

Current Status and Right-Censored Data Structures When Observing a Marker at the Censoring Time.

Mark J. van der Laan and Nicholas P. Jewell

Division of Biostatistics

University of California

Berkeley, CA 94720

May 15, 2001

Abstract

We study nonparametric estimation with two types of data structures. In the first data structure n i.i.d. copies of $(C, N(C))$ are observed, where N is a counting process jumping at time-variables of interest and C a random monitoring time. In the second data structure n i.i.d. copies of $(C \wedge T, I(T \leq C), N(C \wedge T))$ are observed, where N is a counting process with a final jump at time T (e.g. death). This data structure includes observing right-censored data on T and a marker variable at the censoring time.

In these data structures, easy to compute estimators, namely (Weighted)-Pool-Adjacent-Violator estimators for the marginal distributions of the unobservable time variables, and the Kaplan-Meier estimator for the time T till the final observable event, are available. These estimators ignore seemingly important information in the data. In this paper we prove that, at most continuous data generating distributions with compact support, the ad hoc estimators yield asymptotically efficient estimators of \sqrt{n} -estimable parameters,.

Some key words: Asymptotically linear estimator, Asymptotically efficient estimator, Current

status data, Right-censored data, Isotonic regression.

1 Introduction.

In this paper we study nonparametric estimation with two types of data structures. Below we will discuss these two data structures in detail. Subsequently, we will provide an overview of this paper.

1.1 Current status data on a counting process.

Consider a counting process $N(t) = \sum_{j=1}^k I(T_j \leq t)$, $T_1 < \dots < T_k$, where T_j is the time-variable at which an event occurs and where N jumps from value $j - 1$ to j . We consider the data structure $(C, N(C))$ for a single random monitoring time C . The only assumption is that C is independent of N : the distribution G of C , and F of N are unspecified. Such data structures occur in cross-sectional studies where each subject is monitored once. Here $N(t)$ might be the number of times a person of age t has been married, the number of sexual partners a person of age t has had, or the number of children a married couple has with t being the number of years the couple has been married. We refer to Jewell and van der Laan (1995) for additional applications. In many situations it is simply not feasible to obtain more than one measurement on N ; for example, in a carcinogenicity experiment one can only determine a discretized occult tumor size measured by $N(t)$ at time t by sacrificing the mouse at time t .

The distribution of $(C, N(C))$ depends on the distribution of $\vec{T} = (T_1, \dots, T_k)$ only through the marginal distributions F_j of T_j , $j = 1, \dots, k$ (see section 2). In this problem, the NPMLE of the distribution of T_j requires an iterative algorithm. On the other hand, an ad hoc method for estimation of the distribution of T_j is directly available: reduce the observation $(C, N(C))$ to classical current status data $(C, \Delta_s = I(T_j \leq C))$ on T_j . This implies that one can esti-

mate the distribution of T_j with the NPMLE based on the reduced current status observations, which we will refer to as the reduced data NPMLE (RNPMLE). This estimator provides regular and asymptotically linear estimators of pathwise differentiable functionals of F_j such as $\mu_s = \int(1 - F_{T_s})(u)r(u)du$, for a given r , in the nonparametric model under certain conditions (Groeneboom and Wellner, 1992). Previous work and examples of classical current status data on a time variable T can be found in Diamond, McDonald and Shah (1986), Jewell and Shiboski (1990), Diamond and McDonald (1991), Keiding (1991) and Sun and Kalbfleisch (1993). In its nonparametric setting, it is also known as interval censoring, case I (Groeneboom and Wellner, 1992). Current status data commonly arise in epidemiological investigations of the natural history of disease and in animal tumorigenicity experiments. Jewell, Malani and Vittinghoff (1994) give two examples that arise from studies of Human Immunodeficiency Virus (HIV) disease.

Note that the RNPMLE of F_j ignores the value of $N(C)$, beyond information on whether $N(C) \geq j$ or not. For example, if $N(t)$ is tumor size in a carcinogenicity experiment, then the simple current status estimator of the distribution of time till onset of tumor would not distinguish between an observation $(C, N(C))$ with $N(C)$ is large and an observation $(C, N(C))$ with $N(C)$ small but larger than 0, while the latter observation seems to suggest that onset just occurred. Nonetheless, we will establish that the RNPMLE yields efficient estimators of pathwise differentiable parameters at a large class of continuous data generating distributions of interest.

1.2 Current status data on a counting process when the final event is right censored.

We also consider the data structure $(\tilde{T}_k \equiv C \wedge T_k, N(\tilde{T}_k))$ for a general counting process $N(t) = \sum_{j=1}^k I(T_j \leq t)$, where T_k represents the final event (say death) which is right-censored by

the monitoring time C . Note that this observation includes observing the failure indicator $I(\tilde{T}_k = T_k)$. For example, consider a carcinogenicity experiment with mice in which T_1 is time till onset of colon tumor, T_2 time to liver metastasis and T_3 time to death from tumor, where we assume that colon tumors do not cause death except through liver failure secondary to metastasis. Here C is either a sacrificing time or time till death from any unrelated cause.

Let's consider another example. Suppose that we are concerned with estimation of the survival function of the time $T = J - I$ between time I at seroconversion and time J at death of a hemophiliacs patient infected with the HIV-virus. For this purpose we observe n i.i.d. subjects in a fixed time-interval of 10 years. If we assume that the time I at seroconversion of the subject is observed (which is approximately true for hemophiliacs patients), then the subject's survival time T is right-censored by $C \equiv 10 - I$. Let $Z(t)$ be some monotone surrogate process which measures the progression of the disease of the subject t years after seroconversion; for example, $Z(t)$ may be viral load of the subject t years after seroconversion, where it is reasonable to assume that the viral load is a non-decreasing process in the absence of treatment. Suppose that for every subject who did not die before the end of the study C one measures the surrogate $Z(C)$ at time C only. In other words, we just have the failure times for the subjects who failed before end of follow up and for every subject who is alive at end of follow up we also have a marker indicating future prognosis. We again only assume that C is independent of T . By defining a counting process N which jumps each time when viral load achieves its next value (given a set of increasing values) and has a final jump at death, assuming that death always occurs after the highest viral load value measured by the counting process, we represent (i.e. approximate arbitrarily well) the data as $(\tilde{T} = T \wedge C, N(\tilde{T}))$. A seemingly ad hoc estimator of $S(t) = P(T > t)$ is the Kaplan-Meier estimator which simply ignores the marker information. In this example, a natural question is whether one can improve on the Kaplan-Meier estimator

using data on Z . In this paper we prove that the Kaplan-Meier estimator is asymptotically efficient at most compact continuous data generating distributions.

A special case of this data structure has been treated in the literature. Consider a carcinogenicity experiment with $N(t) = \sum_{j=1}^2 I(T_j \leq t)$, T_1 is time till onset of tumor and T_2 is time till death from tumor. Thus one observes $(\tilde{T}_2 \equiv C \wedge T_2, N(\tilde{T}_2))$. This data structure has been considered in Kodell, Shaw and Johnson (1982), Dinse and Lagakos (1982), Turnbull and Mitchell (1984), van der Laan, Jewell and Peterson (1997) and recently Groeneboom (1998). The NPMLE for this data structure requires an iterative algorithm: Turnbull and Mitchell (1984) implemented the NPMLE with the EM-algorithm while Groeneboom (1998) implements the NPMLE with a modern optimization algorithm. In this problem, an ad hoc estimator of the lifetime-distribution of T_2 is the Kaplan-Meier estimator based on the reduced data $(\tilde{T}_2, \Delta_2 = I(\tilde{T}_2 = T_2))$. In Dinse and Lagakos (1982), the Kaplan-Meier estimator of F_2 was proposed and it was suggested that the NPMLE might be more efficient than the Kaplan-Meier estimator. In van der Laan, Jewell and Peterson (1997) it is shown that the Kaplan-Meier is efficient under a weak condition on (F_1, F_2) , and an isotonic regression estimator of F_1 is provided: note that the Pool-Adjacent-Violator estimator for F_1 is not applicable anymore since for some subjects one only observes T_2 and thus that $T_1 < T_2$, where T_2 cannot be viewed as an independent monitoring time for T_1 . In van der Laan, Jewell and Peterson (1997) a simulation study was carried out which uses an incorrect algorithm (instead of EM) for the NPMLE thereby wrongly suggesting that Kaplan-Meier outperforms the NPMLE in practice (specifically the derivation of the score equations for the NPMLE is not valid since it assumes that the NPMLE \hat{F}_1 is strictly smaller than the NPMLE \hat{F}_2). Groeneboom (1998) provides a fast algorithm for computing the NPMLE and future work of Groeneboom will provide a valid simulation study comparing the NPMLE and the ad hoc estimators for root- n -estimable parameters and non-smooth parameters.

1.3 Organization and overview of results.

In section 2 we prove that if the F_j 's are continuous with Lebesgue density bounded away from zero on $[0, \tau_j]$ and zero elsewhere and G is continuous as well, then any estimator of a parameter $\mu = \Phi(F) \in \mathbb{R}$ which is regular and asymptotically linear at $P_{F,G}$ is also asymptotically efficient. It is also explained why the NPMLE is more complex and that it is, in fact, more efficient at many data generating distributions with F_j 's having different dominating measures: e.g. F_1 discrete and F_2 continuous.

