The Annals of Statistics
1996, Vol. 24, No. 2, 596–627

EFFICIENT ESTIMATION IN THE BIVARIATE CENSORING MODEL AND REPAIRING NPMLE

By Mark J. van der Laan

University of California, Berkeley

The NPMLE in the bivariate censoring model is not consistent for continuous data. The problem is caused by the singly censored observations. In this paper we prove that if we observe the censoring times or if the censoring times are discrete, then a NPMLE based on a slightly reduced data set, in particular, we interval censor the singly censored observations, is asymptotically efficient for this reduced data and moreover if we let the width of the interval converge to zero slowly enough, then the NPMLE is also asymptotically efficient for the original data. We are able to determine a lower bound for the rate at which the bandwidth should converge to zero. Simulation results show that the estimator for small bandwidths has a very good performance. The efficiency proof uses a general identity which holds for NPMLE of a linear parameter in convex models. If we neither observe the censoring times nor the censoring times are discrete, then we conjecture that our estimator based on simulated censoring times is also asymptotically efficient.

1. Introduction. In this paper we are concerned with estimation of the bivariate survival function of two dependent survival times. For example, one might be interested in estimation of the bivariate survival function of twins with a certain disease. Suppose that for each twin one observes two calendar times (U_1, U_2) at which the disease started for twin 1 and twin 2 and that one keeps track of the bivariate survival time (T_1, T_2) of the twin measured from (U_1, U_2) until a given calendar point t_0. At t_0 one wants to use the available data to estimate the bivariate survival function of (T_1, T_2). In this setting, T_1 will be potentially (i.e., if $T_1 > C_1$) right randomly censored at the observed censoring time $C_1 = t_0 - U_1$ and similarly T_2 will be potentially right randomly censored at the observed censoring time $C_2 = t_0 - U_2$.

In this paper we propose an estimator for the *bivariate survival function* of $T = (T_1, T_2)$ based on *bivariate right randomly censored data*, assuming that the censoring times $C = (C_1, C_2)$ are always observed, as in the example above, or assuming that the censoring times are discrete. We prove asymptotic efficiency of this estimator. In the case that the censoring times are not observed for the failures and the censoring times are not discrete, then we propose a simulation of the unobserved censoring variables and conjecture

Received November 1993; revised March 1995.

AMS 1991 *subject classifications.* Primary 62G07; secondary 62F12.

Key words and phrases. Bivariate censorship, self-consistency equations, efficiency.

(no proof, but heuristic argument) that our estimator based on these simulated censoring variables will also be asymptotically efficient.

We found it useful not to use a special notation for vectors in $\mathbb{R}^2$; if we do not mean a vector, this will be clear from the context. So if we write T we usually mean $T = (T_1, T_2) \in \mathbb{R}^2_{\geq 0}$ and if we write $\leq$, $\geq$, $<$ and $<$, then this should hold componentwise: for example, if $x, y \in \mathbb{R}^2$, then $x \leq y \Leftrightarrow x_1 \leq y_1$, $x_2 \leq y_2$. We will write T_i, $i = 1, \dots, n$, as notation for n i.i.d. bivariate survival times with the same distribution as T, while we write T_1 and T_2 for the components of T.

Bivariate right randomly censored data can be modelled as follows: T is a positive bivariate lifetime vector with bivariate distribution F_0 and survival function S_0; $F_0(t) \equiv \Pr(T \leq t)$ and $S_0(t) \equiv \Pr(T > t)$. Let C be a positive bivariate censoring vector with bivariate distribution G_0 and survivor function H_0; $G_0(t) \equiv \Pr(C \leq t)$ and $H_0(t) \equiv \Pr(C > t)$. Assume that T and C are independent; $(T, C) \in \mathbb{R}^4$ has distribution $F_0 \times G_0$. Let (T_i, C_i), $i = 1, \dots, n$, be n independent copies of (T, C). We observe the following many-to-one mapping Φ of (T_i, C_i):

$$Y_i \equiv \Phi(T_i, C_i) \equiv (T_i \wedge C_i, I(T_i \leq C_i)) \equiv \left(\tilde{T}_i, D_i\right),$$

with components given by

$$\tilde{T}_{ij} = \min\{T_{ij}, C_{ij}\}, \qquad D_{ij} = I(T_{ij} \leq C_{ij}), \qquad j = 1, 2.$$

In other words, the minimum and indicator are taken componentwise, so that $\tilde{T}_i \in [0, \infty)^2$ and $D_i \in \{0, 1\}^2$ are bivariate vectors. The observations Y_i are elements of $[0, \infty)^2 \times \{0, 1\}^2$ and $Y_i \sim P_{F_0, G_0} = (F_0 \times G_0)\Phi^{-1}$. We are concerned with estimation of S_0.

Each observation Y_i tells us that $(T_i, C_i) \in B(Y_i) \equiv \Phi^{-1}(Y_i) \subset \mathbb{R}^2 \times \mathbb{R}^2$, where $B(Y_i) = B(Y_i)_1 \times B(Y_i)_2$ for the projections $B(Y_i)_1 \subset \mathbb{R}^2$ and $B(Y_i)_2 \subset \mathbb{R}^2$ of $B(Y)$ on the T and C space, respectively. The kind of region $B(Y_i)_1$ for T_i (point, vertical half-line, horizontal half-line, quadrant) generates a classification of the observations $Y_i = (\tilde{T}_i, D_i)$ in four groups:

Uncensored. If $D_i = (1, 1)$, then the observation Y_i is called uncensored and it tells us that $T_i \in B(Y_i)_1 = \{\tilde{T}_i\}$. So $T_i = \tilde{T}_i$.
Singly censored. If $D_i = (0, 1)$ or $D_i = (1, 0)$, then the observation Y_i is called singly censored. If $D_i = (0, 1)$, then it tells us that $T_i \in B(Y_i)_1 = \{(\tilde{T}_{i1}, \infty) \times \{\tilde{T}_{i2}\}\}$ (horizontal half-line), and if $D_i = (1, 0)$, that $T_i \in B(Y_i)_1 = \{\{\tilde{T}_{i1}\} \times (\tilde{T}_{i2}, \infty)\}$ (vertical half-line).
Doubly censored. If $D_i = (0, 0)$, then the observation Y_i is called doubly censored and it tells us that $T_i \in B(Y_i)_1 = \{\tilde{T}_{i1}, \infty) \times (\tilde{T}_{i2}, \infty)\}$ (upper quadrant).

The uncensored observations are the *complete* observations and the singly censored and doubly censored are incomplete observations. An NPMLE solves the self-consistency equation [Efron, (1967); Gill, (1989)] and a solution of the self-consistency can be found with the *EM algorithm* [Dempster, Laird and

Rubin, (1977); Turnbull, (1976)], which does in fact nothing else than iterate the self-consistency equation. In the EM algorithm each observation Y_i gets mass $1/n$, which it needs to redistribute over $B(Y_i)_1$ in a self-consistent way. The incomplete observations Y_i need to get information from the observed T_i about how to redistribute their mass $1/n$ over $B(Y_i)_1$, and for this purpose they need complete observations in $B(Y_i)_1$. The EM algorithm listens only to the observations with a region $B(Y_j)_1$ which has an intersection with $B(Y_i)_1$. It is only possible to have uncensored observations in $B(Y_i)_1$ if $F_0(B(Y_i)_1) > 0$, which is typically not true for the singly censored observations. If F_0 is continuous, then the probability that T falls on a line is zero. Indeed it is well known that the NPMLE for continuous data is not consistent [Tsai, Leurgans and Crowley (1986)].

Many proposals for estimation of the bivariate survival function in the presence of bivariate censored data have been made. Because the usual NPML and self-consistency principle do not lead to a consistent estimator for continuous data, most proposals are explicit estimators based on representations of the bivariate survival function in terms of distribution functions of the data: among these proposals are Tsai, Leurgans and Crowley (1986), Dabrowska (1988, 1989), Burke (1988), the so-called Volterra estimator of P. J. Bickel [see Dabrowska (1988)] and Prentice and Cai (1992a, b).

Prentice and Cai (1992a) proposed a nice estimator which is closely related to Dabrowska's estimator except that it also uses the Volterra structure suggested by Bickel. Dabrowska's multivariate product-limit estimator, based on a very clever representation of a multivariate survival function in terms of its conditional multivariate hazard measure, and the Prentice–Cai estimator have a better practical performance in comparison w.r.t. the Volterra, pathwise estimator and the estimator proposed in Tsai, Leurgans and Crowley (1986) [see Bakker (1990), Prentice and Cai (1992b), Pruitt (1993) and Chapter 8 of van der Laan (1996)]. It is expected that the Dabrowska and Prentice–Cai estimators are certainly better than the other proposed explicit estimators. Besides, these two estimators are smooth functionals of the empirical distributions of the data so that such results as consistency, asymptotic normality, correctness of the bootstrap, consistent estimation of the variance of the influence curve and LIL all hold by application of the functional delta method: see Gill (1992), Gill, van der Laan and Wellner (1993) and van der Laan (1990). In Gill, van der Laan and Wellner (1993), Dabrowska's results about her estimator are reproved and new ones are added by application of the functional delta method, and similar results are proved for the Prentice–Cai estimator. Moreover, it is proved that the Dabrowska and Prentice–Cai estimators are efficient in the case that T_1, T_2, C_1, C_2 are all independent.

All the estimators proposed above are ad hoc estimators which are not asymptotically efficient [except at some special points (F, G)]. This is also reflected by the fact that most of these estimators put a nonnegligible proportion of negative mass to points in the plane [Pruitt, (1991a); Bakker (1990)].

Pruitt (1991b) proposed an interesting implicitly defined estimator which is the solution of an ad hoc modification of the self-consistency equation. Pruitt points out why the original self-consistency equation has a wide class of solutions and his estimator tackles this nonuniqueness problem in a very direct way by estimating conditional densities over the half-lines implied by the singly censored observations. Uniform consistency, $\sqrt{n}$-weak convergence and the bootstrap for his normalized estimator are proved in van der Laan (1993c) and Chapter 7 of van der Laan (1996) under some smoothness assumptions which are due to the fact that his estimator uses kernel density estimators. However, this estimator is not asymptotically efficient (except at some special points) and its practical performance is (somewhat surprisingly) worse, except at the tail where one hardly finds uncensored observations [as shown in Chapter 8 of van der Laan (1996)] than the Dabrowska and Prentice and Cai estimators. In the case that the sampling distribution is smooth, Pruitt's estimator appeared (as expected) to improve by using large bandwidths.

As noticed by Pruitt (1991b), the inconsistency of the NPMLE is due to the fact that the singly censored observations imply half-lines for T which do not contain any uncensored observations. Based on this understanding we propose in Section 2 to (slightly) interval censor the singly censored observations in the sense that we replace the uncensored component (say) T_{1i} of the singly censored observations by the observation that T_{1i} lies in a small predetermined interval around T_{1i}. These intervals are determined by a grid partition π_h with a width $h = h_n$. Now, for these interval censored singly censored observations Y_i^h the regions $B(Y_i^h)_1$ are strips with contain with positive probability uncensored observations.

The interval censoring of the singly censored observations causes one problem. The joint likelihood for F and G does not factorize anymore in an F-term and a G-term, which is due to the fact that the region for (T, C) implied by the interval censored singly censored observations is not rectangular anymore. This tells us that for computing the NPMLE of F we also need to estimate G by maximizing over G. Because of similar reasons as for the NPMLE of F, the NPMLE of G will only be good if we do a symmetric reduction (lines should be strips for C as well as for T). In other words, an extra reduction of the data will be necessary. Because the involvement of G in computing the NPMLE F_n^h certainly complicates the analysis and makes the estimator more computer intensive, we decided to choose a reduction of the data which recovers the orthogonality (i.e., factorization of the likelihood), while at the same time, as will appear, not losing asymptotic efficiency. The further reduction is based on the insight that if G_0 is purely discrete on π^h, then $p^h_{F_0, G_0}(\cdot, d)$ factorizes, as shown in Section 2. Hence if the actual G is discrete, then by choosing π_h (which can be done with probability tending to 1 if the number of observations converges to infinity) so that censoring variables lie on the grid π^h, we still have factorization of the likelihood. If the actual G is not discrete, but we observe $C_1, \dots, C_n$, then we can (1) discretize (to the left) these C_i's to C_i^h on π_h, (2) replace the original Y_i's by $\Phi(T_i, C_i^h)$

and (3) replace the singly censored observations of $\Phi(T_i, C_i^h)$ by interval singly censored observations Y_i^h. In this way, we construct new observations Y_i^h for which the density factorizes in an F and G part.

This further reduction leads also to a good practical estimator as appears in the simulations in Chapter 8 of van der Laan (1996). Its performance for a small value of h is better than the Dabrowska, Prentice and Cai and Pruitt estimators, except at the tail and under complete independence of T_1, T_2, C_1, C_2. We show that if $h_n \to 0$ at a rate slower than $n^{-1/18}$, then the estimator is asymptotically efficient, and if h is fixed, then one still has an asymptotically normal estimator with an asymptotic variance arbitrarily close (small h) to the asymptotic optimal variance. Our derived lower bound is purely of theoretical value since it shows the existence of rates $h = h_n$ for which the estimator is efficient, but quicker rates will also provide efficient estimators. Obtaining theoretical insight about the precise rate at which h_n should converge to zero if $n \to \infty$ is very hard and not very useful because constants are not available. Simulations show that if $n = 200$ and the range of the observations is transformed back to $[0,1] \times [0,1]$, then choosing the width of the strips equal to $h = 0.02$ gives a very good estimator, so a few observations in each strip is already effective. The estimator gets essentially worse if we increase h independent of the smoothness of (F, G). This bandwidth behavior is explained as follows. A large h means a large reduction of the data and hence an increase in asymptotic variance. On the other hand, we needed a $h > 0$ so that the EM algorithm is able to use the uncensored observations in the strips around the singly censored half-lines for obtaining a redistribution of mass $1/n$ over the half-lines. However, our primary interest is not the distribution over the half-line, but the survival function itself (which integrates over the distributions over the half-lines), which explains that a smaller bandwidth than the one advised by density estimation literature will suffice. In practice, a sensible method for programming a sensible grid π^h would be to set the width for the horizontal axis equal to a fixed proportion of the cross-validated bandwidth h_1^* using the observed T_{1i}'s and similarly compute the vertical width.

