

# Efficient Estimation in the Bivariate Censoring Model and Repairing NPMLE

Mark J. van der Laan  
Department of Mathematics  
University of Utrecht  
P.O.Box 80010, 3508 TA Utrecht  
The Netherlands.

May 15, 2001

## Abstract

The NPMLE in the bivariate censoring model is not consistent. The problem is caused by the singly censored observations. In this paper we prove that the NPMLE based on interval censoring the singly censored observations is 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 efficient for the original data. We are able to determine a lower bound for the rate at which the bandwidth should converge to zero.

The efficiency proof uses the general identity which holds for NPMLE of a linear parameter in convex models as proved in van der Laan (1993a).

## 1 Introduction.

We do not 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  $T = (T_1, T_2) \in \mathbb{R}_{\geq 0}^2$  and if we write  $\leq$ ,  $\geq$ ,  $<$ ,  $>$  then this should hold componentwise: for example if  $x, y \in \mathbb{R}^2$  then  $x \leq y \Leftrightarrow x_1 \leq y_1, x_2 \leq y_2$ . Assuming the notation, we will write  $T_i, i = 1, \dots, n$ , as notation for  $n$  i.i.d. bivariate observations with the same distribution as  $T$ , while we write  $T_1$  and  $T_2$  for the components of  $T$ .

A formal description of the model for estimating the *bivariate survival function* based on *right censored* and uncensored observations is the following.  $T$  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  $S_0(t) \equiv \Pr(T > t)$ .  $C$  is a positive bivariate censoring vector with bivariate distribution  $G_0$  and survivor function  $H_0$ ;  $G_0(t) \equiv \Pr(C \leq t)$  and  $H_0(t) \equiv \Pr(C > t)$ .  $T$  and  $C$  are independent;  $(T, C) \in \mathbb{R}^4$  has distribution  $F_0 \times G_0$ .  $(T_i, C_i), i = 1, \dots, n$  are  $n$  independent copies of  $(T, C)$ . We only observe the following many to one mapping  $\Phi$  of  $(T_i, C_i)$ :

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

with components given by:

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

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 estimating  $S_0$ .

The probability measure  $P_{F_0, G_0}$  of the data is indexed by two unknown parameters  $F_0$  and  $G_0$ . Each observation  $Y_i$  tells us that  $(T_i, C_i) \in B(Y_i) \equiv \Phi^{-1}(Y_i) \subset \mathbb{R}^2 \times \mathbb{R}^2$ . Therefore this model is a *missing data model*.

For each region  $B(Y_i)$  we have that  $B(Y_i) = B(Y_i)_1 \times B(Y_i)_2$  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. In other words, observing  $Y_i$  is equivalent with observing that  $T_i \in B(Y_i)_1$  and  $C_i \in B(Y_i)_2$ . Because  $T$  and  $C$  are independent it follows now that  $P(T \in dt \mid C \in B(Y)_2) = P(T \in dt)$ , which means that the observation that  $C_i \in B(Y_i)_2$  does not give any information about  $T_i$  and thereby for estimating  $F_0$ . Formal information calculations indeed show that knowing  $G_0$  does not increase the information for estimating  $F_0$ , as we will see in section 3: the efficient influence function for estimating  $S_0(t)$  in the model with  $G$  unknown equals the efficient influence function for estimating  $S_0(t)$  in the model with  $G$  known.

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 4 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 rest are incomplete observations. In the literature there has been paid a lot of attention to constructing ad hoc explicit estimators. For a description of the literature for this model we refer to the bibliographic remarks (section 7) at the end of this paper.

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 iterating the self-consistency equation. In chapter 3 of van der Laan (1993e) (and van der Laan, 1993d) we analyzed a general class of missing data models. There we found that there are essentially two *crucial* assumptions for efficiency of the NPMLE. The first assumption says for the bivariate censoring model that, given  $T = t$ , the probability that  $T$  will be observed is larger than  $\delta > 0$ ; in other words  $H_0(t) > \delta > 0$  on the support of  $F_0$ , which is an assumption which can be naturally arranged. Assumption 2 says: for each incomplete observation  $Y_i$  we need

$$P(T \in B(Y_i)_1) = F_0(B(Y_i)_1) > \delta > 0.$$

If  $F_0$  is continuous, then this assumption is not satisfied for the singly-censored observations, because then the probability that  $T$  falls on a line is zero. In a specific analysis of a model one might be able to weaken these two assumptions to their version with  $\delta = 0$ , though then the estimators will be unstable. The heuristic behind these assumptions was the following: In the EM-algorithm the incomplete observations  $Y_i$  need to get information from the observed  $X_i$  about how to redistribute their mass  $1/n$  over  $B(Y_i)_1$ , and for this purpose they need complete observations in  $B(Y_i)_1$ . Hence we need that  $F_0(B(Y_i)_1) > \delta > 0$ . Indeed it is well known that the NPMLE for continuous data is not consistent (Tsai, Leurgans and Crowley, 1986).

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  $T_i$  of the singly censored observations by the observation that  $T_i$  lies in a small predetermined interval around  $T_i$ . This interval will have a width of magnitude  $h = h_n$ . Now, for these interval censored singly censored observations  $Y_i^h$  the regions  $B(Y_i^h)_1$  are strips and therefore assumption 2 is satisfied. Because we do not touch the uncensored observations assumption 1 still requires that  $H > \delta > 0$ , which can be easily arranged by reducing the data to  $[0, \tau]$ , where  $\tau$  is chosen so that  $H(\tau) > 0$ .

The interval censoring of the singly censored observations causes one problem. Namely, the conditional density of  $T$  given what we observe about  $C$  does not equal the conditional density of  $T$  anymore. Therefore the joint likelihood for  $F$  and  $G$  does not factorize anymore in a  $F$ -term and  $G$ -term which tells us that for computing the NPMLE of  $F$  we also need to maximize over  $G$ . In section 2 we discuss this problem and give a number of proposals. The estimator we analyze is based on discretized  $C_i$  so that the joint likelihood factorizes, again. In fact, we prove efficiency of the sieved-NPMLE for the case that we observe  $C_1, \dots, C_n$  or if  $G_0$  is known. However, we introduce a method for simulating  $C_i$  so that the estimator can also be computed if neither of these hold and it will be heuristically clear that this estimator will have essentially the same performance as the analyzed one.

After having recovered the orthogonality between  $F$  and  $G$ , the general efficiency theorem 6.2 in van der Laan (1993d) for missing data models tells us that we should expect a good performance of the sieved-NPMLE  $F_n^h$  of  $F_0$  based on  $Y_i^h$  (and reducing the data to  $[0, \tau]$ ) and indeed efficiency for this transformed data can easily be proved by verifying the assumptions of theorem 6.2, which would prove supnorm efficiency of  $F_n^h$  for the transformed data reduced to a rectangle  $[0, \tau]$ . We state this result precisely in our theorem, but leave the verification for the reader; it follows also by keeping  $h$  fixed in the analysis followed in this paper (we do the analysis for  $h_n \rightarrow 0$  which implies results for fixed  $h$ ).

For obtaining efficiency for the original data we have to let the width  $h = h_n$  of the strips converge to zero slowly enough. We will prove this and give a lower bound on the rate at which  $h_n$  should converge to zero.

We will call this sieved-NPMLE based on a reduction, or call it a slight transformation, of the data a ‘‘Sequence of Reductions’’-NPMLE and will abbreviate it with SOR-NPMLE. 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 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-NPMLE 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-NPMLE which holds in general for convex models which are linear in the parameter (van der Laan, 1993a). 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. In section 7 we have some bibliographic remarks. For validity of the nonparametric and semiparametric bootstrap we refer to section 4.7 in van der Laan (1993e); these results follow easily from the identity approach which we follow.

## 2 SOR-NPMLE for the bivariate censoring model.

Our original data is given by:

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

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 with  $D = (0, 1)$ ,  $D = (1, 0)$  and  $D = (0, 0)$ , respectively. Then

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

Let  $f_0 \equiv dF_0/d\mu$  for some measure  $\mu$  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  $\mu$  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-Nykodim derivative and similarly we define  $S_{01}(x_1, x_2) \equiv -S_0(dx_1, x_2)/\mu(dx_1, (x_2, \infty))$ ,  $H_{01}$  and  $H_{02}$ . Then the density  $p_0$  of  $P_{F_0, G_0}$  w.r.t.  $(\mu \times \nu)\Phi^{-1}$  is given by

$$\begin{aligned} p_0(x_1, x_2, d) &= f_0(x)H_0(x)I(d = (1, 1)) + S_{01}(x_1, x_2)H_{02}(x_1, x_2)I(d = (0, 1)) \\ &\quad + S_{02}(x_1, x_2)H_{01}(x_1, x_2)I(d = (1, 0)) + S_0(x)g_0(x)I(d = (0, 0)) \\ &\equiv p_{11}(x)I(d = (1, 1)) + p_{01}(x)I(d = (0, 1)) + p_{10}(x)I(d = (1, 0)) \\ &\quad + p_{00}(x)I(d = (0, 0)) \\ &= \sum_{\delta \in \{1,0\}^2} p_\delta(x)I(D = \delta). \end{aligned} \quad (2)$$

It is important to notice that  $p_0(\cdot, D = d)$  for a fixed  $d$  factorizes in a part which only depends on  $F_0$  and a part which only depends on  $G_0$ . This tells us that the NPMLE of  $F_0$  can be computed by just maximizing that part of the log likelihood which only depends on  $F_0$  and that the ranges of the score operators for  $F_0$  and  $G_0$  are *orthogonal* in  $L_0^2(P_0)$  (see section 3).