In section 3 we prove the same result for the nonparametric model for the data structure $(C \wedge T_k, N(C \wedge T_k))$. This shows that the Kaplan-Meier estimator of the distribution of T_k , based on the reduced data (\tilde{T}_k, Δ_k) , is asymptotically efficient at most continuous data generating distributions, which explains also the result proved in van der Laan, Jewell and Peterson (1998) for the case $k = 2$. Moreover, simple isotonic regression estimators for the distributions F_j , $j = 1, \dots, k - 1$, are proposed which also yield asymptotically efficient estimators of smooth functionals by our general result.

2 Current status data on a counting process.

We will first review classical current status data.

2.1 Classical current status data.

Classical current status data can be viewed as current status data on a simple counting process as follows. Let T be a univariate failure time of interest and define the process $\Delta(t) = I(T \leq t)$ as the counting process with one single jump at point T . Let $Y = (C, \Delta(C))$ represent current status data on Δ at a monitoring time C . We will assume that C is independent of T (i.e. of $\Delta(\cdot)$). The parameter of interest is the distribution F of T .

The properties of the NPMLE F_n of the distribution of T have been established in Groeneboom and Wellner (1992). Here the NPMLE is defined as the maximum likelihood estimator over all discrete distributions with jumps at the monitoring times. Beyond proving a limit distribution result for F_n these authors also established efficiency of smooth functionals of F_n with a closed form expression of the limit variance so that Wald-type confidence intervals are directly available. Huang and Wellner (1995) provide an alternative proof of asymptotic linearity of the NPMLE of smooth functionals of F under weak conditions. The following lemma (is a consequence of their result.

We refer to Bickel, Klaassen, Ritov, Wellner (1993) for definitions of a *regular, asymptotically linear and efficient estimator* and *influence curve* of an estimator. The semiparametric information bound at $P_{F,G}$ is defined as the infimum of parametric information bounds over a specified class of parametric submodels. We choose as parametric one-dimensional submodels

$$\{\epsilon \rightarrow P_{F_{\epsilon,h_1},G_{\epsilon,h_2}} : \|h_j\|_{\infty} < \infty, j = 1, 2, \int h_1 dF = \int h_2 dG = 0\},$$

where $dF_{\epsilon,h_1}(\cdot) = (1 + \epsilon h_1(\cdot))dF(\cdot)$, $dG_{\epsilon,h_2}(\cdot) = (1 + \epsilon h_2(\cdot))dG(\cdot)$ and ϵ is the unknown parameter with parameter space $[-\delta, \delta]$ for some small $\delta > 0$. The tangent space at $P_{F,G}$ is now defined as the closure in $L_0^2(P_{F,G})$ of the linear span of all the scores of these one-dimensional submodels, where for a given measure μ we define $L_0^2(\mu) = \{h : \int h^2 d\mu < \infty, \int h d\mu = 0\}$ as the Hilbert space endowed with inner product $\langle h_1, h_2 \rangle_{\mu} = \int h_1(y)h_2(y)d\mu(y)$. Thus the tangent space at $P_{F,G}$ is a sub-Hilbert space of $L_0^2(P_{F,G})$.

In this paper it is particularly important to realize that efficiency of an estimator is a local property in the sense that a regular estimator can be efficient at a particular $P_{F,G}$ and inefficient at another element of the model. So efficiency of an estimator is defined at a fixed element $P_{F,G}$ of the model.

Lemma 2.1 Consider the nonparametric model for $Y = (C, \Delta(C))$, where $\Delta(\cdot) \equiv I(T \leq \cdot)$, T has unspecified distribution F and C is independent of T with unspecified distribution G . We observe n i.i.d. observations of $Y = (C, \Delta(C))$. Consider the parameter $\mu = \int(1 - F)(u)r(u)du$ for a given function r . Consider the estimator $\mu_n = \int(1 - F_n)(u)r(u)du$, where F_n is the NPMLE of F . We have that μ_n is regular and asymptotically linear at any (F, G) for which F is continuous with density $f_T > 0$ on $[0, M]$ and zero elsewhere ($M < \infty$), $g(x) = dG/dx > 0$ on $[0, M]$, r bounded on $[0, M]$.

The influence curve of μ_n is given by:

$$IC(Y | F, g, r) = \frac{r(C)}{g(C)} (F(C)(1 - \Delta) - (1 - F(C))\Delta). \quad (1)$$

The variance of IC is given by:

$$VAR(IC) = \int \frac{r^2(c)}{g(c)} F(c)(1 - F(c))dc.$$

We can actually prove the following tangent space result as well.

Lemma 2.2 Consider the nonparametric model for $Y = (C, \Delta(C))$, where $\Delta(\cdot) \equiv I(T \leq \cdot)$, T has unspecified distribution F and C is independent of T with unspecified distribution G . We observe n i.i.d. observations of $Y = (C, \Delta(C))$. Suppose that 1) F has a Lebesgue density f with $f > 0$ on $[0, \tau_F)$ and if $\tau_F < \infty$ ($\tau_F = \infty$ is allowed), then $f = 0$ on (τ_F, ∞) and 2) G has a Lebesgue density g . Then the tangent space at $P_{F,G}$ equals $L_0^2(P_{F,G})$. This implies that an estimator of a parameter $\mu(F)$ which is regular and asymptotically linear at $P_{F,G}$ is also asymptotically efficient if F, G satisfy 1) and 2).

In Gill, van der Laan, Robins (1997) it is proved that if one only assumes that the conditional distribution of the observed data Y , given the full data T , satisfies ‘‘coarsening at random’’ (CAR), then the tangent space at $P_{F,G}$ is saturated, i.e. equals $L_0^2(P_{F,G})$. The tangent space generated by $G(\cdot | T)$ under the sole assumption CAR equals $T_{CAR} = \{v(Y) \in L_0^2(P_{F,G}) :$

$E(v(Y) | T) = 0$. Therefore, the main idea of the proof below is to show that under the independent censoring model $G(\cdot | T) = G(\cdot)$ the tangent space of the marginal distribution G still equals T_{CAR} at a $P_{F,G}$ satisfying 1) and 2) of lemma 2.2. The proof below will be an ingredient of the proofs of our two main theorems.

Proof of lemma 2.2. Let $A : L_0^2(F) \rightarrow L_0^2(P_{F,G})$, $A(h)(Y) = E_F(h(T) | Y)$ be the score operator for F and let $A^\top : L_0^2(P_{F,G}) \rightarrow L_0^2(F)$, $A(V)(T) = E_G(V(Y) | T)$ be its adjoint. It is a well known fact that the closure of the range of a Hilbert space operator equals the orthogonal complement of the null-space of its adjoint: i.e. $\overline{R(A)} = N(A^\top)^\perp$, where $\overline{R(A)}$ is the closure of the range of the score operator and $N(A^\top)$ is the nullspace of A^\top . Thus $L_0^2(P_{F,G}) = \overline{R(A)} + N(A^\top)$.

The data generating distribution is indexed by two locally variation independent parameters F and G so that the tangent space at $P_{F,G}$ can be obtained as a sum of two tangent spaces, namely the tangent space for F , which is given by $\overline{R(A)}$, and the tangent space for G . For every $h \in L_0^2(G)$ with finite supremum norm we have that $\epsilon \rightarrow (1 + \epsilon h_2)dG$ is a one-dimensional submodel through G at $\epsilon = 0$. Thus the tangent space corresponding with submodels $\epsilon \rightarrow P_{F,G_\epsilon}$ equals $L_0^2(G)$. Thus we have that the tangent space is given by $\overline{R(A)} + L_0^2(G)$. We conclude that it suffices to show that $N(A^\top) = L_0^2(G)$.

We have

$$A^\top(V)(T) = \int_0^T V(c, 0)dG(c) + \int_T^\infty V(c, 1)dG(c).$$

Thus $\int V(c, \Delta(c))dG(c) = 0$ F -a.e. implies that

$$\int_0^T \{V(c, 0) - V(c, 1)\}g(c)dc = \int_0^\infty V(c, 1)dG(c) \text{ for } T \in [0, \tau_F). \quad (2)$$

Differentiation w.r.t T yields $V(C, 0) = V(C, 1)$ on $[0, \tau_F)$ G -a.e. If $\tau_F < \infty$ and $c > \tau_F$, then $c > T$ and thus $V(c, \Delta(c)) = V(c, 1)$. Thus $V(C, 0) = V(C, 1)$ G -a.e. which proves $N(A^\top) = L_0^2(G)$. \square

It is of interest to note that one can represent $F_T(t)$ as a monotonic regression of Δ on C since $F(t) = E(\Delta(C) | C = t)$. This suggest that one can estimate F_T with the estimator $F_n(t)$ which minimizes

$$\sum_{i=1}^n (\Delta(C_i) - F_T(C_i))^2 \text{ over all distribution functions } F_T.$$

The solution of this problem is obtained with the fast Pool-Adjacent-Violator-Algorithm (see Barlow, Bartholomew, Bremner, Brunk, 1972). This estimator happens to correspond with the NPMLE.

