

Locally Efficient Estimation of the Survival Distribution with
Right Censored Data and Covariates when Collection of Data is
Delayed

MARK J. VAN DER LAAN AND ALAN E. HUBBARD

Division of Biostatistics
University of California
Berkeley, CA 94720
laan@stat.berkeley.edu
hubbard@stat.berkeley.edu

May 15, 2001

SUMMARY

For many sources of survival data, there is a delay between the recording of vital status and its availability to the analyst, and the Kaplan-Meier estimator is typically inconsistent in these situations. In this paper we identify the optimal estimation problem. As a result of the curse of dimensionality, no globally efficient nonparametric estimator exist with a good practical performance at moderate sample sizes. Following the approach of Robins & Rotnitzky (1992), given a correctly specified model for the hazard of censoring conditional on the delay process and T , we propose a closed form one-step estimator of the distribution of T whose asymptotic variance attains the efficiency bound, if we can correctly specify a lower-dimensional working model for the conditional distribution of T given the ascertainment process. The estimator remains consistent and asymptotically normal even if this latter submodel is misspecified. In particular, if we choose as working model independence between T and the ascertainment process, then the estimator is efficient when this holds and remains consistent and asymptotically linear otherwise.

Moreover, we incorporate in our data structure a covariate process that is observed during

the follow-up time and is reported with the same delays. We propose closed form locally efficient estimators of the type described above which use all the data and allow for dependent censoring.

Some key words: Asymptotically efficient; Asymptotically linear estimator; Cox proportional hazards model; Influence curve; Right censored data.

1 INTRODUCTION

1.1 Background

In many clinical trials and epidemiological studies one is often concerned with estimation of the distribution function F of a survival time T , which is subject to right-censoring. In the absence of covariates, the Kaplan-Meier estimator is the standard method for estimating the survival distribution, but it fails to be consistent when the censoring is informative, for example, when report of death is subject to delay.

To illustrate this we follow the notation of Hu & Tsiatis (1996). Let $U_1 < U_2 < \dots < U_{k-1} < T$ be the monitoring times of the subject under study. Let $A_1 < A_2 < \dots < A_{k-1}$ be the corresponding times at which these U_j s with vital status are reported and let A_k be the time at which T is reported. Thus, at time A_j we know that $T > U_j$, $j = 1, \dots, k-1$, and at time A_k we know that $T = t$ for some $t \leq A_k$. For simplicity, assume that the U_j s are reported immediately, i.e. $A_j = U_j$, but that T is reported at a possibly delayed time A_k . Let C be the time at analysis, which is assumed to be independent of T . If death is reported before the censoring time C , i.e. $A_k < C$, then the censoring variable is simply C . Suppose now that at time C death has yet to be reported, i.e. $A_k > C$, and C is between $U_{j-1} = A_{j-1}$ and $U_j = A_j$. Then we cannot be sure that T did not happen between U_{j-1} and C , since all we know is that $T > U_{j-1}$. It is common practice to set $C = U_{j-1}$ and thus to let T be right-censored at U_{j-1} . The censoring variable is now a function of A_k and thus of T , which implies that censoring is no longer independent of T . This can lead to serious bias in the Kaplan-Meier estimator, as nicely illustrated in Hu & Tsiatis (1996).

Hu & Tsiatis (1996) assumed that the process of ascertainment of vital status, i.e. the U_j s and A_j s, is observed during the follow up period. Although they wish not to assume that T is independent of the process of ascertainment, it is assumed that the follow-up time

is independent of T and the ascertainment process. Their estimator of the distribution of T is explicit and they show that it is consistent and asymptotically normal with a closed-form expression for its limiting variance. They make the estimator dependent on an unknown constant $C(x)$ which bounds the maximal reporting delay and show with simulations that, if this constant is chosen large enough, then the estimator performs well. We provide an explanation and representation of their estimator which does not depend on this constant, suggesting that this constant is purely artificial and can simply be deleted in the definition of their estimator. They state that A_1, \dots, A_k are not usually recorded, suspecting that this is due to the presumption that this additional information is not useful. Their work shows that this is misguided.

We extend the work of Hu & Tsiatis (1996) in several directions. First, we identify the optimal semiparametric information bound in their model and we construct estimators achieving these optimal bounds. Moreover, their estimator is inconsistent if censoring, i.e. the follow up time, depends on T through the ascertainment process. Therefore, it remains to construct estimators that allow censoring to depend on T through the ascertainment process. In addition, in many applications one will also observe for every subject time-independent and/or time-dependent covariates. Our estimator (1) incorporates covariates, (2) is locally efficient and (3) allows for dependent censoring.