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 (in  $n$ , order to make martingale arguments work) grid of  $[0, \tau]$  which depends on a scalar  $h = h_n$  in the following way  $\epsilon h_n < u_{k+1} - u_k < M h_n$ , where  $\epsilon$  and  $M$  are independent of  $n, k$ , and

similarly for  $v_{l+1} - v_l$ . In other words, the grid must have a width between  $\epsilon h_n$  and  $Mh_n$ . This tells us that the grid  $\pi^h$  has (in order of magnitude)  $1/h_n^2$  points  $(u_k, v_l)$ .

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

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

where  $\text{Id}^h$  is a many to one mapping on our original data  $(\tilde{T}_i, D_i)$  which is defined as follows.

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

We used the notation  $\text{Id}^h$  (Id from Identity) because for  $h \rightarrow 0$  (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  $G_0$ .

To be more precise, we have

$$Y^h \sim P_{F_0, G_0}^h(\cdot, \cdot),$$

where

$$\begin{aligned} P_{F_0, G_0}^h(\cdot, D = d) &= P_{11}(\cdot)I(d = (1, 1)) + P_{01}^h(\cdot)I(d = (0, 1)) + P_{10}^h(\cdot)I(d = (1, 0)) \\ &\quad + P_{00}(\cdot)I(d = (0, 0)), \end{aligned} \quad (3)$$

where

$$\begin{aligned} p_{01}^h(y_1, v_l) &= \int_{(v_l, v_{l+1}]} p_{01}(y_1, y_2) \mu((y_1, \infty), dy_2) \\ &= \int_{(v_l, v_{l+1}]} S_{02}(y_1, y_2) H_{01}(y_1, y_2) \mu((y_1, \infty), dy_2). \end{aligned}$$

Similarly,  $p_{10}^h(u_k, y_2) = \int_{(u_k, u_{k+1}]} p_{10}(y_1, y_2) \mu(dy_1, (y_2, \infty))$ . We denote the density of  $P_{F_0, G_0}^h$  w.r.t.  $(\mu \times \nu) \Phi_h^{-1}$  by  $p_0^h$ .

In this model  $p_0^h(\cdot, d)$  does not factorize anymore in a part which only depends on  $F_0$  and a part which only depends on  $G_0$ . Thus in order to be able to compute the NPMLE  $F_n^h$  we need to write down the likelihood for  $(F, G)$  and maximize over  $(F, G)$  which provides us with the joint NPMLE  $(F_n^h, G_n^h)$ . Because of similar reasons as for  $F_n$  the NPMLE  $G_n^h$  will only be good if we do a symmetric reduction (lines should be strips for C as well as for T). Therefore a proposal of which we expect a good behavior is the joint NPMLE  $(F_n^h, G_n^h)$  based on rectangle-censored  $((u_k, u_{k+1}] \times (v_l, v_{l+1}])$  singly-censored observations instead of our chosen interval-censored singly-censored observations; so now we reduce the data so that we behave as if we only observe that the singly censored observations  $(\tilde{T}_i, d_i)$ ,  $d_i = (1, 0)$  or  $d_i = (0, 1)$ , fall in the rectangle  $(u_k, u_{k+1}] \times (v_l, v_{l+1}])$  which contains  $\tilde{T}_i$ . We can compute the joint NPMLE  $(F_n^h, G_n^h)$  by iterating the self-consistency equations for  $(F, G)$  jointly.

In practice, one might also just plug a  $G_n$  in the log likelihood, instead of reducing the data, again; write down the joint likelihood for  $F$  and  $G$ , substitute  $G_n$  for  $G$  and

maximize over  $F$ . Because the likelihood does factorize asymptotically, it should suffice to use here an inefficient estimator for  $G$  instead of an efficient estimator (if you had plugged in the NPMLE of  $G$  you would have found the NPMLE of  $F$ ). The last estimator does not require extra randomisation and is clearly less computer intensive.

Because the involvement of  $G$  in computing the NPMLE  $F_n^h$  certainly complicates the analysis and it makes the estimator more computer intensive we decided to do another further reduction of the data which recovers the orthogonality, while at the same, as will appear, not losing asymptotic efficiency. The further reduction is based on the insight that if  $G_0$  is purely discrete on  $\pi^h$ , then  $p_{F_0, G_0}^h(\cdot, d)$  factorizes. This further reduction leads also to a good practical estimator as appears in the simulations in chapter 8 of van der Laan (1993e).

For the further reduction we need to observe the  $n$  i.i.d.  $C_i$  or simulate them. We split up in three cases as follows:

**Case 1.** Suppose that we observe  $C_i$ . Move each  $C_i$  to the left lower corner  $(u_k, v_l)$  of the rectangle  $((u_k, v_l), (u_{k+1}, v_{l+1}))$  of  $\pi^h$  which contains  $C_i$ . Denote these discretized  $C_i$  by  $C_i^h$ . Then  $C_i^h \sim G_h$  where  $G_h$  is the step function with jumps on  $\pi^h$  corresponding with  $G_0$ . Consider now the observations  $Y_i(C_i^h, T_i) = \Phi(C_i^h, T_i) \sim P_{F_0, G_h}$ . Notice that we are able to observe these  $Y_i(C_i^h, T_i)$  because for this we only need to know  $Y_i(C_i, T_i)$ . That is the reason for moving  $C_i$  to the left lower corner which means some extra censoring. Now, we just apply  $\text{Id}^h$  to the i.i.d. data  $Y_i(C_i^h, T_i) \sim P_{F_0, G_h}$ . Then we obtain i.i.d. data  $Y_i^h = \Phi^h(C_i^h, T_i) \sim P_{F_0, G_h}^h$ , where the density  $p_{F_0, G_h}^h(\cdot, D)$  factorizes and therefore the NPMLE based on  $Y_i^h(C_i^h, T_i)$  can be computed without knowing  $G_h$ .

**Case 2.** Assume  $G_0$  is known. Then the following simulation method takes care that our observations are sampled from such a discretized  $G_0$ .

*Simulation method.* Each observation  $Y_i$  tells us that  $C_i \in B(Y_i)_2$  (line, point, or quadrant). Given  $Y_i$  we draw a  $C_i' \sim G_0(\cdot \mid C \in B(Y_i)_2)$ ; in other words we draw an observation from the conditional distribution under  $G_0$  of  $C_i$  given that  $C_i$  falls in  $B(Y_i)_2$ . We have now  $n$  i.i.d.  $C_1', \dots, C_n'$  where  $C_i' \sim G_0' \equiv E_{P_0}(G_0(\cdot \mid Y)) = G_0$  and moreover  $\Phi(T_i, C_i') = \Phi(T_i, C_i)$ . Now, as above we can discretize these  $C_i'$  in order to obtain  $C_i^h \sim G_h$  and its corresponding i.i.d. data  $Y_i^h(C_i^h, T_i) \sim P_{F_0, G_h}^h$ .

**Case 3.** If neither of above holds, then one can estimate  $G_0$  with an estimator  $G_n$  and carry out the simulation method with  $G_n$  instead of  $G_0$ .

Our analysis proves efficiency of the sieved-NPMLE based on  $n$  i.i.d. observations  $Y_i^h(T_i, C_i^h) \sim P_{F_0, G_h}^h$  as obtained in case 1 and 2. Since the estimate  $G_n$  in case 3 uses all the  $C_1, \dots, C_n$ , the observations obtained in the third case are identical but not completely independent (though the dependence is very weak): if  $C_1' = c_1$ , then the value  $c_1$  says something about the estimator  $G_n$  and hence over the distribution of  $C_2'$ . Therefore our analysis does not cover the third case, but it is at least a practical proposal which is less computer intensive than the joint NPMLE  $(F_n^h, G_n^h)$  and which approximates the second case.

