to our problem since is applicable under any CAR-bivariate-censoring mechanism. It turns out that this (in general implicit) estimator simplifies to a very simple form in our special case. Here we propose this estimator.

2.2 The reduced data NPMLE of F

Let $0 = a_0 < a_1 < \ldots < a_k = \tau$ be a partition of the interval $[0, \tau]$, and set $a_{k+1} = \infty$. Define the discretized version X_d of X by:

$$X_d \equiv \begin{cases} a_j & \text{if } X \in (a_j, a_{j+1}] \text{ and } \xi = 0 \\ X & \text{if } \xi = 1. \end{cases}$$

In other words, if Δ is observed, then X is unchanged, but if Δ is missing ($\xi = 1$), then X is interval censored in the sense that we only observe that $X \in (a_j, a_{j+1}]$. Our estimator of F will be the NPMLE based on the reduced data $(X_d, \xi\Delta, \xi)$.

Let $E(x) \equiv (a_j, a_{j+1}]$, where a_j is such that $x \in (a_j, a_{j+1}]$. Let $R_j \equiv (a_j, a_{j+1}] \times \{0, 1\}$ be the regions for (X, Δ) implied by an observation $(X_d = a_j, \Delta_\xi = -1)$ with a missing failure indicator. As in van der Laan (1996b) we restrict the NPMLE of F to be discrete with pointmasses at all complete observations (X_i, Δ_i) and on one (or more) artificially chosen point in each R_j that contains no complete observations. Of course, if the partition does not depend on n, then as $n \to \infty$ all R_j contain complete observations with probability tending to 1.

Let F_n be the NPMLE and denote its marginal distribution in X by $F_{X,n}$. Also, let $f_n(x, \delta) = F_n(\{x\}, \delta)$ be the density of F_n with respect to the counting measure on the above mentioned support points. Since we are dealing with a censored data model, f_n satisfies the self-consistency equation

$$f_n(x, \delta) = E_{f_n}\left(\frac{1}{n} \sum_{i=1}^{n} I(X_{di} = x, \Delta_i = \delta) \,\middle|\, \text{reduced data} \right)$$

(see Gill, 1989, Example 2), which can be written (cf. (7.4) in Efron, 1967) as

$$f_n(x, \delta) = \frac{1}{n} \sum_{i=1}^{n} I(X_i = x, \Delta_i = \delta, \xi_i = 1) + \frac{1}{n} \sum_{i=1}^{n} I(X_i \in E(x), \xi_i = 0) \frac{f_n(x, \delta)}{F_{X,n}(E(x))}. \quad (6)$$

The way to read a self-consistency equation is that each observation gets mass $1/n$ which has to be redistributed over its induced region for (X, Δ) according to the estimate (based on f_n itself) of the conditional distribution over this region. So a point (x, δ) gets mass $1/n$ from each complete observation on (x, δ) and it gets mass $1/n$ times $f_n(x, \delta)/F_{X,n}(E(x))$ from each incomplete observation with an implied region R_j containing (x, δ).

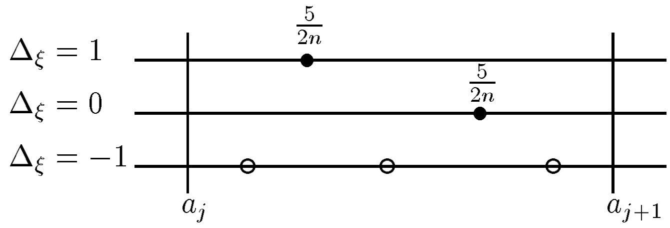


Figure 1: Pointmasses for the reduced data NPMLE of F.

At first sight it seems that (6) is not easily solvable. In this special case, however, $F_{X,n}(E(x))$ is known so we can obtain an explicit solution; incomplete observations with a different X_d do not interact in the sense that their implied regions R_j are disjoint. Hence the mass given to a region R_j is just $1/n$ times the number of observations with $X \in (a_j, a_{j+1}]$; other observations cannot give any mass to R_j. Thus $F_{X,n}(E(x))$ equals this fraction with $R_j = E(x) \times \{0, 1\}$.

Denote the marginal distribution of X by P_X. Let P_0 and P_1 be the sub-distributions of (X, Δ) with $\xi = 0$ and 1, respectively. The marginal sub-distributions of the X components of P_0 and P_1 are written $P_{0,X}$ and $P_{1,X}$. A subscript n added to any of these (sub)-distributions will indicate that we are referring to its empirical counterpart.

Note that $P_{X,n}(E(x))$ is the fraction of $X_i \in E(x)$, and $P_{X,n}(E(x))$ is the empirical distribution of the discretized X. Thus we have $F_{X,n}(E(x)) = P_{X,n}(E(x))$. Also, the first term on the right-hand side of (6) is just $P_{1,n}(\{x\}, \delta)$. For each (x, δ) corresponding to a complete observation $(X_i = x, \Delta_i = \delta)$ we can explicitly solve for $f_n(x, \delta)$, which provides us with:

$$f_n(x, \delta) = \frac{P_{X,n}(E(x))}{P_{1,X,n}(E(x))} P_{1,n}(\{x\}, \delta). \tag{7}$$

The mass $f_n(x, \delta)$ for the artificially chosen points in the R_j that do not contain complete observations is only determined by $F_{X,n}((a_j, a_{j+1}]) = P_{0,n}((a_j, a_{j+1}])$; so incomplete observations with $X_d = a_j$ where R_j does not contain any complete observations can redistribute their mass in an arbitrary manner over R_j. The latter fact is exactly the reason why the interval censoring of the observations with missing failure indicators is essential for estimation of F; regions implied by incomplete observations should contain complete observations with probability tending to 1.

A simple example is helpful for understanding f_n. Figure 1 displays five observations in an interval, three of which have missing failure indicators. The combined mass of these three points ($\frac{3}{n}$) is redistributed to the two complete observations, each of which will have a mass of $\frac{1}{n}+\frac{3}{2n}=\frac{5}{2n}$. This agrees with the answer obtained from (7).

2.3 The estimator of the survival function of T

The estimator (7) provides us with an estimator of $F(dx,1)$ and of $S_X(x-)=P(X\geq x)$. Hence substitution of (7) into (5) provides us with our proposal for estimating $S_T(t)$:

$$\hat{S}_T(t)\equiv \prod_{(0,t]}\left(1-\frac{F_n(dx,1)}{S_{X,n}(x-)}\right), \tag{8}$$

where $S_{X,n}$ is the survival function corresponding to $F_{X,n}$. Notice that if $P(\xi=1)=1$, then $\hat{S}_T$ is just the Kaplan-Meier estimator, as it should be.

