

Efficiency of NPMLE in Nonparametric Missing Data Models

Mark J. van der Laan¹ and Richard D. Gill²

Division of Biostatistics¹

University of California

Berkeley, CA 94720

Department of Mathematics²

Utrecht University

The Netherlands.

May 15, 2001

Abstract

Suppose that a random variable X of interest is grouped or censored or missing so that one only observes a coarsening of X , i.e., a random set containing X with probability 1. It is assumed that the coarsening mechanism has the *coarsening at random* property. Suppose furthermore that the coarsening either equals X itself or that is a set with positive X -probability. We modify the NPMLE of the distribution of X by demanding that its support is the set of observed data points. We provide a general theorem giving sufficient conditions for efficiency of this NPMLE, or efficiency of the NPMLE after a small data reduction. We apply the theorem to a number of examples.

1 Introduction.

Suppose we are concerned with nonparametric estimation of the distribution function F of a random variable X . However, there is a nuisance censoring, missingness or grouping mechanism so that for each of n i.i.d. observations X_i we only observe a coarsening or reduction Y_i of X_i , $i = 1, \dots, n$. A coarsening of X is a random region Y which contains X with probability 1. We assume such measurability that the conditional distribution G of Y , given X , is (almost everywhere) well defined; see Gill, van der Laan and Robins [12] for a convenient set-up, covering as far as we know all examples of interest. The distribution of Y is determined by the marginal of X , together with the conditional of Y given X (i.e. by F and G together); however, though G is not necessarily known, we will denote it by P_F suppressing the dependence on G . This makes good sense under the assumption of *coarsening at random* defined by Gill, van der Laan, and Robins [12] as, for each x, x' :

$$G(dy | X = x) = G(dy | X = x') \text{ on } \{y : x \in y\} \cap \{y : x' \in y\} \quad (1)$$

Intuitively, this means that observing $Y = y$ tells us no more and no less than $X \in y$. The coarsening at random assumption was introduced by Heitjan and Rubin [14] and further studied by Jacobsen and Keiding [15] and Gill, van der Laan and Robins [12]. Very many specific nonparametric CAR models have been studied in the literature. Well known examples are: *the (univariate) random censoring model*—see Andersen et al. [1], Wellner [30], Gill [10]; *the double censoring model*—see Chang and Yang [5], Chang [6], Gu and Zhang [13]; and *the multivariate random censoring model*—see Dabrowska [6,7], Prentice and Cai [22,23], van der Laan [18].

Let $\Delta = I(Y = \{X\})$ and let $P_{F,0}(B) = P(Y \in B, \Delta = 0)$. Thus Δ is the indicator of a complete observation, while $P_{F,0}$ is the law of the incomplete observations. Attention will be restricted to CAR models of laws P_F of Y with

$$\pi_G(x) = P(\Delta = 1 \mid X = x) > 0 \text{ } F\text{-a.e.}, \text{ and } F(Y) > 0 \text{ } P_{F,0}\text{-a.e.} \quad (2)$$

In words, the coarsening mechanism must be ‘at random’; next, given the underlying value $X = x$, x should be exactly observed with positive probability; if however x is coarsened, the coarsening Y should be a set of positive probability for the underlying distribution of X . Note that $\pi_G(x) = G(\{x\} \mid X = x)$ indeed only depends on the censoring mechanism G .

The reason the CAR property is so significant is that, under CAR, the likelihood for the data factors into an F part and a G part, where the F part is the same as if the coarsening had been according to a predetermined, fixed, partition of the underlying sample space. Let μ be a dominating probability measure for F , and define $f = dF/d\mu$. Then the log likelihood for F based on our n observations is

$$L_\mu(F) = \sum_{i=1}^n \log \left(f(X_i)^{\Delta_i} F(Y_i)^{1-\Delta_i} \right). \quad (3)$$

We obtain our modified NPMLE F_n of F by maximizing $L_{\mu_n}(F)$ over all F dominated by a data-dependent, discrete measure μ_n generated as follows. Select a point x_i from each set Y_i , $i = 1, \dots, n$: thus if $\Delta_i = 1$, then $x_i = X_i$ and if $\Delta_i = 0$, then one can select an arbitrary point x_i in Y_i , $i = 1, \dots, n$. It may be that one of the selected points x_i is never alone in the observed Y_j , i.e., each observed set Y_j containing x_i also contains at least one other, different, $x_{i'}$. If so delete then this point x_i from the collection. Repeat till no more deletions are possible. At that stage we have (relabelling the selected points) a collection of points x_1, \dots, x_m such that each x_i is the *only one* of these points in at least one of the Y_j , while each Y_j does contain one or more x_i . Note that the collection includes all exact observations (X_i such that $Y_i = \{X_i\}$ and therefore $\Delta_i = 1$), since such points are forced into the initial collection and can never be deleted as they are alone in their own Y_i . It only contains more points than the exact observations when there are observed regions Y_i which are empty of exact observations. Let μ_n be counting measure on $\{x_1, \dots, x_m\}$ and define $\mathcal{F}(\mu_n)$ as the set of discrete distributions F with $F \ll \mu_n$. Let P_n be the empirical measure of the data Y_1, \dots, Y_n . Finally, define F_n by:

$$\begin{aligned} F_n &= \arg \max_{F \in \mathcal{F}(\mu_n)} L_{\mu_n}(F) = \arg \max_{F \in \mathcal{F}(\mu_n)} \int \log(p_F) dP_n \\ &= \arg \max_{F \in \mathcal{F}(\mu_n)} \frac{1}{n} \sum_{i=1}^n \log(p_F(Y_i)), \end{aligned} \quad (4)$$

where $p_F(Y_i) = f(X_i)^{\Delta_i} F(Y_i)^{1-\Delta_i}$, $i = 1, \dots, n$. Since μ_n is discrete we have that $f(x) = dF(x)/d\mu_n(x)$ is the probability mass of F at x and thus $F(Y_j) = \sum_{i: x_i \in Y_j} f(x_i)$. Since $L_{\mu_n}(F) = -\infty$ if $F(Y_j) = 0$ for any j , and each x_i is the only point in at least one Y_j , it follows that F_n puts positive mass on each of the points in $\{x_1, \dots, x_m\}$. We note that F_n is might not be an actual NPMLE in the sense of Kiefer and Wolfowitz [16]: in other words, there might exist another F'_n with $L_{F_n+F'_n}(F'_n) > L_{F_n+F'_n}(F_n)$. Our modification of the NPMLE leads to an estimator which is easier to compute and easier to study, while it does not lead to a loss in asymptotic efficiency. We will call F_n simply *the* NPMLE though a more correct name would be ‘sieved NPMLE’, where the ‘sieve’ does not only depend on the data but also on the arbitrary choices

in selecting an x_i in a Y_i with $\Delta_i = 0$. In appendix A we prove existence and uniqueness (up to the possible freedom of choice in μ_n) of this NPMLE.

The aim of the present paper is to find sufficient conditions for consistency, asymptotic normality and efficiency of the NPMLE which on the one hand cover many models of interest (both old and new) but on the other hand do not require delicate model-specific calculations. Our claim is that a strengthened version, assumption 1 below, of the condition (2) together with further ‘structural’ conditions on the class of sets Y which can be observed does exactly this. The models which are amenable to our approach are certainly the ‘easier’ examples for the NPMLE in missing data problems; for instance, condition (2) puts us into the realm of root- n rate estimation. However, they include many examples which so far were only analysed using heavy and model-specific calculations. So far, the general theory which is available (Bickel and Ritov [3], van der Vaart [28]) though certainly deep, is so general that much further work has to be done in any specific example before it is clear whether or not it is covered, while in the case of the first paper, the simple NPMLE we have described is replaced by a sieved NPMLE on a larger and finer collection of points, depending on a band-width parameter which remains to be chosen by the statistician.

Of the examples mentioned above, our approach covers univariate random censoring and (under further conditions) double censoring, but not the general bivariate random censoring model. However even when the approach in first instance does not work, it can provide valuable insight, by suggesting how the statistician might make a small reduction of the data after which the assumptions *are* satisfied. This will of course cost a small loss of efficiency but on the other hand it often produces estimators which have much more stable behaviour in small samples (see e.g. van der Laan [19]), as well as being much easier to study. One can consider this data-reduction as a form of regularisation.

Assumption 1 will strengthen (2) by replacing the two positivity assumptions by uniform positivity. Indeed, the existence of points x where $\pi_G(x) = 0$ or in the neighbourhood of which π_G is arbitrarily close to 0, and of points y with $F(y) = 0$ though $\delta = 0$, typically correspond to singularities in the sense of requiring delicate mathematical analysis. If actually $\pi_G(x) = 0$ on a set with positive probability under F , then F cannot be estimated at root- n rate at all, on this set. If $F(Y) = 0$ for $\Delta = 0$ on a set with positive probability the NPMLE may be inconsistent, as several examples show. For example, consider the bivariate right-censoring model where a continuous (T_1, T_2) is independently right-censored by (C_1, C_2) . The coarsening of the singly-censored observations, i.e. T_1 is observed and T_2 is right-censored, are half-lines in the plane and thus assumption (2) fails since $F(Y) = 0$ for the singly-censored observations. The fact that these lines do not contain any observed $X_i = (T_{1i}, T_{2i})$ ’s causes inconsistency of the NPMLE (Tsai, Leurgans and Crowley [24]). In van der Laan [18] the following approach is taken to solve this problem. The censoring times and the observed component T_j , $j = 1, 2$, of the singly-censored observations are interval censored so that the coarsenings of these new reduced observations are now strips. It is shown that the NPMLE based on the reduced data is efficient for the reduced data, and moreover, if one lets the amount of reduction converge to zero slowly enough as the sample size increases to infinity, it is asymptotically efficient for the original data.

As another example of such singularities, consider x equal to the right-hand endpoint of the support of right-censored data, for which our NPMLE is the famous Kaplan-Meier estimator. In Gill [8] the Kaplan-Meier estimator at this point has been shown, using quite delicate martingale arguments, to be asymptotically normal under an integrability condition, but efficiency has still not been proved. Without the integrability condition the estimator is only consistent at a lower rate and probably not even at the optimal rate (see Gill [10], conjecture at end of section 11).

This shows that even in models where the data-structure is well-understood and the NPMLE is explicit it can be very difficult to analyse, when assumption 1 fails. This example also shows that the weakest possible conditions will depend on the particular data structure. In this paper we study general NPMLE in the sense that the data-structure is not even specified.

As we will see, application of our theorem to the classical Kaplan-Meier estimator shows (as is well known) that the Kaplan-Meier estimator is efficient on any compact interval strictly inside the support of the data. Application of our theorem to doubly-censored data shows that the sieved-NPMLE is efficient under the same conditions as used in the (delicate and lengthy) consistency and asymptotic normality proof in Chang and Yang [5] and Chang [4], respectively. In particular applications a specific fine-tuned analysis might prove efficiency of the sieved-NPMLE under weaker or even optimal conditions (see Gu and Zhang [13], for double censoring), but such an analysis has to deal in a very model-specific manner with the points where assumption 1 fails. In our examples, after we have arranged matters so that assumption 1 is satisfied, our further assumptions turn out to hold automatically.