If we do not observe C_i, then we can draw a C_i' from a conditional distribution of C, given $C \in B(Y_i)_2$, and consider these simulated C_i' as the observed C_i's above. For example, if we observe that $C_{1i} \in (T_{1i}, \infty)$, we set $C_{1i}' = T_{1i} + U_i$, where U_i is a realization from a known distribution on $(0, \tau]$. Then $Y_i' = \Phi(T_i, C_i') = Y_i$, but we now observe C_i'. C_i', $i = 1, \dots, n$, are still i.i.d., but C_i' depends on T_i only through Y_i. However, if the density of C, given $T = t$, depends only on T through $Y = \Phi(C, T)$, then the censoring mechanism satisfies *coarsened at random* [see Heitjan and Rubin (1991)], which implies that the density of Y still factorizes, where the F part of the density of Y' is still the same as the F part of the density of Y, that is where C and T are independent. Consequently, we have that the efficient influence function for estimating F based on Y_i' equals the efficient influence function for estimating F based on Y_i. Hence, if we construct an estimator of F based on (C_i', Y_i') which is efficient, then it is also efficient for the original data Y_i. In other words, without any loss we arranged that we have available a set of observed C_i''s. However, because of the dependence between C' and T, the likelihood does not factorize anymore for the data $\Phi(T, C_h')$ based on the discretized C_h', so that our proposed estimator is not a NPMLE for the interval censored $\Phi(T, C_h')$ and hence has a bias. On the other hand, we let h converge to zero when the number of observations converges to infinity so that this bias converges to zero. Therefore, we conjecture (no proof) that our estimator based on these simulated C' is asymptotically efficient if $h = h_n$ converges to zero at an appropriate rate (not too slow and not too quick). In the sequel it will be assumed that the C_i's are observed or that G_0 is discrete.

We will call the MLE based on a reduction (or call it a slight transformation) of the data a "sequence of reductions" MLE and will abbreviate it with SOR-MLE. It is a general way to repair the real NPMLE in problems where the real NPMLE does not work. If one understands why the usual NPMLE does not work, then one can hope to find a natural choice for the transformation of the data. Moreover, if we do not lose the identifiability, we have for a *fixed* transformation, consistency, asymptotic normality and efficiency of the NPMLE among estimators based on the transformed data, while we obtain efficiency by letting the amount of reduction of the data converge to zero slowly enough if n converges to infinity.

In the next section we will define, in detail, the SOR-MLE for the bivariate censoring model. In Section 3 we will give an outline of the efficiency proof, which is based on an *identity* for the SOR-MLE which holds in general for convex models which are linear in the parameter [van der Laan (1993)]. This identity gives a direct link between efficiency of the SOR-MLE and properties of the efficient influence function corresponding with the data Y_h. In Section 4 we prove the ingredients of this general proof; the crucial lemmas of this section are proved in Section 6. We summarize the results in Section 5.

For validity of the nonparametric and semiparametric bootstrap we refer to Section 4.7 in van der Laan (1996); these results follow easily from the identity approach which we follow.

2. SOR-MLE for the bivariate censoring model.

Our original observations are given by

$$\left(\tilde{T}_i, D_i\right) = \Phi(T_i, C_i) \sim P_{F_0, G_0}(\cdot, \cdot), \qquad i = 1, \dots, n.$$

Let $P_{11}(\cdot) = P_{F_0, G_0}(T \leq \cdot, D = (1,1))$ be the subdistribution of the (doubly) uncensored observations and similarly let P_{01}, P_{10} and P_{00} be the subdistributions corresponding to $D = (0,1)$, $D = (1,0)$ and $D = (0,0)$, respectively. Then

$$\begin{aligned} P_{F_0, G_0}(\cdot, D = d) &= P_{11}(\cdot) I(d = (1,1)) + P_{01}(\cdot) I(d = (0,1)) \\ &\quad + P_{10}(\cdot) I(d = (1,0)) + P_{00}(\cdot) I(d = (0,0)). \end{aligned} \tag{1}$$

Let $f_0 \equiv dF_0/d\mu$, for some finite measure μ which dominates F_0. Similarly, let $G_0 \ll \nu$ with density g_0. $S_0(x_1, \cdot)$ generates a measure on $\mathbb{R}_{\geq 0}$. This measure is absolutely continuous w.r.t. $\mu((x_1,\infty),\cdot)$, the marginal of the measure μ restricted to $(x_1,\infty)\times\mathbb{R}_{\geq 0}$. Now, we define $S_{02}(x_1, x_2) \equiv -S_0(x_1, dx_2)/\mu((x_1,\infty), dx_2)$ as the Radon–Nikodym derivative and similarly we define $S_{01}(x_1, x_2) \equiv -S_0(dx_1, x_2)/\mu(dx_1,(x_2,\infty))$, $H_{01}(x_1, x_2) \equiv -H_0(dx_1, x_2)/\nu(dx_1,(x_2,\infty))$ and $H_{02}(x_1, x_2) \equiv -H_0(x_1, dx_2)/\nu(x_1,\infty), dx_2)$. Then the density p_{F_0,G_0} of P_{F_0,G_0} w.r.t. $(\mu\times\nu)\Phi^{-1}$ is given by

$$\begin{aligned}
p_{F_0,G_0}(x_1, x_2, d) &= f_0(x)H_0(x)I(d=(1,1)) \\
&\quad + S_{01}(x_1, x_2)H_{02}(x_1, x_2)I(d=(1,0)) \\
&\quad + S_{02}(x_1, x_2)H_{01}(x_1, x_2)I(d=(0,1)) \\
&\quad + S_0(x)g_0(x)I(d=(0,0)) \\
&\equiv p_{11}(x)I(d=(1,1)) + p_{10}(x)I(d=(1,0)) \\
&\quad + p_{01}(x)I(d=(0,1)) + p_{00}(x)I(d=(0,0)) \\
&= \sum_{\delta\in\{1,0\}^2} p_\delta(x)I(d=\delta).
\end{aligned} \tag{2}$$

Suppose that we observe C_i and $(\tilde{T}_i, D_i)$, $i=1,\dots,n$. We will transform $(\tilde{T}_i, D_i)$ and base our NPMLE on the transformed data. The transformation depends on a *grid*. For this purpose, let $\pi^h = (u_k, v_l)^h$ be a nested grid in $h = h_n$ of $[0,\tau]$ which depends on a scalar $h = h_n$ in the following way: $\varepsilon h_n < u_{k+1} - u_k < Mh_n$, where ε and M are independent of n, k, and similarly for $v_{l+1} - v_l$. With nested we mean that the grid points of π_{h_n} are a subset of the grid points of $\pi_{h_{n+m}}$ (we use this in order to make martingale arguments work for conditional expectations, given increasing sigma fields). In other words, the grid must have a width between εh_n and Mh_n. This tells us that the grid π^h has (in order of magnitude) $1/h_n^2$ points (u_k, v_l). Let $R_{k,l} \equiv (u_k, u_{k+1}]\times(v_l, v_{l+1}]$.

Move each C_i to the left lower corner (u_k, v_l) of the rectangle $R_{k,l}$ of π^h which contains C_i. Denote these discretized C_i with C_i^h. Then $C_i^h \sim G_h$ where G_h is the step function with jumps on π^h corresponding to G_0:

$$P\big(C^h = (u_k, v_l)\big) = \int_{R_{k,l}} dG_0(c).$$

Consider now the n i.i.d. observations

$$Y_i\big(T_i, C_i^h\big) = \Phi\big(T_i, C_i^h\big) \sim P_{F_0,G_h}.$$

Notice that we are able to observe these $Y_i(T_i, C_i^h)$ because we only need to know $Y_i(T_i, C_i)$. If $h = h_n$ converges to zero, then the distribution of $\Phi(T, C^h)$ converges to the distribution of $\Phi(T, C)$.

For convenience we will denote $\Phi(T_i, C_i^h)$ with $Y_i = (\tilde{T}_i, D_i)$, again, and still use the notation p_{11}, p_{10}, p_{01} and p_{00}, suppressing the dependence on h, but we have to realize that all censored $\tilde{T}_{1i}$ equal u_k, for some k, and $\tilde{T}_{2i}$

equal v_l, for some l. Now we can define the *reduced data* $(\tilde{T}_i, D_i)^h$, which we will use for our estimator:

$$Y_i^h = \left(\tilde{T}_i, D_i\right)^h = \Phi^h\left(T_i, C_i^h\right) \equiv \mathrm{Id}^h\left(\left(\tilde{T}_i, D_i\right)\right) = \mathrm{Id}^h\left(\Phi\left(T_i, C_i^h\right)\right),$$

where Id^h is a many-to-one mapping on the data $(\tilde{T}_i, D_i)$, which is defined as follows:

$$\begin{aligned}
&\mathrm{Id}^h(\tilde{T}, D) = (\tilde{T}, D) \quad \text{if } D = (1,1),\\
&\mathrm{Id}^h(\tilde{T}, D) = \left(\left(u_i, \tilde{T}_2\right), D\right) \quad \text{for } u_i \text{ s.t. } \tilde{T}_1 \in \left(u_i, u_{i+1}\right] \text{ if } D = (1,0),\\
&\mathrm{Id}^h(\tilde{T}, D) = \left(\left(\tilde{T}_1, v_j\right), D\right) \quad \text{for } v_j \text{ s.t. } \tilde{T}_2 \in \left(v_j, v_{j+1}\right] \text{ if } D = (0,1),\\
&\mathrm{Id}^h(\tilde{T}, D) = (\tilde{T}, D) \quad \text{if } D = (0,0).
\end{aligned}$$

Notice that Id^h equals the identity for the uncensored and doubly censored observations and it groups all singly censored observations $(T_1, C_2, I(T_1 \leq C_1) = 1, I(T_2 \leq C_2) = 0)$ with $T_1 \in (u_k, u_{k+1}]$ to one observation and similarly with the singly censored observations with $D = (0,1)$. We used the notation Id^h (Id from identity) because, for $h \to 0$ (in other words, if the partition gets finer), this transformation converges to the identity mapping. We will still call the Y^h with $D = (1,0)$ and $D = (0,1)$ singly censored observations, in spite of the fact that they are really censored singly censored observations. Y_i^h are i.i.d. observations with a distribution which is indexed by the (same as for Y_i) parameters F_0 and G_h.

To be more precise, we have

$$Y^h \sim P^h_{F_0, G_h}(\cdot, \cdot),$$

where

$$\begin{aligned}
P^h_{F_0, G_h}(x, D = d) &= P_{11}(\cdot) I(d = (1,1)) + P^h_{01}(\cdot) I(d = (0,1))\\
&\qquad + P^h_{10}(\cdot) I(d = (1,0)) + P_{00}(\cdot) I(d = (0,0)),
\end{aligned} \tag{3}$$

where the density $p^h_{F_0}$ of $P^h_{F_0, G_h}$ w.r.t. $(\mu \times \nu_h)\Phi_h^{-1}$, ν_h being the counting measure on π_h, is given by

$$\begin{aligned}
p_{11}(y_1, y_2) &= f_0(y_1, y_2) H_h(y_1, y_2)\\
p_{00}(v_k, v_l) &= S_0(v_k, v_l) g_h(v_k, v_l)
\end{aligned}$$

and

$$\begin{aligned}
p^h_{01}(v_k, v_l) &= \int_{(v_l, v_{l+1}]} p_{01}(v_k, y_2)\, \mu((v_k, \infty), dy_2)\\
&= \int_{(v_l, v_{l+1}]} S_{02}(v_k, y_2) H^h_{01}(v_k, v_l)\, \mu((v_k, \infty), dy_2)\\
&= F_0\left((v_k, \infty), (v_l, v_{l+1}]\right) H^h_{01}(v_k, v_l).
\end{aligned}$$

Similarly, $p^h_{10}(u_k, y_2) = S_{01}((v_k, v_{k+1}], v_l) H^h_{02}(v_k, v_l)$. Notice that $p^h_0(\cdot, d)$, $d \neq (1,1)$, is discrete on π_h. The independence between C_h and T and the fact that C_h is discrete on π_h implied that the density $p^h_{F_0}(\cdot, d)$ also factorized for $d = (1, 0)$ and $d = (0, 1)$.

Let P^h_n be the empirical distribution function based on n i.i.d. $Y^h_i(T_i, C^h_i) \sim P^h_{F_0, G_h}$, which is the distribution of the data corresponding to $T \sim F_0$, $C \sim G_h$, where G_h is discrete on the grid π^h and the singly censored observations are interval censored by Id^h (i.e., half-lines are grouped to strips). Let $\{x_1, \dots, x_{m(n)}\}$ consist of the uncensored T_i and one point of each $B(Y_j)_1$ which does not contain uncensored T_i. Let μ_n be the counting measure on $\{x_1, \dots, x_{m(n)}\}$. Now, we let $\mathscr{F}(\mu_n)$ be the set of all distributions which are absolutely continuous w.r.t. μ_n.