3 Analysis of the estimator and its influence curve

We will first show that (7) indeed defines a sensible estimator. We need a slightly stronger CAR assumption than the minimal-CAR assumption (1) in order that F be identifiable from the discretized data; we assume

$$P(\xi=1\mid X,\Delta)=P(\xi=1\mid X_D), \tag{9}$$

where $X_D=a_j$ if $X\in(a_j,a_{j+1}]$. Applying this condition to the denominator in (2) we obtain

$$F(dx,\delta)=\frac{P_X(E(x))}{P_{1,X}(E(x))}P_1(dx,\delta), \tag{10}$$

which shows that (7) will provide us with an consistent estimator of F.

Now regard (10) as defining a map $\Phi_1:(D[0,\tau])^3\to(D[0,\tau])^2$ from the distributions $(P_1(x,1),P_1(x,0),P_0(x))$, i.e. the distributions that determine the observation (X_d,Δ_ξ), to the distributions $(F(x,1),F(x,0))$. Then

$$S_T=\Phi(\Phi_1(P_1(\cdot,1),P_1(\cdot,0),P_0)$$

and

$$\hat{S}_T=\Phi(\Phi_1(P_{1,n}(\cdot,1),P_{1,n}(\cdot,0),P_{0,n})).$$

The functional delta-method (Gill, 1989) tells us that for proving weak convergence of $\sqrt{n}(\hat{S}_T-S_T)$ as random elements of $D[0,\tau]$ to a Gaussian process it suffices to prove compact differentiability of Φ and Φ_1. The compact differentiability of Φ has already been established [see Gill and Johansen (1990) and Gill (1989)], and the compact differentiability of Φ_1 only requires compact differentiability of the mapping $(F,G)\to\int FdG$. The latter has been proved in Gill (1989). The assumptions needed here are that the denominators are bounded away from zero; for Φ this means that $S_X(\tau)>0$ and for Φ_1 it means that $P_{1,X}((a_j,a_{j+1}])>0$ for all $j=0,1,\ldots,k$.

As shown in van der Laan (1996b, Theorem 5.1) the estimator F_n of F is efficient for the reduced data. Hence the compact differentiability of Φ and van der Vaart's (1991) result imply that $\hat{S}_T$ is efficient for the reduced data. This proves the following theorem.

Theorem 3.1 *Let the partition* $0 = a_0 < a_1 < \ldots, a_k = \tau$ *be such that* $P(X \in (a_j, a_{j+1}], \xi = 1) > 0$ *for* $j = 0, 1, \ldots, k-1$ *and* $P(X > \tau) > 0$. *Also assume that* $P(\xi = 1 \mid X, \Delta) = P(\xi = 1 \mid X_D)$. *Then* $\sqrt{n}(\hat{S}_T - S_T)$ *converges weakly as a sequence of random elements of* $D[0, \tau]$ *to a Gaussian process. Moreover,* $\hat{S}_T(t)$ *is asymptotically efficient for the reduced data* $(X_d, \xi\Delta, \xi)$.

Here X_D and X_d can be chosen arbitrarily close to X. Of course, if ξ depends on the full X, then the estimator will still be efficient if the mesh of the partition converges to zero at a rate which is not too slow and not too quick (cf. van der Laan, 1996b, Theorem 5.1), but $\hat{S}_T$ would be inconsistent for a fixed partition. On the other hand, if ξ depends on X only through X_D (for some fixed partition), and the mesh of the partition converges to zero slowly enough as $n \to \infty$, then $\hat{S}_T(t)$ is asymptotically efficient for the original data (cf. van der Laan, 1996b, Theorem 5.1). In particular, this holds if ξ is independent of X.

3.1 The influence curve of the estimator

The compact differentiability of Φ implies (see Gill, 1989) that $\hat{S}_T(t)$ is asymptotically linear:

$$\hat{S}_T(t) - S_T(t) = \frac{1}{n}\sum_{i=1}^{n} IC_t(Y_i) + o_P(1/\sqrt{n}),$$

where the i.i.d. random variables $IC_t(Y_i)$ are just the derivative $d\Phi \circ d\Phi_1(P)$ of $\Phi \circ \Phi_1$ at $P \equiv (P_1(\cdot, 1), P_1(\cdot, 0), P_0)$ applied to the empirical distribution of P based on one observation $Y_i = (X_i, \xi_i\Delta_i, \xi_i)$, and evaluated at t. Here IC_t is called the influence curve of $\hat{S}_T(t)$. We have that $\sqrt{n}(\hat{S}_T(t) - S_T(t))$ is asymptotically normal with mean zero and variance equal to the variance of $IC_t(Y)$. Hence an estimator of IC_t will lead to an estimate of the asymptotic variance of $\hat{S}_T(t)$ and a (pointwise) confidence interval for $S_T(t)$ in the usual fashion.

Determining the influence curve comes down simply to finding the derivatives (linear approximations) of Φ and Φ_1 by neglecting all second order terms and substituting the linearization of Φ_1 in the linearization of Φ. Here it means that we need to find the linearization of $F_n(dx, 1) - F(dx, 1)$ and the corresponding linearization of $S_{X,n}(x) - S_X(x)$ and substitute these in the linearization of $\hat{S}_T(t) - S_T(t)$, the product integral mapping, in terms of $F_n(dx, 1) - F(dx, 1)$ and $S_{X,n}(x) - S_X(x)$.