Let  $P_n^h$  be the empirical distribution function based on  $n$  i.i.d.  $Y_i^h(T_i, C_i^h) \sim P_{F_0, G_h}^h$  which is the distribution of the data corresponding with  $X \sim F_0$ ,  $C \sim G_h$ , where  $G_h$  is discrete on the grid  $\pi^h$ , and the singly censored observations are interval censored by  $\text{Id}^h$  (i.e. grouped to strips). Notice that  $P_{F_0, G_h}^h(\cdot, d)$ ,  $d \neq (1, 1)$ , is purely discrete on  $\pi^h$ .

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  $\mu_n$  be the counting measure on  $\{x_1, \dots, x_{m(n)}\}$ . Now, we let  $\mathcal{F}(\mu_n)$  be the set of all distributions which are absolutely continuous w.r.t.  $\mu_n$ .

We define our sieved-SOR-NPMLE  $F_n^h$  of  $F_0$  which we will analyze;

$$F_n^h = \arg \max_{F \in \mathcal{F}(\mu_n)} \int \log(p_{F, G_h}^h) dP_n^h, \quad (4)$$

where the maximum can be determined without knowing  $G_h$  by maximizing the term which only depends on  $F$ . We define  $S_n^h$  as the survival function corresponding with  $F_n^h$ .

**Summary of practical proposals.** In this paper, we prove efficiency of  $F_n^h$  as defined in (4) which assumes that we observe  $C_i$  or that  $G_0$  is known so that the simulation method leads to  $n$  i.i.d.  $Y_i^h \sim P_{F_0, G_h}^h$ ,  $i = 1, 2, \dots, n$ . In chapter 8 of van der Laan (1993e) it appears that  $F_n^h$  has also a very good practical performance. If we are in the third case, then we expect that estimating  $G_n$  and applying the simulation method will lead to an estimator which is very close (second order difference) in behavior to  $F_n^h$  so that our results will also hold for this estimator. Other good practical methods are 1) plug an estimator  $G_n$  for  $G$  in the joint loglikelihood of  $(F, G)$  and maximize over  $F$ , 2) compute the joint NPMLE  $(F_n^h, G_n^h)$  based on rectangular-censored singly censored observations.

## 2.1 Existence and uniqueness of the sieved-SOR-NPMLE and EM- equations.

We refer to lemma 4.1 in van der Laan (1993d) for the general class of missing data models. For application of this lemma we need to verify certain assumptions 1 and 2. Assumption 1 requires that  $H_0 > \delta > 0$   $F_0$  a.e. and assumption 2 requires that  $F_0(B(Y_i^h)_1) > \delta > 0$  for all censored  $Y_i^h$  ( $D = (1, 0)$ ,  $D = (0, 1)$ ,  $D = (0, 0)$ ). This holds if all data lives on a rectangle  $[0, \tau] \subset \mathbb{R}_{\geq 0}$ , where  $\tau$  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  $P_{F_0^\tau, G_h}^h$ , where  $F_0^\tau$  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  $n \rightarrow \infty$ . 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.

Application of lemma 4.1 in van der Laan (1993d) provides us under the stated assumptions and the artificial censoring to  $[0, \tau]$  with the existence and uniqueness (for  $n$  large enough,  $h$  fixed) of  $F_n^h$  and that  $F_n^h$  solves:

$$P_n^h(A_{F_n^h}^h(g - F_n^h(g))) = 0 \text{ for all } g \in L^2(F_n^h) \text{ with } \|g\|_\infty < \infty, \quad (5)$$

where the so called score operator  $A_F^h$  for  $F$  is given by:

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

Moreover, it says that for each set  $A$ :

$$F_n^h(A) \geq P_{11}^n(A). \quad (6)$$

### 3 Outline of the efficiency proof.

Firstly, we define the models corresponding with the data  $Y^h$  and  $Y$ . Let  $\mathcal{F}$  be the set of all bivariate distributions on  $[0, \infty)$  and  $\mathcal{F}_h$  be the set of all possible bivariate distributions  $G_h$  which live on  $\pi^h$ . Then the model corresponding with  $Y^h$  (see (3)) is given by

$$\mathcal{M}_h \equiv \{P_{F,G_h}^h : F \in \mathcal{F}, G_h \in \mathcal{F}_h\}$$

and the model corresponding with  $Y$  (see (1)) by

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

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

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

Similarly, we define

$$\vartheta : \mathcal{M} \rightarrow D : \vartheta(P_{F,G}) = S.$$

To begin with we will prove pathwise differentiability of these parameters (see e.g. Bickel et al., 1993, van der Vaart, 1988).

Let  $\mathcal{S}(F)$  the class of lines  $\epsilon F_1 + (1 - \epsilon)F$ ,  $F_1 \in \mathcal{F}$ , with score  $h = d(F_1 - F)/dF \in L_0^2(F)$ , through  $F$ . By convexity of  $\mathcal{F}$  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  $\mathcal{S}(F)$  with score  $g$  will be denoted with  $F_{\epsilon,g}$ . The score of the one dimensional submodels  $P_{F_{\epsilon,g},G_h}^h \subset \mathcal{M}_h$ ,  $g \in S(F)$ , is given by  $A_F^h(g)$  where  $A_F^h$  is called the score operator:

$$A_F^h : L^2(F) \rightarrow L^2(P_{F,G_h}^h) : A_F^h(g)(Y^h) = E_F(g(X) | Y^h),$$

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

$$A_F : L^2(F) \rightarrow L^2(P_{F,G}) : A_F(g)(Y) = E_F(g(X) | Y).$$

Similarly we find the score operator corresponding with one dimensional submodels  $P_{F,G_h,\epsilon,g_1}^h$ , where  $G_{h,\epsilon,g_1} \subset \mathcal{M}_h$  is a line through  $G_h$  with score  $g_1$ . This score operator is given by:

$$B_G^h : L_0^2(G_h) \rightarrow L_0^2(P_{F,G_h}^h) : B_G^h(g_1) = E_{G_h}(g_1(C) | Y^h).$$

The score of a one dimensional submodel  $P_{F_{\epsilon,g},G_{\epsilon,g_1}}^h$  is now given by  $A_F^h(g) + B_G^h(g_1)$ . Hence the tangent space  $T(P)$  is given by  $\overline{R(A_F^h) + R(B_G^h)}$ , where  $R$  stands for range.

We will now show orthogonality in  $L_0^2(P_{F,G_h}^h)$  of the scores  $A_F^h(g)$  and  $B_G^h(g_1)$  for all pairs  $g, g_1$ . Notice that observing  $Y^h$  is equivalent with knowing  $X \in B(Y^h)_1$  and  $C \in B(Y^h)_2$ . By the independence of  $T$  and  $C$  this tells us that  $V(g)(Y^h) \equiv E(g(X) | Y^h) = E(g(X) | X \in B(Y^h)_1)$  and  $W(g_1)(Y^h) \equiv E(g_1(C) | Y^h) = E(g_1(C) | C \in B(Y^h)_2)$ ,

where the region  $B(Y^h)_1$  does not depend on  $C \in B(Y^h)_2$  and similarly  $B(Y^h)_2$  does not depend on  $X$ . Therefore

$$\begin{aligned} E\left(V(g)(Y^h)W(g_1)(Y^h)\right) &= E\left(E(g(X) \mid X \in B(Y^h)_1)\right) \times E\left(E(g_1(C) \mid C \in B(Y^h)_2)\right) \\ &= 0 \times 0 = 0, \end{aligned}$$

which proves the orthogonality in  $L^2_0(P_{F,G_h}^h)$  of the score operators  $A_F^h$  and  $B_G^h$ . Similarly, we have orthogonality of the score operators  $A_F$  and  $B_G$  for the model  $\mathcal{M}$ .