In section 2 we show that we can compute the NPMLE with the Turnbull- or EM-algorithm (Turnbull [25]), i.e., iterating the well known self-consistency equation, by choosing an initial estimator with support $\{x_1, \dots, x_m\}$. In section 3 we state two theorems which provide sufficient conditions for efficiency of the NPMLE.

The first theorem is for the case that X is a real vector, and has stronger conditions, while the second theorem is more general. The first theorem is applicable (possibly after further data-reduction to remove singularities) in our main examples. In section 4 we provide the proof of these theorems. The outline of the efficiency proof is given in section 4.1. It is based on an identity for the NPMLE in convex linear models (van der Laan [17]) expressing the difference between estimator and estimand as a function-indexed empirical process, indexed by a random point. The function is the efficient influence curve for the estimand under consideration, indexed by the unknown F ; the random index is the NPMLE. The further ingredients of this proof are the invertibility of the information operator, a Donsker class condition and a continuity condition for the efficient influence curve. The supremum-norm invertibility of the information operator is established in section 4.2 under assumption 1 and a second condition. In the remaining subsections of section 4 the Donsker class and continuity condition are covered and convenient sufficient conditions are worked out for the case that X is a real vector. In section 5 the first theorem is applied to right-censored data, doubly censored data, bivariate rectangle-censored data, and censored data supplemented with current status data.

The paper generalizes and improves part of the first author's thesis, published as van der Laan [17].

2 EM-estimating-equations for the NPMLE.

In this section we will show that the NPMLE can be computed by iterating the self-consistency or Turnbull equations. Let $\mathcal{S}(F_n)$ be the class of lines $\epsilon F_1 + (1 - \epsilon)F_n$, $F_1 \in \mathcal{F}(\mu_n)$, through F_n with score $h = d(F_1 - F_n)/dF_n \in L_0^2(F_n)$. By convexity of $\mathcal{F}(\mu_n)$ these lines are submodels of $\mathcal{F}(\mu_n)$. Let $S(F_n)$ be the corresponding tangent cone (collection of scores) and notice that it includes all $h \in L_0^2(F_n)$ with finite supremum norm. Then it is trivial to verify that the tangent space $T(F_n)$ (the closure of the linear span of $S(F_n) \subset L_0^2(F_n)$) equals $L_0^2(F_n)$. The lines $F_{n,\epsilon,h} \in \mathcal{S}(F_n)$ with score h generate one-dimensional submodels $P_{F_{n,\epsilon,h}}$ through P_{F_n} with score $A_{F_n}(h) \in L_0^2(P_{F_n})$, where A_{F_n} called the score operator.

In coarsened data models it is given by

$$A_{F_n} : L_0^2(F_n) \rightarrow L_0^2(P_{F_n}) : A_{F_n}(h)(Y) = E_{F_n}(h(X) | Y)$$

(van der Vaart [26], Gill [9], Bickel et al. [2], section 6.6).

F_n maximizes the log likelihood over $\mathcal{F}(\mu_n)$. By differentiating the log likelihood $\epsilon \rightarrow L(F_n, \epsilon, h)$ for the one-dimensional submodel $P_{F_n, \epsilon, h}$ we obtain:

$$0 = \int A_{F_n}(h) dP_n = \frac{1}{n} \sum_{i=1}^n E_{F_n}(h(X) | Y_i) \text{ for all } h \in S(F_n) \text{ with } \|h\|_\infty < \infty. \quad (5)$$

In particular, this holds for $h = I_E - F_n(E)$ for a collection of measurable events E . Let \mathcal{E} be a collection of measurable sets so that each $F \in \mathcal{F}(\mu_n)$ is uniquely determined by $F(E)$, $E \in \mathcal{E}$. Then (5) reduces to the self-consistency equation:

$$F_n(E) = \int P_{F_n}(X \in E | Y) dP_n(Y) \text{ for all } E \in \mathcal{E} \quad (6)$$

or equivalently, with $f_n = dF_n/d\mu_n$,

$$f_n(x_i) = \int P_{F_n}(X = x_i | Y) dP_n(Y) \text{ for all } x_i, i = 1, \dots, m.$$

A solution of (6) can be computed with the EM-algorithm which corresponds with iterating the self-consistency equation in the following manner. Start with a $F_n^0 \equiv \mu_n$. Now, for $k = 0, 1, \dots$ we compute

$$f_n^{k+1}(x_i) = \int P_{f_n^k}(X = x_i | Y) dP_n(Y) = \frac{1}{n} \sum_{j=1}^n P_{f_n^k}(X = x_i | Y_j), \quad i = 1, \dots, m. \quad (7)$$

This means that each observation Y_j has mass $1/n$ which it redistributes over Y_j as follows: a point $x_i \in Y_j$ gets mass $1/n \times P_{f_n^k}(X = x_i | Y_j)$. We can see why assumption 1 helps the solution to behave well: for large sample sizes each region Y_j contains many exact observations, so the redistribution of mass over Y_j can be done sensibly.

We will now show that an application of Corollary 1 in Wu [31] proves that if one starts the algorithm with an initial point $F_n^0 \equiv \mu_n$, then F_n^k will converge to the unique MLE F_n . Corollary 1 in Wu [31] states that if the likelihood $L(p_1, \dots, p_m)$ is a unimodal function in Ω with $F_n \in \Omega$ being the only stationary point and L and Ω satisfy some differentiability and compactness condition, then for any $F_n^0 \in \Omega$ the sequence F_n^k converges to the unique maximizer F_n . We apply this corollary with $\Omega \equiv \{(p_1, \dots, p_k) : p_j \geq 1/n\}$, where p_j represents the pointmass of F at x_j , $j = 1, \dots, k$. Now note that since the MLE F_n solves the self-consistency equation the mass assigned to the x_i , $i = 1, \dots, m$, will be at least $1/n$ and thus $F_n \in \Omega$. In addition, if we start the algorithm with an initial point $F_n^0 \equiv \mu_n$, then at each step of the algorithm $F_n^k \in \Omega$. If F_n is a local maximum of L , then it follows, as shown above for the MLE F_n , that it solves the self-consistency equations. In Appendix A it is shown that equation (6) has a unique solution in the class $\{F : F \equiv \mu_n\}$, given by the actual NPMLE F_n , and thus that the local maximum equals F_n . Thus that proves that the likelihood L has no local maxima and thus is unimodal. The compactness condition and differentiability conditions stated in Wu [31] (see (5), (6), (7) on page 96 and continuity of his $D_{10}Q$) hold trivially for the functional $(p_1, \dots, p_m) \rightarrow L(p_1, \dots, p_m)$ defined on Ω . This verifies the conditions of Corollary 1 in Wu [31] and thus proves the wished convergence of the EM-algorithm. We summarize the obtained facts in the following lemma (notation: $P_n f = \int f dP_n$):

Lemma 2.1

1. The sieved-NPMLE $F_n \in \mathcal{F}(\mu_n)$ over $\mathcal{F}(\mu_n)$ exists and is unique.
2. The score operator at P_F is given by:

$$A_F : L_0^2(F) \rightarrow L_0^2(P_F) : h \mapsto E_F(h(X) | Y).$$

F_n solves

$$P_n(A_{F_n}(h - F_n(h))) = 0 \text{ for all } h \in L^2(F_n) \text{ with } \|h\|_\infty < \infty. \quad (8)$$

3. F_n can be found by iterating the self-consistency equation from some $F_n^0 \equiv \mu_n$. F_n is the unique solution of (8) in $\{F \in \mathcal{F}(\mu_n) : F \equiv \mu_n\}$.
4. Suppose that for a $F_1 \equiv F$ with $\|dF/dF_1\|_\infty < \infty$

$$P_F A_{F_1}(h - F_1 h) = 0 \text{ for all } h \text{ with } \|h\|_\infty < \infty.$$

Then $F_1 = F$.

The last statement shows that the self-consistency equation identifies F uniquely in the limit, at least up till all distributions which are equivalent with the true F . This result is proved in Appendix B.

3 Theorems.

In this section we will state efficiency theorems for the NPMLE $F_n(E)$ of $F(E)$ for a given set E . For efficiency theory we refer to Bickel, Klaassen, Ritov, Wellner [2]. Firstly, we will need to show that $F(E)$ is pathwise differentiable and provide the formula for the canonical gradient which is also called the efficient influence curve of $F(E)$. Subsequently, we will define supremum-norm efficiency of an estimator over a class of sets \mathcal{E} in terms of this efficient influence curve. Finally, we will state the efficiency theorems for F_n .

3.1 Pathwise differentiability and the efficient influence function.

For each probability law F define $\mathcal{S}(F)$ as all lines $\epsilon F_1 + (1 - \epsilon)F$, $F_1 \ll F$, with scores $h = d(F_1 - F)/dF \in L_0^2(F)$. Let $S(F)$ be the tangent cone of $\mathcal{S}(F)$ and recall that the tangent space $T(F)$ of $\mathcal{S}(F)$ equals $L_0^2(F)$. The score operator at F is given by:

$$A_F : L^2(F) \rightarrow L^2(P_F) : A_F(h)(Y) = E_F(h(X) | Y).$$

Let $S(P_F)$ be the tangent cone at P_F corresponding with the submodels $P_{F_{\epsilon,h}}$, $h \in S(F)$; in other words $S(P_F)$ equals the range of $S(F)$ under A_F . We want to show that $F(E)$ is pathwise differentiable relative to the class of submodels specified above with a canonical gradient in $S(P_F)$ (for a definition of pathwise differentiability we refer to BKRW).

The pathwise differentiability result will be stated in terms of the following operators directly derived from the score operator. The adjoint of A_F is given by:

$$A_F^\top : L^2(P_F) \rightarrow L^2(F) : A_F^\top(v)(X) = E_G(v(Y) | X).$$

The information operator is defined by:

$$I_F = A_F^\top A_F : L^2(F) \rightarrow L^2(F) : I_F(h)(X) = E_G(E_F(h(X) | Y) | X).$$

Application of a theorem in van der Vaart [27] (see also Bickel et al. [2]) to the parameter $F(E)$, for a given set E , and the $L^2(F)$ -invertibility result for the information operator (see lemma 5.1 in appendix C) provides us now with the following result.