We define our SOR-MLE F^h_n of F_0 which we will analyze:

$$F^h_n = \arg \max_{F \in \mathscr{F}(\mu_n)} \int \log\left(p^h_{F, G_h}\right) dP^h_n, \tag{4}$$

where the maximum can be determined without knowing G_h by maximizing the term which only depends on F. We define S^h_n as the survival function corresponding to F^h_n.

2.1. *Existence and uniqueness of the SOR-MLE and EM equations.* In Lemma 4.1 in van der Laan (1995), for a general class of missing data models, it is proved that the MLE over all F with support $\{x_1, \dots, x_{m(n)}\}$ exists and is unique if the following two assumptions hold: $H_0 > \delta > 0$ F_0 a.e. and $F_0(B(Y^h_i)_1) > 0$, for all censored Y^h_i $[D = (1, 0),\ D = (0, 1),\ D = (0, 0)]$. This holds if all data live on a rectangle $[0, \tau] \subset \mathbb{R}_{\geq 0}$, where τ is such that $H_0(\tau) > 0$, $S_0(\tau -) > 0$, $F_0(\tau) = 1$, $F_0(T_1 \in [u_i, u_{i+1}],\ T_2 > \tau_2) > 0$ and $F_0(T_1 > \tau_1,\ T_2 \in [v_j, v_{j+1}]) > 0$, for all grid points (u_i, v_j). By making all observations $\tilde{T}_i \in [0, \tau]^c$ uncensored at the projection point on the edge of $[0, \tau]$, we obtain truncated observations with distribution $P^h_{F^\tau_0, G_h}$, where F^τ_0 equals F_0 on $[0, \tau)$, but puts all ($= 1$) its mass on $[0, \tau]$. This means that our efficiency result proves efficiency for data reduced to $[0, \tau]$. For obtaining full efficiency, we can let $\tau = \tau_n$ converge slowly enough to infinity, for $n \to \infty$. In our analysis this will mean an extra singularity of magnitude $1/H(\tau_n)$ and therefore our analysis can be straightforwardly extended to this case.

Let $g \in L^2(F^h_n)$ have finite supnorm. We will use the notation $F(g) = \int g\, dF$. We have that $dF^h_{n,\varepsilon} = (1 + \varepsilon(g - F^h_n(g)))\, dF^h_n$, $\varepsilon \in (-\delta, \delta)$, $\delta > 0$ small enough, is a one-dimensional submodel through the MLE dF^h_n and hence by definition of F^h_n,

$$\varepsilon \to \int \log\left(p^h_{F^h_{n,\varepsilon}, G_h}\right) dP^h_n$$

is maximized at $\varepsilon = 0$. Consequently, the derivative of this real-valued function on $(-\delta, \delta)$ at $\varepsilon = 0$ equals zero, so that exchanging integration and differentiation provides us with

$$P^h_n\left(A^h_{F_n}\left(g - F^h_n(g)\right)\right) = 0 \quad \text{for all } g \in L^2\left(F^h_n\right) \text{ with } \|g\|_\infty < \infty, \tag{5}$$

where the so-called score operator A_F^h for a distribution function F is given by

$$A_F^h: L^2(F) \to L^2\left(P_{F,G_h}^h\right): g \mapsto E_F(g(T)|Y^h).$$

The form of the score operators follows from the general fact that the score operator in missing data models equals the conditional expectation operator [see Gill (1989); Bickel, Klaassen, Ritov and Wellner (hereafter BKRW) (1993), Section 6.6]. In particular, by setting $g(T) = I_{(0,t]}(T)$ in (5) one obtains the well known self-consistency equation [Efron (1967)]

$$F_n^h(t) = \frac{1}{n}\sum_{i=1}^{n} P_{F_n^h}\left(T \le t|Y_i^h\right), \qquad t \in [0,\tau], \tag{6}$$

where $P_F(T \le t|Y^h) = P_F(T \le t|T \in B(Y^h)_1)$, where $B(Y^h)_1$ is a point, horizontal strip, vertical strip or an upper quadrant, where the strips and quadrants start at the grid points. The SOR-MLE F_n^h is computed by iterating this equation with an initial estimator of F which puts mass on each point of the support of F_n^h. The self-consistency equation tells us that F_n^h puts at least mass $1/n$ on each uncensored observation, which provides us with the following useful bound: for each set A,

$$F_n^h(A) \ge P_{11}^n(A). \tag{7}$$

3. Outline of the efficiency proof. First, we define the models corresponding to the data Y^h and Y. Let $\mathscr{F}$ be the set of all bivariate distributions on $[0,\infty)$ and let $\mathscr{F}_h$ be the set of all possible bivariate distributions G_h which live on π^h. Then the model corresponding to Y^h [see (3)] is given by

$$\mathscr{M}_h \equiv \left\{P_{F,G_h}^h: F \in \mathscr{F}, G_h \in \mathscr{F}_h\right\}$$

and the model corresponding to Y [see (1)] is given by

$$\mathscr{M} \equiv \{P_{F,G}: F, G \in \mathscr{F}\}.$$

Let $D[0,\tau]$ be the space of bivariate cadlag functions on $[0,\tau]$ as defined in Neuhaus (1971). We are interested in estimating the parameter

$$\vartheta_h: \mathscr{M}_h \to D[0,\tau]: \vartheta_h\left(P_{F,G_h}^h\right) = S.$$

Similarly, we define

$$\vartheta: \mathscr{M} \to D[0,\tau]: \vartheta(P_{F,G}) = S.$$

To begin with we will prove pathwise differentiability of these parameters [see, e.g., BKRW (1993), Chapter 3; van der Vaart (1988)].

Let $\mathscr{S}(F)$ the class of lines $\varepsilon F_1 + (1-\varepsilon)F$, $F_1 \in \mathscr{F}$, with score $h = d(F_1 - F)/dF \in L_0^2(F)$, through F. By convexity of $\mathscr{F}$ this is a class of submodels. Let $S(F) \subset L_0^2(F)$ be the corresponding tangent cone (i.e., set of scores). It is easily verified that the tangent space $T(F)$ [the closure of the linear extension of $S(F)$] equals $L_0^2(F)$. Each submodel of $\mathscr{S}(F)$ with score g will be denoted with $F_{\varepsilon,g}$. The score of the one-dimensional submodels $P_{F_{\varepsilon,g},G_h}^h \subset \mathscr{M}_h$, $g \in S(F)$, is given by $A_F^h(g)$, where A_F^h is the score operator,

$$A_F^h: L^2(F) \to L^2\left(P_{F,G_h}^h\right): A_F^h(g)(Y^h) = E_F(g(T)|Y^h),$$

which is a well known result which holds in general for missing data models [van der Vaart (1988) Gill (1989) BKRW (1993) section 6.6]. The score operator A_F for the one-dimensional submodels $P_{F_{\varepsilon,g},G} \subset \mathscr{M}$, $g \in S(F)$, is given by

$$A_F \colon L^2(F) \to L^2(P_{F,G}) \colon A_F(g)(Y) = E_F(g(T)|Y).$$

Let $G_{h,\varepsilon,g_1} \subset \mathscr{M}_h$ be a line through G_h with score g_1. Because of factorization of $p^h_{F,G_h}(y)$ and $p_{F,G}(y)$ the scores $B^h_G(g_1)$ of $P^h_{F,G_{h,\varepsilon,g1}}$ and the scores $B_G(g_1)$ of $p_{F,G_{\varepsilon,g1}}$ are orthogonal to the range of A_F and A^h_F, respectively.

It is easily verified and a well known fact [see BKRW (1993), Section 6.6] that the adjoint of A_F is given by

$$A^T_F \colon L^2(P_{F,G}) \to L^2(F) \colon A^T_F(v)(T) = E_{F,G}(v(Y)|T)$$

and similarly that the adjoint of A^h_F is given by

$$A^{hT}_F \colon L^2\big(P^h_{F,G^h}\big) \to L^2(F) \colon A^{hT}_F(v)(T) = E_{F,G^h}\big(v(Y^h)|T\big).$$

Hence the corresponding information operator is defined by

$$I^h_F = A^{hT}_F A^h_F \colon L^2(F) \to L^2(F) \colon I^h_F(g)(X) = E_{F,G_h}\big(E_{F,G_h}(g(X)|Y^h)|X\big).$$

If $H > \delta > 0$, then it is trivially verified that $\|A_F(h)\|_{P_F} > \sqrt{\delta}\|h\|_F$. Now, application of Lemma 1.3 in van der Laan (1993) tells us that this implies that $I^h_F \colon L^2(F) \to L^2(F)$ has a bounded inverse, uniformly in $F \in \mathscr{F}$ [Lemma 5.2 in van der Laan (1995) formulates this result in general for missing data models]. The same result holds for $I_F \colon L^2(F) \to L^2(F)$. This proves the following lemma.

LEMMA 3.1. *Let $I_{F,G} = A^T_F A_F \colon L^2(F) \to L^2(F)$ be the information operator for $\mathscr{M}$. We have the following. If $H > \delta > 0$ F-a.e., for certain $\delta > 0$, then $I_{F,G}$ has bounded inverse $I^{-1}_{F,G}$ with norm smaller than $1/\delta$ and is onto. The same holds for the information operator $I^h_{F,G_h} \colon L^2(F) \to L^2(F)$, for $\mathscr{M}_h$ with inverse I^{-1}_{h,F,G_h}, where the bound is uniform in h.*

Let $b_t \colon D[0,\tau] \to \mathbb{R}$ be defined by $b_t F = F(t)$. Define $\kappa_t \equiv I_{(t,\infty)} - S(t)$. For each one-dimensional submodel $P_{F_{\varepsilon,g},G_{h,\varepsilon,g1}}$, we have

$$\begin{aligned}
\frac{1}{\varepsilon}\Big(b_t\vartheta_h\big(P^h_{F_{\varepsilon,g},G_{h,\varepsilon,g1}}\big) - b_t\vartheta_h\big(P^h_{F,G_h}\big)\Big) &= \int_{(t,\infty)} g\,dF \\
&= \langle I_{(t,\infty)} - S(t), g\rangle_F \\
&= \langle \kappa_t, g\rangle_F \\
&= \langle I^h_F I^{-1}_{h,F}(\kappa_t), g\rangle_F \\
&= \langle A^h_F I^{-1}_{h,F}(\kappa_t), A^h_F(g)\rangle_{P^h_{F,G_h}} \\
&= \langle A^h_F I^{-1}_{h,F}(\kappa_t), A^h_F(g) + B^h_G(g_1)\rangle_{P^h_{F,G_h}},
\end{aligned}$$

where we used the orthogonality of the scores at the last step. The same holds for ϑ and $P_{F,G}$ without h. This proves by definition [see, e.g., BKRW (1993)] that for each $t \in [0, \tau]$, $b_t \vartheta_h$ is pathwise differentiable at P^h_{F,G_h}, for each one-dimensional submodel $P_{F_{\varepsilon,g}, G_{h,\varepsilon,g1}}$ at P^h_{F,G_h} with efficient influence function (suppressing the G in the notation) given by

$$\tilde{I}^h(F,t)(\cdot) = A^h_F I^{-1}_{h,F}(\kappa_t)(\cdot), \tag{8}$$

and similarly for ϑ at $P_{F,G}$ with

$$\tilde{I}(F,t)(\cdot) = A_F I^{-1}_F(\kappa_t)(\cdot). \tag{9}$$

Notice that these are the same efficient influence curves as we would have found in the models where $G = G_0$ would have been known. In the sequel G_0 does not vary and therefore we can skip the G in the notation; $P^h_F \equiv P^h_{F,G_h}$ and $P_F \equiv P_{F,G_0}$, $I_F \equiv I_{F,G}$ and so forth.

We recall the relevant efficiency and empirical process theory: an estimator $F_n(t)$ is efficient if

$$F_n(t) - F_0(t) = (P_n - P_{F_0})\tilde{I}(F_0, t) + R_{n,t},$$

where $R_{n,t} = o_P(1/\sqrt{n})$. The variable $\sqrt{n}(P_n - P_{F_0})\tilde{I}(F_0, t)$ is a sum of n i.i.d. mean zero random variables which converges by the CLT to a normal distribution with mean zero and variance $P_{F_0}\tilde{I}(F_0, t)^2$. By varying $t \in [0, \tau]$ we obtain an empirical process $(\sqrt{n}(P_n - P_{F_0})\tilde{I}(F_0, t)\colon t \in [0, \tau])$, which can be considered as a random element of $l^\infty(\mathscr{G}) \equiv \{H\colon \mathscr{G} \to \mathbb{R}\colon \sup_{g \in \Phi}|H(g)| < \infty\}$, where $\mathscr{G} = \{\tilde{I}(F_0, t)\colon t \in [0, \tau]\}$ and where $l^\infty(\mathscr{G})$ is endowed with the Borel sigma algebra. Empirical process theory investigates if the empirical process indexed by some class converges in distribution to a tight Gaussian process corresponding with the covariance structure of the empirical process. Here convergence in distribution (i.e., weak convergence) is defined in the Hoffman–Jørgensen sense, making measurability questions (for finite n) irrelevant [see, e.g., Hoffmann–Jørgensen (1984); van der Vaart and Wellner (1995); Pollard (1990)]. A class for which this weak convergence holds is called a Donsker class. If $\mathscr{G}$ is Donsker and $\sup_{t \in [0,\tau]}|R_{n,t}| = o_P(1/\sqrt{n})$, then we say that F_n is sup-norm efficient.