It is easily verified (see Bickel et al., 1993, or Lemma 3.2 in van der Laan, 1993e) that the adjoint of  $A_F^h$  is given by:

$$A_F^{h\top} : L^2(P_{F,G_h}^h) \rightarrow L^2(F) : A_F^{h\top}(v)(X) = E_{F,G_h}(v(Y^h) \mid X)$$

and the corresponding information operator is defined by:

$$I_F^h = A_F^{h\top} A_F^h : L^2(F) \rightarrow L^2(F) : I_F^h(g)(X) = E_{F,G_h}\left(E_{F,G_h}(g(X) \mid Y^h) \mid X\right).$$

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, 1993a, tells us that this implies that  $I_F^h : L^2(F) \rightarrow L^2(F)$  has a bounded inverse, uniformly in  $F \in \mathcal{F}$  (lemma 5.2 in van der Laan (1993d) formulates this result in general for missing data models). And the same result holds for  $I_F : L^2(F) \rightarrow L^2(F)$ . This proves:

**Lemma 3.1** *Let  $I_{F,G} = A_F^\top A_F : L^2(F) \rightarrow L^2(F)$  be the information operator for  $\mathcal{M}$ . We have: If  $H > \delta > 0$   $F$ -a.e., for certain  $\delta > 0$  then  $I_{F,G}$  has bounded inverse  $I_{F,G}^{-1}$  with norm smaller than  $1/\delta$  and is onto. The same holds for the information operator  $I_{F,G_h}^h : L^2(F) \rightarrow L^2(F)$  for  $\mathcal{M}_h$  with inverse  $I_{h,F,G_h}^{-1}$ .*

Let  $b_t : D[0, \tau] \rightarrow \mathbb{R}$  be defined by  $b_t F = F(t)$ . Define  $\kappa_t \equiv I_{(t,\infty)} - S(t)$ . For each one dimensional submodel  $P_{F_{\epsilon,g}, G_{h,\epsilon,g_1}}^h$ , we have

$$\begin{aligned} \frac{1}{\epsilon} \left( b_t \vartheta_h(P_{F_{\epsilon,g}, G_{h,\epsilon,g_1}}^h) - b_t \vartheta_h(P_{F,G_h}^h) \right) &= \int_{(t,\infty)} g dF \\ &= \langle I_{(t,\infty)} - S(t), g \rangle_F \\ &= \langle \kappa_t, g \rangle_F \\ &= \langle I_F^h I_{h,F}^{-1}(\kappa_t), g \rangle_F \\ &= \langle A_F^h I_{h,F}^{-1}(\kappa_t), A_F^h(g) \rangle_{P_{F,G_h}^h} \\ &= \langle A_F^h I_{h,F}^{-1}(\kappa_t), A_F^h(g) + B_G^h(g_1) \rangle_{P_{F,G_h}^h}, \end{aligned}$$

where we used the orthogonality of the scores at the last step. The same holds for  $\vartheta$  and  $P_{F,G}$  without  $h$ . This proves by definition (see e.g. Bickel et al., 1993) that for each  $t \in [0, \tau]$   $b_t \vartheta_h$  is pathwise differentiable at  $P_{F,G_h}^h$  for each one dimensional submodel  $P_{F_{\epsilon,g}, G_{h,\epsilon,g_1}}^h$  at  $P_{F,G_h}^h$  with efficient influence function (suppressing the  $G$  in the notation) given by :

$$\tilde{I}^h(F, t)(\cdot) = A_F^h I_{h,F}^{-1}(\kappa_t)(\cdot). \quad (7)$$

And similarly for  $\vartheta$  at  $P_{F,G}$  with

$$\tilde{I}(F, t)(\cdot) = A_F I_F^{-1}(\kappa_t)(\cdot). \quad (8)$$

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;  $P_F^h \equiv P_{F,G_h}^h$  and  $P_F \equiv P_{F,G_0}$ ,  $I_F = I_{F,G}$  etc.

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  $R_{n,t} = o_P(1/\sqrt{n})$ .  $\sqrt{n}(P_n - P_{F_0})\tilde{I}(F_0, t)$  is a sum of  $n$  i.i.d. mean zero random variables which converges by the C.L.T. 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  $(\sqrt{n}(P_n - P_{F_0})\tilde{I}(F_0, t) : t \in [0, \tau])$ , which can be considered as a random element of  $\ell^\infty(\mathcal{G}) \equiv \{H : \mathcal{G} \rightarrow \mathbb{R} : \sup_{g \in \mathcal{G}} |H(g)| < \infty\}$ , where  $\mathcal{G} = \{\tilde{I}(F_0, t) : t \in [0, \tau]\}$ , and where  $\ell^\infty(\mathcal{G})$  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 Hoffmann-Jørgensen sense, making measurability-questions (for finite  $n$ ) irrelevant (see e.g. Hoffmann-Jørgensen, 1984, van der Vaart and Wellner, 1993, Pollard, 1990). A class for which this weak convergence holds is called a Donsker class. If  $\mathcal{G}$  is Donsker and  $\sup_{t \in [0, \tau]} |R_{n,t}| = o_P(1/\sqrt{n})$ , then we say that  $F_n$  is supnorm efficient.

Our goal is to prove efficiency of  $S_n^h$  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 (1993d) provides us under the assumptions as stated in section 2.1, by simple verification, with efficiency of  $S_n^h$ , among estimators based on the data  $Y_i^h$ ,  $i = 1, \dots, n$ , as an estimator of  $\vartheta_h(P_{F_0}^h) = S_0$ . 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 what happens if  $h_n \rightarrow 0$  when the number of observation converges to infinity.

It works as follows: The model  $\mathcal{M}_h$  is convex and the  $F \rightarrow P_F^h$  is linear. Theorem 1.1 in van der Laan (1993a) 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  $dF_0/dF_1 \in L_0^2(F_1)$ . Hence by verification of a straightforward extension condition as verified in general by lemma 5.12 in van der Laan (1993d), it follows that this identity holds also for  $F_1 = F_n$ :

$$S_n^h(t) - S_0(t) = - \int \tilde{I}^h(S_n^h, t) dP_{F_0}^h. \quad (9)$$

It remains to verify:

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

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

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

The efficient score equation and the identity (9) provide us with the crucial identity

$$S_n^h(t) - S_0(t) = \int \tilde{I}^h(F_n^h, t) d(P_n^h - P_0^h). \quad (10)$$

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

$$\sup_{t \in [0, \tau]} \left| \int (\tilde{I}^h(F_n^h, t) - \tilde{I}^h(F_0, t)) d(P_n^h - P_0^h) \right| = o_{P_0^h} (1/\sqrt{n}).$$

This condition requires a lot of hard work (done in section 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  $L_0^2(P_0)$ . Therefore  $\tilde{I}^h(F_n^h, t)$  cannot be shown to be an element of a fixed Donsker-class when  $h_n \rightarrow 0$ . In other words the  $P$ -Donsker class and  $\rho_P$ -consistency condition of as used in the proof for the general class of missing data models (van der Laan, 1993d) 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  $h_n \rightarrow 0$  and then to use (10) in order to obtain a rate for  $\|S_n^h - S_0\|_\infty$  so that terms can be shown to converge to zero if  $h_n \rightarrow 0$  slowly enough.

The empirical process condition provides us with:

$$S_n^h(t) - S_0(t) = \int \tilde{I}^h(F_0, t) d(P_n^h - P_0^h) + o_{P_0^h} (1/\sqrt{n}),$$

where the remainder holds uniformly in  $t$ .

**Approximation condition.** Finally, we need to show

$$\int \tilde{I}^h(F_0, t) d\sqrt{n}(P_n^h - P_0^h) \xrightarrow{D} N(0, \sigma^2(\tilde{I}(F_0, t))).$$

This is shown by application of a lemma in Bickel and Freedman (1981).

We are able to show this condition pointwise and 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 sieved-SOR-NPMLE.

Recall the assumptions made in section 2.1: in particular  $F_0(\tau) = 1$  and hence  $P_0^h(\cdot, d)$  lives on  $[0, \tau]$ . In all statements the width (of grid)  $h$  converges to zero for  $n \rightarrow \infty$ ; 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 $h_n \rightarrow 0$ .

The starting point of the analysis is (10). The indicators are a uniform Donsker class. This tells us that  $\sup_h \|P_n^h - P_0^h\|_\infty = O_P(1/\sqrt{n})$ .

A real valued function on  $[0, \tau] \subset \mathbb{R}^2$  is called to be of bounded *uniform sectional variation* if the variations of all sections ( $s \rightarrow f(s, t)$  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  $\|\cdot\|_\vee^*$ . In van der Laan (1993e, 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  $\|1/f\|_v^* \leq M\|f\|_v^*$  for some  $M < \infty$  which does not depend on  $f$  (Gill, 1993). We have:

**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  $\pi^h$  and  $E_{k,l}^h(0,1)$  be the horizontal strips. Suppose that  $F_0(E_{k,i}^h) > \delta h_n$  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*

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

**Proof.** See section 7.

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, 1993e) 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 van der Laan, 1993e) we can apply this to (10) with  $F_1 = \tilde{I}^h(F_n^h, t)(\cdot, d)$  and  $H_1 = (P_n^h - P_0^h)(\cdot, d)$  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  $h \rightarrow 0$  slower than  $n^{-1/3}$ , then  $F_n^h$  is uniformly consistent (also for  $h$  is fixed).