Lemma 3.1 *If $\pi_G(x) > \delta > 0$ for some $\delta > 0$ F -a.e., then the real valued parameter $F(E)$ is pathwise differentiable at P_F relative to $S(P_F)$ with efficient influence function (i.e. canonical gradient) given by:*

$$\tilde{I}(F, E) = A_F I_F^{-1} (I_E - F(E)) \in L_0^2(P_F). \quad (9)$$

Here $I_F^{-1} : L_0^2(F) \rightarrow L_0^2(F)$ is the inverse of I_F which is given by:

$$I_F^{-1} = \sum_{k=0}^{\infty} (I - I_F)^k,$$

where I is the identity operator. The operator norm of I_F^{-1} is bounded by $1/\delta$:

$$\| I_F^{-1}(h) \|_F \leq \frac{1}{\delta} \| h \|_F.$$

We will now recall the relevant efficiency and empirical process theory (see BKRW). $F_n(E)$ is an efficient estimator of $F(E)$ if

$$F_n(E) - F(E) = (P_n - P_F)\tilde{I}(F, E) + R_{n,E},$$

where $R_{n,E} = o_P(1/\sqrt{n})$. Let \mathcal{E} be a collection of measurable sets. $\sqrt{n}(P_n - P_F)\tilde{I}(F, E)$ is a sum of n i.i.d. mean zero random variables which converges by the central limit theorem to a normal distribution with mean zero and variance $P_F\tilde{I}(F, E)^2$. By varying $E \in \mathcal{E}$ we obtain an empirical process $(\sqrt{n}(P_n - P_F)\tilde{I}(F, E) : E \in \mathcal{E})$, which can be considered as an element of $\ell^\infty(\mathcal{G}) = \{H : \mathcal{G} \rightarrow \mathbb{R} : \sup_{g \in \mathcal{G}} |H(E)| < \infty\}$ where $\mathcal{G} = \{\tilde{I}(F, E) : E \in \mathcal{E}\}$ and $\ell^\infty(\mathcal{E})$ is endowed with the Borel sigma-algebra. Empirical process theory investigates if the empirical process $(\sqrt{n}(P_n - P_F)g : g \in \mathcal{G})$, indexed by some class \mathcal{G} of measurable functions, converges in distribution to a tight Gaussian process corresponding with the covariance structure of the empirical process. Here weak convergence is defined in the Hoffmann-Jørgensen sense (see e.g. van der Vaart and Wellner [29], Pollard [21]). A class for which this weak convergence holds is called a Donsker class. If \mathcal{G} is Donsker and $\sup_{E \in \mathcal{E}} |R_{n,E}| = o_P(1/\sqrt{n})$, then we call F_n \mathcal{E} -supremum-norm efficient.

3.2 Efficiency of the NPMLE when X is multivariate.

The next theorem provides sufficient conditions for efficiency of the NPMLE F_n for the case that $\mathcal{X} = \mathbb{R}^k$ and $\mathcal{E} = \{(-\infty, t] : t \in \mathbb{R}^k\}$.

These conditions are expressed in terms of the uniform sectional variation norm of a multivariate real valued cadlag function as introduced in Gill, van der Laan, Wellner [11], where ‘‘cadlag’’ is defined in Neuhaus [20]. We will say that a function f is cadlag F -a.e. if there exists a function f' which equals f F -a.e. and which is right-continuous with left-hand limits in the sense of Neuhaus [20]. A real valued cadlag function is said to be of bounded uniform sectional variation if the variations of all sections ($s \rightarrow f(s, t)$ is a section of the function f) and of the function itself is uniformly (in all sections) bounded. The corresponding norm, i.e. the maximum of all the variations, is denoted by $\| \cdot \|_v^*$. Here the variation is taken over a support K of F , where K is a set satisfying $\int_K h dF = \int h dF$ for all $h \in L^2(F)$: the theorem only needs to hold for one such K . When we write $\| f \|_v^*$ for a function f which is cadlag F -a.e., then we mean the variation norm of its version f' which is cadlag. Similarly, if we say that f is of uniform sectional variation, then we mean that its cadlag version f' is of uniform sectional variation.

In assumption 2 of the theorem $P_{1n}(A) = 1/n \sum_{i=1}^n I(X_i \in A, \Delta_i = 1)$ is the empirical distribution of the distribution $P_{F,1}(A) = P(X \in A, \Delta = 1)$ of the observed X_i 's. Furthermore, in assumption 2 one needs to recall that a collection of sets \mathcal{A} is called $P_{F,1}$ -Glivenko-Cantelli if

$$\sup_{A \in \mathcal{A}} |P_{1n}(A) - P_{F,1}(A)| \rightarrow 0 \text{ in probability.}$$

We also define $\mathcal{F}(\delta) \equiv \{F : \inf_{\{Y: \Delta=0\}} F(Y) > \delta > 0\}$.

Theorem 3.1 *Assume that X is a \mathbb{R}^k -valued random variable, and for some $\delta > 0$*

Assumption 1: $\pi_G(x) = P(\Delta = 1 | X = x) > \delta > 0$ for $x \in K$ and $F(Y) > \delta > 0$ $P_{F,0}$ -a.e.

Assumption 2: $\{y : \Delta = 0\}$ is a $P_{F,1}$ -Glivenko-Cantelli class of sets.

Assumption 3*:

(1) $x \rightarrow \pi_G(x)$ is a multivariate cadlag function F -a.e., it is of uniform sectional variation,

$$(2) \sup_{\|h\|_v^* \leq 1, F_1 \in \mathcal{F}(\delta)} \|x \rightarrow \int_{\{y: x \in y, \Delta(y)=0\}} \frac{\int_y h(u) dF_1(u)}{F_1(y)} dG(y | x)\|_v^* < \infty, \quad (10)$$

where the function is cadlag F -a.e. for each $h, F_1 \in \mathcal{F}(\delta)$ and

$$(3) \left\{ (1 - \Delta) \frac{\int_y h dF}{F(y)} : F \in \mathcal{F}(\delta), \|h\|_v^* < 1 \right\} \quad (11)$$

is a P_F -Donsker class.

Then F_n is asymptotically supremum norm efficient.

If assumption 1 holds and the support points $\{x_1, \dots, x_m\}$ of F_n are in K (which holds with probability converging to 1), then

$$F_n(E) - F(E) = (P_n - P_F) \tilde{I}(F_n, E). \quad (12)$$

Since assumption 3 in this theorem substitutes for the weaker assumptions 3 and 4 in the general theorem below we denoted this assumption with 3*.

The identity (12) in this theorem assumes that the support points $\{x_1, \dots, x_m\}$ of F_n are contained in the chosen support K of F . We can show that with probability tending to 1 the set $\{x_1, \dots, x_m\}$ consists only of the exact observations. This is shown as follows. Consider a possible censored region y . By assumption 1 we know that $F(y) > \delta > 0$. Since $\pi_G(x) > \delta > 0$ this implies that $P_{F,1}(y) > \delta_1 > 0$ for some δ_1 . Now, by assumption 2 we have $P_{1,n}(y) - P_{F,1}(y)$ converges to zero in probability uniformly in $\{y : \Delta = 0\}$. Thus the probability that there are no exact observations in y converges to zero uniformly in y when n converges to infinity.

As a consequence the requirement $\{x_1, \dots, x_m\} \subset K$ holds with probability tending to 1 if K contains the exact observations among $\{x_1, \dots, x_m\}$.

3.3 Efficiency of the NPMLE in general.

The following theorem provides sufficient conditions for \mathcal{E} -supremum-norm efficiency of F_n for a general outcome space of X . Recall the definition $P_{F,1}$ -Glivenko-Cantelli class given in the preceding subsection.

Theorem 3.2 *Suppose that Assumption 1 and 2 as in Theorem 3.1 and the next two Assumptions 3 and 4 hold.*

Assumption 3: *For a given collection of measurable sets \mathcal{E} we have that $\tilde{I}(F_n, E), E \in \mathcal{E}$ falls in a P_F -Donsker class \mathcal{G} with probability tending to 1.*

Assumption 4: *If $\|F_n - F\|_{\mathcal{E}} \rightarrow 0$ in probability, then*

$$\sup_{E \in \mathcal{E}} \|\tilde{I}(F_n, E) - \tilde{I}(F, E)\|_{P_F} \rightarrow 0 \text{ in probability.}$$

Then F_n is a \mathcal{E} -supremum-norm asymptotically efficient estimator of F .

If assumption 1 holds and $\{x_1, \dots, x_m\} \subset K$ (which holds with probability converging to 1), then

$$F_n(E) - F(E) = (P_n - P_F)\tilde{I}(F_n, E).$$

If assumptions 1, 2 and 3 hold, then $\sup_{E \in \mathcal{E}} |F_n - F|(E) = O_P(1/\sqrt{n})$.

A sufficient condition for assumption 3 is the following:

$$\mathcal{G} \equiv \cup_{F \in \mathcal{F}(\delta)} A_F(\mathcal{G}(F)) \text{ is a } P_F\text{-Donsker class,}$$

where

$$\mathcal{G}(F) = \left\{ \frac{1}{\pi_G(\cdot)} \left(I_E(\cdot) - \int_{\Delta=0} A_F(g)(y) dG(y | X = \cdot) \right) : \|g\|_{\infty} < 1, E \in \mathcal{E} \right\}.$$

4 Proof of efficiency theorems.

In the next subsection we will outline a proof of theorems 3.1 and 3.2. The proof consists of a number of steps, concerning respectively: the supremum norm invertibility of the information operator, the validity of the identity expressing the NPMLE in terms of empirical processes, a Donsker class condition, and a consistency condition on the estimated optimal influence curve. The remaining subsections are devoted to these separate steps.

4.1 Outline of proof of efficiency.

Since the model \mathcal{M} , with G known, is convex and $F \rightarrow P_F$ is linear Theorem 2.2 in van der Laan [17] yields the following identity:

$$F_1(E) - F(E) = - \int \tilde{I}(F_1, E) dP_F,$$

for all F_1 with $F \ll F_1$ and $dF/dF_1 \in L_0^2(F_1)$. We want to apply this identity to $F_1 = F_n$. Usually F_n does not dominate F so this identity cannot be directly applied. However, notice that the identity holds in particular for $F_1 = F_n(\alpha) = (1 - \alpha)F_n + \alpha F$ for any $\alpha \in (0, 1]$. Hence if $\tilde{I}(F_n(\alpha), E)$ converges to $\tilde{I}(F_n, E)$ in $L^1(P_{F,G})$ for $\alpha \rightarrow 0$, then the identity also holds for F_n . We refer to this condition as the *identity condition* which is proved by lemma 4.4 under assumption 1. The proof of this identity condition uses the in subsection 4.2 established supremum norm invertibility of the information operator $I_{F_n(\alpha)}$ uniformly in $\alpha \in [0, 1]$ and for fixed n . The supremum norm invertibility requires the condition $\{x_1, \dots, x_m\} \subset K$ which holds with probability tending to 1. Since this condition is asymptotically true we do not need to state the condition as an assumption for the asymptotic efficiency result.

If $\tilde{I}(F_n, E)$ is a score of $P_{F_n, \epsilon, h}$ for a certain one dimensional line $F_{n, \epsilon, h}$ through F_n with score h with finite supremum norm, then by lemma 2.1 (iii) we have the *efficient score equation*:

$$P_n \tilde{I}(F_n, E) = 0.$$

By (9), for this it suffices to show that $h_E^n = I_{F_n}^{-1}(I_E)$ has finite supremum norm on \mathcal{X} , which is proved by lemma 4.1 in the next subsection 4.2 under assumptions 1 and $\{x_1, \dots, x_m\} \subset K$.

Then combining the last two identities provides us with the following crucial identity: for each E

$$(F_n - F)(E) = \int \tilde{I}(F_n, E) d(P_n - P_F).$$

Note that this identity is proved under assumptions 1 and $\{x_1, \dots, x_m\} \subset K$ only and does thus hold for general sample spaces.