By using telescoping (see Gill, van der Laan, Wellner, 1995) it follows that:

$$\begin{aligned} F_n(dx, \delta) - F(dx, \delta) &= P_{1,n}(dx, \delta)\frac{P_{1,X,n}(E(x))}{P_{1,n}(E(x))} - P_1(dx, \delta)\frac{P_X(E(x))}{P_{1,X}(E(x))} \\ &\approx \frac{P_X(E(x))}{P_{1,X}(E(x))}(P_{1,n} - P_1)(dx, \delta) + (P_{X,n} - P_X)(E(x))\frac{P_1(dx, \delta)}{P_{1,X}(E(x))} \end{aligned}$$

$$-(P_{1,X,n}-P_{1,X})(E(x))\frac{P_X(E(x))}{P_{1,X}(E(x))^2}P_1(dx,\delta).$$

This provides us with the linearization of $F_n(dx,\delta)-F(dx,\delta)$ in terms of the empirical distribution of the data. We have

$$\begin{aligned}
\frac{P_X(E(x))}{P_1(E(x))} &= \frac{1}{P(\xi=1\mid X\in E(x))}\\
\frac{P_1(dx,\delta)}{P_1(E(x))} &= \frac{F(dx,\delta)}{F_X(E(x))}\\
\frac{P_X(E(x))}{P_1(E(x))^2}P_1(dx,\delta) &= \frac{1}{P(\xi=1\mid E(x))}\frac{F(dx,\delta)}{F_X(E(x))}.
\end{aligned}$$

For notational convenience we define $\pi_D(x)\equiv P(\xi=1\mid X\in E(x))$. Substitution of these expressions in the linearization of $F_n(dx,\delta)-F(dx,\delta)$ provides us with:

$$\begin{aligned}
F_n(dx,\delta)-F(dx,\delta) &\approx \frac{1}{\pi_D(x_d)}(P_{1,n}-P_1)(dx,\delta)+(P_{X,n}-P_X)(E(x))\frac{F(dx,\delta)}{F_X(E(x))}\\
&\quad -(P_{1,X,n}-P_{1,X})(E(x))\frac{1}{\pi_D(x)}\frac{F(dx,\delta)}{F_X(E(x))}. \qquad (11)
\end{aligned}$$

The linearization of $F_{X,n}(dx)-F_X(dx)$ is now simply obtained by adding the linearizations of $F_n(dx,1)-F(dx,1)$ and $F_n(dx,0)-F(dx,0)$. Hence

$$\begin{aligned}
S_{X,n}(x)-S_X(x) &= \int_x^\infty \frac{1}{\pi_D(x)}(P_{1,X,n}-P_{1,X})(dx)+\int_x^\infty (P_{X,n}-P_X)(E(x))\frac{F_X(dx)}{F_X(E(x))}\\
&\quad -\int_x^\infty (P_{1,X,n}-P_{1,X})(E(x))\frac{1}{\pi_D(x)}\frac{F_X(dx)}{F_X(E(x))}. \qquad (12)
\end{aligned}$$

Now, it remains to find the linearization of $\hat{S}_T-S_T$ in terms of the linearizations (11) and (12). The linearization of the product integral $\Lambda_T\to \prod(1-d\Lambda_T)=S_T$ is given in Gill and Johansen (1990) and follows directly from the Duhamel equation. The linearization of $(F(dx,\delta),S_X(x))\to F(dx,1)/S_X(x)=\Lambda_T(dx)$ is trivial. We have:

$$\begin{aligned}
\hat{S}_T(t)-S_T(t) &\approx S_T(t)\int_0^t \frac{1}{1-\Lambda_T(\{x\})}\left(\frac{F_n(dx,1)-F(dx,1)}{S_X(x)}\right.\\
&\quad \left.+\frac{(S_{X,n}-S_X)(x)}{S_X(x)^2}F(dx,1)\right). \qquad (13)
\end{aligned}$$

Substitute now for $S_{X,n}-S_X$ and $F_n(dx,1)-F(dx,1)$ in (13) the linearizations (11) and (12) to obtain:

$$\frac{\hat{S}_T(t)-S_T(t)}{S_T(t)} \approx \int_0^t \frac{1}{1-\Lambda_T(\{x\})}\frac{1}{S_X(x)}\left((P_{X,n}-P_X)(E(x))\frac{F(dx,1)}{F_X(E(x))}\right. \qquad (14)$$

$$
\begin{aligned}
&+\frac{(P_{1,n}-P_1)(dx,1)}{\pi_D(x)}-\frac{(P_{1,X,n}-P_{1,X})(E(x))F(dx,1)}{\pi_D(x)F_X(E(x))}\Bigg)\\
&+\int_0^t \frac{1}{1-\Lambda_T(\{x\})}\frac{1}{S_X(x)^2}F(dx,1)\left(\int_x^\infty \frac{1}{\pi_D(s)}(P_{1,X,n}-P_{1,X})(ds)\right.\\
&\left.+\int_x^\infty \left((P_{X,n}-P_X)(E(s))-\frac{(P_{1,X,n}-P_{1,X})(E(s))}{\pi_D(s)}\right)\frac{F_X(ds)}{F_X(E(s))}\right).
\end{aligned}
$$

If we have only one observation (i.e. $n=1$), then $P_{1,n}(dx,\delta)=I(X\in dx,\Delta=\delta,\xi=1)$, $P_{X,n}(E(x))=I(X\in E(x))$ and $P_{1,n}(E(x))=I(X\in E(x),\xi=1)$. Substitution of these indicators into (15) provides us with the influence curve of $\hat{S}_T(t)$:

$$
\begin{aligned}
\frac{IC_t(Y)}{S_T(t)} &= \frac{I(X\le t,\Delta=1,\xi=1)}{(1-\Lambda(\{X\}))S_X(X)\pi_D(X)} \qquad (15)\\
&+\int_0^t \frac{1}{1-\Lambda_T(\{x\})}\left(I(X\in E(x))-\frac{I(X\in E(x),\xi=1)}{\pi_D(x)}\right)\frac{F(dx,1)}{S_X(x)F_X(E(x))}\\
&+\int_0^t \frac{1}{1-\Lambda_T(\{x\})}\frac{1}{S_X(x)^2}\frac{I(X>x,\xi=1)}{\pi_D(X)}F(dx,1)\\
+\int_0^t \frac{1}{1-\Lambda_T(\{x\})}&\left(\int_x^\infty\left(I(X\in E(s))-\frac{I(X\in E(s),\xi=1)}{\pi_D(s)}\right)\frac{F_X(ds)}{F_X(E(s))}\right)\frac{F(dx,1)}{S_X(x)^2}.
\end{aligned}
$$

Estimation of IC_t requires estimation of S_T and F, and is simply carried out by plugging-in our proposals, and an estimate of $\pi_D(x)$. If it is known that ξ is independent of X, then one simply estimates $P(\xi=1)$ by the fraction of uncensored (X_i,Δ_i). If dependence between ξ and X is expected, then one estimates $\pi_D(x)$ by the fraction of completely observed (X_i,Δ_i) with $X_i\in E(x)$.