## 4.2 Empirical process condition.

Define  $Z_n^h \equiv \sqrt{n}(P_n^h - P_0^h)$  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 that:

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

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

**Proof.** See appendix.

**Analysis of the uncensored term.** Let's 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 absolute 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  $P_{11}^d$ . Firstly 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  $\mu_k$ , 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

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

where the bound does not depend on  $t$ . Consequently, if  $nh_n^9 \rightarrow \infty$ , 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 nested in  $\pi^h$ : so  $\pi^h \subset \pi^{a_n}$ . Now

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

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:

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

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  $d$ ):

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

In order to show that  $\int f_{nt}^h(y, d) dZ_n(y, d) = o_P(1)$  for a rate  $h_n \rightarrow 0$ , 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:

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

In other words this is the *modulus of continuity* of a bivariate empirical process. Firstly, we will bound the two terms in  $W_{i,j}^n(a_n)$ .

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

$$\begin{aligned} P(\|Z_n^{a_n} - Z_n\|_\infty > \epsilon) &\leq P\left(\max_{i,j} W_{i,j}^n(a_n) > \epsilon\right) \\ &\leq \sum_{i,j} P\left(W_{i,j}^n(a_n) > \epsilon\right). \end{aligned} \quad (11)$$

Furthermore we have

$$\|Z_n^{a_n}\|_v^* \leq \sum_{i,j} W_{i,j}^n(a_n). \quad (12)$$

**Analysis of the modulus of continuity.** Let  $W_n(a_n) \equiv \sup_{\max_{l=1,2}|s_l-t_l|\leq a_n} |Z_n(s) - Z_n(t)|$ . It is clear that we have  $W_{i,j}^n(a_n) \leq W_n(a_n)$ . Einmahl's (1987) inequality 6.4, for  $W_n(a_n)$  holds for a continuous density which is bounded away from zero and infinity on  $[0, \tau]$  and is given by:

$$P(W_n(a_n) > \lambda) \leq \frac{C}{a_n} \exp\left(\frac{-c_1 \lambda^2}{a_n} \Psi\left(\frac{\lambda}{\sqrt{na_n}}\right)\right) \text{ for any } \lambda > 0, \quad (13)$$

where  $\Psi(x) \geq 1/(1 + 1/3x)$ .  $p_{11}^c$  is bounded away from zero and infinity on  $[0, \tau]$  (it has only jumps on  $\pi^h$ ) and is continuous on the vertical and horizontal strips containing  $A_{i,j}(a_n)$  (here we use the nesting of  $\pi^{h_n}$  in  $\pi^{a_n}$ ) and hence for the modulus of continuity on the sets  $A_{i,j}(a_n)$  the discontinuities on  $\pi_h$  play no role. Consequently, (13) holds also for  $W_{i,j}^n(a_n)$ .

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

$$P\left(\frac{W_{i,j}^n(a_n)}{a_n^{0.5-\epsilon}} > \delta_n\right) \leq \frac{C}{a_n} \exp(-C_1/a_n^{\epsilon'}). \quad (14)$$

So  $W_{i,j}^n(a_n)/a_n^{0.5-\epsilon}$  converges to zero in probability exponentially fast.

Assume  $na_n \rightarrow \infty$  at a polynomial rate. Applying (14) to (11) provides us with:

$$\begin{aligned} P\left(\frac{\|Z_n^{a_n} - Z_n\|_\infty}{a_n^{0.5-\epsilon}} > \epsilon\right) &\leq \sum_{i,j} \frac{C}{a_n} \exp(-C_1/a_n^{\epsilon'}) \\ &\leq \frac{C'}{a_n^3} \exp(-C_1/a_n^{\epsilon'}) = o(1). \end{aligned}$$

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

Furthermore, applying (14) to (12) provides us with:

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

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

$$\int f_{nt}^h(y, 1, 1) dZ_n(y) = o_P\left(r_1(h_n)a_n^{0.5-\epsilon}\right) + o_P\left(\frac{r_1(h_n)r_2(h_n)}{\sqrt{na_n^{1.5+\epsilon}}}\right). \quad (15)$$

For the first term we need that  $a_n$  converges quicker to zero than  $h_n^3$ . Substituting this in the second term tells us that we need that  $h_n$  converges 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  $P_0^h I(D \neq (1, 1))$  is purely discrete on the grid  $\pi^h$ , which contains  $O(1/h_n^2)$  points. Let  $p_0^h$  and  $p_n^h$  be the densities of  $P_0^h$  and  $P_n^h$  w.r.t.  $\nu_h$ , 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  $\|p_n^h - p_0^h\|_\infty = O_P(1/\sqrt{n})$ . In the following result for  $h_n \rightarrow 0$  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**

$$\begin{aligned} \|p_{01}^{hn} - p_{01}^h\|_{L_1(\nu_h)} &= O_p\left(\frac{1}{\sqrt{h^4 n}}\right) \\ \|p_{10}^{hn} - p_{10}^h\|_{L_1(\nu_h)} &= O_p\left(\frac{1}{\sqrt{h^4 n}}\right) \\ \|p_{00}^{hn} - p_{00}^h\|_{L_1(\nu_h)} &= O_p\left(\frac{1}{\sqrt{h^4 n}}\right). \end{aligned}$$

**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  $\pi^h$  we have that  $E(p_{01}^{nh}(u_k, v_l)) = p_{01}^h(u_k, v_l)$  and  $\text{Var}(p_{01}^{nh}(u_k, v_l)) = \frac{1}{n} p_{01}^h(u_k, v_l)(1 - p_{01}^h(u_k, v_l))$ .  $\pi^h$  has  $O(h_n^2)$  grid points  $(u_k, v_l)$  by definition of  $\pi^h$ . Now, we have

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

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:

$$\begin{aligned} \int f_{nt}^h dZ_n^h &= \sqrt{n} \int f_{nt}^h (p_n^h - p_0^h) d\nu_h \\ &\leq \sqrt{n} \|f_{nt}^h\|_\infty \|p_n^h - p_0^h\|_{L_1(\nu_h)} \\ &= \sqrt{n} O_P\left(\frac{1}{\sqrt{h_n^9 n}}\right) O_P\left(\frac{1}{\sqrt{nh_n^4}}\right) \\ &= O_P\left(\frac{1}{\sqrt{h_n^{13} n}}\right). \end{aligned}$$

This proves the following lemma:

**Lemma 4.6** *If  $h_n$  converges to zero slower than  $n^{-1/13}$ , then  $\int f_{nt}^h I(D = d) dZ_n^h = o_P(1)$  for  $d \in \{(1, 0), (0, 1), (0, 0)\}$ .*

Lemma 4.4 and lemma 4.6 prove the empirical process condition for a rate of  $h_n$  slower than  $n^{-1/18}$ .

### 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  $h = h_n \rightarrow 0$   $X_i^h(t) \xrightarrow{D} X_i(t)$  and  $\text{Var}(X_i^h(t)) \rightarrow \text{Var}(X_i(t))$ , then this sum converges weakly to a normal distribution with mean zero and variance equal to  $\text{Var}(X_i(t))$ . We will prove these two conditions:

**Lemma 4.7** *For  $F_0$  almost each  $T$  we assume that if  $F_{01}(T_1, \Delta v) > 0$  (i.e. if it has an atom), then  $H_0((T_1, \infty), \Delta v) = 0$ . Similarly for  $F_0$  almost each  $T$  we assume that if  $F_{02}(\Delta u, T_2) > 0$ , then  $H_0(\Delta u, (T_2, \infty)) = 0$ .*

*Define the following real valued random variables  $X^h(t) \equiv \tilde{I}^h(F_0, t)(Y^h)$ ,  $Y^h \sim P_0^h$  and  $X(t) \equiv \tilde{I}(F_0, t)(Y)$ ,  $Y \sim P_0$ . We have for each  $t \in [0, \tau]$  that for  $h_n \rightarrow 0$*

$$E((X^{h_n}(t) - X(t))^2) \rightarrow 0$$

and

$$E(X^{h_n}(t)X^{h_n}(s)) \rightarrow E(X(t)X(s)) \text{ uniformly in } s, t \in [0, \tau].$$

**Proof.** See section 7.

We already assumed that  $F_0 = F_0^d + F_0^c$ . Therefore this assumption means for us that if  $F_0$  has an atom at  $t = (t_1, t_2)$ , then  $H$  should not have atoms on the vertical and horizontal lines starting at  $t$ . The assumption should be interpreted as follows: the EM-algorithm tells us that one needs to be able to estimate the conditional densities  $P(T_2 > v_2 \mid T = t_1)$ . Suppose that this density has an atom at  $v_2$ , then if one draws observations from this conditional density one needs uncensored observations at  $v_2$  and therefore you do not like to have a positive probability of being censored at  $v_2$ . So you want that  $H_0((t_1, \infty), \Delta v_2) = 0$ . Because of our convention that if  $T = C$ , then the observation is uncensored, it seems to be an unnecessary assumption.

Lemma 4.7 has the following corollary

**Corollary 4.1** *Under the assumptions of lemma 4.7  $\int \tilde{I}^{h_n}(F_0, t)(y) dZ_n^{h_n}(y)$  converges in distribution to a normal distribution with mean zero and variance equal to  $\text{Var}_{P_0}(\tilde{I}^0(F_0, t))$ .*

#### 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  $P_0$ -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, p. 153).