In the sequel of the proof we can assume that $\{x_1, \dots, x_m\} \subset K$ since this holds with probability tending to 1. Suppose now that with probability tending to 1 $\tilde{I}(F_n, E)$ lies in a P_F -Donsker class for all $E \in \mathcal{E}$. This is just the general assumption 3 of theorem 3.2. Lemma 4.5 in subsection 4.3 proves that assumptions 1,2 and the sufficient condition for assumption 3 of theorem 3.2 imply this *P_F -Donsker class condition*. In addition, lemma 4.7 in subsection 4.3 proves that assumptions 1,2 and 3* of theorem 3.1 are sufficient for the P_F -Donsker class condition for the case that $\mathcal{X} = \mathbb{R}^k$.

We need to remark here that the proof of the Donsker class condition (lemma 4.7) uses that the cadlag and uniform sectional variation requirements of assumptions (1) and (2) in assumption 3* hold for the functions $f_1(x) \equiv \pi_G(x)$ and $f_2(x) \equiv \int_{\{y: x \in y, \Delta(y)=0\}} \int_y h(u) dF_1(u) / F_n(y) dG(y | x)$. The assumptions (1) and (2) only require the existence of an F -a.e equal version which satisfies these requirements. We have that h_E^n is a simple function of f_1, f_2 , given in (14), and $\tilde{I}(F_n, E)(Y) = A_{F_n}(h_E^n)(Y) = E_{F_n}(h_E^n(X) | Y)$. Let f'_1, f'_2 be the F -a.e. equal versions of f_1, f_2 which are cadlag and for which the uniform sectional variation requirements hold. Let $h_E^{n'}$ be the corresponding version of h_E^n and let $\tilde{I}(F_n, E)' = A_{F_n}(h_E^{n'})$ the corresponding version of $\tilde{I}(F_n, E)$. So lemma 4.7 actually only proves that $\tilde{I}(F_n, E)'$ falls in a P_F -Donsker class with probability tending to 1. Thus it remains to show that this also implies that $\tilde{I}(F_n, E)$ falls in a

P_F -Donsker class with probability tending to 1. Let B be the set of points on which f'_j differs from f_j , $j = 1, 2$. Then $P(X \in B) = F(B) = 0$ and (because B is a given set independent of the data) the probability that F_n has a support point in B is zero as well. As a consequence $h_{E'}^n(X)\Delta = h_E^n(X)\Delta$ as functions of $Y \sim P_F$ (i.e. as random variables being given functions of Y) and with probability one $(1 - \Delta) \int_Y h_{E'}^n(x) dF_n(x) = (1 - \Delta) \int_Y h_E^n(x) dF_n(x)$ as functions of $Y \sim P_F$. Thus with probability one (i.e. when the support points of F_n do not fall in B) $A_{F_n}(h_{E'}^n)(Y) = A_{F_n}(h_E^n)(Y)$ as functions of $Y \sim P_F$, which shows that lemma 4.7 also proves that $\tilde{I}(F_n, E)$ falls in a P_F -Donsker class with probability tending to 1.

By well known empirical process theory (e.g. Pollard [21]) the P_F -Donsker class condition proves $\|F_n - F\|_{\mathcal{E}} = O_P(1/\sqrt{n})$. Again, well known empirical process theory (see equivalence of tightness condition for an empirical process, Pollard [21]) says that if

$$\sup_{E \in \mathcal{E}} \|\tilde{I}(F_n, E) - \tilde{I}(F, E)\|_{P_F} \rightarrow 0 \text{ in probability,}$$

then $\sup_{E \in \mathcal{E}} (P_n - P_F)(\tilde{I}(F_n, E) - \tilde{I}(F, E)) = o_P(1/\sqrt{n})$. We call this the $L^2(P_F)$ -consistency condition, which is simply assumption 4 of theorem 3.2. The latter provides us with supremum-norm-efficiency of F_n . This completes the proof of theorem 3.2. Lemma 4.8 proves that assumptions 1,2 and 3* imply the $L^2(P_F)$ -consistency condition for the case $\mathcal{X} = \mathbb{R}^k$. This completes the proof of theorem 3.1.

In the the following subsections we prove the identity condition (lemma 4.4), the P_F -Donsker class condition (lemmas 4.5 and 4.7) and the $L^2(P_F)$ -consistency condition for the Euclidean case (lemma 4.8). But first we will establish the needed supremum norm invertibility of the information operator uniformly in $F \in \mathcal{F}(\delta)$, and show that $F_n \in \mathcal{F}(\delta)$ for some $\delta > 0$ with probability tending to 1.

4.2 Supremum norm invertibility of the information operator.

In this subsection F represents an element of $\{F_0, F_n : n = 1, 2, \dots\}$, where F_0 represents the true distribution. The score operator at F is given by:

$$\begin{aligned} A_F(h)(Y) &= E_F(h(X) | Y)\Delta + E_F(h(X) | Y)(1 - \Delta) \\ &= h(X)\Delta + \frac{\int_Y h(x)F(dx)}{F(Y)}(1 - \Delta). \end{aligned}$$

To obtain the information operator we need to take the conditional expectation of Y , given $X = x$. Thus the information operator is given by:

$$I_F(h)(x) = \pi_G(x)h(x) + \int_{\Delta=0} \frac{\int_y h(u)F(du)}{F(y)} dG(y | x). \quad (13)$$

If $P(\Delta = 0 | x) = 0$, then we define the integral as zero. We remark that the conditional distribution $dG(\cdot | x)$ is only almost everywhere unique in x . As a consequence, the supremumnorm invertibility established in this subsection is true relative to a particular choice of the conditional distribution. In the sequel we let $dG(\cdot | x)$ be one such choice which is fixed throughout the paper.

Consider the equation $I_F(h)(x) = f(x)$ for some pointwise well defined f with finite supremum norm. For $x \in K$ we have $\pi_G(x) > \delta > 0$ so that we can divide both sides of $I_F(h)(x) = f(x)$ by $\pi_G(x)$ which gives for each $x \in K$

$$h(x) = \frac{1}{\pi_G(x)} \left\{ f(x) - \int_{\Delta=0} \frac{\int_y h(u)F(du)}{F(y)} dG(y | x) \right\}. \quad (14)$$

For the moment denote the right-hand side by $C_F(h, f)(x)$: i.e., we consider the equation $h(x) = C_F(h, f)(x)$. If we assume that f lies in the range of I_F (so in particular if assumption 1 holds), then we know by lemma 5.1 (Appendix C) that there exists a $h' \in L^2(F)$, which is unique in $L^2(F)$, with $\| I_F(h') - f \|_F = 0$: i.e. $\| h' - C_F(h', f) \|_F = 0$. Notice that if $\| h - g \|_F = 0$, then for each x $C_F(h - g, f)(x) = 0$. So even if h' is only uniquely determined in $L^2(F)$, then $C_F(h', f)(x)$ is uniquely determined for each x . Now, we can define $h(x) = C_F(h', f)(x)$. Then $\| h - h' \|_F = \| C_F(h', f) - h' \|_F = 0$. So in this way we have found a solution h of (14) which holds for each $x \in K$ instead of only in $L^2(F)$ sense. This proves that I_F is onto in supremum-norm sense.

Now, suppose that $I_F(h)(x) = 0$, $x \in K$. By the $L^2(F)$ -invertibility of I_F this implies $\| h \|_F = 0$. If $I_F(h) = 0$ pointwise and $\| h \|_F = 0$, then by the argument given above this implies (see (13)) that $\pi_G(x)h(x) = 0$ pointwise. This proves that $h = 0$ pointwise.

Let $(B(K), \|\cdot\|_\infty)$ be the Banach space of functions on K with finite supremum norm $\|\cdot\|_\infty$. We have proved the following lemma.

Lemma 4.1 *Let $F \in \{F_0, F_n : n = 1, 2, \dots\}$, where F_0 is the true distribution of X , and $\{x_1, \dots, x_m\} \subset K$. If assumption 1 (at F_0) holds, then $I_F : (B(K), \|\cdot\|_\infty) \rightarrow (B(K), \|\cdot\|_\infty)$ is 1-1 and onto.*

From now on, when we talk about $I_F^{-1}(f)$, we mean this pointwise well defined solution.

For the Donsker class condition we have to consider the solution h_E^n of $I_{F_n}(h_E^n) = I_E$ for $E \in \mathcal{E}$. We will show that $\| h_E^n \|_\infty < M \| I_E \|_\infty$ for some $M < \infty$. So we want a uniform (in n) bound on the norm of the mapping $I_{F_n}^{-1}$ w.r.t. the supremum norm.

For this purpose we consider $I_{F_n}(h^n) = f$. The approach to be followed is to bound the right-hand side of (14) with $F = F_n$ in the supremum norm of f and $\| h^n \|_{F_n}$, where we can use that the latter is uniformly bounded by the $L^2(F_n)$ norm of f (by lemma 5.1).

Firstly, we need to bound $F_n(y)$ away from zero for y with $\Delta = 0$. Assumption 2 will take care of this.

Lemma 4.2 *Assume $\{x_1, \dots, x_m\} \subset K$. If assumptions 1,2 hold, then $F_n(y) > \delta_1 = \delta^2 > 0$ $P_{F_0,0}$ -a.e. with probability tending to 1, (where δ is the δ in assumption 1).*

Proof. The completely observed X_i receives full mass $1/n$ from Y_i and therefore $f_n^k(X_i) \geq \#\{j : X_j = X_i\}/n$, $i = 1, \dots, m$. This tells us that for each measurable A we have $F_n(A) \geq P_{1n}(A)$. In particular, $F_n(y) \geq P_{1n}(y)$. Assumption 2 tells us that this converges uniformly in $\{y : \Delta = 0\}$ to $P_{F_0,1}(y)$. Since $\pi_G > \delta > 0$ we have $dP_{F_0,1}(x) \geq \delta dF_0(x)$ and thus for y with $\Delta(y) = 0$ we have $P_{F_0,1}(y) \geq \delta F_0(y)$. This proves that if $\Delta(y) = 0$, then $F_n(y) \geq \delta F_0(y)$ with probability tending to 1. Finally use that $F_0(y) > \delta$ by assumption 1. \square

Thus with probability tending to 1 $F_n \in \mathcal{F}(\delta)$ for some $\delta > 0$. Assume that $F_n \in \mathcal{F}(\delta)$. Then the denominators $F_n(y)$ in (14) are uniformly bounded away from 0 by $\delta > 0$. We can bound $\int h^n(u)F_n(du)$ by $\| h^n \|_{F_n}$. By bounding $1/\pi_G$ and the denominator by $1/\delta$ we obtain:

$$h^n(x) = C_n(h^n, f)(x) \leq \frac{1}{\delta} \left(\| f \|_\infty + \frac{\| h^n \|_{F_n}}{\delta} \int_{\Delta=0} dG(y | x) \right),$$

where $\int_{\Delta=0} dG(y | x) = 1 - \pi_G(x) \leq 1$.

By lemma 5.1 (the uniform in n bounded invertibility of I_{F_n} w.r.t. $L^2(F_n)$) we have

$$\| h^n \|_{F_n} \leq M_1 \| f \|_{F_n} \leq M \| f \|_\infty \text{ for some } M < \infty.$$

Note that the last step uses that $\{x_1, \dots, x_m\} \subset K$ so that indeed the $L^2(F_n)$ -norm can be bounded by the supremum norm over K . Consequently, we have $\| h^n \|_\infty = \| I_{F_n}^{-1}(f) \|_\infty \leq M \| f \|_\infty$ for some $M < \infty$.