3.2 The efficient influence curve

The influence curve (15) is the influence curve of $\hat{S}_T$ for a fixed grid. If we let the mesh of the grid converge to zero slowly with n, then $\hat{S}_T$ is efficient; so it is asymptotically linear with influence curve equal to the efficient influence curve. Hence the efficient influence curve must be the limit of (15) for $\max_i \mid a_i-a_{i-1}\mid\to 0$. If the mesh of the grid converges to zero, then an integral with integrand $I(X\in E(x))$ only integrates over an infinitesimal interval $(a_i,a_{i+1}]$ and hence

$$
\int_0^t \frac{I(X\in E(x))F(dx,1)}{S_X(x)F_X(E(x))}\to\frac{I(X\le t)}{S_X(X)}\lim\frac{\int_{E(X)}F(dx,1)}{F_X(E(X))}=\frac{I(X\le t)}{S_X(X)}\frac{dF(\cdot,1)}{dF_X}(X),
$$

where

$$
k(x)\equiv\frac{dF(\cdot,1)}{dF_X}(x)=\frac{P(X\in dx,C>x)}{P(T\wedge C\in dx)}=\frac{\Lambda_T(dx)}{\Lambda_T(dx)+\Lambda_C(dx)}.
$$

Furthermore, we have that $\pi_D(X) \to \pi(X)$ and $\int I(X \in E(x))F_X(dx)/F_X(E(x)) \to 1$. Hence the efficient influence curve is given by:

$$\begin{aligned}\frac{IC_t^*(Y)}{S_T(t)} &= \frac{1}{1-\Lambda(\{X\})}\left(I(X \le t) - \frac{I(X \le t, \xi = 1)}{\pi(X)}\right)\frac{k(X)}{S_X(X)} \\ &+ \frac{I(X \le t, \Delta = 1, \xi = 1)}{(1-\Lambda(\{X\}))S_X(X)\pi(X)} + \int_0^t \frac{1}{1-\Lambda(\{x\})}\frac{1}{S_X(x)^2}\frac{I(X > x, \xi = 1)}{\pi(X)}F(dx, 1) \\ &+ \left(1 - \frac{I(\xi = 1)}{\pi(X)}\right)\int_0^{t\wedge X}\frac{1}{1-\Lambda(\{x\})}\frac{F(dx,1)}{S_X(x)^2}. \end{aligned} \tag{16}$$

Estimation of the efficient influence curve requires nonparametric estimation of the density k and hence requires smoothing, which explains why a standard NPMLE is not consistent for this problem. Moreover, if ξ depends fully on X, then it requires nonparametric estimation of the binary regression function $\pi(X)$. Although we do not pursue it here, the efficient influence curve can be used to construct one-step efficient estimators. If the efficient influence curve is estimated consistently, then the one-step estimator will be efficient (see van der Laan, 1996a, Corollary 2.2).

Gill, van der Laan and Robins (1997) have shown for general nonparametric models under minimal-CAR that there exists only one influence curve; in our case the weakest possible CAR assumption is (1), and (16) is the unique influence curve. Consequently, any inefficient estimator cannot be asymptotically linear. This explains why inefficient estimators can only be constructed under stronger assumptions than just minimal-CAR; the missing failure indicator model is an interesting example (another one is the bivariate censoring model) were many estimators have been proposed, all being inconsistent under minimal CAR.

4 An alternative approach

Results of Robins and Rotnitzky (1992) make it possible to construct an alternative efficient estimator for S_T using the general theory of semiparametric efficiency bounds [see, e.g., Newey (1991) and Bickel et al. (1993)].

Suppose we observe a random vector Y having distribution $P \in \{P_\theta\}$, which is identified by an unknown (possibly infinite dimensional) parameter θ. Let $L_0^2(P)$ denote the Hilbert space of P-square integrable functions with mean zero. Consider a smooth one-dimensional (SOD) submodel $\{P^\epsilon\} \subset \{P_\theta\}$ passing through P and having score function $k(Y) \in L_0^2(P)$ at $\epsilon = 0$, see Bickel et al.'s (1993) definition of a 'regular parametric' submodel. The tangent space $\mathbf{T}(P)$ is the $L_0^2(P)$-closure of the linear span of all such score functions k. For example, if nothing is known about P, then $P^\epsilon(dy) = (1 + \epsilon k(y))P(dy)$ is a SOD submodel for any bounded function k with mean zero (provided ϵ is sufficiently small), so $\mathbf{T}(P)$ is seen to be the whole of $L_0^2(P)$ in this case.

Let $\mu = \mu(\theta) = \mu(P_\theta)$ be a real parameter that is pathwise differentiable at P: there exists $g \in L_0^2(P)$ such that $\lim_{\epsilon \to 0}(\mu(P^\epsilon) - \mu(P))/\epsilon = \langle g, k\rangle$, for any SOD submodel $\{P^\epsilon\}$

with score function k, where $\langle \cdot, \cdot \rangle$ is the inner product in $L_0^2(P)$. The function g is called a gradient (or influence curve) for μ; the projection IC_μ^* of any gradient on the tangent space is unique and is known as the canonical gradient (or efficient influence curve). The supremum of the Cramér–Rao bounds for all SOD submodels (the information bound) is given by the second moment of $IC_\mu^*(Y)$.

In the present context, we observe $Y = (X, \xi\Delta, \xi)$ and the distribution P of Y is identified by $\theta = (F, \pi)$. Consider the parameter $\mu = \mu(F) = F(x, \delta)$ for given x and $\delta \in \{0, 1\}$. In the full model that only assumes a CAR missingness process, we have $\mathbf{T}(P) = L_0^2(P)$, see Gill, van der Laan and Robins (1997). Thus, *any* gradient for μ is necessarily its canonical gradient. The canonical gradient of μ can be found as a gradient in the submodel with π known minus its projection on the space of missingness scores (i.e., the tangent space for the submodel in which only π is unknown). Robins and Rotnitzky (1992, Theorem 4.2) provide closed form expressions for the projection on a space of missingness scores when the missingness process is monotone. In our case, however, the missingness process is very simple (it acts on only one component of the complete data vector), so it is easy to find the projection without reference to this general theorem, see the Appendix.

The following proposition expresses the canonical gradient of μ explicitly in terms of the functions $\pi(x)$ and $p(x) = P(\Delta = \delta \mid \xi = 1, X = x)$.