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

Let  $V_n(t) = 1/\sqrt{n} \sum_{i=1}^n X_i(t)$ . Firstly, 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

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

Firstly, 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:

$$\begin{aligned} \text{Var} \left( \langle e_j, V_n^h - V_n \rangle \right) &= \text{Var} \left( \langle e_j, \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i^h - X_i) \rangle \right) \\ &= \text{Var} \left( \langle e_j, X^h - X \rangle \right) \\ &\leq E \left( \langle e_j, X^h - X \rangle^2 \right) \\ &\leq \langle e_j, e_j \rangle E \langle X^h - X, X^h - X \rangle. \end{aligned}$$

Assume now that  $\mathcal{H} = L^2(\lambda)$  for a certain finite measure  $\lambda$ . By lemma 4.7 we have  $\text{Var}(X^{hn}(t))$  converges to  $\text{Var}(X(t))$  and  $E((X^{hn}(t) - X(t))^2) \rightarrow 0$ , both uniformly in  $t$ . Therefore,

$$\begin{aligned} E \langle X^h - X, X^h - X \rangle &= E \int (X^h - X)^2(s) d\lambda(s) \\ &\leq \sup_{s \in [0, \tau]} |E((X^h - X)(s)^2)| \int d\lambda(s) \rightarrow 0, \end{aligned}$$

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

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

which is just a sum of i.i.d. mean zero random variables. By the C.L.T., for showing that this converges in distribution to  $\langle e_j, V_0 \rangle$  it suffices to have that  $\text{Var}(\int e_j(s) X_i(s) d\lambda(s)) < \infty$ . 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:

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

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

At the first, second, third equality we used Fubini's theorem, then we use the uniform convergence of  $\mathbb{E}(V_n^h(s)V_n^h(t))$  to  $\mathbb{E}(V_0(s)V_0(t))$ , 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  $N \rightarrow \infty$ , then both (tail) series converge to zero.

Application of lemma 4.8 provides us now with:

**Lemma 4.9** *Suppose the same assumption as in lemma 4.7. If  $\lambda$  is a finite measure and  $h_n \rightarrow 0$ , then  $V_n^{h_n} \xrightarrow{D} V_0$  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  $\pi^h$ -interval-censoring of the singly censored observations. We defined  $E_{k,l}^h(1, 0)$  and  $E_{k,l}^h(0, 1)$  as the vertical and horizontal strips of  $\pi^h$  starting at  $(u_k, v_l)$ . We defined  $Z_n^h \equiv \sqrt{n}(P_n^h - P_{F_0, G_h}^h)$  as the empirical process corresponding with 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 efficiency (among all estimators based on the reduced data) under the assumptions as stated in subsection 2.1 provides us with the following theorem:

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

**Fixed grid efficiency.** *Suppose that we do not change the grid  $\pi^h$  for  $n \rightarrow \infty$  and that for each grid point  $F_0(E_{k,l}^h(1, 0)) > 0$  and  $F_0(E_{k,l}^h(0, 1)) > 0$ .*

*Then  $S_n^h$  is a supnorm-efficient estimator of  $S_0$  for the data  $Y_i^h$ ,  $i = 1, 2, \dots, n$ :*

$$\sqrt{n}(F_n^h - F_0)(t) = \int \tilde{I}^h(F_0, t) dZ_n^h + R_n^h(t),$$

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

$$E(N_h(s)N_h(t)) = E_{P_0^h}(\tilde{I}^h(F_0, s)\tilde{I}^h(F_0, t)).$$

**Uniform consistency.** *Suppose that  $F_0(E_{k,l}^{h_n}(1, 0)) > \delta h_n$  and  $F_0(E_{k,l}^{h_n}(0, 1)) > \delta h_n$  for some  $\delta > 0$ .*

Then for any rate  $h_n \rightarrow 0$

$$\|S_n^{h_n} - S_0\|_\infty = O_P\left(1/\sqrt{nh_n^3}\right).$$

**Efficiency.** Suppose  $F_0 = F_0^d + F_0^c$ , where  $F_0^d$  is purely discrete with finite support and  $F_0^c$  is absolutely continuous w.r.t. Lebesgue measure with continuous density uniformly bounded away from zero on  $[0, \tau]$ . Moreover, assume that if  $F_0$  has an atom at  $T = (T_1, T_2)$ , then  $H$  puts no mass on the vertical and horizontal lines starting at  $T$ , going upwards and to the right.

We have that for  $h_n \rightarrow 0$

$$E_{P_0^h} \left( \tilde{I}^h(F_0, s)(Y^h) \tilde{I}^h(F_0, t)(Y^h) \right) \rightarrow E_{P_0} \left( \tilde{I}(F_0, s)(Y) \tilde{I}(F_0, t)(Y) \right)$$

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

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

$$\text{Var}(N_0(t)) = \text{Var}\left(\tilde{I}(F_0, t)\right).$$

Moreover, for any finite measure  $\lambda$   $V_n^h$  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  $nh_n^3 \rightarrow \infty$ , then  $F_n^{h_n}$  converges uniformly to  $F_0$ . Therefore, we hope that this will also be a good rate for obtaining an efficient estimator, though we did not prove this.

## 6 Technical lemmas.

In formulas the score operator  $A_0^h$  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  $\pi^h$ ):

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

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 stepfunctions on  $\pi^h$ .

In formulas  $I_0^h$  is given by:

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

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

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

## 6.1 Proof of lemma 4.1.

**Lemma 6.1** *Let  $E_{k,l}^h(1, 0) \equiv (u_k, u_{k+1}] \times [v_l, \infty)$  be the vertical strips of  $\pi^h$  and  $E_{k,l}^h(0, 1)$  be the horizontal strips. Suppose that  $H_0(\tau) > 0$  and  $F_0(E_{k,l}^{h_n}) > \delta h_n$  for certain  $\delta > 0$ .*

*Then there exists an  $\epsilon > 0$  so that for any sequence  $h_n$  which converges to zero slower than  $1/\sqrt{n}$  we have*

$$\min_{k,l} F_n^{h_n}(E_{k,l}^{h_n}(1, 0)) \geq \epsilon h_n, \text{ with probability tending to 1.}$$

*Similarly, for  $E_{k,l}^{h_n}(0, 1)$ .*

**Proof.** We use the notation  $E_{k,l}^h$  for both strips. Firstly, by the EM-equations (see (6)) we have

$$F_n^h(E_{k,l}^h) \geq P_{11}^n(E_{k,l}^h), \tag{16}$$

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

$$P_{11}(E_{k,l}^{h_n}) \geq H_0(\tau)F_0(E_{k,l}^{h_n}) > \delta_1 h_n \text{ for some } \delta_1 > 0. \tag{17}$$

Furthermore,  $\{I_{E_{k,l}^h} : h \in (0, 1], k, l\}$ , the collection of indicators of  $E_{k,l}^h$  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  $\epsilon > 0$  and rate  $r(n)$  slower than  $\sqrt{n}$  that

$$P \left( \sup_{k,l} | (P_{11}^n - P_{11})(E_{k,l}^{h_n}) | > \frac{\epsilon}{r(n)} \right) \rightarrow 0. \tag{18}$$

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

$$\limsup_{n \rightarrow \infty} P \left( \min_{k,l} P_{11}^n(E_{k,l}^{h_n}) \leq \epsilon h_n \right) > \delta > 0 \text{ for some } \delta > 0. \tag{19}$$

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

$$\lim_{n \rightarrow \infty} P \left( \min_{k,l} P_{11}^n(E_{k,l}^{h_n}) \geq \epsilon h_n \right) = 1,$$

which combined with (16) proves the lemma. So it remains to prove the contradiction. We have by (17) and (19), respectively,

$$\begin{aligned} \limsup_{n \rightarrow \infty} P \left( \sup_{k,l} |(P_{11}^n - P_{11})(E_{k,l}^{h_n})| > \delta_1 h_n - \epsilon h_n \right) &\geq P \left( \min_{k,l} P_{11}^n(E_{k,l}^{h_n}) \leq \epsilon h_n \right) \\ &> \delta > 0. \end{aligned}$$

However, we also have (18). These two contradict if  $h_n$  converges slower to zero than  $1/\sqrt{n}$ .  $\square$

For obtaining 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 with the censored observations. Then this equation is equivalent with the following equation:

$$g(x) = \frac{1}{H_h(x)} \{f(x) - K_F^h(g)(x)\}. \quad (20)$$

For the moment denote the right-hand side with  $C_F^h(g, f)(x)$ : i.e. 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  $\|I_F^h(g') - f\|_F = 0$ : i.e.  $\|g' - C_F^h(g', f)\|_F = 0$ . 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 (20) which holds for each  $x$  instead of only in  $L^2(F)$  sense.