This proves the following lemma.

Lemma 4.3 *Assume $\{x_1, \dots, x_m\} \subset K$. If assumptions 1 holds and $F_n \in \mathcal{F}(\delta)$ for some $\delta > 0$, then $I_{F_n} : (B(K), \| \cdot \|_\infty) \rightarrow (B(K), \| \cdot \|_\infty)$ is onto and has bounded inverse satisfying: $\sup_{\|h\|_\infty \leq 1} \| I_{F_n}^{-1}(h) \|_\infty \leq M$ for certain $M < \infty$ depending on F_n only through δ . If assumption 2 holds, then for some $\delta > 0$ $F_n \in \mathcal{F}(\delta)$ with probability tending to 1.*

4.3 Identity condition.

Let n be given and define $F_n(\alpha) = (1 - \alpha)F_n + \alpha F$. Denote $h_{\alpha, E} = I_{F_n(\alpha)}^{-1}(I_E)$ and $h_E = I_{F_n}^{-1}(I_E)$, as defined pointwise in the preceding subsection. We have

$$\begin{aligned} A_{F_n(\alpha)} I_{F_n(\alpha)}^{-1}(I_E) - A_{F_n} I_{F_n}^{-1}(I_E) &= (A_{F_n(\alpha)} - A_{F_n}) I_{F_n(\alpha)}^{-1}(I_E) - A_{F_n(\alpha)} I_{F_n(\alpha)}^{-1}(I_{F_n(\alpha)} - I_{F_n}) I_{F_n}^{-1}(I_E) \\ &= (A_{F_n(\alpha)} - A_{F_n})(h_E) - A_{F_n(\alpha)} I_{F_n(\alpha)}^{-1}(I_{F_n(\alpha)} - I_{F_n})(h_E). \end{aligned} \quad (15)$$

By telescoping, the first term is written as a sum of two differences. The first difference is given by:

$$\frac{1 - \Delta(Y)}{F_n(Y)} \int_Y h_E(x) d(F_n(\alpha) - F_n)(x) \quad (16)$$

while the second difference is very similar. The other terms are very similar and can be dealt with in the same manner as term (16).

Consider term (16). Firstly, note that $F_n(\alpha) - F_n = \alpha(F_n - F)$. Secondly, if assumption 1 holds, then we know that $F_n, F \in \mathcal{F}(\delta)$ for some $\delta > 0$: note that n is fixed here and thus for $\delta = 1/n$ we have $F_n \in \mathcal{F}(\delta)$. Then by lemma 4.3 $I_{F_n(\alpha)}^{-1}$ has a bounded supremum norm uniformly in $\alpha \in [0, 1]$, assuming that $\{x_1, \dots, x_m\} \subset K$. As a consequence, h_E and $h_{E, \alpha}$ have a supremum norm bounded by a $M < \infty$ independent of $E \in \mathcal{E}$ and $\alpha \in [0, 1]$. Thus (16) is bounded by α times

$$\frac{M}{\delta} \left\{ \int_Y dF_n + \int_Y dF \right\} \text{ with probability tending to 1.}$$

Since this term is bounded by $2M/\delta$ this proves that (16) converges uniformly to zero when $\alpha \rightarrow 0$. This proves the following lemma

Lemma 4.4 *Let $F_n(\alpha) = (1 - \alpha)F_n + \alpha F$ and $\{x_1, \dots, x_m\} \in K$ (which is true with probability converging to 1). If assumption 1 holds, then*

$$\int | \tilde{I}(F_n(\alpha), t) - \tilde{I}(F_n, t) | dP_{F, G} \rightarrow 0 \text{ if } \alpha \rightarrow 0. \quad (17)$$

4.4 The P-Donsker class condition of theorem 3.2.

We will now prove the sufficient condition for assumption 3 of theorem 3.2. Lemma 4.3 tells us that, with probability tending to 1, for each $E \in \mathcal{E}$ $h_E^n = I_{F_n}^{-1}(I_E)$ exists pointwise, $M \equiv \sup_{E,n} \|h_E^n\|_\infty < \infty$ and we know that it solves

$$h_E^n(x) = C_n(h_E^n, I_E)(x) = \frac{1}{\pi_G(x)} \left(I_E(x) - \int_{\Delta=0} A_{F_n}(h_E^n)(y) dG(y | X = x) \right). \quad (18)$$

Define

$$\mathcal{G}(F) = \left\{ \frac{1}{\pi_G(x)} \left(I_E(\cdot) - \int_{\Delta=0} A_F(g)(y) dG(y | X = \cdot) \right) : \|g\|_\infty < 1, E \in \mathcal{E} \right\}.$$

Thus $h_n^E/M \in \mathcal{G}(F_n)$ with probability tending to 1 so that $\tilde{I}(F_n, E)/M \in A_{F_n}(\mathcal{G}(F_n))$. Since $F_n \in \mathcal{F}(\delta)$ with probability tending to 1 this implies that

$$\tilde{I}(F_n, E)/M \in \cup_{F \in \mathcal{F}(\delta)} A_F(\mathcal{G}(F)) \text{ with probability tending to 1.}$$

By the sufficient condition for assumption 3 of theorem 3.2 the right-hand side is a P_F -Donsker class. This proves the following lemma

Lemma 4.5 *If assumptions 1-2 and the sufficient condition of theorem 3.2 hold, then $\{\tilde{I}(F_n, E) : E \in \mathcal{E}\}$ falls in a P_F -Donsker class with probability tending to 1.*

4.5 The P-Donsker class condition of theorem 3.1.

Let $\mathcal{X} = \mathbb{R}^k$ for certain $k \in \mathbb{N}$ and let $\mathcal{E} = \{(-\infty, t] : t \in \mathbb{R}^k\}$. In Gill, van der Laan and Wellner [11] and van der Laan ([17], example 1.2) it is proved that the the class of multivariate cadlag functions with uniform sectional variation smaller than $M < \infty$ is a Donsker class. As above, let K be a support of F with $\{x_1, \dots, x_m\} \subset K$.

We showed in the outline of the proof of theorem 3.1 that for proving the Donsker class conditions we can assume that π_G and $x \rightarrow \int_{\Delta=0} h(y)/F(y) dG(y | x)$ are cadlag on K and that the uniform sectional variation requirements of Assumption 3* hold for these functions themselves (instead of for their versions). Suppose that $\|x \rightarrow \pi_G(x)\|_v^* < \infty$ which is guaranteed by assumption 3*. Let $f : \mathbb{R}^k \rightarrow \mathbb{R}$ be a cadlag function with $\|f\|_v^* < \infty$ and consider the equation $I_{F_n}(h^n)(x) = f(x)$, $x \in K$, or equivalently $h^n(x) = C_{F_n}(h^n, f)(x)$, $x \in K$. Since f is cadlag and π_G and $x \rightarrow \int_{\Delta=0} h(y) dG(y | x)$ are cadlag it follows that h^n is cadlag as well. Another fact is that if f is a multivariate cadlag function with $f > \delta > 0$, then $\|1/f\|_v^* \leq M \|f\|_v^*$ for some $M < \infty$ which does not depend on f (Gill [10]). Since the uniform sectional variation norm of π_G is bounded this fact implies that the uniform sectional variation norm of h^n over K is bounded by a constant times the uniform sectional variation norm of f and $x \rightarrow \int_{\Delta=0} \int_y h^n dF_n/F_n(y) dG(y | x)$, where we know that $F_n(y) > \delta > 0$ and $\|h^n\|_\infty \leq M < \infty$ for some $\delta > 0$, $M < \infty$, with probability tending to 1. The latter is bounded by the second condition in assumption 3*.

We showed that:

Lemma 4.6 *Assume $\{x_1, \dots, x_m\} \subset K$. Consider the case where $\mathcal{X} = \mathbb{R}^k$. If assumptions 1, 2 and condition (1), (2) of assumption 3* hold for the functions themselves (instead of F -a.e.), then for any multivariate cadlag function f we have that $I_{F_n}^{-1}(f)$ is cadlag with $\|I_{F_n}^{-1}(f)\|_\infty < M \|f\|_\infty$ and $\|I_{F_n}^{-1}(f)\|_v^* < M \|f\|_v^*$, where $M < C \inf_{\{y:\Delta(y)=0\}} 1/F_n(Y)$ for a universal constant C and the supremum and uniform sectional variation norm are taken over K .*

Consequently, since $\|I_{(-\infty, t]}\|_v^* = 2^k$ this tells us that $\|I_{F_n}^{-1}(I_{(-\infty, t]})\|_v^* < M$ with probability tending to 1, as required. Now, we want to prove that with $h_t^n \equiv I_{F_n}^{-1}(I_{(-\infty, t]})$

$$A_{F_n}(h_t^n) = h_t^n(x)\Delta + \frac{\int_y h_t^n(u)dF_n(u)}{F_n(y)}(1 - \Delta)$$

falls in a P_F -Donsker class with probability tending to 1, where we can use by lemma 4.6 that $\|h_t^n\|_v^* < M < \infty$. We already established that h_t^n falls in the Donsker class of multivariate cadlag functions with uniform sectional variation norm bounded by some universal constant. Therefore we only need to show the Donsker condition for the $1 - \Delta$ -term. Thus it suffices to assume that

$$\left\{y \rightarrow \frac{\int_y h(u)dF(u)}{F(y)}(1 - \Delta) : \|h\|_v^* < \infty, F \in \mathcal{F}(\delta)\right\}$$

is a P_0 -Donsker class, where $\mathcal{F}(\delta) = \{F : \inf_{\{Y:\Delta(Y)=0\}} F(Y) > \delta > 0\}$, which is guaranteed by (11) in assumption 3*. We can now state the following lemma:

Lemma 4.7 *Consider the case where $\mathcal{X} = \mathbb{R}^k$. If assumptions 1, 2 and 3* hold, then there exists a P_F Donsker class $\mathcal{G} \subset L^2(P_F)$ such that*

$$\{\tilde{I}(F_n, t) : t \in \mathbb{R}^k\} \subset \mathcal{G} \text{ with probability tending to 1.}$$

4.6 The P-consistency condition for theorem 3.1.

In the general theorem 3.2 we just assumed the P_F -consistency condition. In this subsection we will show that for the case $\mathcal{X} = \mathbb{R}^k$ the P_F -consistency condition holds under assumption 1,2 and 3*.

Denote $h_{nE} = I_{F_n}^{-1}(I_E)$ and $h_E = I_F^{-1}(I_E)$, as defined in the invertibility subsection. As in (15) we have

$$A_{F_n}I_{F_n}^{-1}(I_E) - A_F I_F^{-1}(I_E) = (A_{F_n} - A_F)(h_E) - A_{F_n}I_{F_n}^{-1}(I_{F_n} - I_F)(h_E).$$

To verify the P_F -consistency condition one needs to show that the P_F -norm of these terms converges to zero in probability, uniformly in $E \in \mathcal{E}$. Recall the score operator A_F and that the denominator $F_n(y)$ is uniformly bounded away from zero by lemma 4.2. By telescoping the first term, it is written as a sum of two differences. The first difference is given by:

$$\int \left(\frac{1 - \Delta(Y)}{F(Y)} \int_Y h_E(x)d(F_n - F)(x) \right)^2 dP_F(Y) \quad (19)$$

while the second difference is very similar. The other terms are very similar and can be dealt with in the same manner as term (19).