2.2 Current status data on a counting process.

Let the process of interest be a counting process $N(t) = \sum_{j=1}^k I(T_j \leq t)$, $T_1 < \dots < T_k$, where T_j is the time-variable at which an event occurs and where N jumps from value $j - 1$ to j . Let C be a monitoring time and consider the data structure $Y = (C, N(C))$. We observe n i.i.d. copies of Y . We will only assume that C is independent of N .

The distribution of $(C, N(C))$ depends on the distribution of \vec{T} only through the marginal distributions F_j of T_j , $j = 1, \dots, k$. To be precise we have (denoting $S_i = 1 - F_i$): for $j \in \{0, \dots, k\}$

$$\begin{aligned} P_{F,G}(dc, N(C) = j) &= I(j = 0)S_1(c)dG(c) + I(j = k)F_k(c)dG(c) \\ &+ I(j = 1)\{S_2(c) - S_1(c)\}dG(c) + \dots + I(j = k - 1)\{S_k(c) - S_{k-1}(c)\}dG(c). \end{aligned}$$

This shows that the distribution of $Y = (C, N(C))$ only identifies the marginal distributions of T_j , $j = 1, \dots, k$.

The NPMLE does not exist in closed form and can only be computed with an iterative algorithm For a given j , we can reduce the observation $(C, N(C))$ to simple current status data $(C, \Delta_j = I(T_j \leq C))$ on T_j and estimate F_j with the RNPMLE. Under the conditions stated in

lemma 2.1 with $F = F_j$ and $G = G$, this estimator provides regular and asymptotically linear estimators of smooth functionals of the type $\mu_j = \int(1 - F_{T_j})(u)r(u)du$, for a given r in the nonparametric model. The following theorem proves that at a data generating distribution of Y satisfying a mild condition any regular asymptotically linear estimator will provide asymptotically efficient estimators of smooth functionals of F_{T_j} . We decided to state a condition which is easy to understand and practical, but our proof shows that our condition can be weakened.

Theorem 2.1 *Let $T_1 < T_2 < \dots < T_k$ be time-variables corresponding with chronological events of interest. Define the counting process with jumps of size 1 at these T_j 's by:*

$$N(t) = \sum_{j=1}^k I(T_j \leq t).$$

Let $Y = (C, N(C))$. Consider the following semiparametric model for Y : Let $C \sim G$ be independent of $\vec{T} \sim F$, but leave G and F unspecified. Then the distribution of Y only depends on the multivariate distribution F of $\vec{T} = (T_1, \dots, T_k)$ through the marginal distributions F_1, \dots, F_k of T_1, \dots, T_k .

Consider a data generating distribution $P_{F,G}$ in the model above satisfying the following condition (3): For certain $\tau_1 < \dots < \tau_k < \infty$, let F_j have Lebesgue density f_j with

$$\begin{aligned} f_j > 0 & \quad \text{on } [0, \tau_j] \text{ and } f_j = 0 \text{ on } (\tau_j, \infty), j = 1, \dots, k \\ F_j > F_{j+1} & \quad \text{on } (0, \tau_j], j = 1, \dots, k - 1 \\ G & \quad \text{has Lebesgue density } g. \end{aligned} \tag{3}$$

Then the tangent space at $P_{F,G}$ equals $L_0^2(P_{F,G})$ and is thus saturated.

This implies that any estimator of a real valued parameter of F which is regular and asymptotically linear estimator at $P_{F,G}$ is also asymptotically efficient if $P_{F,G}$ satisfies (3). In particular, given $j \in \{1, \dots, k\}$, if $P_{F,G}$ satisfies (3) and F_j, G satisfies the conditions of lemma 2.1 for the

RNPMLE of μ_{F_j} based on $(C, I(T_j \leq C))$ (thus with $F = F_j$ and $G = G$), then the RNPMLE of μ_{F_j} is asymptotically efficient.

2.2.1 Heuristic understanding of the difference between NPMLE and RNPMLE.

To understand the difference between NPMLE and RNPMLE we consider the special case $k = 2$ in detail. In this case N can have three possible values:

$$N(C) = \begin{cases} 0 & \text{if } C < T_1 \\ 1 & \text{if } T_1 < C < T_2 \\ 2 & \text{if } C > T_2 \end{cases}$$

Let's assume that C has a Lebesgue density g . The likelihood of $(C, N(C))$ is given by:

$$p_{F_1, F_2, G}(c, N(c) = j) = S_1(c)^{I(j=0)}(S_2 - S_1)(c)^{I(j=1)}F_2(c)^{I(j=2)}g(c).$$

We note that the density $p_{F_1, F_2, G}$ can be reparametrized as:

$$p_{R, F_2, G}(c, \delta) = R(c)^{I(j=0)}(1 - R(c))^{I(j=1)}S_2(c)^{I(j \in \{0,1\})}F_2(c)^{I(j=2)}g(c),$$

where $R(t) \equiv S_1(t)/S_2(t)$. Thus, if we ignore the relation between F_2 and R , then the NPMLE of F_2 of the likelihood corresponding with $P_{R, F_2, G}$ would actually be equal to the reduced data NPMLE based on the reduced data $(C, I(T_2 \leq C))$. However, F_2 and R are related since S_2R has to be a survival function. Therefore it is not possible to determine the NPMLE by separate maximization w.r.t. F_2 and R , which explains that the actual NPMLE of F_2 will differ from the RNPMLE of F_2 .

Theorem 2.1 shows that this relation between F_2 and R is not informative for estimation of smooth functionals of F_2 at a large class of data generating distributions since the RNPMLE, ignoring this relation, is still asymptotically efficient for estimation of root- n -estimable parameters. Our proof of theorem 2.1 for $k = 2$ shows that the efficient score operator (for the definition

of an efficient score operator see the proof) of F_2 equals the efficient score operator for F_2 in the reduced data model just observing (C, Δ_2) . This implies that at (F_1, F_2) satisfying (3) the efficient influence curve for any smooth functional of F_2 equals the influence curve of the RNPMLE as given in lemma 2.1. Closer inspection of the proof for $k = 2$ also proves that if (e.g.) F_2 is continuous while F_1 is discrete on $[0, \tau_1]$ or F_2 is discrete with support not containing the support of a discrete F_1 , then the efficient score operator for F_2 actually differs from the efficient score operator for F_2 in the reduced data model so that, in particular, the efficient influence curves (and information bounds) differ for the two models. Thus at such (F_1, F_2) the RNPMLE of smooth functionals of F_2 is inefficient.

Here we will provide a likelihood-based explanation of this fact that the true NPMLE will be more efficient than the RNPMLE at such (F_1, F_2) . Let R_n be the NPMLE of R . The NPMLE of F_2 maximizes the likelihood corresponding with p_{R_n, F_2} over all F_2 for which $S_2 R_n$ is a survival function, while the RNPMLE maximizes the likelihood unrestrictedly over all distributions F_2 . Suppose now that the model consists of F_1 's discrete and F_2 's being continuous. This model, though smaller than the model with F_1, F_2 being fully unspecified, has the same semiparametric efficiency bound at a (F_1, F_2) in this smaller model as the efficiency bound in the original model. This follows from the fact that the class of one dimensional submodels as needed to compute the tangent space can still be chosen the same. In this smaller model an $R = S_1/S_2$ will be discrete at the support points of F_1 and the shape of R between the support points equals the shape of $1/S_2$. As a consequence, since R determines the shape of F_2 between the support points, knowing R in the smaller model helps enormously in estimating S_2 . In particular, for a given R_n , maximizing the likelihood corresponding with p_{R_n, F_2} over F_2 with $S_2 R_n$ being a survival function, is very different from maximizing this likelihood over all possible distributions F_2 . This shows that the RNPMLE in the smaller model is inefficient at such (F_1, F_2) . Since the efficiency

bound in the smaller model is the same as the efficiency bound in the original model this also shows that the RNPMLE will be inefficient as well at such (F_1, F_2) .