Our goal is to prove efficiency of S^h_n as an estimator of $\vartheta(P_{F_0}) = S_0$. It should be remarked that, for fixed h, application of Theorem 6.2 for a general class of missing data models in van der Laan (1995) provides us (under the assumptions as stated in Section 2.1, by simple verification) with efficiency of S^h_n among estimators based on the data Y^h_i, $i = 1, \dots, n$, as an estimator of $\vartheta_h(P^h_{F_0}) = S_0$. However, we want more than efficiency for a fixed reduction. For this purpose we will follow the same analysis as followed for the general class of missing data models, except that we look carefully at what happens if $h_n \to 0$ when the number of observations diverges to infinity.

It works as follows: The model $\mathscr{M}_h$ is *convex* and the $F \to P_F^h$ is *linear*. Theorem 1.1 in van der Laan (1993) says now that we have the following identity: for each $t \in [0, \tau]$, we have

$$S_1(t) - S_0(t) = -\int \tilde{I}^h(S_1, t)\, dP_{F_0}^h,$$

for all F_1 with $F_0 \ll F_1$ and $dF_0/dF_1 \in L_0^2(F_1)$. So in particular this identity holds for

$$S_n^h(\alpha) = \alpha S_0 + (1-\alpha) S_n^h, \qquad \alpha \in (0, 1],$$

which provides us with the identity

$$S_n^h(\alpha)(t) - S_0(t) = -\int \tilde{I}^h\big(S_n^h(\alpha), t\big)\, dP_{F_0}^h, \qquad \alpha \in (0,1]. \tag{10}$$

Notice now that $S_n^h(\alpha) - S_n^h = \alpha(S_n^h - S_0)$. If $\alpha \to 0$, the left-hand side of (10) converges to $S_n^h(t) - S_0(t)$ and it has been verified for the general class of missing data models [Lemma 5.12 in van der Laan (1995)] that the right-hand side converges to $-\int \tilde{I}^h(S_n^h, t)\, dP_{F_0}^h$. In fact in our proof we show that $\int (I^h(S_h^h, t) - I^h(S_0, t))^2\, dP_{F_0}^h \to 0$ which basically proves this much weaker result [notice that $S_n^h(\alpha)$ converges to S_n^h w.r.t. each norm]. It follows that we have the following identity:

$$S_n^h(t) - S_0(t) = -\int \tilde{I}^h\big(S_n^h, t\big)\, dP_{F_0}^h. \tag{11}$$

It remains to verify the following items:

Efficient score equation. For all $t \in [0, \tau]$,

$$\int \tilde{I}^h\big(F_n^h, t\big)\, dP_n^h = 0.$$

The score equations (5) tell us that it suffices to prove that $I_{F_n^h}^{-1}(I_{(t,\infty)})$ has finite sup-norm. This is proved by Lemma 6.2 in Section 6 of this paper.

The efficient score equation and the identity (11) provide us with the crucial identity

$$S_n^h(t) - S_0(t) = \int \tilde{I}^h\big(F_n^h, t\big) d\big(P_n^h - P_{F_0}^h\big). \tag{12}$$

Empirical process condition. Now, we will show, for an appropriate rate $h_n \to 0$, that

$$\sup_{t \in [0,\tau]} \left| \int \big(\tilde{I}^h\big(F_n^h, t\big) - \tilde{I}^h(F_0, t)\big) d\big(P_n^h - P_{F_0}^h\big) \right| = o_{P_{F_0}^h}(1/\sqrt{n}).$$

This condition requires a lot of hard work (done in Sections 4 and 7). The reason for this is that we are not able to prove that $\tilde{I}(F_0, t)$ has any nice properties, except that it exists as an element in $L_0^2(P_{F_0})$, due to the very complicated form of the information operator I_{F_0}. Therefore $\tilde{I}^h(F_n^h, t)$ cannot be shown to be an element of a fixed Donsker class when $h_n \to 0$. In other words, the P-Donsker class and ρ_P-consistency condition as used in the proof for the general class of missing data models [van der Laan (1995)] do not help us here. More sophisticated conditions are needed. The technique will be to determine how quickly $\tilde{I}^h(F_n^h, t)$ loses its Donsker class properties, for $h_n \to 0$, and then to use (12) in order to obtain a rate for $\|S_n^h - S_0\|_\infty$ so that terms can be shown to converge to zero if $h_n \to 0$ slowly enough.

The empirical process condition provides us with [see, e.g., Pollard (1990)]

$$S_n^h(t) - S_0(t) = \int \tilde{I}^h(F_0, t)\, d\big(P_n^h - P_{F_0}^h\big) + o_{P_{F_0}^h}(1/\sqrt{n}),$$

where the remainder holds uniformly in t.

Approximation condition. Finally, we need to show

$$\int \tilde{I}^h(F_0, t)\, d\sqrt{n}\big(P_n^h - P_{F_0}^h\big) \Rightarrow_D N\big(0, \sigma^2\big(\tilde{I}(F_0, t)\big)\big).$$

Notice that the left-hand side is a sum of i.i.d. random variables given by $1/\sqrt{n}\sum_{i=1}^n X_i^h(t)$, where $X_i^h(t) \equiv \tilde{I}^h(F_0, t)(Y_i^h)$. By Bickel and Freedman (1981) we have that if, for $h = h_n \to 0$, $X_i^h(t) \Rightarrow_D X_i(t)$ and $\mathrm{Var}(X_i^h(t)) \to \mathrm{Var}(X_i(t))$, then this sum converges weakly to a normal distribution with mean zero and variance equal to $\mathrm{Var}(X_i(t))$. These two conditions are proved by Lemma 4.7.

We also show the approximation condition for the case that we consider the left- and right-hand side as a random element of a L^2-space of functions in t, which provides us with pointwise and L^2-efficiency.

4. Proof of efficiency of SOR-MLE.

Recall the assumptions made in Section 2.1: In particular $F_0(\tau) = 1$ and hence $P_{F_0}^h(\cdot, d)$ lives on $[0, \tau]$. In all statements the width (of grid) h converges to zero for $n \to \infty$. The problem is to find a lower bound for the rate at which h should converge to zero.

4.1. *Uniform consistency of F_n^h for $h_n \to 0$.* The starting point of the analysis is (12). The indicators are a uniform Donsker class. This tells us that $\sup_h \|P_n^h - P_{F_0}^h\|_\infty = O_P(1/\sqrt{n})$.

A real-valued function on $[0, \tau] \subset \mathbb{R}^2$ is said to be of bounded *uniform sectional variation* if the variations of all sections [$s \to f(s, t)$ is a section of the bivariate function f] and of the function itself is uniformly (in all sections) bounded. The corresponding norm is denoted with $\|\cdot\|_v^*$. In van der Laan [(1996), Example 1.2)] it is proved that the class of functions with uniform sectional variation smaller than $M < \infty$ is a uniform Donsker class (it is well known that the real-valued functions with variation smaller than $M < \infty$ form a uniform Donsker class, so this is a generalization of this one-dimensional result). Another fact is that if $f > \delta > 0$, then $\|1/f\|_v^* \le M\|f\|_v^*$, for some $M < \infty$, which does not depend on f [Gill (1994)]. We have the following lemma:

LEMMA 4.1 (Uniform sectional variation of efficient influence curve). *Let $E_{k,l}^h(1,0) \equiv (u_k, u_{k+1}] \times [v_l, \infty)$ be the vertical strips of π^h and let $E_{k,l}^h(0,1)$*

be the horizontal strips. Suppose that the grid π^h *is so that* $F_0(E^{h_n}_{k,l}) > \delta h_n$, *for certain* $\delta > 0$. *Let* $r_1(h_n) = 1/h_n^{3/2}$.

For all $d \in \{0,1\}^2$ *we have that, for some* $M < \infty$, $\tilde{I}^h(F_n^h, t)(\cdot, d) \in D[0,\tau]$ *and*

$$\sup_{t\in[0,\tau]} \|\tilde{I}^h\big(F_n^h, t\big)(\cdot, d)\|_v^* \le M r_1(h) \quad \textit{with probability tending to } 1.$$

For the proof, see Section 6.

Consider an integral $\int F_1\, dH_1$, where $F_1 \in D[0,\tau]$ and $H_1 \in D[0,\tau]$ are bivariate real-valued cadlag functions which are of bounded uniform sectional variation. By integration by parts [see Gill (1992) or Lemma 1.3 in van der Laan (1996)] we can bound it by $C\|H_1\|_\infty \|F_1\|_v^*$. Because $\tilde{I}^h(F_n^h, t)(\cdot, d)$ generates a signed measure [see Lemma 1.2 in van der Laan (1996)], we can apply this to (12) with $F_1 = \tilde{I}^h(F_n^h, t)(\cdot, d)$ and $H_1 = (P_n^h - P_{F_0}^h)(\cdot, d)$ and apply Lemma 4.1 to F_1. This proves the following lemma:

LEMMA 4.2 (Uniform consistency). *Under the assumption of Lemma* 4.1 *we have*

$$\|F_n^{h_n} - F_0\|_\infty = O_P\left(\frac{r_1(h_n)}{\sqrt{n}}\right) = O_P\left(\frac{1}{\sqrt{nh_n^3}}\right).$$

So if $h \to 0$ *slower than* $n^{-1/3}$, *then* F_n^h *is uniformly consistent* (*also for h is fixed*).

4.2. *Empirical process condition.* Define $Z_n^h \equiv \sqrt{n}\,(P_n^h - P_{F_0}^h)$ and $f_{nt}^h \equiv \tilde{I}^h(F_n^h, t) - \tilde{I}^h(F_0, t)$. We will show that $\int f_{nt}^h\, dZ_n^h$ converges to zero uniformly in t with probability tending to 1. By using that $\|F_n^h - F_0\|_\infty = O_P(r_1(h_n)/\sqrt{n}\,)$ (Lemma 4.2) we are able to show the following lemma:

LEMMA 4.3 (Sup-norm convergence of efficient influence curve). *Under the assumption of Lemma* 4.1 *we have, for all* $d \in \{1,0\}^2$, *with* $r_2(h_n) = 1/h_n^3$,

$$\|f_{nt}^h(\cdot, d)\|_\infty = O_P\big(r_1(h_n) r_2(h_n)/\sqrt{n}\,\big) = O_P\Big(1/\sqrt{nh_n^9}\,\Big).$$

For the proof, see the Appendix.

Analysis of the uncensored term. Let us first analyze $\int f_{nt}^h I(d = (1,1))\, dZ_n^h$. Recall that $Z_n^h I(d = (1,1)) = Z_n I(d = (1,1)) = \sqrt{n}\,(P_{11}^n - P_{11})$, where $p_{11} = f_0 H_h$. We will assume that $F_0 = F_0^d + F_0^c$, where F_0^c is absolutely continuous w.r.t. the Lebesgue measure with continuous density which is bounded away from zero and F_0^d is purely discrete with finite support. Then we can decompose $P_{11} = P_{11}^d + P_{11}^c$, where $p_{11}^d = f_0^d H_h$ is purely discrete on the finite number of support points of F_0^d, and P_{11}^c is absolutely continuous w.r.t. Lebesgue measure with density bounded away from zero.

For P_{11}^n we have a corresponding decomposition $P_{11}^n = P_{11}^{nd} + P_{11}^{nc}$, where P_{11}^{nd} only counts the number of observations coming from P_{11}^d. First consider the integral w.r.t. $\sqrt{n}(P_{11}^{nd} - P_{11}^{d})$. Let p_{11}^{d} be the density of P_{11}^{d} w.r.t. the counting measure, say μ_k, which lives on the support of P_{11}^{d}. We have that $\int |p_{11}^{nd} - p_{11}^{d}|\, d\mu_k = O_P(1/\sqrt{n})$. Therefore, with $Z_{nd} \equiv \sqrt{n}(P_{11}^{nd} - P_{11}^{d})$ we have

$$\begin{aligned}\int f_{nt}^{h} I(d = (1,1))\, dZ_{nd} &= \sqrt{n}\int f_{nt}^{h} I(d = (1,1))\big(p_{11}^{nd} - p_{11}^{d}\big)\, d\mu_k \\ &\leq \sqrt{n}\, \|f_{nt}^{h} I(d = (1,1))\|_\infty \int |\big(p_{11}^{nd} - p_{11}^{d}\big)|\, d\mu_k \\ &= \sqrt{n}\, O_P\left(\frac{1}{\sqrt{nh_n^9}}\right) O_P\left(\frac{1}{\sqrt{n}}\right) \\ &= O_P\left(\frac{1}{\sqrt{nh_n^9}}\right),\end{aligned}$$

where the bound does not depend on t. Consequently, if $nh_n^9 \to \infty$, then $\int f_{nt}^{h} I(d = (1,1)\, dZ_{nd} = o_P(1)$.

Consider now $\int f_{nt}^{h} I(d = (1,1))\, dZ_n^c$, where $Z_n^c I(d = 1,1) = \sqrt{n}(P_{11}^{nc} - P_{11}^{c})$. For convenience, we denote Z_n^c with Z_n, again. We construct a lattice grid $\pi^{a_n} = (t_i, t_j)$, with maximal mesh $a_n < h_n$, on $[0, \tau] = [0, \tau_1] \times [0, \tau_2]$, which we force to be such that $\pi^{h_n} \subset \pi^{a_n}$. Now

$$[0,\tau] = \bigcup_{i,j} A_{i,j}(a_n), \quad \text{where } A_{i,j}(a_n) \equiv \big((t_i, t_{i+1}] \times (t_j, t_{j+1}]\big) \cap [0,\tau]$$

and the union is over all partition elements $A_{i,j}(a_n)$, $i = 1, \dots, n_1(a_n)$, $j = 1, \dots, n_2(a_n)$. The number of partition elements will be denoted by $n(a_n)$ and it is clear that $n(a_n) = O(1/a_n^2)$. Now we define an approximation of Z_n as follows:

$$Z_n^{a_n}(t) \equiv Z_n(t_i, t_j) \quad \text{if } t \in A_{i,j}(a_n).$$

So $Z_n^{a_n}$ is constant on each $A_{i,j}(a_n)$ with value $Z_n(t_i, t_j)$.