Proposition 4.1 *Suppose the CAR assumption (1) holds and $\pi(X)$ is bounded away from zero. Then the canonical gradient of μ is*

$$IC_\mu^*(Y) = \frac{\xi}{\pi(X)} \left[I(X \le x, \Delta = \delta) - \mu \right] - \left[I(X \le x) p(X) - \mu \right] \left(\frac{\xi}{\pi(X)} - 1 \right).$$

This result can be used to obtain a closed form efficient estimator $\hat{\mu}$ of μ by solving the estimating equation $\sum_{i=1}^n \widehat{IC_\mu^*}(Y_i) = 0$, where $\widehat{IC_\mu^*}$ is a plug-in estimate of IC_μ^* in which π and p are replaced by suitable estimates $\hat{\pi}$ and $\hat{p}$. The solution is

$$\hat{\mu} = n^{-1} \sum_{\imath=1}^n \hat{\pi}(X_i)^{-1} I(X_i \le x) \{ \xi_i I(\Delta_i = \delta) - [\xi_i - \hat{\pi}(X_i)] \hat{p}(X_i) \},$$

which is a special case of an estimator that has been studied by Robins and Ritov (1997, Sections 7 and 8). Under our assumption that $P(\xi = 1 \mid X) = P(\xi = 1 \mid X_D)$, a natural estimator of $\pi(x) = \pi(x_D)$ is the empirical proportion of subjects with $\xi = 1$ among the subjects whose discretized value of $X = x_D$. A kernel estimator can be used for $p(x)$.

The above representation of the efficient influence curve for μ implies that if one estimates p inconsistently, then $\hat{\mu}$ is still consistent and asymptotically linear. This build-in protection against misspecification of p allows the construction of estimators of μ that are efficient at a chosen submodel for p and are always consistent and asymptotically linear. Such estimators typically have a better finite sample performance at the chosen submodel.

The estimator $\hat{\mu} = \hat{\mu}(x, \delta)$ could be used in place of $F_n(x, \delta)$ in the product integral formula (8) to provide an alternative efficient estimator of $S_T(t)$ having the efficient influence curve given by (16). Another way of computing (16) would be to apply the results of Robins and Rotnitzky (1992) directly to the known influence curve of $S_T(t)$ in the case of complete data (X, Δ). This would lead to yet another efficient estimator of $S_T(t)$ in terms of $\hat{\pi}$ and $\hat{p}$, along the lines used to construct $\hat{\mu}$.

The benefit of the Robins–Rotnitzky approach is that the computation of the efficient influence curve is relatively simple, and leads directly to an efficient estimator via standard estimating equation technology, without the need for the artificial reduced data model. Our approach, on the other hand, provides an understanding of the role of nonparametric maximum likelihood through the link to the bivariate censoring model. Moreover, the reduced data NPMLE has the attractive property that it 'solves' the efficient estimating equation at every π, and a consistent estimator of π is not required. Our reduced data model plays a similar role to the data reduction inherent in the kernel estimator $\hat{p}$ used in the Robins–Rotnitzky approach.

5 Numerical results

We now report the results of a small simulation study comparing the performance of the proposed estimator with that of Lo's (1991) estimator. The comparison is made in terms of mean integrated squared error (MISE).

The survival time T and the censoring time C are taken to be exponentially distributed with parameters 1.4 and 0.6 for 30% censoring, and 0.6 and 1.4 for 70% censoring, respectively. The sample size is set at $n = 100$. The coarsening mechanism is MCAR (required for Lo's estimator to be consistent), and the probability $\pi(x) = \pi$ that a failure indicator is non-missing is taken as $0.1, 0.2$ or 0.3. The partition consists of k points on a regular grid, with $k = 10, 30$ or 50.

Two artificial points are used in each region R_j having no complete observations: $(x_j, 0)$ and $(x_j, 1)$, where x_j is the midpoint of the interval $(a_j, a_{j+1}]$. The mass redistributed to these points is divided according to the proportions of censored and uncensored observations in the complete data, respectively.

The results are given in Table 1. The proposed estimator improves considerably upon Lo's estimator, the greatest gains being obtained when the proportion of missing failure indicators is high ($\pi = 0.1$). This was to be expected, of course, since the strength of our approach comes from the way it handles the missing data. Note also that although the performance of $\hat{S}_T$ is relatively insensitive to the choice of the partition when $k \geq 30$, it is significantly degraded under the coarsest grid ($k = 10$). The best performance is obtained using the finest grid, irrespective of the degree of censorship.

The effect of the partition is also readily seen by comparing Figures 1–3, which give plots of $\hat{S}_T$ based on simulated data, for different values of k. The closest fit to the underlying survival function (exponential with parameter 1.4) clearly corresponds to the finest partition. Figure 1 illustrates how the proposed estimator can adapt more efficiently to the missing data than Lo's estimator; see especially the region between $t = 0.1$ and $t = 0.5$. Figures 2 and 3 show how $\hat{S}_T$ deteriorates as the partition becomes coarser.

Table 1. Mean integrated squared error of the proposed estimator $\hat{S}_T$ and of Lo's estimator under the MCAR model $P(\xi = 1 \mid X, \Delta) \equiv \pi$.

	30% Censoring			70% Censoring		
	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
$\hat{S}_T$ $(k = 50)$	1.105	0.620	0.469	1.659	0.918	0.695
$\hat{S}_T$ $(k = 30)$	1.194	0.623	0.478	1.647	0.972	0.747
$\hat{S}_T$ $(k = 10)$	1.215	0.705	0.515	2.007	1.200	0.866
Lo	1.932	0.916	0.621	2.482	1.253	0.849
Ratio	1.75	1.48	1.32	1.50	1.36	1.22

NOTE: The MISE is expressed in units of 0.01 and is calculated over the interval $[0, 1]$. "Lo" refers to the second estimator of Lo (1991). "Ratio" refers to the ratio of the MISE of Lo's estimator to the MISE of $\hat{S}_T$ $(k = 50)$. Each MISE is based on 10,000 samples.

[Insert Figures 1–3 about here]

Figure 1. The estimator $\hat{S}_T$ compared with Lo's estimator for simulated data. The data were generated using the MCAR model in Table 1 with 30% censoring and $\pi = 0.2$. The partition uses $k = 100$ equispaced points over the interval $[0, 2]$. Sample size $n = 100$. The data points with missing failure indicators are represented by $\times$ and the remaining data points by $+$.

Figure 2. The estimator $\hat{S}_T$ based on $k = 60$ grid points over the interval $[0, 2]$. See the caption for Figure 1.