2.3 Proof of theorem 2.1.

We need to prove that assumption (3) implies that the tangent space at $P_{F,G}$ equals $L_0^2(P_{F,G})$ and is thus saturated. The data generating distribution $P_{F,G}$ is indexed by F and G , where the dependence on F is only through the marginals F_j , $j = 1, \dots, k$. Thus the tangent space at $P_{F,G}$ can be obtained as a sum of two tangent spaces, namely the tangent space for F and the tangent space for G , where the latter equals $L_0^2(G)$. Let F, G be given and satisfy (3). We now claim that the tangent space for F is given by the closure of the sum of the k tangent spaces for F_j calculated as if the F_j 's are variation independent parameters, $j = 1, \dots, k$. We will show this now. Let $h_j \in L_0^2(F_j)$ have finite supremum norm and let F_{j,ϵ,h_j} be the one-dimensional perturbation $\epsilon \rightarrow \int_0^{\cdot} (1 + \epsilon h_j) dF_j$ through F_j at $\epsilon = 0$, $j = 1, \dots, k$. Firstly, note that the support of F_{j,ϵ,h_j} equals the support $[0, \tau_j]$ of F_j , $j = 1, \dots, k$. Since $F_j > F_{j+1}$ (strictly) on $(0, \tau_j]$ we have that, given an arbitrarily small $\delta_1 > 0$, there exists a neighborhood $\epsilon \in (-\delta, \delta)$ so that $F_{j,\epsilon,h_j} \geq F_{j+1,\epsilon,h_{j+1}}$ on $(\delta, \tau_j]$ for all $j = 1, \dots, k-1$. Thus $P_{F_{j,\epsilon,h_j}, j=1, \dots, k, G}$ does satisfy the constraints $F_j \geq F_{j+1}$, $j = 1, \dots, k-1$ of our model except on an arbitrarily small neighborhood of 0. Thus by modifying h_j on an arbitrarily small neighborhood of 0 we can arrange that $\epsilon \rightarrow P_{F_{j,\epsilon,h_j}, j=1, \dots, k, G}$ is a true one-dimensional submodel. Since a tangent space for F is obtained as the *closure* in $L_0^2(F)$ of the linear span of scores of all possible 1-d-submodels it follows that the score of the unmodified $\epsilon \rightarrow P_{F_{j,\epsilon,h_j}, j=1, \dots, k, G}$ belongs to the tangent space as well. This proves our claim.

Let $j \in \{1, \dots, k\}$ be given. For a given $h_j \in L_0^2(F_j)$ we consider the one-dimensional submodel $F_{j,\epsilon}$ given by $\epsilon \rightarrow (1 + \epsilon h_j(t)) dF_j(t)$ which goes through F_j at $\epsilon = 0$. For notational

convenience we define the random variable $R = N(C) + 1 \in \{1, \dots, k + 1\}$ and let F_{-j} be the $k - 1$ dimensional vector of cdf's excluding F_j . This one-dimensional submodel $F_{j,\epsilon}$ implies a score for $P_{F_j,\epsilon,F_{-j},G}$ given by:

$$\begin{aligned} A_1(h_1) &= I(R = 1) \frac{\int_c^\infty h_1 dF_1}{S_1(c)} - I(R = 2) \frac{\int_c^\infty h_1 dF_1}{(S_2 - S_1)(c)} \text{ if } j = 1. \\ A_j(h_j) &= I(R = j) \frac{\int_c^\infty h_j dF_j}{(S_j - S_{j-1})(c)} - I(R = j + 1) \frac{\int_c^\infty h_j dF_j}{(S_{j+1} - S_j)(c)}, \text{ if } j \in \{2, \dots, k - 1\}. \\ A_k(h_k) &= I(R = k) \frac{\int_c^\infty h_k dF_k}{(S_k - S_{k-1})(c)} - I(R = k + 1) \frac{\int_c^\infty h_k dF_k}{F_k(c)} \text{ if } j = k. \end{aligned}$$

If we define $S_0 \equiv 0$ and $S_{k+1} \equiv 1$, then for $j = 1, \dots, k$

$$A_j(h_j) = I(R = j) \frac{\int_c^\infty h_j dF_j}{(S_j - S_{j-1})(c)} - I(R = j + 1) \frac{\int_c^\infty h_j dF_j}{(S_{j+1} - S_j)(c)},$$

where we use that $S_1 - S_0 = S_1$ and $S_{k+1} - S_k = F_k$. Here $A_j : L_0^2(F_j) \rightarrow L_0^2(P_{F,G})$ is called the score operator of F_j , $j = 1, \dots, k$. The tangent space for F_j is given by the closure of the range of A_j denoted by $\overline{R(A_j)}$. Let's define $A_F : L_0^2(F_1) \times \dots \times L_0^2(F_k) \rightarrow L_0^2(P_{F,G})$ by $A_F(h_1, \dots, h_k) = A_1(h_1) + \dots + A_k(h_k)$. Thus the tangent space for F equals $\overline{R(A_F)}$ so that the tangent space at $P_{F,G}$ is given by $\overline{R(A_F) + L_0^2(G)}$. Thus to prove the theorem it suffices to show that $\overline{R(A_F) + L_0^2(G)} = L_0^2(P_{F,G})$ at any F, G satisfying (3).

Thus the remaining task is to understand the range of A_F . We will decompose A_F as a sum of efficient score operators A_j^* , where A_j^* is defined as A_j minus its projection on the sum-space spanned by the ranges of the other score operators $A_1, \dots, A_{j-1}, A_{j+1}, \dots, A_k$, $j = 1, \dots, k$. We will be able to prove that the efficient score operator of F_j at a $P_{F,G}$ satisfying (3) equals $A_j^*(h_j) = E(h_j(T_j) \mid (C, \Delta_j = I(T_j \leq C)))$, which is the score operator for the reduced current status data structure (C, Δ_j) , $j = 1, \dots, k$. Since the information bounds for smooth functionals of F_j are in both models solely expressed in terms of the efficient score operator for F_j , the latter result proves that an efficient estimator of μ_j based on (C, Δ_j) , $j = 1, \dots, k$, like the RNPMLE, is also efficient in the model for the more informative data structure $(C, N(C))$ (e.g

Bickel, Klaassen, Ritov, Wellner, 1993). This proves that the RNPMLE actually yields efficient estimators. Subsequently, we will prove that this special structure of the efficient score operators proves that the tangent space at a $P_{F,G}$ satisfying (3) is saturated which proves the more general statement of theorem 2.1 and thus completes the proof.

Derivation of the efficient score operators of F_j . Since $E(A_l(h_l)A_m(h_m)(Y)) = 0$ if $|l - m| \geq 2$, it will follow that the efficient score operators mainly involves calculation of projections of the type $\Pi(A_j | \overline{R(A_{j-1})})$ and $\Pi(A_j | \overline{R(A_{j+1})})$. Therefore we will first obtain closed form expressions, in general, for these projection operators.

If the projection $\Pi(A_j(h_j) | \overline{R(A_{j-1})})$ is actually an element of $R(A_{j-1})$, then this projection is given by (compare with the formula $X(X'X)^{-1}X'Y$ for the least squares estimator):

$$\Pi(A_j(h_j) | \overline{R(A_{j-1})}) = A_{j-1}(A_{j-1}^\top A_{j-1})^{-1} A_{j-1}^\top A_j(h_j), \quad (4)$$

where $A_{j-1}^\top : L_0^2(P_{F,G}) \rightarrow L_0^2(F_{j-1})$ is the adjoint of $A_{j-1} : L_0^2(F_{j-1}) \rightarrow L_0^2(P_{F,G})$ and $(A_{j-1}^\top A_{j-1})^{-1}$ stands for the generalized inverse of $A_{j-1}^\top A_{j-1} : L_0^2(F_{j-1}) \rightarrow L_0^2(F_{j-1})$. Similarly,

$$\Pi(A_j(h_j) | \overline{R(A_{j+1})}) = A_{j+1}(A_{j+1}^\top A_{j+1})^{-1} A_{j+1}^\top A_j(h_j). \quad (5)$$

The adjoint A_l^\top is defined by:

$$\langle A_l(h_l), \eta \rangle_{P_F} = \langle h_l, A_l^\top(\eta) \rangle_{F_l} \text{ for all } h_l \in L_0^2(F_l) \text{ and } \eta \in L_0^2(P_{F,G}).$$

It is easily shown that for $l \in \{1, \dots, k\}$

$$A_l^\top(V)(T_l) = \int_0^{T_l} \{(V(c, l) - V(c, l+1))\} dG(c).$$

We have that

$$A_l^\top A_l(h_l)(T_l) = \int_0^{T_l} \phi_l(c) \int_c^\infty h_l dF_l dG(c),$$

where

$$\begin{aligned}\phi_1 &= \frac{S_2}{S_1(S_2 - S_1)} \\ \phi_l &= \frac{S_{l+1} - S_{l-1}}{(S_{l+1} - S_l)(S_l - S_{l-1})} \quad l = 2, \dots, k-1 \\ \phi_k &= \frac{F_{k-1}}{(S_k - S_{k-1})F_k},\end{aligned}$$

or, in fact, with our convention of $S_0 = 0$ and $S_{k+1} = 1$

$$\phi_l = \frac{S_{l+1} - S_{l-1}}{(S_{l+1} - S_l)(S_l - S_{l-1})}, \quad l = 1, \dots, k.$$

Thus, given a K with $K \ll G$, a solution (if it exists) of $A_l^\top A_l(h_l) = K$ has to satisfy: for G -a.e.