Consider now the case where $\mathcal{X} = \mathbb{R}^k$. The term $\int h_t d(F_n - F)$ can be bounded by integration by parts by $C \|F_n - F\|_\infty \|h_t\|_v^*$ for some constant $C < \infty$ (see Gill, van der Laan and Wellner [11]). By lemma 4.6 we know that h_t is of bounded uniform sectional variation. Consequently, the P_F -consistency condition follows from assumptions 1, 2 and 3*.

Lemma 4.8 *Consider the case that $\mathcal{X} = \mathbb{R}^k$ and $\mathcal{E} = \{(-\infty, t] : t \in \mathbb{R}^k\}$. If assumption 1,2 and 3* hold, then*

$$\sup_{t \in \mathbb{R}^k} \|\tilde{I}(F_n, t) - \tilde{I}(F, t)\|_{P_F} \rightarrow 0 \text{ a.s.}$$

5 Examples.

In the next subsection we verify the conditions of theorem 3.1 for a number of examples, possibly after some data reduction. Subsequently, we will summarize the results obtained.

5.1 Verification of the assumptions of theorem 3.1.

We will refer to the first, second and third condition in assumption 3* as assumption 3.1, assumption 3.2 and assumption 3.3, respectively.

Example 5.1 Univariate censoring.

Model. We have n i.i.d. copies $X_i \in \mathbb{R}_{\geq 0}$ of $X \sim F$, where F is completely unknown. We have n i.i.d. copies $C_i \in \mathbb{R}_{\geq 0}$ of $C \sim G$, where G is completely unknown. X and C are independent. Denote the survival functions of F and G by S and H , respectively. We observe:

$$W_i = (Z_i, D_i) = \Phi(X_i, C_i) = (\min(X_i, C_i), \Delta_i = I(X_i < C_i)), \quad i = 1, \dots, n.$$

It is well known and easily verified that the coarsening $\{X\}$ if $\Delta = 1$ and $[C, \infty)$ if $\Delta = 0$ satisfies (1), i.e. that it is a coarsening at random.

Assumption 1. $P(\Delta = 1 \mid X = x) = H(x)$. So assumption 1 requires that $H(x) > \delta > 0$ F a.e. Furthermore, it requires that $S(z) > \delta > 0$ for all possible censored z .

How to arrange assumption 1? Fix $\tau < \infty$ so that $S(\tau) > \delta > 0$ and $H(\tau) > \delta > 0$. Make each observation $z > \tau$ uncensored at τ . This does not influence the NPMLE on $[0, \tau)$; by the EM-algorithm (as explained in the next section) we know that all uncensored and right-censored observations after τ put only mass on (τ, ∞) . Then these truncated observations have distribution P_{F^δ} , where F^δ equals F on $[0, \tau)$, but has an atom at τ so that $F^\delta(\tau) = 1$. Now, $S^\delta(z) > \delta$ for all censored $(z, 0)$ and $H(\tau) > \delta > 0$ and thereby assumption 1 is satisfied for the truncated data.

Assumption 2. This requires that the set $\{x \rightarrow I_{[C, \infty)}(x) : C\}$ of indicator functions is a Glivenko-Cantelli class which is a well known result (van der Vaart, Wellner [29]).

Assumption 3.1 We have that $\pi_G(t) = H(t) = 1 - G(t)$. So assumption 3.1 requires that G is cadlag and of bounded variation which is trivially satisfied since G is a distribution function.

Assumption 3.2. We need to show that

$$T \rightarrow \int_{(0, T]} \frac{\int_{[z, \infty)} h dF}{S(z)} dG(z)$$

is cadlag F -a.e. and has a uniformly (in F with $S > \delta$ and $\|h\|_\infty \leq 1$) bounded uniform sectional variation norm. If f is uniformly bounded and G is a cumulative distribution function, then $t \rightarrow \int_{(0, t]} f(s) dG(s)$ is cadlag and its variation is bounded by the supremum norm of f .

Assumption 3.3. We have

$$A_F(h)(z, \delta = 0) = \frac{\int_{[z, \infty)} h dF}{S(z)}.$$

We assume that for each discontinuity point x of F $P(C = x) = 0$. Then the random variable $A_F(h)(Z, \Delta = 0)$ equals the random variable $\frac{\int_{(Z, \infty)} h dF}{S(Z)}$ since either $\int_{(z, \infty)} h dF = \int_{[z, \infty)} h dF$ (if $F(\{z\}) = 0$) or if this actually changes the integral at z , then the change has only an effect on

the outcome of the random variable if $C = z$ which happens with probability zero. We conclude that for proving the Donsker class condition we can replace $A_F(h)(Z, \delta = 0)$ by its version:

$$A_F(h)(z, \delta = 0) = \frac{\int_{(z, \infty)} hdF}{S(z)}.$$

Now, if f is uniformly bounded and F is a cumulative distribution function, then $t \rightarrow \int_{(t, \infty]} f(s)dF(s)$ is cadlag and its variation is bounded by the supremum norm of f . Thus $\int_{(z, \infty)} hdF$ is cadlag and its variation is uniformly bounded in the supremum norm of h . We arranged that $S(z) > \delta$ and therefore the denominator is uniformly bounded away from zero. It follows that the variation of $A_F(h)(z, 0)$ is uniformly (in F and h) bounded. This proves assumption 3.3.

Example 5.2 Double censoring.

Model. We have n i.i.d. copies X_i of $X \sim F_X$, F_X unknown. We have n i.i.d. copies (C_{1i}, C_{2i}) of $(C_1, C_2) \sim G_{C_1, C_2}$ unknown, with $P(C_2 > C_1) = 1$. Let $C_1 \sim G_1$ and $C_2 \sim G_2$. We assume that X and (C_1, C_2) are independent. Let $W = \min(\max(C_1, X), C_2)$ and $D = 1$ if $W = C_1$ and $X > C_1$, $D = 2$ if $W = X$ and $C_2 > X$ and $D = 3$ if $W = C_2$. We observe: (W, D) . So if $C_1 \leq X < C_2$, then X is completely observed and if $X \geq C_2$, then X is right censored at C_2 and if $X < C_1$, then X is left censored at C_1 . The coarsening at random of X implied by the observation is $[0, C_1]$ if $D = 1$, $\{X\}$ if $D = 2$ and $[C_2, \infty)$ if $D = 3$.

Assumption 1. $P(\Delta = 1 \mid X = x) = G_{C_1, C_2}([0, x] \times (x, \infty)) > \delta > 0$. Furthermore, assumption 1 requires $F_X(C_1) > \delta > 0$ for all C_1 with $D = 1$ and $S_X(C_2) > \delta > 0$ for all C_2 with $D = 3$.

How to arrange assumption 1? We assume that for a $\tau < \infty$ with $F_X(\tau) < 1$, we have $G_1(\tau) = 1$ and $G_2(\tau) > 0$. This means that after τ we only have uncensored ($d = 2$) and right-censored ($d = 3$) observations. Then as in the univariate censoring model we can make all observations after τ uncensored and by the same reason this does not influence the NPMLE on $[0, \tau)$. Then these truncated observations have distribution $P_{F_X^\delta}$ where F_X^δ equals F_X on $[0, \tau)$, but has an atom at τ so that $F_X^\delta(\tau) = 1$. Now, $S_X^\delta(C_2) > \delta$ for all C_2 which right-censor X .

Now, we assume that there exists a δ_1 with $F_X(\delta_1) > 0$ so that $G_1[0, \delta_1] = G_1(\{0\}) > 0$, i.e., it has an atom at 0 and no mass immediately after this atom. Notice that if $C_1 = 0$, then $D \neq 1$ with probability 1. Thus for C_1 with $D = 1$ we have $F_X(C_1) \geq F_X(\delta_1) > 0$. We also have

$$G_{C_1, C_2}(C_1 \in [0, w], C_2 \in (w, \infty)) \geq G_1(\{0\}) > \delta.$$

This verifies assumption 1 for the truncated data under the mentioned assumptions.

To summarize: by assuming

(i) There exists a $\tau < \infty$ for which $G_1(\tau) = 1$, $\bar{G}_2(\tau) > 0$ and $F_X(\tau) < 1$.

(ii): $G_1[0, \delta_1] = G_1(\{0\}) > 0$ and $F_X(\delta_1) > 0$ for certain $\delta_1 > 0$,

assumption 1 holds for data artificially truncated at τ . The assumptions (i) and (ii) are the same as the assumptions used in Chang [4] for proving asymptotic normality. Gu and Zhang [13] succeeded, by a specific analysis, to weaken these conditions.

Assumption 2. This requires that the sets $\{x \rightarrow I_{(0, C_1)}(x) : C_1\}$ and $\{x \rightarrow I_{[C_2, \infty)}(x) : C_2\}$ of indicator functions are Glivenko-Cantelli classes which is a well known fact.

Assumption 3.1. Here $\pi_G(t) = G_{C_1, C_2}((0, t] \times (t, \infty)) = \bar{G}_2(t) - \bar{G}_1(t)$ is a difference of two survival functions. Thus π_G is cadlag and of bounded variation.

Assumption 3.2 We need to show that

$$T \rightarrow \int_{(0, T]} \frac{\int_{c_2}^{\infty} hdF}{S(c_2)} dG_2(c_2) + \int_{(T, \infty)} \frac{\int_0^{c_1} hdF}{S(c_1)} dG_1(c_1)$$

is cadlag and has a uniformly (in F with $S > \delta$ and $\|h\|_\infty \leq 1$) bounded uniform sectional variation norm. This is proved as in the univariate censoring example.

Assumption 3.3.

$$A_F(h)(w, d) = h(w)I(d=1) + I(d=2) \frac{\int_{(w, \infty)} hdF_X}{F_x(w)} + I(d=3) \frac{\int_{(0, w)} hdF_X}{F_X(w)}.$$

If we assume that for each discontinuity point x of F_X $P(C_j = x) = 0$, $j = 1, 2$, then the same proof as in the univariate censoring example can be applied.

Example 5.3 Rectangle-censored data in the plane.

Let $[A, B] = [A_1, B_1] \times [A_2, B_2]$ be a random rectangle in the plane; we will denote the distribution of (A, B) with G . Let $X \sim F$ be a $\mathbb{R}_{\geq 0}^2$ -valued random variable of interest. It is assumed that X is independent of (A, B) . We observe the following coarsening of Y of X : $Y = \{X\}$ if $X \in [A, B)$ and $Y = [A, B)^c$ (complement) if $X \notin [A, B)$. We are concerned with estimation of the distribution F of X .

Assumption 1. $\pi_G(x) = P(x \in [A, B) \mid X = x) = P(A \leq x, B > x) = G((0, x] \times (x, \infty)) = \bar{G}_B(x) - \bar{G}_A(x) > \delta > 0$ F -a.e.

Furthermore, assumption 1 requires that $F([A, B)^c) = 1 - F([A, B]) > \delta > 0$ G -a.e.