Figure 3. The estimator $\hat{S}_T$ based on $k = 20$ grid points over the interval $[0, 2]$. Same data as in Figure 1.

Appendix

Proof of Proposition 4.1. The main step of the proof is find the canonical gradient of μ in the submodel $\mathcal{M}(F)$ in which only F is unknown (π known). This submodel has tangent space

$$\mathbf{T}(F) = \overline{\mathrm{sp}}\{E(h(X, \Delta)|Y) \colon Eh(X, \Delta) = 0\}$$

formed as the closed linear span of the conditional expectations of all complete data scores $h(X, \Delta)$ given the observed data Y. Also, the orthogonal complement of $\mathbf{T}(F)$ is given by

$$\mathbf{T}(\pi) = \overline{\mathrm{sp}}\{\phi(Y) \colon E(\phi(Y)|X, \Delta) = 0\},$$

which is the tangent space in the submodel in which only π is unknown (F known), see Robins and Rotnitzky (1992) and Gill, van der Laan and Robins (1997).

Note that a gradient for μ in the submodel $\mathcal{M}(F)$ is

$$IC_\mu(Y) = \frac{\xi}{\pi(X)}\left(I(X \leq x, \Delta = \delta) - \mu\right).$$

To see this, consider a SOD submodel $\{P^{F_\epsilon}\} \subset \mathcal{M}(F)$ with score function $k(Y) = E(h(X,\Delta)|Y)$, where h is a bounded complete data score function and $F_\epsilon(du,\delta) = (1+\epsilon h(u,\delta))F(du,\delta)$. Then

$$\begin{aligned}\lim_{\epsilon\to 0}\left(\mu(F_\epsilon) - \mu(F)\right)/\epsilon &= E(h(X,\Delta)I(X \leq x, \Delta = \delta)) \\ &= E(h(X,\Delta)IC_\mu(Y)) \\ &= \langle IC_\mu, k\rangle,\end{aligned}$$

where the second equality above uses $E(IC_\mu(Y)|X,\Delta) = I(X \leq x, \Delta = \delta) - \mu$, which is a consequence of the CAR assumption.

The canonical gradient of μ in the submodel $\mathcal{M}(F)$ is the projection of IC_μ on $\mathbf{T}(F)$, which can be expressed as $IC_\mu - \Pi_{nu}(IC_\mu)$, where Π_{nu} is the projection on the 'nuisance' tangent space $\mathbf{T}(\pi)$, the orthogonal complement of $\mathbf{T}(F)$. For any function $\psi(Y) \in L_0^2(P)$ we have

$$\Pi_{nu}(\psi) = E(\psi(Y) \mid \xi, X) - E(\psi(Y) \mid X). \tag{17}$$

Indeed, the right hand side is a function of Y with conditional mean zero, given the complete data, so it belongs to $\mathbf{T}(\pi)$. Also, $\psi - \Pi_{nu}(\psi)$ is orthogonal to $\mathbf{T}(\pi)$, which can be seen by first taking the conditional expectation given (X,ξ) and then the conditional expectation given X, establishing (17). The application of (17) to IC_μ is straightforward, resulting in the expression given by IC_μ^*.

The final step is to show that the canonical gradient of μ in the submodel $\mathcal{M}(F)$ is also a gradient for μ in the full model. Let $k \in L_0^2(P)$ be bounded and consider the SOD submodel $P^\epsilon(dy) = P^{F_\epsilon,\pi_\epsilon}(dy) = (1+\epsilon k(y))P(dy)$. The score k can be expressed uniquely in the form $k = k_1 + k_2$, where $k_1 \in \mathbf{T}(F)$ and $k_2 \in \mathbf{T}(\pi)$. Then $\{F_\epsilon\}$ defines a SOD submodel for $\mathcal{M}(F)$ having score k_1, because k_1 does not depend on $\{\pi_\epsilon\}$ and k_2 does not depend on $\{F_\epsilon\}$. Thus, using the fact that μ does not depend on π, we have

$$\lim_{\epsilon\to 0}\left(\mu(P^\epsilon) - \mu(P)\right)/\epsilon = \lim_{\epsilon\to 0}\left(\mu(F_\epsilon) - \mu(F)\right)/\epsilon = \langle IC_\mu^*, k_1\rangle = \langle IC_\mu^*, k\rangle,$$

as required.

Acknowledgements. We thank Jamie Robins for pointing out the alternative approach based on Robins and Rotnitzky (1992). We also thank the Editor, Larry Brown, and the Associate Editor for helpful comments.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Bijlsma, F. (1994). Mortality among young people: causes and background. *Ned. Tijdschr. Geneeskd.*, **138** 2439–2443. (In Dutch)

Dinse, G. E. (1982). Nonparametric estimation for partially-complete time and type of failure data. *Biometrics* **38** 417–431.

Efron, B. (1967). The two sample problem with censored data. In *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4** 831–853.

Gijbels, I., Lin, D. Y. and Ying, Z. (1993). Non- and semi-parametric analysis of failure time data with missing failure indicators. *Preprint.*

Gill, R. D. (1989). Non- and semi- parametric maximum likelihood estimators and the von Mises method (part 1). *Scand. J. of Stat* **16** 97–124.

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

Gill, R. D., van der Laan, M. J. and Robins, J. M. (1997). *Coarsening at Random*, to appear in *Proceedings of the First Seattle Symposium on Biostatistics* (D.-Y. Lin, ed.), Springer, New York.

Gill, R. D., van der Laan, M. J. and Wellner, J. A. (1995). Inefficient estimators of the bivariate survival function for three models. *Ann. Inst. Henri Poincaré* **31** 545–597.

Heitjan D. F. and Rubin D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253.

Jacobsen, M. and Keiding, N. (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Statist.* **23** 774–786.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

Van der Laan, M. J. (1996a). Efficient and inefficient estimation in semiparametric models. CWI-tract 114, Centre of Computer Science and Mathematics, Amsterdam.

Van der Laan, M. J. (1996b). Efficient estimation in the bivariate censoring model and repairing NPMLE. *Ann. Statist.* **24**, 596–627

Lapidus G., Braddock M., Schwartz R., Banco L., and Jacobs L. (1994). Accuracy of fatal motorcycle-injury reporting on death certificates. *Accid. Anal. Prev.*, **26** 535–542.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

Lo, S.-H. (1991). Estimating a survival function with incomplete cause-of-death data. *J. Multvariate Anal.* **39** 217–235.