$c \in [0, \tau]$

$$\int_c^\infty h_l dF_l = \frac{dK}{dG} \frac{1}{\phi_l(c)}, \quad l = 1, \dots, k.$$

We have for $l \in \{1, \dots, k-1\}$

$$\begin{aligned}A_l^\top A_{l+1}(h_{l+1}) &= \int_0^{T_l} [A_{l+1}(h_{l+1})I(R=l) - A_{l+1}(h_{l+1})I(R=l+1)] dG(c) \\ &= - \int_0^{T_l} A_{l+1}(h_{l+1})I(R=l+1) dG(c) \\ &= - \int_0^{T_l} \frac{1}{S_{l+1} - S_l} \int_c^\infty h_{l+1} dF_{l+1} dG(c).\end{aligned}$$

We note that this element is indeed absolutely continuous w.r.t. G . Similarly, it follows that for

$l \in \{1, \dots, k-1\}$

$$A_{l+1}^\top A_l(h_l) = - \int_0^{T_{l+1}} \frac{1}{S_{l+1} - S_l} \int_c^\infty h_l dF_l dG(c).$$

Thus $h_{j-1,j} \equiv (A_{j-1}^\top A_{j-1})^{-1} A_{j-1}^\top A_j(h_j)$ is the h satisfying

$$\int_c^\infty h dF_{j-1} = - \frac{1}{S_j - S_{j-1}} \int_c^\infty h_j dF_j \frac{1}{\phi_{j-1}} \quad \text{for } G\text{-a.e. } c \in [0, \tau_{j-1}] \quad (6)$$

and $h_{j+1,j} \equiv (A_{j+1}^\top A_{j+1})^{-1} A_{j+1}^\top A_j(h_j)$ is the h satisfying

$$\int_c^\infty h dF_{j+1} = - \frac{1}{S_{j+1} - S_j} \int_c^\infty h_j dF_j \frac{1}{\phi_{j+1}} \quad \text{for } G\text{-a.e. } c \in [0, \tau_{j+1}]. \quad (7)$$

If we can take a derivative of the right-hand sides in (6) and (7) w.r.t. F_{j-1} and F_{j+1} , then the equations (6) and (7) in terms of h have a solution. This is possible if $F_j \ll F_l$ (i.e. F_j is absolute continuous w.r.t. F_l) on $[0, \tau_l]$, $l \in \{j-1, j+1\}$, which thus holds under assumption (3) since we assumed that all F_j have positive Lebesgue density. The efficient score operator A_j^* also involves projections requiring existence of solutions $h_{l-1,l}, h_{l+1,l}$ for l different from j . Therefore we assumed condition (3) to include (with an easy to understand condition) the necessary and sufficient conditions for the existence of $h_{l-1,l}, h_{l+1,l}$ for all possible l , as needed below.

This gives us the following closed form expressions for the projections (4) and (5) by simply replacing $\int_c^\infty h dF_l$ in $A_l(h)$ by the expressions above. We have for $j = 1, \dots, k-1$,

$$\begin{aligned} \Pi(A_j(h_j) | \overline{R(A_{j+1})}) &= A_{j+1}(h_{j+1,j}) \\ &= -\frac{\int_c^\infty h_j dF_j}{(S_{j+1} - S_j)^2 \phi_{j+1}} I(R = j+1) + \frac{\int_c^\infty h_j dF_j}{(S_{j+2} - S_{j+1})(S_{j+1} - S_j) \phi_{j+1}} I(R = j+2) \end{aligned} \quad (8)$$

and for $j = 2, \dots, k$

$$\begin{aligned} \Pi(A_j(h_j) | \overline{R(A_{j-1})}) &= A_{j-1}(h_{j-1,j}) \\ &= -\frac{\int_c^\infty h_j dF_j}{(S_j - S_{j-1})(S_{j-1} - S_{j-2}) \phi_{j-1}} I(R = j-1) + \frac{\int_c^\infty h_j dF_j}{(S_j - S_{j-1})^2 \phi_{j-1}} I(R = j) \end{aligned} \quad (9)$$

For the sake of simplicity we will derive the efficient score operators for the case $k = 3$. The proof generalizes to the general case, but it will be easier for the reader to follow this special case. Firstly, we define

$$A_j^l = A_j - \Pi(A_j | \overline{R(A_l)}).$$

The efficient score operators $A_j^* : L_0^2(F_j) \rightarrow L_0^2(P_F)$ are given by

$$\begin{aligned} A_3^* &= A_3 - \Pi(A_3 | \overline{R(A_1 + A_2)}) \\ &= A_3 - \Pi(A_3 | \overline{R(A_2^1)}) \\ A_2^* &= A_2 - \Pi(A_2 | \overline{R(A_1 + A_3)}) \end{aligned}$$

$$\begin{aligned}
&= A_2 - \Pi(A_2 \mid \overline{R(A_1)}) - \Pi(A_2 \mid \overline{R(A_3)}) \\
A_1^* &= A_1 - \Pi(A_1 \mid \overline{R(A_2 + A_3)}) \\
&= A_1 - \Pi(A_1 \mid \overline{R(A_2^3)}).
\end{aligned}$$

Calculation of A_2^* . Let's first calculate A_2^* . Applying (8) and (9) with $j = 2$ gives us

$$\Pi(A_2(h_2) \mid \overline{R(A_1)}) = -\frac{\int_c^\infty h_2 dF_2}{(S_2 - S_1)(S_1 - S_0)\phi_1} I(R = 1) + \frac{\int_c^\infty h_2 dF_2}{(S_2 - S_1)^2 \phi_1} I(R = 2)$$

and

$$\Pi(A_2(h_2) \mid \overline{R(A_3)}) = -\frac{\int_c^\infty h_2 dF_2}{(S_3 - S_2)^2 \phi_3} I(R = 3) + \frac{\int_c^\infty h_2 dF_2}{(S_4 - S_3)(S_3 - S_2)\phi_3} I(R = 4).$$

Thus

$$\begin{aligned}
A_2^*(h_2) &= \frac{\int_c^\infty h_2 dF_2}{(S_2 - S_1)\phi_1} I(R = 1) + \left\{ \frac{1}{S_2 - S_1} - \frac{1}{(S_2 - S_1)^2 \phi_1} \right\} \int_c^\infty h_2 dF_2 I(R = 2) \\
&+ \left\{ \frac{1}{(S_3 - S_2)^2 \phi_3} - \frac{1}{S_3 - S_2} \right\} \int_c^\infty h_2 dF_2 I(R = 3) - \frac{\int_c^\infty h_2 dF_2}{(S_4 - S_3)(S_3 - S_2)\phi_3} I(R = 4).
\end{aligned}$$

Now, notice that

$$\begin{aligned}
(S_2 - S_1)S_1\phi_1 &= S_2 \\
(S_4 - S_3)(S_3 - S_2)\phi_3 &= S_4 - S_2 = F_2 \\
\frac{1}{(S_3 - S_2)^2 \phi_3} - \frac{1}{S_3 - S_2} &= -\frac{1}{F_2} \\
\frac{1}{S_2 - S_1} - \frac{1}{(S_2 - S_1)^2 \phi_1} &= \frac{1}{S_2}.
\end{aligned}$$

Thus (using that $\int_0^\infty h_2 dF_2 = 0$)

$$A_2^*(h_2) = \frac{\int_c^\infty h_2 dF_2}{S_2(c)} I(R \in \{1, 2\}) + \frac{\int_0^c h_2 dF_2}{F_2(c)} I(R \in \{3, 4\}).$$

Calculation of A_1^* . Formula (8) with $j = 1$ gives us:

$$\Pi(A_2(h_2) \mid \overline{R(A_3)}) = -\frac{\int_c^\infty h_2 dF_2}{(S_3 - S_2)^2 \phi_3(c)} I(R = 3) + \frac{\int_c^\infty h_2 dF_2}{(S_4 - S_3)(S_3 - S_2)\phi_3(c)} I(R = 4).$$

Thus

$$\begin{aligned}
A_2^3(h_2) &= A_2(h_2) - \Pi(A_2(h_2) | \overline{R(A_3)}) \\
&= \frac{\int_c^\infty h_2 dF_2}{(S_2 - S_1)(c)} I(R = 2) + \left\{ \frac{1}{(S_3 - S_2)^2 \phi_3(c)} - \frac{1}{(S_3 - S_2)(c)} \right\} \int_c^\infty h_2 dF_2 I(R = 3) \\
&\quad - \frac{\int_c^\infty h_2 dF_2}{(S_4 - S_3)(S_3 - S_2) \phi_3(c)} I(R = 4)
\end{aligned}$$

We now note that

$$\begin{aligned}
(S_4 - S_3)(S_3 - S_2) \phi_3 &= F_2 \\
\frac{1}{(S_3 - S_2)^2 \phi_3(c)} - \frac{1}{(S_3 - S_2)(c)} &= -\frac{1}{F_2}.
\end{aligned}$$

Thus

$$A_2^3(h_2) = \frac{\int_c^\infty h_2 dF_2}{(S_2 - S_1)(c)} I(R = 2) - \frac{\int_c^\infty h_2 dF_2}{F_2(c)} I(R \in \{3, 4\}).$$

It is easily verified that the adjoint $A_2^{3\top} : L_0^2(P_F) \rightarrow L_0^2(F_2)$ is given by:

$$A_2^{3\top}(V) = \int_0^{T_2} \left\{ V(c, 2) - \frac{(S_3 - S_2)(c)}{F_2(c)} V(c, 3) - \frac{F_3(c)}{F_2(c)} V(c, 4) \right\} dG(c).$$