By using integration by parts it is clear that we have, for $d = (1,1)$ (the integral is over $y \in [0,\tau]$, fixed d),

$$\begin{aligned}&\int f_{nt}^{h}(y,d)\, dZ_n(y,d) \\ &\quad = \int f_{nt}^{h}(y,d) d(Z_n - Z_n^{a_n})(y,d) + \int f_{nt}^{h}(y,d)\, dZ_n^{a_n}(y,d) \\ &\quad \leq C \|f_{nt}^{h}(\cdot,d)\|_v^* \|(Z_n - Z_n^{a_n})(\cdot,d)\|_\infty + \|f_{nt}^{h}(\cdot,d)\|_\infty \|Z_n^{a_n}(\cdot,d)\|_v^* \\ &\quad \leq O_P(r_1(h_n)) \|(Z_n - Z_n^{a_n})(\cdot,d)\|_\infty \\ &\qquad + O_P\left(\frac{r_1(h_n) r_2(h_n)}{\sqrt{n}}\right) \|Z_n^{a_n}(\cdot,d)\|_v^*.\end{aligned}$$

In order to show that $\int f_{nt}^{h}(y,d)\, dZ_n(y,d) = o_P(1)$, for a rate $h_n \to 0$, it suffices to show that there exists a rate a_n for which the last two terms converge to zero in probability.

For convenience we will neglect the d in our notation. Define

$$W^h_{i,j}(a_n) \equiv \sup_{s,t \in A_{i,j}(a_n)} |Z_n(s) - Z_n(t)|$$

and

$$W_n(a_n) \equiv \max_{i,j} W^n_{i,j}(a_n).$$

It other words, $W_n(a_n)$ is a *modulus of continuity* of a bivariate empirical process. First, we will bound the two terms in $W^n(a_n)$.

We have $\|Z_n^{a_n} - Z_n\|_\infty \leq \max_{i,j} W^n_{i,j}(a_n)$. Therefore,

$$P(\|Z_n^{a_n} - Z_n\|_\infty > \varepsilon) \leq P(W_n(a_n) > \varepsilon). \tag{13}$$

Furthermore we have

$$\|Z_n^{a_n}\|_v^* \leq \sum_{i,j} W^n_{i,j}(a_n) \leq \frac{c}{a_n^2} W_n(a_n). \tag{14}$$

Analysis of the modulus of continuity. For a rectangular R we define $Z_n(R)$ as the measure of R assigned by the bivariate signed measure Z_n. Define $W_{n,R}(a_n) \equiv \sup_{R:\ R| \leq a_n} |Z_n(R)|$. Einmahl's (1987) inequality 6.4, for $W_{n,R}(a_n)$, holds for an empirical process from a sample of a continuous density which is bounded away from zero and infinity on $[0, \tau]$ and is given by

$$P(W_{n,R}(a_n) > \lambda) \leq \frac{C}{a_n} \exp\left(\frac{-c_1 \lambda^2}{a_n} \Psi\left(\frac{\lambda}{\sqrt{n a_n}} \right) \right) \quad \text{for any } \lambda > 0, \tag{15}$$

where $\Psi(x) \geq 1/(1 + 1/3x)$. Notice that $W_n(a_n)$ is a bound on the measure assigned by Z_n to strips instead of rectangles. However, the strips are a union of at most c/a_n rectangles $A_{i,j}(a_n)$ and on each rectangle, $A_{i,j}(a_n)$ of these strips, p^c_{11} is bounded away from zero and infinity and is continuous (here we use the nesting of π^{h_n} in π^{a_n}), and hence for the modulus of continuity on the sets $A_{i,j}(a_n)$ the discontinuities on π_h play no role. Consequently, (15) can be applied to each rectangle $A_{i,j}(a_n)$ in the strips. So the bound (15) implies the following bound for $W_n(a_n)$:

$$\begin{aligned} P(W_n(a_n) > \lambda) &\leq \frac{c}{a_n} P(W_{n,R}(a_n) > \lambda) \\ &\leq \frac{C}{a_n^2} \exp\left(\frac{-c_1 \lambda^2}{a_n} \Psi\left(\frac{\lambda}{\sqrt{n a_n}} \right) \right) \quad \text{for any } \lambda > 0, \end{aligned}$$

where $\Psi(x) \geq 1/(1 + 1/3x)$ and where the C is now different from the preceding one.

By using this inequality with $\lambda = a_n^{0.5-\varepsilon}$ it is trivial to see that if $na_n \to \infty$ at an arbitrarily small polynomial rate (n^{ε}), then for each $\varepsilon > 0$ there exists a sequence $\delta_n \to 0$ and an $\varepsilon' > 0$ so that

$$(16) \qquad P\left(\frac{W_n(a_n)}{a_n^{0.5-\varepsilon}} > \delta_n\right) \le \frac{C}{a_n^2}\exp\left(-\frac{C_1}{a_n^{\varepsilon'}}\right).$$

So $W_n(a_n)/a_n^{0.5-\varepsilon}$ converges to zero in probability exponentially fast.

Assume $na_n \to \infty$ at a polynomial rate. Applying (16) to (13) provides us with

$$P\left(\frac{\|Z_n^{a_n} - Z_n\|_\infty}{a_n^{05-\varepsilon}} > \varepsilon\right) \le \frac{C}{a_n^2}\exp\left(-\frac{C_1}{a_n^{\varepsilon'}}\right) = o(1).$$

So $\|Z_n^{a_n} - Z_n\|_\infty = o_P(a_n^{0.5-\varepsilon})$. This proves that $r_1(h_n)\|(Z_n - Z_n^{a_n})(\cdot, d)\|_\infty = o_P(r_1(h_n)a_n^{0.5-\varepsilon})$, for any $\varepsilon > 0$.

Furthermore, applying (16) to (14) provides us with

$$\|Z_n^{a_n}\|_v^* = O\big(1/a_n^2\big)o_P\big(a_n^{0.5-\varepsilon}\big) = o_P\big(a_n^{-(1.5+\varepsilon)}\big).$$

Consequently, this tells us that for each $\varepsilon > 0$ we have: If $na_n \to \infty$ (at least at a polynomial rate), then

$$(17) \quad \int f_{nt}^h(y,1,1)\,dZ_n(y) = o_P\big(r_1(h_n)a_n^{0.5-\varepsilon}\big) + o_P\left(\frac{r_1(h_n)r_2(h_n)}{\sqrt{n}\,a_n^{1.5+\varepsilon}}\right).$$

For the first term it suffices that a_n converges to zero more quickly than h_n^3. Substituting this in the second term tells us that it suffices to let h_n converge to zero slower than $n^{-1/18}$. This proves the following lemma:

LEMMA 4.4. *Suppose that $F_0 = F_0^d + F_0^c$, where F_0^c is absolutely continuous w.r.t. Lebesgue measure with continuous density which is bounded away from zero on $[0, \tau]$ and F_0^d is purely discrete with finite support on $[0, \tau]$. If h_n converges to zero slower than $n^{-1/18}$, then $\int f_{nt}^h I(D = (1,1))\,dZ_n^h = o_P(1)$.*

Analysis of the censored terms. We will now analyze the terms $\int f_{nt}^h I(D \neq (1,1))\,dZ_n^h$. Recall that $P_{F_0}^h I(D \neq (1,1))$ is purely discrete on the grid π^h, which contains $O(1/h_n^2)$ points. Let $p_{F_0}^h$ and p_n^h be the densities of $P_{F_0}^h$ and P_n^h w.r.t. ν_h, respectively. So $p_{00}^{h,n}(v_i, v_j) \equiv p_n^h(v_i, v_j, 0, 0)$ is the fraction of doubly censored observations which falls on (v_i, v_j) and similarly for $D = (1,0)$ and $D = (0,1)$. It is clear that, for fixed h_n, we have $\|p_n^h - p_{F_0}^h\|_\infty = O_P(1/\sqrt{n})$. In the following result for $h_n \to 0$ we do not make any assumptions. Under weak assumptions, the rate would be $O_p(1/\sqrt{h_n^2 n})$, but this improvement is not interesting because of the slow rate in Lemma 4.4.

LEMMA 4.5. *We have that*

$$\|p_{01}^{hn} - p_{01}^h\|_{L_1(\nu_h)} = O_P\left(\frac{1}{\sqrt{h^4 n}}\right)$$

and we have the same rate result for p_{10}^{hn} and p_{00}^{hn}.

PROOF. We give the proof for the first term; the others are dealt with similarly. Because we are just dealing with a multinomial distribution on the grid π^h, we have that $E(p_{01}^{nh}(u_k, v_l)) = p_{01}^h(u_k, v_l)$ and $\mathrm{Var}(p_{01}^{nh}(u_k, v_l)) = (1/n)p_{01}^h(u_k, v_l)(1 - p_{01}^h(u_k, v_l))$. π^h has $O(h_n^2)$ grid points (u_k, v_l) by definition of π^h. Now, we have

$$\begin{aligned} E\Big(\sum_{k,l}\big|\big(p_{01}^{hn} - p_{01}^h\big)(u_k, v_l)\big|\Big) &= \sum_{k,l} E\big(\big|\big(p_{01}^{hn} - p_{01}^h\big)(u_k, v_l)\big|\big) \\ &\leq \frac{1}{\sqrt{n}} \sum_{k,l} \sqrt{p_{01}^h(u_k, v_l)\big(1 - p_{01}^h(u_k, v_l)\big)} \\ &\leq \frac{1}{\sqrt{n}} \frac{1}{h^2}. \qquad \square \end{aligned}$$

Again, we will neglect the d in our notation, but the reader should remember that we only integrate over the singly censored and doubly censored observations. Now we have

$$\begin{aligned} \int f_{nt}^h \, dZ_n^h &= \sqrt{n} \int f_{nt}^h\big(p_n^h - p_{F_0}^h\big)\, d\nu_n \\ &\leq \sqrt{n}\, \|f_{nt}^h\|_\infty \|p_n^h - p_{F_0}^h\|_{L_1(\nu_h)} \\ &= \sqrt{n}\, O_P\left(\frac{1}{\sqrt{h_n^9 n}}\right) O_P\left(\frac{1}{\sqrt{n h_n^4}}\right) \\ &= O_P\left(\frac{1}{\sqrt{h_n^{13} n}}\right). \end{aligned}$$

This proves the following lemma:

LEMMA 4.6. *If h_n converges to zero slower than $n^{-1/13}$, then*

$$\int f_{nt}^h I(D = d)\, dZ_n^h = o_P(1) \quad \text{for } d \in \{(1,0), (0,1), (0,0)\}.$$

Lemmas 4.4 and 4.6 prove the empirical process condition for a rate of h_n slower than $n^{-1/18}$. Recall that all the derived lower bounds are derived without any knowledge about $\tilde{I}(F_0, t)$, except that it has a finite variance, and therefore they only have a theoretical value. □

4.3. *Approximation condition.*

4.3.1. *Pointwise convergence.* Let $t \in [0, \tau]$ be fixed. Define $V_n^h(t) \equiv \int \tilde{I}^h(F_0, t)(y)\, dZ_n^h(y)$. $V_n^h(t)$ is a sum of i.i.d. mean zero random variables given by $(1/\sqrt{n})\sum_{i=1}^n X_i^h(t)$, where $X_i^h(t) \equiv \tilde{I}^h(F_0, t)(Y_i^h)$. By Bickel and Freedman (1981) we have that if, for $h = h_n \to 0$, $X_i^h(t) \Rightarrow_D X_i(t)$ and

$\mathrm{Var}(X_i^h(t)) \to \mathrm{Var}(X_i(t))$, then this sum converges weakly to a normal distribution with mean zero and variance equal to $\mathrm{Var}(X_i(t))$. We will prove these two conditions:

LEMMA 4.7. *Define the following real-valued random variables* $X^h(t) \equiv \tilde{I}^h(F_0, t)(Y^h)$, $Y^h \sim P_{F_0}^h$, *and* $X(t) \equiv \tilde{I}(F_0, t)(Y)$, $Y \sim P_{F_0}$. *We have for each* $t \in [0, \tau]$ *that for* $h_n \to 0$,

$$E\Big(\big(X^{h_n}(t) - X(t)\big)^2\Big) \to 0$$

and

$$E\big(X^{h_n}(t) X^{h_n}(s)\big) \to E\big(X(t)X(s)\big) \quad \textit{uniformly in } s, t \in [0, \tau].$$

For the proof, see Section 6. Lemma 4.7 has the following corollary:

COROLLARY 4.1. *The empirical process* $\int \tilde{I}^{h_n}(F_0, t)(y)\, dZ_n^{h_n}(y)$ *converges in distribution to a normal distribution with mean zero and variance equal to* $\mathrm{Var}_{P_{F_0}}(\tilde{I}^0(F_0, t))$.

4.3.2. *Hilbert space convergence.* For showing that V_n^h converges weakly as a process in $(D[0, \tau], \|\cdot\|_\infty)$ we need to show at least that $\{\tilde{I}(F_0, t): t \in [0, \tau]\}$ is a P_{F_0}-Donsker class. We have not been able to do this. Therefore we concentrate on proving weak convergence as a process in a Hilbert space. We use the following result, which can be found in Parthasarathy [(1967), page 153].