McKeague, I. W. and Subramanian, S. (1996). Product-limit estimators and Cox regression with missing censoring indicators. *Preprint.*

Newey, W. K. (1990). Semiparametric efficiency bounds. *J. Applied Econometrics* **5** 99–135.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.* **1** 502–527.

Pruitt, R. C. (1991). On negative mass assigned by the bivariate Kaplan-Meier estimator. *Ann. Statist.* **19** 443–453.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16** 285–319.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In: *AIDS Epidemiology – Methodological Issues* (N. Jewell, K. Dietz, V. Farewell, eds.), 279–331, Birkhäuser, Boston.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–590.

Van der Vaart, A. W. (1991). Efficiency and Hadamard differentiable functionals. *Scand. J. Statist.* **18** 63–75.

Wellner, J. A. (1982). Asymptotic optimality of the product limit estimator. *Ann. Statist.* **10** 595–602.

Figure 1

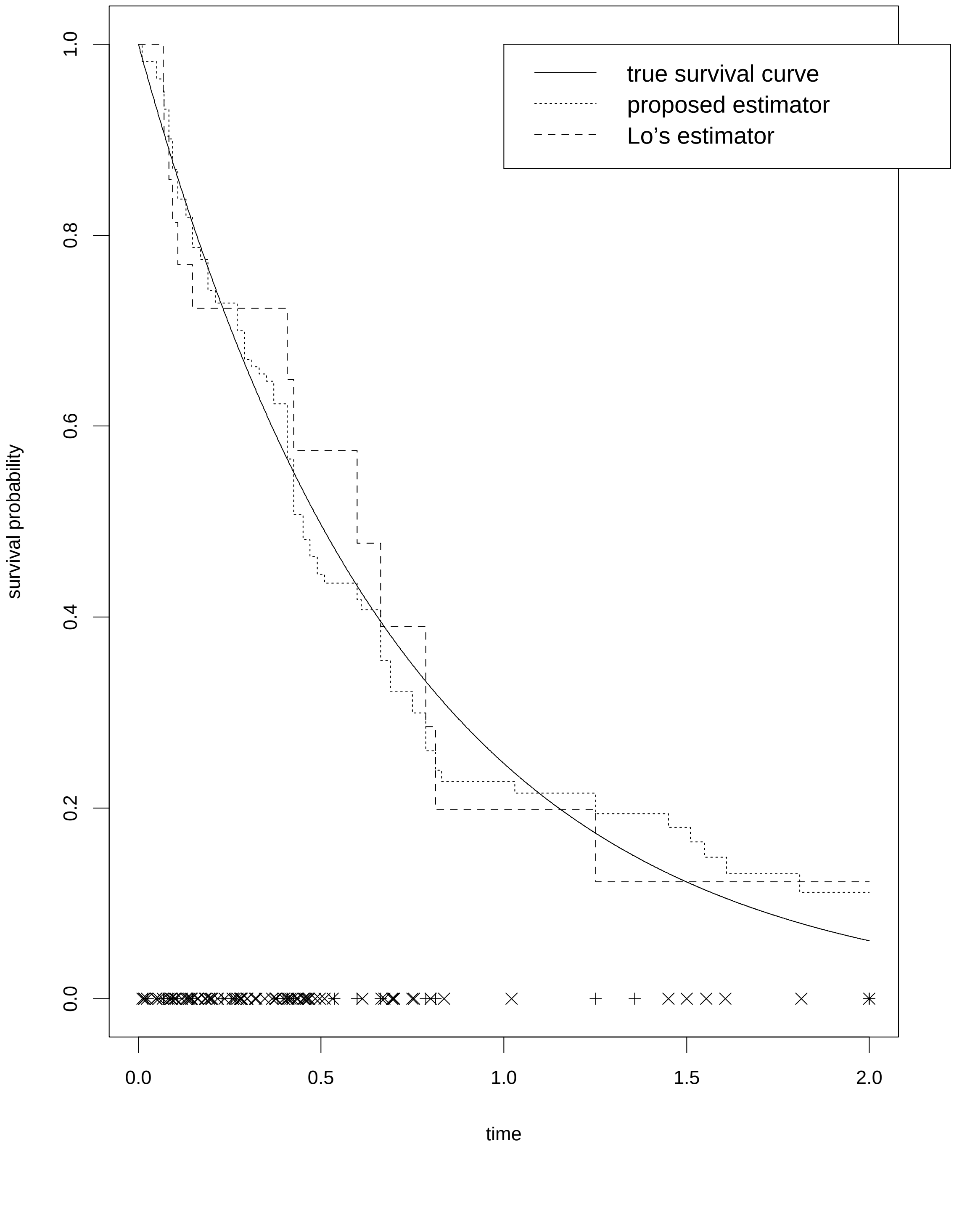

true survival curve
proposed estimator
Lo's estimator
survival probability
time
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0