Subsequently, we can now verify that

$$A_2^{3\top} A_2^3(h_2) = \int_0^{T_2} \phi_2^3(c) \int_c^\infty h_2 dF_2 dG(c),$$

where

$$\phi_2^3 = \frac{F_1}{F_2(S_2 - S_1)}.$$

We need to find $h_{23,1} \equiv (A_2^{3\top} A_2^3)^-(K)$ with

$$K = A_2^{3\top} A_1(h_1) = - \int_0^{T_2} \frac{\int_c^\infty h_1 dF_1}{(S_2 - S_1)(c)} dG(c).$$

This solution has to satisfy:

$$\int_c^\infty h_{23,1} dF_2 = \frac{dK}{dG}(c) \frac{1}{\phi_2^3(c)} = -\frac{F_2}{F_1}(c) \int_c^\infty h_1 dF_1.$$

We note that $h_{23,1}$ exists under the assumption $F_j \equiv F_k$ (i.e. $F_j \ll F_k$ and $F_k \ll F_j$) on $[0, \tau_j]$, $j = 1, \dots, k-1$ which holds under (3). We conclude that

$$\begin{aligned} \Pi(A_1(h_1) \mid \overline{R(A_2^3)}) &= A_2^3(h_{23,1}) \\ &= -\frac{F_2}{F_1(S_2 - S_1)} \int_c^\infty h_1 dF_1 I(R=2) + \frac{\int_c^\infty h_1 dF_1}{F_1} I(R \in \{3, 4\}). \end{aligned}$$

Using that $F_2/(F_1(S_2 - S_1)) - 1/(S_2 - S_1) = -1/F_1$ and $\int_c^\infty h_1 dF_1 = -\int_0^c h_1 dF_1$ provides us now with

$$\begin{aligned} A_1^*(h_1) &= A_1(h_1) - \Pi(A_1(h_1) \mid \overline{R(A_2^3)}) \\ &= \frac{\int_c^\infty h_1 dF_1}{S_1(c)} I(R=1) + \frac{\int_0^c h_1 dF_1}{F_1(c)} I(R \in \{2, 3, 4\}). \end{aligned}$$

Calculation of A_3^* . This calculation is very similar to the one above for A_1^* and is left out here.

We have

$$A_3^*(h_3) = \frac{\int_0^c h_3 dF_3}{F_3(c)} I(R=4) + \frac{\int_c^\infty h_3 dF_3}{S_3(c)} I(R \in \{1, 2, 3\}).$$

Proving that the tangent space is saturated. Given the expressions for the efficient score operators derived above we now want to prove that the tangent space at a $P_{F,G}$ satisfying (3) is saturated. Under our assumption (3), the tangent space equals $L_0^2(G)$ (scores generated by G) plus the closure of the range of $A^* : L_0^2(F_1) \times \dots \times L_0^2(F_k) \rightarrow L_0^2(P_F)$ defined by

$$(h_1, \dots, h_k) \rightarrow A_1^*(h_1) + \dots + A_k^*(h_k),$$

where the marginal efficient score operators are given by $A_j^*(h_j) = E(h_j(T_j) \mid (C, \Delta_j = I(T_j \leq C)))$, $j = 1, \dots, k$. It is a well known fact that the closure of the range of a Hilbert space operator equals the orthogonal complement of the null-space of its adjoint: i.e. $\overline{R(A^*)} = N(A^{*\top})^\perp$. Thus we will need to show that $N(A^{*\top}) = L_0^2(G)$. The adjoint $A^{*\top} : L_0^2(P_F) \rightarrow L_0^2(F_1) \times \dots \times L_0^2(F_k)$ is given by:

$$A^{*\top}(V) = (A_1^{*\top}(V), \dots, A_k^{*\top}(V)),$$

where it is easily verified that the adjoint $A_j^{*\top} : L_0^2(P_F) \rightarrow L_0^2(F_j)$ of A_j^* is given by

$$A_j^{*\top}(V) = E(E(V(C, R) \mid C, \Delta_j) \mid T_j).$$

Consider the operator $B_j^\top : L_0^2(C, \Delta_j) \rightarrow L_0^2(F_j)$ given by $B_j^\top(\eta) = E(\eta(C, \Delta_j) \mid T_j)$, where $L_0^2(C, \Delta_j)$ is the space of functions of (C, Δ_j) with finite variance and zero mean (both taken w.r.t. $P_{F,G}$). Using precisely the same proof as the proof of lemma 2.2 it follows that if F_j has a Lebesgue density $f_j > 0$ on its support $[0, \tau_j]$, then the null-space $N(B_j^\top) = L_0^2(G)$, i.e. it consists of functions independent of Δ_j . Thus under condition (3) $A_j^{*\top}(V) = 0$ implies that $E(V(C, R) \mid C, \Delta_j) = E(V(C, R) \mid C) \equiv \phi(C)$, $j = 1, \dots, k$.

Setting $\Delta_1 = 0$ yields $\phi(C) = E(V(C, R) \mid C, \Delta_1 = 0) = V(C, 1)$. Now, we note that

$$P(R = m \mid \Delta_j = 1, C = c) = I(m \geq j + 1) \frac{P(R = m \mid c)}{F_j(c)}, \quad j = 1, \dots, k,$$

where $P(R = m \mid c) = (S_m - S_{m-1})(c)$. Thus $E(V(C, R) \mid C, \Delta_j = 1)$ is given by

$$\sum_{m \geq j+1} V(c, m) \frac{(S_m - S_{m-1})(c)}{F_j(c)} = \phi(c), \quad j = 1, \dots, k.$$

For $j = k$, this equality gives $V(c, k + 1) = \phi(c)$. For $j = k - 1$, this equality gives then

$$V(c, k) \frac{(S_k - S_{k-1})(c)}{F_{k-1}(c)} = \left(1 - \frac{F_k(c)}{F_{k-1}(c)}\right) \phi(c) = \frac{(S_k - S_{k-1})(c)}{F_{k-1}(c)} \phi(c)$$

so that $V(c, k) = \phi(c)$. In this manner we subsequently find $\phi(c) = V(c, k + 1) = V(c, k) = \dots = V(c, 2)$. This proves that $V(C, R)$ does not depend on R . and thus completes the proof.

3 Current status data on a counting process when final event is right censored.

The following theorem proves efficiency of any regular asymptotically linear estimator at a specified rich sub-model.

Theorem 3.1 *Let $N(t)$ be a counting process $N(t) = \sum_{j=1}^k I(T_j \leq t)$ for random variables $T_1 < \dots < T_k$. Let C be a random censoring time. For every subject we observe the following data structure*

$$Y = (\tilde{T} = T_k \wedge C, \Delta = I(T_k \leq C), N(\tilde{T})).$$

We assume that C is independent of (T_1, \dots, T_k) . The distribution of Y only depends on the multivariate distribution F of (T_1, \dots, T_k) through the marginal distributions F_1, \dots, F_k of (T_1, \dots, T_k) .

Consider a data generating distribution $P_{F,G}$ in the model above satisfying the following condition (10): For certain $\tau_1 < \dots < \tau_k < \infty$, let F_j have Lebesgue density f_j with

$$\begin{aligned} f_j > 0 & \quad \text{on } [0, \tau_j] \text{ and } f_j = 0 \text{ on } (\tau_j, \infty), \quad j = 1, \dots, k \\ F_j > F_{j+1} & \quad \text{on } (0, \tau_j], \quad j = 1, \dots, k-1 \\ G & \quad \text{has Lebesgue density } g. \end{aligned} \tag{10}$$

Then the tangent space at $P_{F,G}$ equals $L_0^2(P_{F,G})$ and is thus saturated. This implies that an estimator of a real valued parameter of the distribution F which is regular and asymptotically linear at $P_{F,G}$ is also asymptotically efficient if $P_{F,G}$ satisfies (10). In particular, if $\bar{G}(t) > 0$ and F, G satisfies (10), then the Kaplan-Meier estimator $S_{k,KM}(t)$ of $S_k(t) = P(T_k > t)$ based on the i.i.d. data on (\tilde{T}, Δ) is asymptotically efficient.

3.1 Regular and asymptotically linear estimators.

The important implication of theorem 3.1 is that if we can construct an estimator of root- n -estimable parameters of F_j which is regular, then this estimator will be asymptotically efficient at any F satisfying (10), $j = 1, \dots, k$. In this subsection we will provide relatively simple regular and asymptotically linear estimators.

Firstly, let's consider estimation of $S_k(t) = P(T_k > t)$. It is well known that $S_{k,KM}(t)$ is a regular asymptotically linear estimator of $S_k(t)$ whenever $\bar{G}(t) > 0$. Secondly, let's consider estimation of $S_j(t) = P(T_j > t)$, $j = 1, \dots, k-1$. Let $\Delta_j \equiv I(T_j \leq C)$. Under independent censoring (we can weaken this till non-informative censoring of T_k) we have

$$E(1 - \Delta_j \mid C = c, T_k > c) = \frac{S_j(c)}{S_k(c)} \equiv R_j(c). \quad (11)$$

So

$$S_j(c) = S_k(c)E(1 - \Delta_j \mid C = c, T_k > c) = E(S_k(c)(1 - \Delta_j) \mid C = c, T_k > c). \quad (12)$$

In other words, estimating S_j can be viewed as estimating a monotonic regression of $S_k(C)(1 - \Delta_j)$ on the observed C 's. This suggests replacing S_k by the efficient Kaplan-Meier estimator $S_{k,KM}$ and minimizing

$$\frac{1}{n} \sum_{i=1}^n w_i \{S_{k,KM}(C_i)(1 - \Delta_{ji}) - S_j(C_i)\}^2 I(C_i \leq T_{ki}) \quad (13)$$

over the vector $(S_j(C_i) : i = 1, \dots, n)$ under the constraint that S_j is monotone, where w_i , $i = 1, \dots, n$ is a given set of weights possibly assigning more mass to observations with smaller variance. The solution $S_{j,n}$ of this problem can be obtained with the ‘‘Pool Adjacent Violator Algorithm’’ (PAVA) (see e.g. Barlow, Bartholomew, Bremner and Brunk, 1972).