LEMMA 4.8. *Let* Z_n, Z_0 *be random processes in a Hilbert space* $\mathscr{H}$ *endowed with the Borel sigma algebra* $\mathscr{B}$. *Let* $e_1, e_2, \dots$ *be an orthonormal basis of* $\mathscr{H}$. *If* $\langle e_j, Z_n \rangle \Rightarrow_D \langle e_j, Z_0 \rangle$, *for all* j, *and* $\lim_{N \to \infty} \sup_n E(\sum_{j=N+1}^{\infty} \langle e_j, Z_n \rangle^2) = 0$, *then* $Z_n \Rightarrow_D Z_0$ *in* $\mathscr{H}$.

Let $V_n(t) = (1/\sqrt{n})\sum_{i=1}^n X_i(t)$. First, we will prove the first condition of Lemma 4.8 with $Z_n = V_n^h$ and $Z_0 = V_0$, the optimal Gaussian process. We have

$$\langle e_j, V_n^h \rangle = \langle e_j, V_n^h - V_n \rangle + \langle e_j, V_n \rangle.$$

First, we will show that $\langle e_j, V_n^h - V_n \rangle = o_P(1)$. The fact that V_n^h and V_n are sums of i.i.d. random variables X_i^h and X_i, respectively, and the Cauchy–Schwarz inequality tell us that

$$\begin{aligned}\mathrm{Var}\big(\langle e_j, V_n^h - V_n \rangle\big) &= \mathrm{Var}\big(\langle e_j, X^h - X \rangle\big)\\ &\le E\big(\langle e_j, X^h - X \rangle^2\big)\\ &\le \langle e_j, e_j \rangle E\langle X^h - X, X^h - X \rangle.\end{aligned}$$

Assume now that $\mathscr{H} = L^2(\lambda)$ for a certain finite measure λ. By Lemma 4.7 we have $\mathrm{Var}(X^{h_n}(t))$ converges to $\mathrm{Var}(X(t))$ and $E((X^{h_n}(t) - X(t))^2) \to 0$, both uniformly in t. Therefore,

$$E\langle X^h - X, X^h - X\rangle \leq \sup_{s\in[0,\tau]} \left|E\big((X^h - X)(s)^2\big)\right| \int d\lambda(s) \to 0,$$

which proves the convergence of $\langle e_j, V_n^h - V_n\rangle$ to zero in probability. Furthermore, we have

$$\langle e_j, V_n\rangle = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int e_j(s) X_i(s)\, d\lambda(s),$$

which is just a sum of i.i.d. mean zero random variables. By the CLT, to show that this converges in distribution to $\langle e_j, V_0\rangle$, it suffices to have that $\mathrm{Var}(\int e_j(s) X_i(s)\, d\lambda(s)) < \infty$. This follows immediately from the fact that $\|E(X^2(s))\|_\infty < \infty$. This proves the weak convergence of $\langle e_j, V_n^h\rangle$ to $\langle e_j, V_0\rangle$.

We will now verify the tightness condition. We have

$$\begin{aligned}
E&\left(\sum_{i=N+1}^{\infty} \langle e_i, V_n^h\rangle^2\right)\\
&= \sum_{i=N+1}^{\infty} E\big(\langle e_i, V_n^h\rangle^2\big)\\
&= \sum_{i=N+1}^{\infty} E\left(\int\!\!\int e_i(s) e_i(t) V_n^h(s) V_n^h(t)\, d\lambda(s)\, d\lambda(t)\right)\\
&= \sum_{i=N+1}^{\infty} \int\!\!\int e_i(s) e_i(t) E\big(V_n^h(s) V_n^h(t)\big)\, d\lambda(s)\, d\lambda(t)\\
&= \sum_{i=N+1}^{\infty} \int\!\!\int e_i(s) e_i(t) \big(E(V_0(s) V_0(t)) + o(1)\big)\, d\lambda(s)\, d\lambda(t)\\
&= o(1)\left(\sum_{i=N+1}^{\infty} (\langle e_i, 1\rangle)^2\right) + \sum_{i=N+1}^{\infty} \langle e_i, V_0\rangle^2.
\end{aligned}$$

In the first, second and third equality we used Fubini's theorem, then we use the uniform convergence of $E(V_n^h(s)V_n^h(t))$ to $E(V_0(s)V_0(t))$ (by Lemma 4.7) and finally we again apply Fubini's theorem but now in the reversed order. The last bound does not depend on n anymore. Because $\|V_0\|^2 = \sum_{i=1}^{\infty}\langle V_0, e_i\rangle^2$ and similarly for the function 1, it follows that if we take the limit for $N \to \infty$, then both (tail) series convergence to zero.

Application of Lemma 4.8 provides us now with the following lemma.

LEMMA 4.9. *Suppose the same assumption as in Lemma* 4.7. *If λ is a finite measure and $h_n \to 0$, then $V_n^{h_n} \Rightarrow_D V_0$ as random elements in $L^2(\lambda)$.*

5. Results. We will summarize the necessary notation for the theorem. Recall the reduced i.i.d. data $Y_i^h \sim P_{F_0, G_h}^h$, obtained by generating n i.i.d.

$C_i \sim G_h$ and the π^h-interval censoring of the singly censored observations. We defined $E^h_{k,l}(1,0)$ and $E^h_{k,l}(0,1)$ as the vertical and horizontal strips of π^h starting at (u_k, v_l). We defined $Z^h_n \equiv \sqrt{n}(P^h_n - P^h_{F_0,G_h})$ as the empirical process corresponding to the reduced data, $\tilde{I}^h(F_0, t)$ as the efficient influence function for estimating $F_0(t)$ using the reduced data and $\tilde{I}(F_0, t)$ as the efficient influence function for estimating $F_0(t)$ using the original data.

We have proved all ingredients of the general efficiency proof of Section 3 in Section 4. Recalling Lemma 4.2 (uniform consistency) and that for fixed h, we have that efficiency (among all estimators based on the reduced data) under the assumptions as stated in Section 2.1 provides us with the following theorem:

THEOREM 5.1. *Let $[0,\tau] \subset \mathbb{R}_{\geq 0}$ be a rectangle so that $H(\tau) > 0$, $S_0(\tau -) > 0$ and $F_0(\tau) = 1$ (data reduced to $[0,\tau]$).*

Fixed grid efficiency. *Suppose that we do not change the grid π^h, for $n \to \infty$, and that for each grid point (u_k, v_l), $F_0(E^h_{k,l}(1,0)) > 0$ and $F_0(E^h_{k,l}(0,1)) > 0$. Then S^h_n is a sup-norm-efficient estimator of S_0 for the data Y^h_i, $i = 1, 2, \dots, n$:*

$$\sqrt{n}\left(F^h_n - F_0\right)(t) = \int \tilde{I}^h(F_0, t)\, dZ^h_n + R^h_n(t),$$

where $\|R^h_n\|_\infty = o_P(1)$ and $\int \tilde{I}^h(F_0, t)\, dZ^h_n$ converges weakly in $(D[0,\tau], \mathscr{B}, \|\cdot\|_\infty)$ to a Gaussian process N_h with mean zero finite-dimensional distributions and covariance structure given by

$$E\big(N_h(s)N_h(t)\big) = E_{P^h_{F_0}}\Big(\tilde{I}^h(F_0, s)\tilde{I}^h(F_0, t)\Big).$$

Uniform consistency. *Suppose that the grid π^h is such that $F_0(E^{h_n}_{k,l}(1,0)) > \delta h_n$ and $F_0(E^{h_n}_{k,l}(0,1)) > \delta h_n$, for some $\delta > 0$. Then, for any rate $h_n \to 0$,*

$$\|S^{h_n}_n - S_0\|_\infty = O_P\Big(1/\sqrt{nh^3_n}\Big).$$

Efficiency. *Suppose $F_0 = F^d_0 + F^c_0$, where F^d_0 is purely discrete with finite support and F^c_0 is absolutely continuous w.r.t. Lebesgue measure with continuous density uniformly bounded away from zero on $[0,\tau]$.*

We have that, for $h_n \to 0$,

$$E_{P^h_{F_0}}\Big(\tilde{I}^h(F_0, s)(Y^h)\tilde{I}^h(F_0, t)(Y^h)\Big) \to E_{P_{F_0}}\Big(\tilde{I}(F_0, s)(Y)\tilde{I}(F_0, t)(Y)\Big)$$

uniformly in $s, t \in [0,\tau]$.

If h_n converges to zero, but slower than $n^{-1/18}$, then we have that $\|R^h_n\|_\infty = o_P(1)$ and, for each $t \in [0,\tau]$, $V^h_n(t) \equiv \int \tilde{I}^h(F_0, t)\, dZ^h_n$ converges in distribution to the normal distribution $N_0(t)$ with mean zero and variance

$$\operatorname{Var}\big(N_0(t)\big) = \operatorname{Var}\Big(\tilde{I}(F_0, t)\Big).$$

Moreover, for any finite measure λ, V^h_n converges weakly as a process in $L^2(\lambda)$ to N_0.

This implies that $F_n^{h_n}(t)$ is an efficient estimator of $F_0(t)$, pointwise and as an element in $L^2(\lambda)$.

We see that if $nh_n^3 \to \infty$, then $F_n^{h_n}$ converges uniformly to F_0. Therefore, we think that $n^{-1/3}$ can also be used as a lower bound for asymptotic efficiency, though we did not prove this.

6. Technical lemmas. In formulas the score operator $A_{F_0}^h$ evaluated at observation $Y^h = (\tilde{T}, D)^h$ is given by [recall that $\tilde{T}$ for $D \neq (1,1)$ lives on the grid π^h]

$$\begin{aligned}
&A_{F_0}^h(g)(\tilde{T}, D)^h \\
&\qquad = g(\tilde{T})I(D = (1,1)) \\
&\qquad\quad + \int_{(u_k, u_{k+1}]}\int_{(v_l, \infty)} g(s_1, s_2)\frac{F_0(ds_1, ds_2)}{F_0((u_k, u_{k+1}], [v_l, \infty))} I(D = (1,0)) \\
&\qquad\quad + \int_{(u_k, \infty)}\int_{(v_l, v_{l+1}]} g(s_1, s_2)\frac{F_0(ds_1, ds_2)}{F_0([u_k, \infty), (v_l, v_{l+1}])} I(D = (0,1)) \\
&\qquad\quad + \int_{(u_k, \infty)}\int_{(v_l, \infty)} g(s_1, s_2)\frac{F_0(ds_1, ds_2)}{F_0([u_k, \infty), (v_l, \infty))} I(D = (0,0)).
\end{aligned}$$

Recall that (u_k, v_l) is a function of $\tilde{T}$ and therefore it is natural to consider v_l as a function in $\tilde{T}_2$: $v_l(\tilde{T}_2) = v_l$, if $\tilde{T}_2 \in (v_l, v_{l+1}]$, and similarly for u_k. In this way all four terms can be considered as functions on $[0, \tau]$, where the last three are step functions on π^h.

In formulas, I_0^h is given by

$$\begin{aligned}
I_{F_0, G_h}^h(g)(T) &= g(T)H_h(T) \\
&\quad + \int_0^{T_2}\left(\int_{(u_k, u_{k+1}]}\int_{(v_l, \infty)} g(s_1, s_2)\frac{F_0(ds_1, ds_2)}{F_0((u_k, u_{k+1}], [v_l, \infty))}\right) \\
&\qquad \times G_h((u_k, \infty), \{v_l\}) \\
&\quad + \int_0^{T_1}\left(\int_{(u_k, \infty)}\int_{(v_l, v_{l+1}]} g(s_1, s_2)\frac{F_0(ds_1, ds_2)}{F_0([u_k, \infty]), (v_l, v_{l+1}])}\right) \\
&\qquad \times G_h(\{u_k\}, (v_l, \infty)) \\
&\quad + \int_{(0, T]}\left(\int_{(u_k, \infty)}\int_{v_l, \infty)} g(s_1, s_2)\frac{F_0(ds_1, ds_2)}{F_0([u_K, \infty), (v_l, \infty))}\right) \\
&\qquad \times G_h(\{u_k\}, \{v_l\}).
\end{aligned}$$

We will write down the singly censored term (second term above) of I_{F_0, G_0}: $L^2(F_0) \to L^2(F_0)$:

$$\int_0^{T_2}\left(\int_{(v_2, \infty)} h(T_1, s_2)\frac{F_{01}(T_1, ds_2)}{F_{01}(T_1, [v_2, \infty))}\right) H_0(T_1, dv_2).$$

6.1. *Proof of Lemma* 4.1.

LEMMA 6.1. *Let* $E^h_{k,l}(1,0) \equiv (u_k, u_{k+1}] \times [v_l, \infty)$ *be the vertical strips of* π^h *and let* $E^h_{k,l}(0,1)$ *be the horizontal strips. Suppose that* $H_0(\tau) > 0$ *and* $F_0(E^{h_n}_{k,l}) > \delta h_n$, *for certain* $\delta > 0$. *Then there exists an* $\varepsilon > 0$ *so that, for any sequence* h_n *which converges to zero slower than* $1/\sqrt{n}$, *we have*

$$\min_{k,l} F_n^{h_n}\big(E^{h_n}_{k,l}(1,0)\big) \geq \varepsilon h_n \quad \textit{with probability tending to } 1.$$

Similarly, for $E^{h_n}_{k,l}(0,1)$.