Discussion of assumption 1. Suppose that F has compact support $[0, \tau] \subset \mathbb{R}_{\geq 0}^2$ with $F\{\tau\} > 0$ and $F\{0\} > 0$. Assume also that $P(A = 0, B > \tau) > 0$. The latter assumption implies $\pi_G(x) \geq P(A = 0, B > \tau) > 0$. Consider a censored $[A, B)$. If $A > 0$, then $F([A, B)^c) > F(\{0\}) > 0$, if $B < \tau$, then $F([A, B)^c) > F(\{\tau\}) > 0$ while if $A = 0$ and $B \geq \tau$, then $X \in [A, B)$ so that the observation is uncensored. This proves assumption 1.

Assumption 2. Note that $[A, B)^c$ is a union of 4 rectangles. Thus assumption 2 only requires that the indicators of rectangles in the plane constitute a Glivenko-Cantelli class. This is a well known empirical process result (van der Vaart, Wellner [29]).

Assumption 3.1 Here $\pi_G(x) = P(B > x) - P(A > x) = \bar{G}_B(x) - \bar{G}_A(x)$ is a difference of two survival functions. Thus π_G is bivariate cadlag and of bounded uniform sectional variation.

Assumption 3.2. We need to show that

$$x \rightarrow \int_{R(x)} \frac{\int_{[a, b)^c} hdF}{1 - F([a, b))} dG(a, b),$$

where $R(x) = \{(a, b) \in \mathbb{R}^4 : x \notin (a, b)\}$, is cadlag F -a.e. and that it (i.e. its cadlag version) has a uniformly (in F with $1 - F([a, b)) > \delta$ and $\|h\|_\infty \leq 1$) bounded uniform sectional variation norm. We have that $R(x) = \cup_{i=1}^4 E_i(x)$, where $E_1(x) = \{A_1 > x_1\}$, $E_2(x) = \{A_2 > x_2\}$, $E_3(x) = \{B_1 \leq x_1\}$ and $E_4(x) = \{B_2 \leq x_2\}$. By writing this union as a union of disjoint regions we obtain:

$$R(x) = \cup_{i=1}^4 E'_i(x),$$

where

$$\begin{aligned} E'_1(x) &\equiv E_1(x) \\ E'_2(x) &\equiv \{A_2 > x_2, A_1 \leq x_1\} \\ E'_3(x) &\equiv \{B_1 \leq x_1, A_1 \leq x_1, A_2 \leq x_2\} \\ E'_4(x) &\equiv \{B_2 \leq x_2, B_1 > x_1, A_1 \leq x_1, A_2 \leq x_2\} \end{aligned}$$

Thus we can represent the integral over $R(x)$ as a sum of four integrals $\int_{E'_i(x)} \frac{\int_{[a,b]^c} h dF}{1-F([a,b])} dG(a, b)$, $i = 1, \dots, 4$. These integrals are of the form $x \rightarrow \int_{E'_i(x)} f(a, b) dG(a, b)$, where $\|f\|_\infty \leq 1$. Since each of the four components is integrated over intervals $[0, x_j]$ or (x_j, ∞) , $j = 1, 2$, such functions are cadlag and have a uniform sectional variation norm bounded by a constant times the supremum norm of f . This proves assumption 3.2.

Assumption 3.3. Here we need to show that the four-variate functions (using that for $h \in L_0^2(F)$ $\int_{[a,b]^c} h dF = -\int_{[a,b]} h dF$)

$$(a, b) \rightarrow \frac{\int_{[a,b]} h(x) dF(x)}{1 - F([a, b])} \quad (20)$$

fall in a Donsker class, where we can use that $\|h\|_v^* < 1$ and $1 - F([a, b]) > \delta > 0$. Assume that if x is an atom of F , then $P(A = x) = P(B = x) = 0$. In that case, for proving the Donsker class condition we can replace $\int_{[a,b]} h(x) dF(x) / (1 - F([a, b]))$ by its version

$$f(a, b) = \frac{\int_{(a,b]} h(x) dF(x)}{1 - F((a, b])}.$$

The latter function is four-variate cadlag. In addition, the denominator $(a, b) \rightarrow F((a, b])$ has uniform sectional variation bounded by a universal constant and it is known to be bounded away from 0 by $\delta > 0$. Similarly, $(a, b) \rightarrow \int_{(a,b]} h(x) dF(x)$ has a uniform sectional variation norm bounded by a universal constant times the supremum norm of h , which is smaller than 1. It is left to the reader to verify that the property of bounded sectional variation carries over, in the present case, to the ratio of these two functions. Here techniques of Gill, van der Laan and Wellner [11] are useful. Thus $(a, b) \rightarrow f(a, b)$ is cadlag and has a uniform sectional variation norm bounded uniformly in h with $\|h\|_\infty \leq 1$ and $F \in \mathcal{F}(\delta)$.

Example 5.4 A mixture of right-censored and current status data.

Consider a study where at recruitment of a subject the subject is either monitored once or the subject is selected to be followed up. Let T be the (e.g. survival) time variable of interest. Let C be the monitoring time or right-censoring time which is always observed and let $\xi \in \{0, 1\}$ be the indicator for the selected sample. We assume that (C, ξ) is independent of T and we denote the subdistributions of $(C, 0)$ and $(C, 1)$ with G_0 and G_1 , respectively; here G_0 represents the distribution of the monitoring times at entrance in the study while G_1 represents the right-censoring distribution (e.g. due to the end of the experiment). We observe (C, ξ) and if $\xi = 1$ (i.e. the subjects belongs to the selected sample) we observe $(T \wedge C, \Delta = I(T < C))$ and if $\xi = 0$ we observe Δ only. The distribution of the data Y is given by:

$$\begin{aligned} P(C \in dc, \xi = 1, T \in dt, \Delta = 1) &= G_1(dc)F(dt) \\ P(C \in dc, \xi = 1, \Delta = 0) &= G_1(dc)S(c) \\ P(C \in dc, \xi = 0, \Delta = 0) &= G_0(dc)S(c) \\ P(C \in dc, \xi = 0, \Delta = 1) &= G_0(dc)F(c). \end{aligned}$$

Without any loss of information for F we can pool together the 2 types of observations corresponding with $\Delta = 0$. We define a new δ to indicate the remaining three types of observations: let $\delta = 1$ if $\xi = 1, \Delta = 1$, let $\delta = 2$ if $\Delta = 0$ and let $\delta = 3$ if $\xi = 0, \Delta = 1$. Let $G(c) = G_0(c) + G_1(c)$. Then the distribution of the data is given by:

$$I(\delta = 1)F(dt)G_1(dc) + I(\delta = 2)S(c)G(dc) + I(\delta = 3)F(c)G_0(dc).$$

We are concerned with proving efficiency of the NPMLE of $F(t)$.

Assumption 1. $\pi_G(t) = P(\xi = 1, \Delta = 1 \mid T = t) = \bar{G}_1(t) > \delta > 0$. Furthermore, assumption 1 requires that $S(c) > \delta > 0$ G -a.e. and $F(c) > \delta > 0$ G_0 a.e.

How to arrange assumption 1? This is similar to the doubly-censored data model: replace here G_1 by G_0 , G_2 by G_1 . By assuming

(i) There exists a $\tau < \infty$ for which $G_0(\tau) = 1$, $\bar{G}_1(\tau) > 0$ and $(\tau) < 1$.

(ii): $G_0[0, \delta_1] = G_0(\{0\}) > 0$ and $F(\delta_1) > 0$ for certain $\delta_1 > 0$,

assumption 1 holds for artificially truncated data at τ in the sense that we make every observation with $\delta = 1$ and $T > \tau$ and every observation with $\delta = 2$ and $C > \tau$ uncensored at τ .

Assumption 2. This requires that the class of indicator functions $\{x \rightarrow I_{(0,C)}(x) : C\}$ and $\{x \rightarrow I_{[C,\infty)}(x) : C\}$ form a Glivenko-Cantelli class which is well known.

Assumption 3.1. Here $\pi_G(t) = \bar{G}_1(t)$ is cadlag and of bounded variation.

Assumption 3.2. and Assumption 3.3. This proof is the same as given in the doubly-censored data example, where at assumption 3.3 we need to assume that if $P(T = t) > 0$, then $G_0(\{t\}) = G_1(\{t\}) = 0$.

5.2 Results for examples.

We applied theorem 3.1 to the examples and obtained the following results.

Result 5.1 Univariate Censoring.

Let $[0, \tau] \subset \mathbb{R}_{\geq 0}$ be an interval such that $H(\tau) > 0$ and $S(\tau) > 0$. We assume that for each discontinuity point x of F $P(C = x) = 0$. Then F_n is supremum norm asymptotically efficient on $[0, \tau]$.

Result 5.2 Double Censoring.

We assume that for each discontinuity point x of F_X $P(C_j = x) = 0$, $j = 1, 2$. If $G((0, x] \times (x, \infty)) > \delta > 0$ F_X -a.e. and $F_X(c_1) > \delta > 0$ G_1 a.e. and $S_X(c_2) > \delta > 0$ G_2 -a.e., then F_n is supremum norm asymptotically efficient.

In particular, let $[0, \tau] \subset \mathbb{R}_{\geq 0}$ be an interval such that $G_1(\tau) = 1$, $\bar{G}_2(\tau) > 0$. Furthermore, assume that there exists a $\delta > 0$ so that $G_1(\delta) = G_1(\{0\}) > 0$ and $F_X(\delta_1) > 0$. Then (still assuming the discontinuity condition) F_n is supremum norm asymptotically efficient on $[0, \tau]$.

This assumption would be satisfied if F has compact support $[0, \tau]$ while the censoring survival function is bounded away from zero on $[0, \tau]$. In case F does not have compact support we described in the example above a reduction of the data so that the assumptions are met for the reduced data. Now, the NPMLE based on the reduced data is asymptotically efficient for the reduced data (and the loss of efficiency is small if the reduction is small).

Result 5.3 Rectangle censoring.

Assume that $\bar{G}_B(x) - \bar{G}_A(x) > \delta > 0$ for F -almost every x and that $F([A, B]^c) > \delta > 0$ for P_0 -almost every censored observation $[A, B]^c$. A sufficient condition for this is that F has compact support $[0, \tau] \subset \mathbb{R}^2$, $F\{0\} > 0$, $F(\{\tau\}) > 0$, $P(A = 0, B > \tau) > 0$. In addition, assume that if $F(\{x\}) > 0$, then $P(A = x) = P(B = x) = 0$.

Then F_n is supremum-norm efficient.

Result 5.4 Mixture of right-censored and current status data.

Assume that if for a point t $P(T = t) > 0$, then $G_0(\{t\}) = G_1(\{t\}) = 0$. If $\bar{G}_1 > \delta > 0$ F -a.e., $S(c) > \delta > 0$ G -a.e. and $F(c) > \delta > 0$ G_0 -a.e., then F_n is supremum norm asymptotically efficient.

In particular we have the following. Let $[0, \tau] \subset \mathbb{R}_{\geq 0}$ be an interval such that $G_0(\tau) = 1$, $G_1(\tau) > 0$. Furthermore, assume that there exists a $\delta > 0$ so that $G_0(\delta) = G_0(\{0\}) > 0$ and $F(\delta_1) > 0$. Then (still assuming the discontinuity condition) F_n is supremum norm asymptotically efficient on $[0, \tau]$.