A simple calculation shows that

$$\begin{aligned} \text{VAR}\{S_k(C)(1 - \Delta_j) \mid C = c, T_k > c\} &= S_k(c)^2 \text{VAR}\{1 - \Delta_j \mid C = c, T_k > c\} \\ &= S_k^2(c) R_j(c) \{1 - R_j(c)\}. \end{aligned} \quad (14)$$

Since R_j is not identified from the data at a better rate than S_j a good set of weights is $w_i = 1/S_{k,KM}^2(C_i)$, $i = 1, \dots, n$ (see van der Laan, Jewell, Peterson, 1997).

It would go beyond the purpose and scope of this paper to prove that smooth functionals of $S_{j,n}$ are regular and asymptotically linear. Since it is straightforward to prove such a theorem

for a standard histogram regression estimator of the regression of $S_k(C)(1 - \Delta_j)$ on the observed C 's one expects that the more sophisticated isotonic regression estimate $S_{j,n}$ (which only differs because it selects its bins data adaptively) is regular and asymptotically linear under the same conditions. We note that the choice of weights w_i , $i = 1, \dots, n$, has no effect on the limit distribution of smooth functionals of $S_{j,n}$.

3.2 Proof of theorem 3.1.

In the first part of the proof we establish that, if condition (10) holds, then the efficient score operator of F_k equals the efficient score operator of F_k in the reduced data model for (\tilde{T}_k, Δ_k) , hereby establishing a proof of the efficiency of the Kaplan-Meier estimator $S_{KM}(t)$. Subsequently, exploiting this special form of the efficient score operator of F_k , we prove saturation of the tangent space and thus theorem 3.1.

Consider the data structure $(\tilde{T}_k = T_k \wedge C, N(\tilde{T}_k))$, where $N(t) = \sum_{j=1}^k I(T_j \leq t)$ and $T_1 < T_2 < \dots < T_k$ are ordered random variables. Let $R = N(\tilde{T}_k) + 1$. The density of the data is given by:

$$P(d\tilde{T}_k, R = j) = \prod_{m=1}^k (S_m - S_{m-1})(\tilde{T}_k)^{R=m} dF_k(\tilde{T}_k)^{R=k+1} dG(t)^{R \neq k+1} \bar{G}(t)^{R=k},$$

where $S_0 \equiv 0$ and $S_{k+1} \equiv 1$. We refer to the beginning of the proof of theorem 2.1 for showing that the tangent space at a $P_{F,G}$ satisfying condition (10) is the closure of the sum of the tangent spaces generated by F_j , $j = 1, \dots, k$ and the tangent space of G , treating F_j as locally variation independent. We have that the score operators $A_j : L_0^2(F_j) \rightarrow L_0^2(P_{F,G})$ for F_j , $j = 1, \dots, k-1$, are given by:

$$A_j(h_j) = \frac{\int_c^\infty h_j dF_j}{(S_j - S_{j-1})(c)} I(R = j) - \frac{\int_c^\infty h_j dF_j}{(S_{j+1} - S_j)(c)} I(R = j + 1)$$

and

$$A_k(h_k) = h_k(\tilde{T}_k)I(R = k + 1) + \frac{\int_c^\infty h_k dF_k}{(S_k - S_{k-1})(c)}I(R = k).$$

Derivation of efficient score operator of F_k . We will first determine the efficient score operator for F_k . For notational convenience we will consider the case $k = 3$. We have

$$A_3^*(h_3) = A_3(h_3) - \Pi(A_3(h_3) \mid \overline{R(A_3^1)})$$

where

$$A_2^1 = A_2 - \Pi(A_2 \mid \overline{R(A_1)}).$$

Applying formula (9) gives

$$\Pi(A_2(h_2) \mid \overline{R(A_1)}) = -\frac{\int_c^\infty h_2 dF_2}{S_2(c)}I(R = 1) + \frac{\int_c^\infty h_2 dF_2}{(S_2 - S_1)^2 \phi_1(c)}I(R = 2),$$

where we need to assume that $F_2 \ll F_1$ on $[0, \tau_1]$. Thus an easy calculation shows that

$$A_2^1(h_2) = \frac{\int_c^\infty h_2 dF_2}{S_2(c)}I(R \in \{1, 2\}) - \frac{\int_c^\infty h_2 dF_2}{(S_3 - S_2)(c)}I(R = 3).$$

Another straightforward calculation shows that the adjoint $A_2^{1\top} : L_0^2(P_{F,G}) \rightarrow L_0^2(F_2)$ of $A_2^1 : L_0^2(F_2) \rightarrow L_0^2(P_{F,G})$ is given by:

$$A_2^{1\top}(V) = \int_0^{T_2} \left\{ V(c, 1) \frac{S_1}{S_2}(c) + \frac{(S_2 - S_1)}{S_2}(c) V(c, 2) - V(c, 3) \right\} dG(c).$$

A straightforward calculation shows now that

$$A_2^{1\top} A_2^1(h_2) = \int_0^{T_2} \int_c^\infty h_2 dF_2 \frac{S_3}{S_2(S_3 - S_2)}(c) dG(c).$$

We also have

$$A_2^{1\top} A_3(h_3) = - \int_0^{T_2} \frac{\int_c^\infty h_3 dF_3}{(S_3 - S_2)(c)} dG(c).$$

This shows that $h_{21,3} \equiv (A_2^{1\top} A_2^1)^{-1} A_2^{1\top} A_3(h_3)$ satisfies

$$\int_c^\infty h_{21,3} dF_2 = -\frac{S_2}{S_3}(c) \int_c^\infty h_3 dF_3,$$

where we need to assume that this equation indeed can be solved in $h_{21,3}$. This is true if $F_3 \ll F_2$ on $[0, \tau_2]$. Then

$$\begin{aligned}\Pi(A_3(h_3) \mid \overline{R(A_2^1)}) &= A_2^1(h_{21,3}) \\ &= -\frac{\int_c^\infty h_3 dF_3}{S_3(c)} I(R \in \{1, 2\}) + \frac{S_2(c)}{S_3(S_3 - S_2)(c)} \int_c^\infty h_3 dF_3 I(R = 3).\end{aligned}$$

This proves that

$$\begin{aligned}A_3^*(h_3) &= h_3(\tilde{T}_3)I(R = 4) + \left\{ \frac{1}{S_3 - S_2} - \frac{S_2}{(S_3 - S_2)S_3} \right\} (c)I(R = 3) + \frac{\int_c^\infty h_3 dF_3}{S_3(c)} I(R \in \{1, 2\}) \\ &= h_3(\tilde{T}_3)I(R = 4) + \frac{\int_c^\infty h_3 dF_3}{S_3(c)} I(R \in \{1, 2, 3\}).\end{aligned}$$

Thus we have proved that if $F_k \equiv F_j$ on $[0, \tau_j]$, $j = 1, \dots, k - 1$, then the efficient score $A_k^*(h_k) = E(h_k(T_k) \mid \tilde{T}_k, \Delta_k)$. The latter condition holds, in particular, if (10) holds. This proves the statement of theorem 3.1 stating efficiency of the Kaplan-Meier estimator S_{KM} .

Saturated tangent space result. Note that for a random variable Y we define $L_0^2(Y) = \{h(Y) : Eh^2(Y) < \infty, Eh(Y) = 0\}$. For simplicity, we will prove this result for the case that $k = 3$. Let $A : L_0^2(F_1) \times L_0^2(F_2) \rightarrow L_0^2(P_{F,G})$ be defined by $A(h_1, h_2) = A_1(h_1) + A_2(h_2)$. Then the tangent space of F is given by $\overline{R(A_1) + R(A_2) + R(A_3)} = \overline{R(A_1) + R(A_2)} \oplus \overline{R(A_3^*)}$. Thus the tangent space at $P_{F,G}$ is given by $\overline{R(A)} \oplus \overline{R(A_3^*)} \oplus \overline{R(B)}$, where $B : L_0^2(G) \rightarrow L_0^2(\tilde{T}_3, \Delta_3)$ is the score operator for the censoring mechanism G , given by $B(h) = E(h(C) \mid \tilde{T}_3, \Delta_3)$. By factorization of the likelihood in a F and G -part, we indeed have that $R(B)$ is orthogonal to F -scores. It is a well known result that $\overline{R(A_3^*)} \oplus \overline{R(B)} = L_0^2(\tilde{T}_3, \Delta_3)$. The latter result just states that the tangent space for the nonparametric right-censored data model for (\tilde{T}_3, Δ_3) , only assuming that C is independent of T , is saturated which is a well known fact (e.g. Bickel, Klaassen, Ritov, Wellner, 1993). Thus we need to prove that $\overline{R(A)} \oplus L_0^2(\tilde{T}_3, \Delta_3) = L_0^2(P_{F,G})$ which is equivalent with proving $N(A^\top) = L_0^2(\tilde{T}_3, \Delta_3)$, where $A^\top : L_0^2(P_{F,G}) \rightarrow L_0^2(F_1) \times L_0^2(F_2)$ is the adjoint of A and $N(A^\top)$ denotes its null-space.