PROOF. We use the notation $E^h_{k,l}$ for both strips. First, by the EM equations [see (7)], we have

$$F_n^h\big(E^h_{k,l}\big) \geq P^n_{11}\big(E^h_{k,l}\big), \tag{18}$$

where P^n_{11} is the empirical distribution of the uncensored observations of $Y_i^h \sim P^h_{F_0, G_h}$. We have

$$P_{11}\big(E^{h_n}_{k,l}\big) \geq H_0(\tau) F_0\big(E^{h_n}_{k,l}\big) > \delta_1 h_n \quad \text{for some } \delta_1 > 0. \tag{19}$$

Furthermore, $\{I_{E^h_{k,l}}: h \in (0,1], k, l\}$, the collection of indicators of $E^h_{k,l}$ over all $(u_k, v_l) \in \pi^h$ and for all $h \in (0,1]$, is a uniform Donsker class. Consequently, we have for any $\varepsilon > 0$ and rate $r(n)$ slower than $\sqrt{n}$ that

$$P\left(\sup_{k,l} |(P^n_{11} - P_{11})(E^{h_n}_{k,l})| > \frac{\varepsilon}{r(n)}\right) \to 0. \tag{20}$$

Assume that there exists an $\varepsilon < \delta_1$ so that

$$\limsup_{n \to \infty} P\Big(\min_{k,l} P^n_{11}\big(E^{h_n}_{k,l}\big) \leq \varepsilon h_n\Big) > \delta > 0 \quad \text{for some } \delta > 0. \tag{21}$$

We will prove that this leads to a contradiction if h_n converges to zero slower than $1/\sqrt{n}$. The contradiction proves that, for each $\varepsilon < \delta_1$ and h_n slower than $\sqrt{n}$,

$$\lim_{n \to \infty} P\Big(\min_{k,l} P^n_{11}\big(E^{h_n}_{k,l}\big) \geq \varepsilon h_n\Big) = 1,$$

which combined with (18) proves the lemma. So it remains to prove the contradiction. We have by (19) and (21), respectively,

$$\begin{aligned} \limsup_{n \to \infty} P\Big(\sup_{k,l} |(P^n_{11} - P_{11})\big(E^{h_n}_{k,l}\big)| > \delta_1 h_n - \varepsilon h_n\Big) \\ \geq P\Big(\min_{k,l} P^n_{11}\big(E^{h_n}_{k,l}\big) \leq \varepsilon h_n\Big) > \delta > 0. \end{aligned}$$

However, we also have (20). These two contradict if h_n converges to zero slower than $1/\sqrt{n}$. □

To obtain a bound for the uniform sectional variation norm of the efficient influence function, consider the equation $I_F^h(g)(x) = f(x)$, for certain $f \in$

$L^2(F)$. We can write $I_F^h(g) = H_h g + K_F^h(g)$, where $K_F^h(g)$ is the sum of the three terms corresponding to the censored observations. Then this equation is equivalent to the following equation:

$$g(x) = \frac{1}{H_h(x)}\{f(x) - K_F^h(g)(x)\}. \tag{22}$$

For the moment denote the right-hand side by $C_F^h(g, f)(x)$; that is, we consider the equation $g(x) = C_F^h(g, f)(x)$.

We know by Lemma 3.1 that, for each f, there exists a $g' \in L^2(F)$, which is unique in $L^2(F)$, with $\|I_F^h(g') - f\|_F = 0$, that is, $\|g' - C_F^h(g', f)\|_F = 0$. Notice that if $\|g_1 - g\|_F = 0$, then for each x, $C_F^h(g_1 - g, f)(x) = 0$. So even if g' is only uniquely determined in $L^2(F)$, then $C_F^h(g', f)(x)$ is uniquely determined for each x. Now, we can define $g(x) \equiv C_F^h(g', f)(x)$. Then $\|g - g'\|_F = \|C(g', f) - g'\|_F = 0$, so in this way we have found a solution g of (22) which holds for each x instead of only in the $L^2(F)$ sense.

To summarize, we have $g_h = I_{h,F}^{-1}(f)$ is given by $g_h(x) = C_F^h(g_h', f)(x)$, where $g_h' = I_{h,F}^{-1}(f)$ in the $L^2(F)$ sense. Moreover, by the bounded invertibility of I_F^h w.r.t. the $L^2(F)$-norm, we have that $\|g_h'\|_F \le C\|f\|_F$, where $C \le 1/\delta$ does not depend on the width h.

Assume that $\|f\|_v^* < 1$. Now we can conclude that $\|g_h\|_\infty \le M\|K_F^h(g_h)\|_\infty$ and $\|g_h\|_v^* \le M\|K_F^h(g_h)\|_v^*$, for certain $M < \infty$.

Therefore it remains to bound the *sup-norm* and *uniform sectional variation norm* of $K_F^h(g)$ and find out how this bound depends on the width h_n. It suffices to do this for one of the singly censored terms of $K_F^h(g_h)$. We take the $D = (1, 0)$ term which is given by

$$W(T) \equiv \int_0^{T_2}\left(\int_{(u_k, u_{k+1}]}\int_{(v_l, \infty)} g_h(s_1, s_2)\frac{F(ds_1, ds_2)}{F((u_k, u_{k+1}], [v_l, \infty))}\right)G_h(u_k, \{v_l\}).$$

For convenience, we will often denote $E_{k,l}(1, 0)$ by $E_{k,l}$.

Supnorm. Recall that $\|f\|_\infty \le 1$. By the Cauchy–Schwarz inequality and $\|g_h\|_F \le C\|f\|_F$, we have

$$\begin{aligned}
\int_{(u_k, u_{k+1}]}\int_{(v_l, \infty)} & g_h(s_1, s_2)\frac{F(ds_1, ds_2)}{F((v_k, u_{k+1}], [v_l, \infty))} \\
&= \int I_{E_{k,l}}(s_1, s_2) g_h(s_1, s_2)\frac{F(ds_1, ds_2)}{F(E_{k,l})} \\
&\le \frac{1}{\sqrt{F(E_{k,l})}}\|g_h\|_F \\
&\le \frac{C}{\sqrt{F(E_{k,l})}}.
\end{aligned}$$

By Lemma 6.1 we can assume that $F_n^{h_n}(E_{k,l}) > \varepsilon h_n$, for certain $\varepsilon > 0$. This proves, by replacing F (above) by F_n^h:

LEMMA 6.2. *There exists a* $C < \infty$ *so that*

$$\sup_{\|f\|_\infty = 1} \|I^{-1}_{h,F_n^h}(f)\|_\infty \le \frac{C}{\sqrt{h_n}} \quad \textit{with probability tending to } 1.$$

Uniform sectional variation norm over $[0,\tau]$. Notice that W is purely discrete with jumps at the grid points (u_k, v_l). Therefore the uniform sectional variation norm of W equals the sum of the absolute values of all jumps. We have

$$W(T_1, \{v_l\}) = \int_{(u_k, u_{k+1}]} \int_{(u_l, \infty)} g_h(s_1, s_2) \frac{F(ds_1, ds_2)}{F((v_k, u_{k+1}], [v_l, \infty))} G_h(u_k, \{v_l\}).$$

So

$$\begin{aligned}\Delta W(u_k, v_l) = {}& \Delta H_h(u_k, v_l) \int_{E_{k,l}} g_h(s_1, s_2) \frac{F(ds_1, ds_2)}{F(E_{k,l})} \\ &+ \frac{-\int_{E_{k,l}} g_h(s_1, s_2) F(ds_1, ds_2)}{F(E_{k,l})^2} (F(E_{k+1,l}) - F(E_{k,l})) \\ &\times H_h(u_k, \{v_l\}) \\ &+ \frac{\left(\int_{E_{k+1,l}} g_h(s_1, s_2) F(ds_1, ds_2) - \int_{E_{k,l}} g_h(s_1, s_2) F(ds_1, ds_2)\right)}{F(E_{k,l})} \\ &\times H_h(u_k, \{v_l\}).\end{aligned}$$

Now doing nothing more sophisticated than (we use Lemma 6.2 in the first inequality and Lemma 6.1 in the second)

$$\frac{\int_{E_{k,L}} g_h \, dF}{F(E_{k,l})} \le \|g_h\|_\infty \le M/\sqrt{h_n} \quad \text{and} \quad F(E_{k,l}) > \varepsilon h_n, \tag{23}$$

we obtain the following bound:

$$\begin{aligned}\Delta W(u_k, v_l)| \le {}& |\Delta H_h(u_k, v_l)| \frac{M}{\sqrt{h_n}} \\ &+ \frac{C}{h_n^{3/2}} (F_n^h(E_{k,l}) + F_n^h(E_{k+1,l})) |H_h(u_k, \Delta v_l)|.\end{aligned}$$

Consequently, we have, for the variation of W with F replaced by F_n^h,

$$\begin{aligned}\sum_{k,l} |\Delta W(u_k, v_l)| &\le \frac{1}{\sqrt{h_n}} \sum_{k,l} |\Delta H_h(u_k, v_l)| + \frac{C}{h_n^{3/2}} \sum_{k,l} F_n^h(E_{k,l}) |H_h(u_k, \Delta v_l)| \\ &\le \frac{1}{\sqrt{h_n}} + \frac{C}{h_n^{3/2}} = O\left(\frac{1}{h_n^{3/2}}\right),\end{aligned}$$

where the bounds hold with probability tending to 1. So we have proved the following lemma:

LEMMA 6.3. *There exists a $C < \infty$ so that*

$$\sup_{\|f\|_v^*=1} \|I_{h,F_n^h}^{-1}(f)\|_v^* \leq \frac{C}{h_n^{3/2}} \quad \text{with probability tending to } 1. \tag{24}$$

Let $g = I_{h,F_n^h}^{-1}(f)$. The uniform sectional variation of the uncensored term of $A_{F_n^h}(g)$ is bounded by a constant times the uniform sectional variation of g and the uniform sectional variation of the censored terms can be bounded as above using (23) by $C/h_n^{3/2}$. Therefore the uniform sectional variation of the efficient influence curve is also bounded by the rate given in (24). This completes the proof of Lemma 4.1 (the cadlag property follows also trivially). □

6.2. *Proof of Lemma* 4.3. We will suppress the d in our notation. We have

$$\begin{aligned}
\|f_{nt}^h\|_\infty &= \|\tilde{I}^h(F_n^h, t) - \tilde{I}^h(F_0, t)\|_\infty \\
&\leq |(S_n^h - S_0)(t)| + \|A_n^h I_{h,n}^{-1}(\kappa_t) - A_0^h I_{h,0}^{-1}(\kappa_t)\|_\infty.
\end{aligned}$$

We know that $\|F_n^h - F_0\|_\infty = O_P(1/(\sqrt{nh_n^3}))$. The rate will be determined by the second term. Let $g_{0t}^h \equiv I_{h,0}^{-1}(\kappa_t)$. We rewrite the second term as a sum of two differences:

$$\begin{aligned}
&A_n^h I_{h,n}^{-1}(\kappa_t) - A_0^h I_{h,0}^{-1}(\kappa_t) \\
&\qquad = (A_n^h - A_0^h) I_{h,0}^{-1}(\kappa_t) + A_n^h I_{h,n}^{-1}(I_n^h - I_0^h) I_{h,0}^{-1}(\kappa_t) \\
&\qquad = (A_n^h - A_0^h)(g_{0t}^h) + A_n^h I_{h,n}^{-1}(I_n^h - I_0^h)(g_{0t}^h).
\end{aligned} \tag{25}$$

We will consider the first term. It suffices to do the analysis for one of the singly censored terms. We consider the $d = (1,0)$ term. We have, by telescoping,

$$\begin{aligned}
&(A_n^h - A_0^h)(g_{0t}^h)(u_k, v_l, d) \\
&\quad = \frac{\int_{E(k,l)} g_{0t}^h \, dF_n^h}{F_n^h(E_{k,l})} - \frac{\int_{E(k,l)} g_{0t}^h \, dF_0}{F_0(E_{k,l})} \\
&\quad = \frac{\int_{E(k,l)} g_{0t}^h d(F_n^h - F_0)}{F_0(E_{k,l})} + \frac{(F_n^h - F_0)(E_{k,l}) \int_{E(k,l)} g_{0t}^h \, dF_n^h}{F_n^h(E_{k,l}) F_0(E_{k,l})}.
\end{aligned}$$

In the first term, we can apply integration by parts. So the first term is bounded by

$$C \|F_n^h - F_0\|_\infty \frac{\|g_{0t}^h\|_v^*}{F_0(E_{k,l})}.$$

By Lemma 6.3 we have $\|g_{0t}^h\|_v^* = O(1/\sqrt{h_n^3})$ and we have $F_0(E_{k,l}) > \delta h$. Therefore the first term is bounded by

$$O_P\left(\frac{1}{\sqrt{nh_n^3}}\right) O_P\left(\frac{1}{\sqrt{h_n^3}}\right) O\left(\frac{1}{h_n}\right) = O_P\left(\frac{1}{\sqrt{nh_n^8}}\right).$$

The second term is bounded by

$$C\|F_n^h - F_0\|_\infty \|g_{0t}^h\|_\infty \frac{1}{F_0(E_{k,l})} = O_P\left(\frac{1}{\sqrt{nh_n^6}}\right).$$

This proves that

$$\|(A_n^h - A_0^h)(g_{0t}^h)\|_\infty = O_P\left(\frac{1}{\sqrt{nh_n^8}}\right).$$

Consider now the second term of (25). Because A_0^T only depends on G, we have, for the term $(I_n^h - I_0^h)(g_{0t}^h)$

$$(I_n^h - I_0^h)(g_{0t}^h) = A_0^{hT}(A_n^h - A_0^h)(g_{0t}^h).$$

Because A_0^{hT} is just a conditional expectation, we have that $\|A_0^{hT}(g)\|_\infty \leq \|g\|_\infty$. Therefore, we also have that $\|(I_n^h - I_0^h)(g_{0t}^h)\|_\infty = O(1/\sqrt{nh_n^8})$. Now, we apply Lemma 6.2 which tells us that $\|I_{h,n}^{-1}(g)\|_\infty \leq (1/\sqrt{h_n})\|g\|_\infty$. This tells us that

$$\|A_n^h I_{h,n}^{-1}(I_n^h - I_0^h)(g_{0t}^h)\|_\infty = O\left(\frac{1}{\sqrt{nh_n^9}}\right).$$

This completes the proof of Lemma 4.3. □

6.3. *Proof of Lemma* 4.7. Lemma 4.7 will be proved as a corollary of the next lemma.