Appendix A: Existence and uniqueness of NPMLE and uniqueness of solution of self-consistency equation.

Recall that we have introduced a set of points x_1, \dots, x_m such that each observed region Y_1, \dots, Y_n contains at least one point x_i , and each point x_i is the *only* such point in some region Y_j . Define $p_i = F(x_i)$ and $P_j = F(Y_j)$ where F is a probability measure on x_1, \dots, x_m . We want to show existence and uniqueness of the maximizer over F of $\sum \log P_j$.

We can rephrase the problem as maximization of $\sum \log P_j$ over $P_j \geq 0$ such that there exist $p_i \geq 0$ with $\sum p_i = 1$, $P_j = \sum_{i: x_i \in Y_j} p_i$. The target function is continuous and strictly concave, taking values in $[-\infty, \infty)$, and defined on a convex, compact set. It takes finite values on the (non-empty) relative interior of its domain. Hence its supremum is uniquely achieved. Since for each i there is a j with $P_j = p_i$ the maximizer is also unique in terms of the original variables p_i .

In terms of the p_i and introducing a Lagrange multiplier λ for the constraint $\sum p_i = 1$ we know, by differentiating $\sum \log P_j - \lambda \sum p_i$ with respect to p_i , that if λ and $p_i \geq 0$ satisfy

$$\sum_{j: x_i \in Y_j} 1/P_j - \lambda = 0, \quad (21)$$

for each i , and $\sum p_i = 1$, then the p_i are the unique solution of our optimization problem. Multiplying (21) by p_i and adding over i we find that $\lambda = n$. But (21) with $\lambda = n$ is nothing but the self-consistency equation (6). Hence these equations have a unique solution.

Appendix B: Identifiability of the self-consistency equation.

Let \mathcal{F}_0 be the set of all distributions on \mathcal{X} which are equivalent with F_0 and for which $\|dF_0/dF\|_{\infty} < \infty$.

Theorem 5.1 *Suppose that for a $F \equiv F_0$ with $\|dF_0/dF\|_{\infty} < \infty$*

$$P_{F_0} A_F(h - Fh) = 0 \text{ for all } h \text{ with } \|h\|_{\infty} < \infty.$$

Then $F = F_0$.

Proof. Let μ dominate F_0 and denote the density $dF/d\mu$ by f . Assume $F \neq F_0$, $F \equiv F_0$, $\|dF_0/dF\|_{\infty} < M$ and $P_{F_0}(A_F(h - Fh)) = 0$ for all h with $\|h\|_{\infty} < \infty$. We want to get a contradiction (then we have to conclude that $F = F_0$). Set $h = (f_0 - f)/f$ and notice that now $Fh = 0$ and $\|h\|_{\infty} < \infty$. Define for this h a function $\Phi_h : I \subset \mathbb{R} \rightarrow \mathbb{R}$ on a closed interval I around zero given by:

$$\Phi_h(\epsilon) = \int \log(p_{F_{\epsilon, h}}) dP_{F_0},$$

$F_{\epsilon, h}$ is a line through F with score h ; in terms of densities it is given by $f_{\epsilon, h} = (1 + \epsilon h)f$. $P_{F_0}(A_F(h)) = 0$ tells us that $\frac{d}{d\epsilon} \Phi_h(\epsilon) |_{\epsilon=0} = 0$. By (2) we have $p_F \neq p_{F_0}$ P_F a.e. Now, by linearity of $f \rightarrow p_f$, the strict concavity of ‘log’, and the latter fact we have:

$$\begin{aligned}
\Phi_h(\epsilon) &= \int \log \left(p_{(1+\epsilon(f_0-f)/f)f} \right) dP_{F_0,0} \\
&= \int \log \left((1-\epsilon)p_f + \epsilon p_{f_0} \right) dP_{F_0,0} \\
&> (1-\epsilon) \int \log(p_f) dP_{F_0,0} + \epsilon \int \log(p_0) dP_{F_0,0} \\
&= (1-\epsilon)\Phi_h(0) + \epsilon \int \log(p_0) dP_{F_0,0}
\end{aligned}$$

and hence

$$\Phi_h(\epsilon) - \Phi_h(0) > \epsilon \left(\int \log(p_0) dP_{F_0,0} - \int \log(p_f) dP_{F_0,0} \right) = \epsilon\delta,$$

where by the Jensen inequality $\delta > 0$, using that $p_f \neq p_0$ $P_{F_0,0}$ a.e. So

$$\frac{1}{\epsilon} (\Phi_h(\epsilon) - \Phi_h(0)) > \delta > 0,$$

which contradicts that $\frac{d}{d\epsilon} \Phi_h(\epsilon) |_{\epsilon=0} = 0$. \square

Appendix C: L2-invertibility of the information operator.

Lemma 5.1 *If for some $\delta > 0$ $\pi_G(x) > \delta > 0$ on the support of F , then $I_F : L^2(F) \rightarrow L^2(F)$ is onto and has a bounded inverse given by*

$$I_F^{-1} = \sum_{k=0}^{\infty} (I - I_F)^k,$$

where I denotes the identity operator. Moreover, the bound does not depend on F :

$$\| I_F^{-1}(h) \|_F \leq \frac{1}{\delta} \| h \|_F.$$

If $\pi_G(x) > 0$ on the support of F , then I_F is 1-1, but not necessarily onto.

Proof. For the onto and bounded invertibility of I_F and the expression for the inverse it suffices to show that $\| A_F(h) \|_{P_F} \geq \sqrt{\delta} \| h \|_F$ (see van der Laan [17]). We have

$$A_F(h)(Y) = \Delta h(X) + (1 - \Delta)E(h(X) | X \in Y).$$

Thus

$$\begin{aligned}
EA_F^2(h)(Y) &\geq E(h^2(X)\Delta) \\
&= E(h^2(X)E(\Delta | X)) \\
&= E(h^2(X)\pi_G(X)) \\
&\geq \delta E h^2(X). \square
\end{aligned}$$

References

- [1] P.K. Andersen, Ø. Borgan, R.D. Gill and N. Keiding (1993), *Statistical models based on counting processes*, Springer, New York.
- [2] P.J. Bickel, C.J. Klaassen, Y. Ritov and J.A. Wellner (1993), *Efficient and adaptive inference in semi-parametric models*, Johns Hopkins University Press, Baltimore.
- [3] Bickel and Ritov (1994), Efficient estimation using both direct and indirect observations, *Theory Probab. Appl.* **38** 194–213.
- [4] M.N. Chang (1990), Weak convergence of a self-consistent estimator of the survival function with doubly censored data, *Ann. Statist.* **18** 391–404.
- [5] M.N. Chang and G. Yang (1987), Strong consistency of a nonparametric estimator of the survival function with doubly censored data, *Ann. Statist.* **15** 1536–1547.
- [6] D.M. Dabrowska (1988), Kaplan Meier Estimate on the Plane, *Ann. Statist.* **16**, 1475–1489.
- [7] D.M. Dabrowska (1989), Kaplan Meier Estimate on the Plane: Weak Convergence, LIL, and the Bootstrap, *J. Multivar. Anal.* **29**, 308–325.
- [8] R.D. Gill (1983), Large sample behavior of the product-limit estimator on the whole line, *Ann. Statist.*, **11**, 49–58.
- [9] R.D. Gill (1989), Non-and Semi-parametric Maximum Likelihood Estimators and the von Mises Method (Part 1), *Scand. J. Statist.* **16**, 97–128.
- [10] R.D. Gill (1994), *Lectures on survival analysis*, pp. 115–241 in: *Lectures on Probability Theory (Ecole d't de Probabilits de Saint Flour XXII - 1992)*, D. Bakry, R.D. Gill and S.A. Molchanov, ed. P. Bernard, Springer Lecture Notes in Mathematics 1581, Springer-Verlag, Berlin.
- [11] R.D. Gill, M.J. van der Laan and J.A. Wellner (1995), Inefficient estimators of the bivariate survival function for three models, *Annales de L'I.H.P. Probabilites et Statistiques* **31**, 545–597.
- [12] R.D. Gill, M.J. van der Laan and J.M. Robins (1997), Coarsening at random: characterizations, conjectures and counter-examples, pp. 255–294 in: *Survival Analysis*, D.-Y. Lin and T.R. Fleming (eds), Springer Lecture Notes in Statistics 123.
- [13] M.G. Gu and C.H. Zhang (1993), Asymptotic properties of self-consistent estimators based on doubly censored data, *Ann. Math. Statist.* **21** 611–614.
- [14] D.F. Heitjan and D.B. Rubin (1991), Ignorability and coarse data, *Ann. Statist.* **19**, 2244–2253.
- [15] M. Jacobsen and N. Keiding (1995), Coarsening at random in general sample spaces and random censoring in continuous time, *Ann. Statist.* **23**, 774–786.
- [16] J. Kiefer and J. Wolfowitz (1956), Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Ann. Statist.*, **27**, 887–906.
- [17] M.J. van der Laan, M.J. (1996), *Efficient and Inefficient Estimation in Semiparametric Models*, CWI-tract # 114, Centre for Mathematics and Computer Science.
- [18] M.J. van der Laan (1996), Efficient estimator of the bivariate survival function and repairing NPMLE, *Ann. Statist.*, **24**, No. 2, 596–627.
- [19] M.J. van der Laan (1997), Nonparametric Estimators of the Bivariate Survival Function under Random Censoring. *Statistica Neerlandica* **51**, No. 2, 178–200.
- [20] G. Neuhaus (1971), On weak convergence of stochastic processes with multidimensional time parameter, *Ann. Math. Statist.* **42** 1285–1295.
- [21] D. Pollard (1990), *Empirical Processes: Theory and applications*, Regional conference series in probability and statistics **2**, Inst. Math. Statist., Hayward, California.
- [22] R.L. Prentice and J. Cai (1992), Covariance and survivor function estimation using censored

- multivariate failure time data. *Biometrika* **79**, 495–512.
- [23] R.L. Prentice and J. Cai (1992), 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.
- [24] 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.
- [25] B.W. Turnbull (1976), The empirical distribution with arbitrarily grouped censored and truncated data, *J. Roy. Statist. Soc. ser. B* **38**, 290–295.
- [26] A.W. van der Vaart (1988), *Statistical Estimation in Large Parameter Spaces.*, CWI tract **44**, Centre for Mathematics and Computer Science, Amsterdam.
- [27] A.W. van der Vaart (1991), Efficiency and Hadamard differentiable functionals, *Scand. J. Statist.* **18**, 63–75.
- [28] A.W. van der Vaart (1994), Maximum likelihood estimation with partially censored data, *Ann. Statist.* **22**, 1896–1916.
- [29] A.W. van der Vaart and J.A. Wellner (1996), *Weak convergence and empirical processes*, Springer Verlag.
- [30] J.A. Wellner (1982), Asymptotic optimality of the product limit estimator, *Ann. Statist.* **10**, 595–602.
- [31] C.F.J. Wu (1983), On the convergence of the EM-algorithm, *Ann. Stat.* **11** 95–103.