Moreover, there is only one such a pointwise solution for each  $f$  and a different (in supremum norm sense)  $f$  gives a different solution. So we have 1)  $I_F^h : (D[0, \tau], \|\cdot\|_\infty) \rightarrow (D[0, \tau], \|\cdot\|_\infty)$  is 1-1 and onto and we know that 2)  $g_h = I_{h,F}^{-1}(f)$  is given by  $g_h(x) = C_F^h(g'_h, f)(x)$ , where  $g'_h = I_{h,F}^{-1}(f)$  in  $L^2(F)$  sense. Moreover we can use that 3)  $\|g'_h\|_F \leq C\|f\|_F$ , where  $C \leq 1/\delta$  does not depend on the width  $h$ .

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

Therefore it remains to bound the *supnorm* 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:

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

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

**Supnorm.** Recall that  $\|f\|_\infty \leq 1$ . By the Cauchy-Schwarz inequality and  $\|g_h\|_F \leq C\|f\|_F$  we have:

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

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

**Lemma 6.2** *There exists a  $C < \infty$  so that:*

$$\sup_{\|f\|_\infty=1} \|I_{h, F_n^h}^{-1}(f)\|_\infty \leq \frac{C}{\sqrt{h_n}} \text{ with probability tending to 1.}$$

**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

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

So

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

Now, doing nothing more sophisticated than

$$\frac{\int_{E_{k,l}} g_h dF}{F(E_{k,l})} \leq \|g_h\|_\infty \leq M/\sqrt{h_n} \text{ and } F(E_{k,l}) > \epsilon h_n \quad (21)$$

we obtain the following (bad) bound:

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

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

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

So we proved the following:

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

$$\sup_{\|f\|_v^* = 1} \|I_{h, F_n^h}^{-1}(f)\|_v^* \leq \frac{C}{h_n^{3/2}} \text{ with probability tending to 1.} \quad (22)$$

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 (21) by  $C/h_n^{3/2}$ . Therefore the uniform sectional variation of the efficient influence curve is also bounded by the rate given in (22). 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:

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

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:

$$\begin{aligned} A_n^h I_{h, n}^{-1}(\kappa_t) - A_0^h I_{h, 0}^{-1}(\kappa_t) &= (A_n^h - A_0^h) I_{h, 0}^{-1}(\kappa_t) + A_n^h I_{h, n}^{-1}(I_n^h - I_0^h) I_{h, 0}^{-1}(\kappa_t) \\ &= (A_n^h - A_0^h)(g_{0t}^h) + A_n^h I_{h, n}^{-1}(I_n^h - I_0^h)(g_{0t}^h). \end{aligned} \quad (23)$$

Firstly, 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:

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

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

$$O_P\left(\frac{1}{\sqrt{nh_n^3}}\right) O_P\left(\frac{1}{\sqrt{h_n^3}}\right) O\left(\frac{1}{h_n}\right) = O_P\left(\frac{1}{\sqrt{nh_n^6}}\right).$$

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 (23). Because  $A_0^\top$  does only depend on  $G$ , we have for the term  $(I_n^h - I_0^h)(g_{0t}^h)$ :

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

Because  $A_0^{h\top}$  is just a conditional expectation we have that  $\|A_0^{h\top}(g)\|_\infty \leq \|g\|_\infty$ . Therefore, we also have that  $\|(I_n^h - I_0^h)(g_{0t}^h)\|_\infty = O(1/\sqrt{nh_n^4})$ . 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

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

We proved:

$$\|\tilde{I}^h(F_n^h, t) - \tilde{I}^h(F_0, t)\|_\infty = O\left(\frac{1}{\sqrt{nh_n^9}}\right).$$

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** *For  $F_0$  almost every  $t$  we assume that if  $F_{01}(t_1, \Delta v) > 0$ , then  $H_0(t_1, \Delta v) = 0$ . Similarly for  $F_0$  almost every  $t$  we assume that if  $F_{02}(\Delta u, t_2) > 0$ , then  $H_0(\Delta u, t_2) = 0$ .*

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

$$\sup_{g \in C} \|(I_0^h - I_0)(g)\|_{F_0} \rightarrow 0, \tag{24}$$

and

$$\sup_{g \in C} E\left(A_0^h(g) - A_0(g)\right)^2 \rightarrow 0 \text{ for } h = h_n \rightarrow 0.$$

**Proof.** By the compactness of  $C$  and the continuity of  $I_0^h : L^2(F_0) \rightarrow L^2(F_0)$  the supremum in (24) is attained by some  $g_0 \in C$ . Let  $g_k$  be a sequence so that  $\|g_k - g_0\|_{F_0} \rightarrow 0$  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  $k \rightarrow \infty$ . Therefore it suffices now to show that  $\|(I_0^{h_n} - I_0)(g_k)\|_{F_0} \rightarrow 0$  for each fixed  $k$ . Now, we have:

$$(I_0^h - I_0)(g_k) = A_0^{h\top}(A_0^h - A_0)(g_k) + (A_0^{h\top} - A_0^\top)(A_0(g_k)).$$

The difference 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. Firstly, 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:

$$(A_0^{h\top} - A_0^\top)(f_k)(T_1, T_2) = \int_0^{T_2} f_k(T_1, v)(H_h - H_0)(T_1, dv).$$

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  $v \rightarrow f_k(T_1, v)$  is bounded and continuous  $H_0(T_1, \cdot)$  a.e., then  $(A_0^{h\top} - A_0^\top)f_k(T_1, T_2) \rightarrow 0$  for this  $T$ . The boundedness follows from:  $\|f_k\|_\infty \leq \|g_k\|_\infty < \infty$ . We have that  $v \rightarrow f_k(T_1, v)$  is given by:

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

This function is continuous at  $v$  if  $v \rightarrow F_{01}(T_1, v)$  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. That is what we assumed. This proves the pointwise convergence of  $f_h \equiv (A_0^{h\top} - A_0^\top)(f_k)$  to zero  $F$ -a.e. We need to show that  $\int f_h^2 dF_0 \rightarrow 0$ . However, we also have  $\|f_h\|_\infty \leq 2\|g_k\|_\infty$  and therefore the dominated convergence theorem provides us with  $\int f_h^2 dF_0 \rightarrow 0$ .

Let's now consider the first term  $A_0^{h\top}(A_0^h - A_0)(g_k)$ . 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  $E_{X,C}((A_0^h - A_0)(g_k))^2 \rightarrow 0$  for  $h \rightarrow 0$ , 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  $(\tilde{T}, D)^h$ : 1) we have a nested sequence of partitions  $\pi^h$  and we simulated 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  $\pi^h$ . This provides us with data  $(\tilde{T}, D)_h \sim P_{F_0, G_h}$ . 3) Finally we discretized  $(\tilde{T}, D)_h$  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  $\mathcal{A}^h$ . Because  $\pi^h$  is nested and the sigma field generated by  $\pi^h$  converges to the Borel sigma-field on  $[0, \tau]$  we have that  $\mathcal{A}^h \uparrow \mathcal{A}^\infty$  for  $h \rightarrow 0$ , where  $\mathcal{A}^\infty$  is the sigma field generated by  $Y = (\tilde{T}, D)$ ,  $Y \sim P_{F_0, G_0}$ .

Consequently  $M_{h_n} \equiv E_{X,C}(g_k(T) | \mathcal{A}^{h_n})$  is a martingale in  $n$  and it is well known that if  $\sup_h E(M_h^2) < \infty$ , then  $E((M_h - M_0)^2) \rightarrow 0$ . We have  $\sup_h E(E(g_k(T) | \mathcal{A}^h)^2) \leq \|g_k\|_\infty < \infty$  and consequently we have  $\|(A_0^h - A_0)(g_k)\|_{F_0 \times G_0} \rightarrow 0$ . This also proves the second statement in lemma 6.4.  $\square$

**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  $h \rightarrow 0$ :*

$$\sup_{g \in C} \|(I_0^{-1} - I_{h,0}^{-1})(g)\|_{F_0} \rightarrow 0. \quad (25)$$

*This implies*

$$\sup_{g, g_1 \in C} \left| \langle A_0^h(I_{h,0}^{-1}(g)), A_0^h(I_{h,0}^{-1}(g_1)) \rangle_{P_0^h} - \langle A_0(I_0^{-1}(g)), A_0(I_0^{-1}(g_1)) \rangle_{P_0} \right| \rightarrow 0.$$

*Moreover, we have*

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

**Proof.** We have:

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

Firstly, 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  $\sup_{g \in C} \|(I_0^h - I_0) I_0^{-1}(g)\|_{F_0} \rightarrow 0$ . 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:

$$\begin{aligned} \langle A_0^h I_{h,0}^{-1}(g), A_0^h I_{h,0}^{-1}(g_1) \rangle_{P_0^h} &= \langle I_{h,0}^{-1}(g), g_1 \rangle_{F_0} \\ &= \langle I_{h,0}^{-1}(g) - I_0^{-1}(g), g_1 \rangle_{F_0} + \langle I_0^{-1}(g), g_1 \rangle_{F_0}. \end{aligned}$$

The first term converges to zero by the Cauchy-Schwarz inequality and (25). The second term equals  $\langle A_0 I_0^{-1}(g), A_0 I_0^{-1}(g_1) \rangle_{P_0}$ .