LEMMA 6.4. *Let $C \subset L^2(F_0)$ be any compact set in $L^2(F_0)$. Then we have*

$$\sup_{g \in C} \|(I_0^h - I_0)(g)\|_{F_0} \to 0 \tag{26}$$

and

$$\sup_{g \in C} E\big(A_0^h(g) - A_0(g)\big)^2 \to 0 \quad \textit{for } h = h_n \to 0.$$

PROOF. By the compactness of C and the continuity of I_0^h: $L^2(F_0) \to L^2(F_0)$, the supremum in (26) is attained by some $g_0 \in C$. Let g_k be a sequence so that $\|g_k - g_0\|_{F_0} \to 0$ and $\|g_k\|_\infty < \infty$, for $k = 1, 2, \dots$. We have

$$\|(I_0^h - I_0)(g_0)\|_{F_0} \leq \|(I_0^h - I_0)(g_0 - g_k)\|_{F_0} + \|(I_0^h - I_0)(g_k)\|_{F_0};$$

$\|(I_0^h - I_0)(g_0 - g_k)\|_{F_0} \leq 2\|g_0 - g_k\|_{F_0}$, which converges to zero for $k \to \infty$. Therefore it suffices now to show that $\|(I_0^{h_n} - I_0)(g_k)\|_{F_0} \to 0$, for each fixed k. Now, we have

$$\left(I_0^h - I_0\right)(g_k) = A_0^{hT}\left(A_0^h - A_0\right)(g_k) + \left(A_0^{hT} - A_0^T\right)\left(A_0(g_k)\right).$$

The differences in the first term are comparable because all can be considered as functions of (C, T) and thereby are defined on the same probability space. First, we will consider the second term. It suffices to deal with one of the singly censored terms. Let $d = (1, 0)$ and $f_k \equiv A_0(g_k)I(D = d)$. We have

$$\left(A_0^{hT} - A_0^T\right)(f_k)(T_1, T_2) = \int_0^{T_2} f_k(T_1, v)(G_h - G_0)((T_1, \infty), dv).$$

Let $T = (T_1, T_2)$ be fixed and let T_2 be a point where $H_0(T_1, \Delta T_2) = 0$. By definition of weak convergence of $H_h(T_1, dv)$ to $H_0(T_1, dv)$ we have now that if $v \to f_k(T_1, v)$ is bounded and continuous $H_0(T_1, \cdot)$ a.e., then $(A_0^{hT} - A_0^T)f_k(T_1, T_2) \to 0$, for this T. The boundedness follows from $\|f_k\|_\infty \leq \|g_k\|_\infty < \infty$. We have that $v \to f_k(T_1, v)$ is given by

$$v \to \frac{\int_v^\infty g_k(T_1, k\, v_2) F_{01}(T_1, dv_2)}{F_{01}(T_1, (v, \infty))}.$$

This function is continuous at v if $v \to F_{01}(T_1, v)$ is continuous at v. Consequently, we need that $F_{01}(T_1, dv)$ puts no mass at a point where $H_0(T_1, dv)$ puts mass. By our convention that if $T = C$, then the observation is uncensored, this is satisfied. This proves the pointwise convergence of $f_h \equiv (A_0^{hT} - A_0^T)(f_k)$ to zero F-a.e. We need to show that $\int f_h^2\, dF_0 \to 0$. However, we also have $\|f_h\|_\infty \leq 2\|g_k\|_\infty$ and therefore the dominated convergence theorem provides us with $\int f_h^2\, dF_0 \to 0$.

Let us now consider the first term $A_0^{hT}(A_0^h - A_0)(g_k)$. Because A_0^h is a conditional expectation, its second moment is bounded by the second moment of $(A_0^h - A_0)(g_k)$. Therefore it suffices to show that $E_{X,C}((A_0^h - A_0)(g_k)^2) \to 0$ for $h \to 0$, where we consider A_0^h and A_0 as functions in (T, C) via Y^h and Y, respectively.

Recall how we constructed the data $(\tilde{T}, D)^h$: $\subset 1)$ We have a nested sequence of partitions π^h and we observed i.i.d. $C_1, \dots, C_n \sim G$. (2) Now we discretize C_i such that $C_i^h \sim G_h$, where G_h lives on π^h. This provides us with data $(\tilde{T}, D)_h \sim P_{F_0, G_h}$. (3) Finally, we discretize $(\tilde{T}, D)_h$ in order to obtain $Y^h = (\tilde{T}, D)^h \sim P_{F_0, G_h}^h$. Denote the sigma field generated by Y^h with $\mathscr{A}^h$. Because π^h is nested and the sigma field generated by π^h converges to the Borel sigma field on $[0, \tau]$, we have that $\mathscr{A}^h \uparrow \mathscr{A}^\infty$, for $h \to 0$, where $\mathscr{A}^\infty$ is the sigma field generated by $Y = (\tilde{T}, D))$, $Y \sim P_{F_0, G_0}$.

Consequently $M_{h_n} \equiv E_{X,C}(g_k(T)|\mathscr{A}^{h_n})$ is a martingale in n and it is well known that if $\sup_h E(M_h^2) < \infty$, then $E((M_h - M_0)^2) \to 0$. We have

$$\sup_h E\left(E\left(g_h(T)|\mathscr{A}^h\right)^2\right) \leq \|g_k\|_\infty < \infty$$

and consequently we have $\|(A_0^h - A_0)(g_k)\|_{F_0 \times G_0} \to 0$. This also proves the second statement in Lemma 6.4. □

COROLLARY 6.1. *We make the same assumptions as in Lemma* 6.4. *For each set $C \subset L^2(F_0)$ which is compact w.r.t. $\|\cdot\|_{F_0}$, we have, for $h \to 0$,*

$$\sup_{g \in C} \|\big(I_0^{-1} - I_{h,0}^{-1}\big)(g)\|_{F_0} \to 0. \tag{27}$$

This implies

$$\sup_{g, g_1 \in C} \Big| \langle A_0^h I_{h,0}^{-1}(g), A_0^h I_{h,0}^{-1}(g_1) \rangle_{P_0^h} - \langle A_0 I_0^{-1}(g), A_0 I_0^{-1}(g_1) \rangle_{P_{F_0}} \Big| \to 0.$$

Moreover, we have

$$\sup_{g \in C} E\big(A_0^h I_{h,0}^{-1}(g) - A_0 I_0^{-1}(g)\big)^2 \to 0.$$

PROOF. We have

$$\begin{aligned}\big(I_{h,0}^{-1} - I_0^{-1}\big)(g) &= I_{h,0}^{-1}\big(I_0 I_0^{-1} - I_0^h I_0^{-1}\big)(g)\\ &= -I_{h,0}^{-1}\big(I_0^h - I_0\big) I_0^{-1}(g).\end{aligned}$$

First, notice that by the bounded L^2-invertibility of I_0 (Lemma 3.1), $I_0^{-1}(C)$ is compact in $L^2(F_0)$. Now, by the preceding lemma we have that

$$\sup_{g \in C} \|\big(I_0^h - I_0\big) I_0^{-1}(g)\|_{F_0} \to 0.$$

Finally, we know by Lemma 3.1 that $\sup_h \|I_{h,0}^{-1}\|_{F_0} < \infty$. This proves the first statement. For the second statement notice that

$$\begin{aligned}\langle A_0^h I_{h,0}^{-1}(g), A_0^h I_{h,0}^{-1}(g_1) \rangle_{P_{F_0}^h} &= \langle I_{h,0}^{-1}(g), g_1 \rangle_{F_0}\\ &= \langle I_{h,0}^{-1}(g) - I_0^{-1}(g), g_1 \rangle_{F_0} + \langle I_0^{-1}(g), g_1 \rangle_{F_0}.\end{aligned}$$

The first term converges to zero by the Cauchy–Schwarz inequality and (27). The second term equals $\langle A_0 I_0^{-1}(g), A_0 I_0^{-1}(g_1) \rangle_{P_{F_0}}$.

It remains to prove the last statement. By the compactness of C and continuity of $A_0 I_0^{-1}$ and $A_0^h I_{h,0}^{-1}$ it suffices to show the statement for a fixed $g \in L_0^2(F_0)$. We have

$$A_0 h I_{h,0}^{-1}(g) - A_0 I_0^{-1}(g) = \big(A_0^h - A_0\big) I_0^{-1}(g) + A_0^h \big(I_{h,0}^{-1} - I_0^{-1}\big)(g).$$

The first term converges to zero by the second statement of Lemma 6.4.

For the second term we have

$$\|A_0^h\big(I_{h,0}^{-1} - I_0^{-1}\big)(g)\|_{P_0^h} \le \|\big(I_{h,0}^{-1} - I_0^{-1}\big)(g)\|_{F_0} \to 0 \quad \text{by (27)}. \qquad □$$

Notice that $C \equiv \{I(0,t]: t \in [0,\tau]\} \subset L^2(F_0)$ is a compact set. Application of the corollary to this set C provides us with Lemma 4.7. □

Acknowledgments. I owe thanks to my thesis adviser Richard Gill and second promotor Peter Bickel for numerous discussions and helpful comments. I thank the referees for their improvements.

REFERENCES

BAKKER, D. M. (1990). Two nonparametric estimators of the survival function of bivariate right censored observations. Report BS-R9035, Centrum Wisk. Inform., Amsterdam.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BICKEL, P. J., KLAASSEN, A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semi-Parametric Models*. Johns Hopkins Univ. Press.

BURKE, M. D. (1988). Estimation of a bivariate survival function under random censorship. *Biometrika* **75** 379–382.

DABROWSKA, D. M. (1988). Kaplan–Meier estimate on the plane. *Ann. Statist.* **16** 1475–1489.

DABROWSKA, D. M. (1989). Kaplan–Meier estimate on the plane: weak convergence, LIL, and the bootstrap. *J. Multivariate Anal.* **29** 308–325.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

EFRON, B. (1967). The two sample problem with censored data. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* 831–853. Univ. California Press, Berkeley.

EINMAHL, J. H. H. (1987). *Multivariate Empirical Processes. CWI Tract* **32**. Centrum Wisk. Inform., Amsterdam.

GILL, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 1). *Scand. J. Statist.* **16** 97–128.

GILL, R. D. (1992). Multivariate survival analysis. *Theory Probab. Appl.* **37** 18–31 and 284–301. (English translation.)

GILL, R. D. (1994). Lectures on survival analysis. *Ecole d'Eté de Probabilités de Saint Flour XXII. Lecture Notes in Math.* **1581** 115–241. Springer, Berlin.

GILL, R. D., VAN DER LAAN, M. J. and WELLNER, J. A. (1993). Inefficient estimators of the bivariate survival function for three models. *Ann. Inst. H. Poincaré Probab. Statist..* **31** 547–597.

HEITJAN, D. F. and RUBIN, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253.

HOFFMANN-JØRGENSEN, J. (1984). Stochastic processes on Polish spaces. Unpublished manuscript.

NEUHAUS, G. (1971). On weak convergence of stochastic processes with multidimensional time parameter. *Ann. Math. Statist.* **42** 1285–1295.

PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces*. Academic Press, New York.

POLLARD, D. (1990). *Empirical Processes: Theory and Applications*. IMS, Hayward, CA.

PRENTICE, R. L. and CAI, J. (1992a). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika* **79** 495–512.

PRENTICE, R. L. and CAI, J. (1992b). Marginal and conditional models for the analysis of multivariate failure time data. In *Survival Analysis State of the Art* (Klein, J. P. and Goel, P. K., eds.). Kluwer, Dordrecht.

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

PRUITT, R. C. (1991b). Strong consistency of self-consistent estimators: general theory and an application to bivariate survival analysis. Technical Report 543, Univ. Minnesota.

PRUITT, R. C. (1993). Small sample comparisons of six bivariate survival curve estimators. *J. Statist. Comput. Simulation.* **45** 147–167.

TSAI, W-Y., LEURGANS, S. and CROWLEY, J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *Ann. Statist.* **14** 1351–1365.

TURNBULL, B. W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data. *J. Roy. Statist. Soc. Ser. B* **38** 290–295.

VAN DER LAAN, M. J. (1990). Dabrowska's multivariate product limit estimator and the delta-method. Master's dissertation, Dept. Mathematics, Univ. Utrecht, The Netherlands.

VAN DER LAAN, M. J. (1993). General identity for linear parameters in convex models with application to efficiency of the (NP)MLE. Preprint 765, Dept. Mathematics, Univ. Utrecht, The Netherlands.

VAN DER LAAN, M. J. (1994). Modified EM-estimator of the bivariate survival function. **3** 213–243. *Math. Methods Statist.*.

VAN DER LAAN, M. J. (1995). Efficiency of the NPMLE in a general class of missing data models. Unpublished manuscript.

VAN DER LAAN, M. J. (1996). Efficient and inefficient estimation in semiparametric models. Technical Report, CWI, Amsterdam.

VAN DER VAART, A. W. (1988). Statistical estimation in large parameter spaces. *CWI Tract* **44**. Centrum Wisk. Inform. Amsterdam.

VAN DER VAART, A. W. AND WELLNER, J. A. (1995). *Weak Convergence and Empirical Processes*. IMS, Hayward, CA.

DIVISION OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
EARL WARREN HALL
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY , CALIFORNIA 94720