1.2 The data structure

First, as in Hu & Tsiatis (1996), let $R(t) = I(T \leq t)$, which represents the vital status at time t . Two functions describe the collection of vital status process, and the data structure will be expressed in terms of both of them. The first function V_1 reports at time t up until when the process R has been observed: if at time t R has been observed up until time $s \leq t$, then $V_1(t) = s$. In particular, if at time t , T is already observed, then $V_1(t) = t$. Thus V_1 is an increasing function with $V_1(t) \leq t$. In the context of reporting delays as described above we have

$$V_1(t) = \begin{cases} U_j & \text{if } t \in [A_j, A_{j+1}) \\ t & \text{if } t \geq A_k, \end{cases}$$

which takes the form of an increasing step function until $t = A_k$.

Let $W(t) \in \mathbb{R}^k$, $t \in \mathbb{R}_{\geq 0}$, be a covariate process which is assumed to have the same reporting delays as does the vital status of T . The function V_1 provides us with a natural

definition of the data observed on a subject up until time t . Consider the process

$$X(t) = (R(V_1(t)), V_1(t), W(V_1(t))).$$

Thus, observing the process X up until time t corresponds to observing R , V_1 and W up until time $V_1(t)$. Let $\bar{X}(t) = \{X(s) : s \leq t\}$ represent the sample path of X up until time t . The data-analyst observes $X(t)$ in the sense that at time t the computer contains the process X up until time t , assuming censoring occurs after t .

Let $V(T)$ be the time at which T is reported. In the context of Hu & Tsiatis (1996) we have $V(T) = A_k$. Note that at time $V(T)$ the computer contains the full data $\bar{X}(V(T))$, which corresponds to observing T , $\bar{V}_1(T)$ and $\bar{W}(T)$. As a result of limited follow-up time or other reasons, one observes the process X up until the minimum of C and $V(T)$ and one knows whether this minimum is either the censoring time or $V(T)$. Thus the observed data structure can be represented as

$$Y = \left\{ \tilde{T} \equiv C \wedge V(T), \Delta \equiv I(V(T) \leq C), \bar{X}(C \wedge V(T)) \right\}.$$

We observe n independent and identically distributed observations Y_1, \dots, Y_n of Y .

Robins (1993) and Robins & Rotnitzky (1992) proposed locally efficient estimators of the distribution of T based on n observations of $(T \wedge C, \Delta = I(T \leq C), \bar{X}(T \wedge C))$ for some process X related to T . Since we can represent the data structure in the same way, we can apply their methods for construction of closed-form locally efficient estimators.

1.3 The model for Y

Let X represent the full data $\bar{X}(V(T))$. Since Y is a function of X and C , its distribution is indexed by the distribution of X and the conditional distribution of C , given X . The distribution F_X of X will be completely unspecified and we assume that the conditional distribution $G(\cdot | X)$ of C , given X , satisfies ‘coarsening at random’, in that censoring is not informative, given the observed covariates, as was originally formulated in Heitjan & Rubin (1991) and generalized in Jacobsen & Keiding (1995) and Gill, van der Laan & Robins (1997). It follows that $G(\cdot | X)$ satisfies coarsening at random if for $c < V(T)$

$$\lambda_C(c | X) = m(c, \bar{X}(c)) \text{ for some function } m \text{ of } (c, \bar{X}(c)), \quad (1)$$

where $\lambda_C(c | X)$ is the Lebesgue hazard corresponding to $G(dc | X)$ (Robins, 1993). The importance of the coarsening at random assumption in estimation of F in the presence of a time-dependent surrogate process has been argued by Robins & Rotnitzky (1992).

As a result of the curse of dimensionality, asymptotically efficient estimators, such as a smoothed nonparametric maximum likelihood estimator, perform poorly in this context. Even without covariate process W , the data structure still includes a time-dependent process V_1 , which makes any fully efficient estimator impractical. Gill, et al. (1997) show that, if (1) is the only assumption, then the model is saturated so that every regular and asymptotically linear estimator of $F(t)$ is asymptotically equivalent and thus efficient. Therefore, it is only possible to construct ad hoc sensible estimators if one makes a stronger assumption than (1); see §4. Hu & Tsiatis (1996) constructed an ad hoc estimator by assuming that C is independent of T and V_1 .