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

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

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

For the second term we have:

$$\|A_0^h (I_{h,0}^{-1} - I_0^{-1})(g)\|_{P_0^h} \leq \|(I_{h,0}^{-1} - I_0^{-1})(g)\|_{F_0} \rightarrow 0 \text{ by (25).} \square$$

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.

## 7 Bibliographic remarks.

Many proposals for estimation of the bivariate survival function in the presence of bivariate censored data have been made. The usual NPML and self-consistency principle do not lead to a consistent estimator: an NPMLE does not have to be consistent (Tsai, Leurgans and Crowley, 1986). Therefore most proposals are explicit estimators based on representations of the bivariate survival function in terms of distribution functions of the data: among them Muñoz (1980), Campbell (1981), Campbell and Földes (1982), Langberg and Shaked (1982), Hanley and Parnes (1983), Tsai, Leurgans and Crowley (1986), Dabrowska (1988, 1989), Burke (1988), the so called Volterra estimator of P.J. Bickel (see Dabrowska, 1988), Prentice and Cai (1992a, 1992b). Three of them (the Dabrowska, Volterra and Prentice-Cai estimators) are studied in Gill, van der Laan, Wellner (1993).

Prentice and Cai (1992a) proposed a nice estimator which is closely related to Dabrowska's estimator except that this one also uses the Volterra structure of Bickel's suggestion. 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, 1992,

and chapter 8 of van der Laan, 1993e). It is expected that Dabrowska's and Prentice-Cai's 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, LIL, all hold by application of the functional delta method: see Gill (1992) and 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 estimator 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 non negligible 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. This is the first implicitly defined estimator. He derives and illustrates intuitively nice properties of his estimator. He points out why the original self-consistency equation has a wide class of solutions and his estimator tackles this non-uniqueness problem in a very direct way. Uniform consistency,  $\sqrt{n}$ -weak convergence, and the bootstrap for his normalized estimator is proved in van der Laan (1991, 1993c) (chapter 7 of van der Laan, 1993e). However this estimator is not asymptotically efficient (except at some special points).

In van der Laan (1992, 1993b, and this paper) an efficient estimator is proposed depending on width of strips. Simulations in chapter 8 of van der Laan (1993e) show indeed that this asymptotically efficient estimator has excellent practical behavior if one does not choose too wide strips.

## References

- D.M. Bakker (1990), *Two nonparametric estimators of the survival function of bivariate right censored observations*, Report BS-R9035, Centre for mathematics and computer science, Amsterdam.
- P.J. Bickel and D.A. Freedman (1981), Some asymptotic theory for the bootstrap, *Ann.Stat.* **9** 1196–1217.
- P.J. Bickel, A.J. Klaassen, Y. Ritov and J.A. Wellner (1993), *Efficient and adaptive estimation for semi-parametric models*, Johns Hopkins University Press, Baltimore.
- M.D. Burke (1988), Estimation of a bivariate survival function under random censorship, *Biometrika* **75**, 379–382.
- D.M. Dabrowska (1988), Kaplan Meier Estimate on the Plane, *Ann. Statist.* **16**, 1475–1489.
- D.M. Dabrowska (1989), Kaplan Meier Estimate on the Plane: Weak Convergence, LIL, and the Bootstrap, *J. Multivar. Anal.* **29**, 308–325.
- A.P. Dempster, N.M. Laird and D.B. Rubin (1977), Maximum likelihood from incomplete data via the EM-algorithm, *J. Roy. Statist. soc. sec.* **39** 1–38.
- B. Efron (1967), The two sample problem with censored data, *Proc. 5th. Berkeley Symp.*

- on *Math. Statist. Prob.*, 831–853, Berkeley, University of California Press.
- J.H.J. Einmahl (1987), *Multivariate empirical processes*, CWI tract **32**, Centre for Mathematics and Computer Science, Amsterdam.
- R.D. Gill (1989), Non- and Semi-parametric Maximum Likelihood Estimators and the von Mises Method (Part 1), *Scand. J. Statist.* **16**, 97–128.
- R.D. Gill (1992), Multivariate survival analysis, *Theory Prob. Appl.* (English Translation) **37**, 18–31 and 284–301.
- R.D. Gill (1993), *Lectures on survival analysis*, In: D. Bakry, R.D. Gill and S. Molchanov, École d’Été de Probabilités de Saint Flour XXII–1992, ed. P. Bernard; Springer Lecture Notes in Mathematics (to appear).
- R.D. Gill, M.J. van der Laan and J.A. Wellner (1993), *Inefficient estimators of the bivariate survival function for three models*, to appear in *Annales de L’I.H.P. Probabilités et Statistiques*.
- D.F. Heitjan and D.B. Rubin (1991), Ignorability and coarse data, *Ann. Statist.* **19**, 2244–2253.
- J. Hoffmann-Jørgensen (1984), *Stochastic processes on Polish Spaces*, Unpublished manuscript.
- M.J. van der Laan (1990), *Dabrowska’s multivariate product limit estimator and the delta-method*, Master’s Thesis, Department of Mathematics, Utrecht, the Netherlands.
- M.J. van der Laan (1992), *Efficient Estimator of the Bivariate Survival Function for Right Censored Data*, Technical Report No. 337, Department of Statistics, University of California, Berkeley.
- M.J. van der Laan (1993a), *General identity for linear parameters in convex models with application to efficiency of the (NP)MLE*, Preprint nr. 765, Department of Mathematics, Utrecht, the Netherlands. Submitted for publication in *Ann. Math. Statist.*
- M.J. van der Laan (1993b), *Efficient estimator of the bivariate survival function and repairing NPMLE*, Preprint nr. 788, Department of Mathematics, Utrecht, the Netherlands.
- M.J. van der Laan (1993c), *Modified EM-estimator of the bivariate survival function*, Preprint nr. 768, Department of Mathematics, Utrecht, the Netherlands. To appear in *Mathematical Methods in Statistics*.
- M.J. van der Laan (1993d), *Efficiency of the NPMLE in a general class of missing data models*, Preprint, Department of Mathematics, Utrecht, the Netherlands. Submitted for publication in *Bernoulli*.
- M.J. van der Laan (1993e), *Efficient and inefficient estimation in semiparametric models*, thesis, ISBN nr. 90-393-0339-8, Department of Mathematics, Utrecht, the Netherlands.
- G. Neuhäus (1971), On weak convergence of stochastic processes with multidimensional time parameter, *Ann. Math. Statist.* **42** 1285–1295.
- K.R. Parthasarathy (1967), *Probability Measures on Metric Spaces*, Academic Press, New York.
- D. Pollard (1990), *Empirical Processes: Theory and applications*, Regional conference series in probability and statistics **2**, Inst. Math. Statist., Hayward, California.
- R.L. Prentice and J. Cai (1992a), Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika* **79**, 495–512.
- R.L. Prentice and J. Cai (1992b), Marginal and conditional models for the analysis of

- multivariate failure time data. Klein, J.P. and Goel, P.K., editors, *Survival Analysis State of the Art*. Kluwer, Dordrecht.
- R.C. Pruitt (1991a), On negative mass assigned by the bivariate Kaplan-Meier estimator, *Ann. Stat.* **19**, 443–453.
- R.C. Pruitt (1991b), *Strong consistency of self-consistent estimators: general theory and an application to bivariate survival analysis*, Technical Report nr. 543, University of Minnesota.
- R.C. Pruitt (1993), Small sample comparisons of six bivariate survival curve estimators, *J. Statist. Comput. Simul.*, **45** 147–167.
- W-Y. Tsai, S. Leurgans and J. Crowley (1986), Nonparametric estimation of a bivariate survival function in the presence of censoring, *Ann. Statist.* **14** 1351–1365.
- B.W. Turnbull (1976), The empirical distribution with arbitrarily grouped censored and truncated data, *J.R. Statist. Soc.* **B38**, 290–5.
- A.W. van der Vaart and J.A. Wellner (1995), *Weak convergence and empirical processes*. IMS Lecture Notes-Monograph Series.