Figure 2

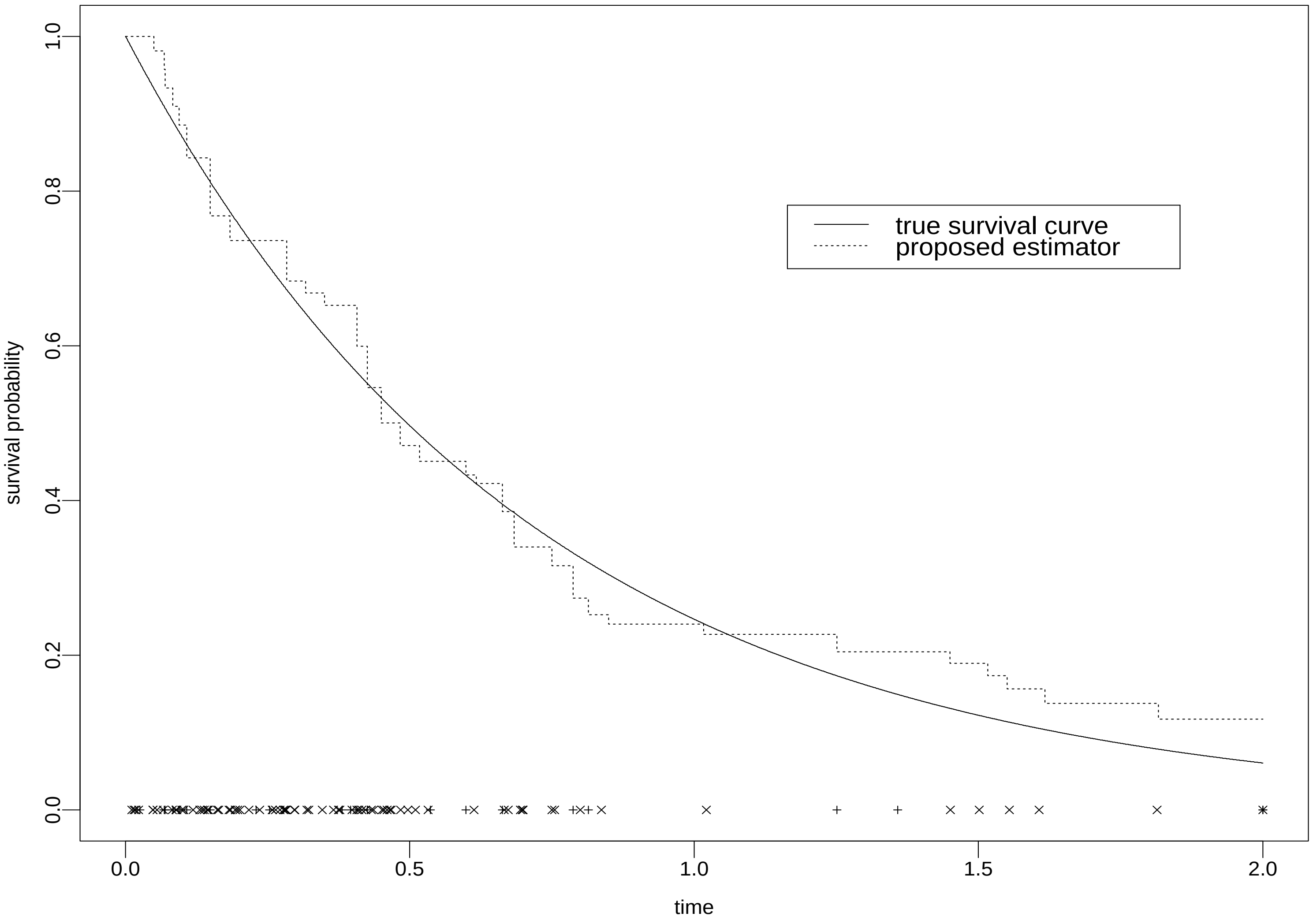
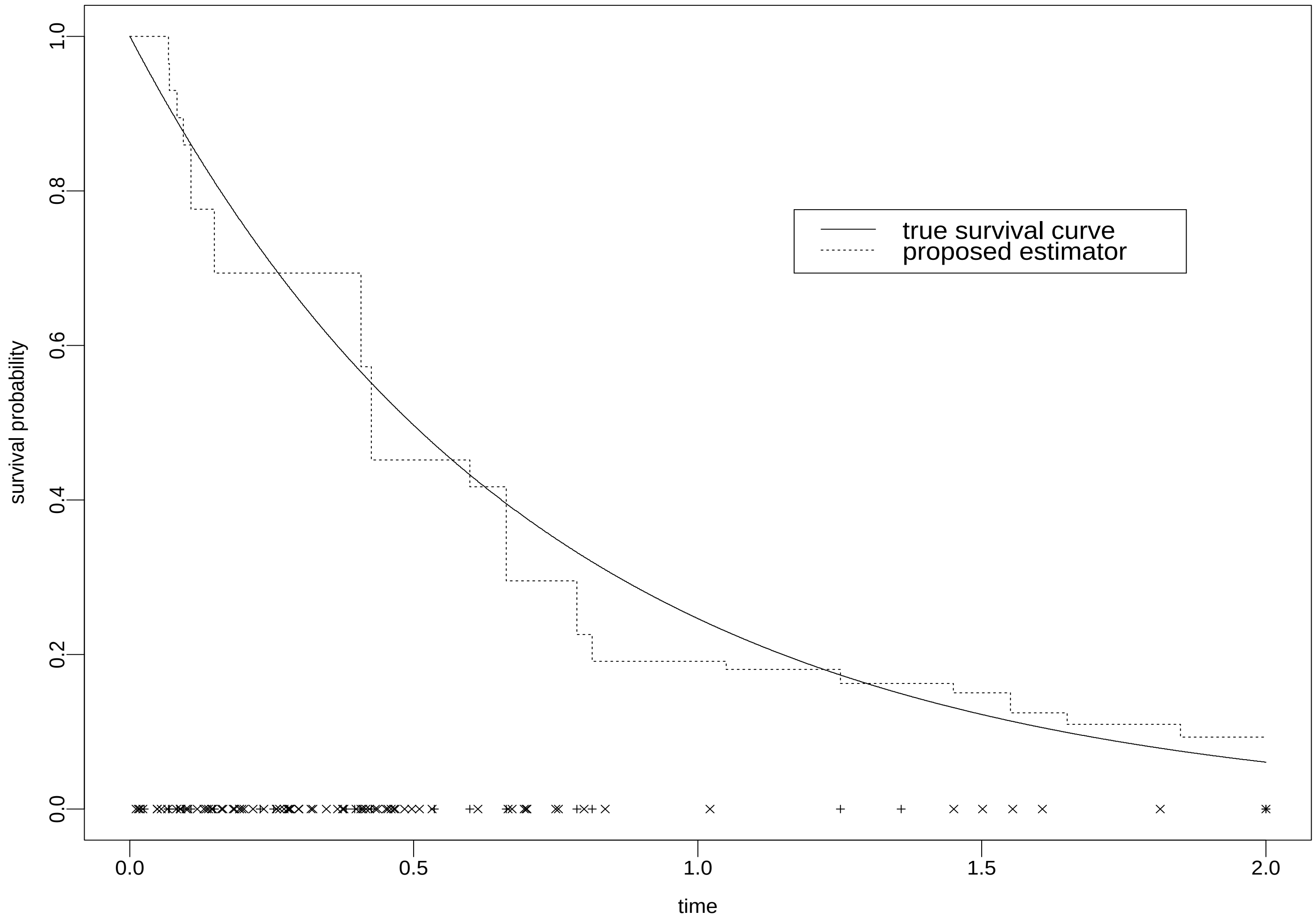

true survival curve
proposed estimator
survival probability
time
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0

Figure 3


true survival curve
proposed estimator
survival probability
0.0
0.2
0.4
0.6
0.8
1.0
time
0.0
0.5
1.0
1.5
2.0