Firstly, we will decompose $A_1 + \dots + A_{k-1}$ into a sum of orthogonal operators (efficient score operators in the model with F_k known). Let $A'_1 = A_1 - \Pi(A_1 | \overline{R(A_2)})$ and $A'_2 = A_2 - \Pi(A_2 | \overline{R(A_1)})$. By (4) it follows that

$$\begin{aligned} A'_1(h_1) &= \frac{\int_c^\infty h_1 dF_1}{S_1(c)} I(R=1) - \frac{\int_c^\infty h_1 dF_1}{(S_3 - S_1)} I(R \in \{2, 3\}) \\ A'_2(h_2) &= \frac{\int_c^\infty h_2 dF_2}{S_2(c)} I(R \in \{1, 2\}) + \frac{\int_0^c h_2 dF_2}{(S_3 - S_2)(c)} I(R=3), \end{aligned}$$

where we need the equivalence assumptions $F_j \equiv F_{j+1}$ on $[0, \tau_j]$ for $j = 1, \dots, k$, again. A more compact manner of representing these operators $A'_j : L_0^2(F_j) \rightarrow H(C, R) \equiv \{V(C, R)I(R < 4) \in L_0^2(P_{F,G}) : V\}$ is

$$A'_j(h_j) = E(h_j(T_j) | C, \Delta_j, T_3 > C) I(T_3 > C), \quad j = 1, 2. \quad (15)$$

Consider the operator $A' : L_0^2(F_1) \times L_0^2(F_2) \rightarrow H(C, R)$ defined by $A'(h_1, h_2) = A'_1(h_1) + A'_2(h_2)$. Proving $N(A^\top) = L_0^2(\tilde{T}_3, \Delta_3)$ is equivalent with proving $N(A'^\top) = L_0^2(\tilde{T}_3, \Delta_3)$, where A'^\top is the adjoint of A' .

From the representation (15) it follows that the adjoint $A_j'^\top : H(C, R) \rightarrow L_0^2(F_j)$ is given by

$$A_j'^\top(V) = E(E(V(C, R)I(T_3 > C) | C, \Delta_j, T_3 > C) | T_j), \quad j = 1, 2$$

and thus $N(A'^\top) = N(A_1'^\top) \cap N(A_2'^\top)$.

Consider now a solution $VI(T_3 > C) \in H(C, R)$ satisfying $A_j'^\top(VI(T_3 > C)) = 0$, $j = 1, 2$. In order to prove $V \in L_0^2(\tilde{T}_3, \Delta_3)$ it suffices to show $I(T_3 > C)V = I(T_3 > C)\phi(C)$ for some ϕ . Using precisely the same proof as the proof of lemma 2.2 it follows that if F_j has a Lebesgue density $f_j > 0$ on its support $[0, \tau_j]$ and G has a Lebesgue density, then for any function $I(T_3 > C)\eta(C, \Delta_j) E(I(T_3 > C)\eta(C, \Delta_j) | T_j) = 0$ implies $\eta(C, 1) = \eta(C, 0)$. This proves that $E(V(C, R)I(T_3 > C) | C, \Delta_j, T_3 > C) = E(V(C, R)I(T_3 > C) | C, T_3 > C) \equiv I(T_3 > C)\phi(C)$ does not depend on Δ_j , $j = 1, 2$.

Setting $\Delta_1 = 0$ yields $I(T_3 > C)\phi(C) = E(V(C, R)I(T_3 > C) \mid C, \Delta_j, T_3 > C) = V(C, 1)I(T_3 > C)$. Now, we note that

$$P(R = m \mid \Delta_j = 1, C = c, T_3 > c) = I(m \geq j + 1, m < 4) \frac{(S_m - S_{m-1})(c)}{(S_3 - S_j)(c)}, \quad j = 1, 2.$$

Thus $E(V(C, R)I(T_3 > C) \mid C, \Delta_j = 1, T_3 > C)$ is given by

$$I(T_3 > C) \sum_{m \geq j+1, m < 4} V(C, m) \frac{(S_m - S_{m-1})(C)}{(S_3 - S_j)(C)} = I(T_3 > C)\phi(C), \quad j = 1, 2.$$

For $j = 2$, this equality gives $I(T_3 > C)V(C, 3) = I(T_3 > C)\phi(C)$. For $j = 1$, this equality gives then

$$I(T_3 > C) \left\{ V(C, 2) \frac{(S_2 - S_1)(c)}{(S_3 - S_1)(C)} + \phi(C) \frac{(S_3 - S_2)(C)}{(S_3 - S_1)(C)} \right\} = I(T_3 > C)\Phi(C)$$

so that $I(T_3 > C)V(C, 2) = I(T_3 > C)\phi(C)$. We have shown $I(T_3 > C)V(C, 1) = I(T_3 > C)V(C, 2) = I(T_3 > C)V(C, 3)$ which proves that $V = I(T_3 < C)V_1(T_3) + I(T_3 > C)\phi(C)$ for some functions V_1 and ϕ and thus that $V \in L_0^2(\tilde{T}_3, \Delta_3)$. This completes the proof.

Acknowledgements

This research was supported by a FIRST award (GM53722) from the National Institute of General Medical Sciences, National Institute of Health. The authors thank the referees and associate editor for their helpful comments.

References

- Barlow, R.E., Bartholomew, R.J., Bremner, J.M. and Brunk, H.D. (1972), *Statistical Inference under Order Restrictions*, Wiley, New York.
- Bickel, P.J., Klaassen, A.J., Ritov, Y., and Wellner, J.A. (1993), *Efficient and adaptive estimation in semi-parametric models*, Johns Hopkins University Press, Baltimore.

- Diamond, I.D. McDonald, J.W. and Shah, I.H. (1986), “Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan,” *Demography* **23**, 607–620.
- Diamond, I.D. and McDonald, J.W. (1991), “The analysis of current status data,” *Demographic Applications of Event History Analysis*, J. Trussell, R. Hankinson, and J. Tilton (eds.), Oxford: Oxford University Press.
- Dinse, G.E. and Lagakos, S.W. (1982), Nonparametric estimation of lifetime and disease onset distributions from incomplete observations, *Biometrics*, **38**, 921–932.
- Groeneboom, P.J. (1998), *Special Topics Course 593C: Nonparametric Estimation for Inverse Problems; Algorithms and Asymptotics*, Technical report # 344, Department of Statistics, University of Washington. (Related software: <http://www.stat.washington.edu/jaw/RESEARCH/SOFTWARE/software.list.html>).
- Huang, J. and Wellner, J.A. (1995), “Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I,” *Statistica Neerlandica* **49**, 153–163.
- Groeneboom, P. and Wellner, J.A. (1992), *Information bounds and nonparametric maximum likelihood estimation*, Birkhäuser Verlag, Basel.
- Jongbloed, G. (1995), *Three statistical inverse problems*, Ph.D. dissertation, Delft University of Technology, Delft.
- Kodell, R.L., Shaw, G.W. and Johnson, A.M. (1982), Nonparametric joint estimators for disease resistance and survival functions in survival/sacrifice experiments, *Biometrics*, **38**, 43–58.
- Turnbull, B.W. and Mitchell, T.J. (1984), Nonparametric estimation of the distribution of time to onset for specific diseases in survival/sacrifice experiments, *Biometrics*, **40**, 41–50
- Jewell, N.P. and van der Laan, M.J. (1995), *Generalizations of Current Status Data with Ap-*

- plications. Lifetime Data Analysis*, **1**, 101–109. (1994).
- Jewell, N.P. Malani, H.M. and Vittinghoff, E. (1994), Nonparametric estimation for a form of doubly censored data with application to two problems in AIDS, *JASA*, **89**, 7–18.
- Jewell, N.P. and Shiboski, S.C. (1990), “Statistical analysis of HIV infectivity based on partner studies,” *Biometrics*, **46**, 1133–1150.
- van der Laan, M.J., Jewell, N.P. and Peterson, D.R. (1997), Efficient estimation of the lifetime and disease onset distribution, *Biometrika* **84**, 539–554.
- Keiding, N. (1991) “Age-specific incidence and prevalence (with discussion),” *Journal of the Royal Statistical Society Ser. A*, **154**, 371–412.
- Sun, J. and Kalbfleisch, D. (1993), The analysis of current status data on point processes, *Journal of the American Statistical Association*, **88**, 1449–1454.