To construct our estimators, we assume a parametric or semiparametric submodel of (1) for $G(\cdot | X)$. Let

$$\lambda_C(c | X) = m_\eta(c, \bar{X}(c)) \text{ for some model } m_\eta, \eta \in \Gamma. \quad (2)$$

We model λ_C with the Cox proportional hazards model using summary measures of $\bar{X}(C)$ as time-dependent and time-independent covariates:

$$\lambda_C(c | x) = \lambda_0(c) \exp \left\{ \alpha_0^\top W_1(c) \right\}, \quad (3)$$

where $\alpha_0 \in \mathbb{R}^k$ and $W_1(c) = f(\bar{X}(c)) \in \mathbb{R}^k$ is a vector of functions of $\bar{X}(c)$. This model for censoring includes the independent censoring model of Hu & Tsiatis (1996). Finally, note that our results require that

$$Pr(\Delta = 1 | X) > 0 \text{ } F_X \text{ almost everywhere.} \quad (4)$$

1.4 Organisation of the rest of the paper

In §2, following the terminology of Robins (1993) and Robins & Rotnitzky (1992), we discuss ‘inverse probability of censoring weighted’ (IPCW) estimators of $F(t)$. First, we provide an IPCW representation of the Hu and Tsiatis estimator. Then, we propose a new IPCW estimator of F that works in more general data situations. In §3, we define a locally efficient one-step estimator in terms of our IPCW estimator plus an empirical mean of an estimator of the efficient influence curve. This curve is a function of the conditional distribution $F(t | \bar{X}(u), \tilde{T} > u)$ of T , given $\bar{X}(u), \tilde{T} > u$, and we provide a generic method for estimating this conditional distribution. The one-step estimator for the marginal data structure of Hu & Tsiatis (1996) is presented in §4. In this case there is no covariate process W , so estimation

of $F(t \mid \bar{X}(u), \tilde{T} > u)$ requires estimation of the conditional distribution of T , given $\bar{V}_1(u)$. In §5, we present a simulation study comparing the naive Kaplan-Meier estimator, the estimator proposed by Hu & Tsiatis (1996), our IPCW estimator, the locally efficient one-step estimator without a covariate and the locally efficient estimator with covariate.

In the Appendix, we state the local efficiency theorem for this one-step estimator. Assuming the ‘regularity’ conditions to be true and a correctly specified model for the censoring mechanism, the theorem states that, if we estimate the conditional distribution $F(\cdot \mid \bar{X}(u))$ consistently, then the resulting one-step estimator of $F(t)$ is efficient, and otherwise it is still consistent and asymptotically normal. This allows us to guess a working model for the conditional distribution of T , given $\bar{X}(u)$. As a result, the one-step estimator is efficient if the working model contains the true $F(\cdot \mid \bar{X}(u))$ and remains consistent and asymptotically normal if it is misspecified, that is, the estimator is locally efficient at the working model.

2 INVERSE PROBABILITY OF CENSORING WEIGHTED ESTIMATORS

2.1 The Hu and Tsiatis estimator

Let $V^*(t)$ be the earliest time at which $R(t) = I(T \leq t)$ was known. So, if $t \in (U_j, U_{j+1}]$, then $V^*(t) = A_{j+1}$ and, if $t \geq A_k$, then $V^*(t) = A_k$. Note that this definition is slightly different from our definition of $V(t)$. Let t be given. Consider now the data without covariates:

$$Z(t) = V^*(t) \wedge C, \Delta(t) = I(V^*(t) \leq C) \text{ and if } \Delta(t) = 1 \text{ we also observe } I(T \leq t).$$

A key identity is that, given (4),

$$E \left\{ \frac{I(T \leq t)\Delta(t)}{\bar{G}(Z(t))} \right\} = F(t), \tag{5}$$

where $\bar{G}(x) = pr(C \geq x)$. This identity follows directly from

$$E \{ \Delta(t) \mid T, V^*(t) \} = pr \{ C \geq V^*(t) \mid T, V^*(t) \} = \bar{G}(V^*(t)),$$

by the assumption of independence between C and $(V^*(t), T)$. This shows that the conditional expectation given T and $V^*(t)$ of the left-hand side of (5) equals $I(T \leq t)$, which suggests the following ad hoc estimator of $F(t)$:

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n \frac{I(T_i \leq t)\Delta_i(t)}{\bar{G}_n(Z_i(t))},$$

where \bar{G}_n is an estimator of \bar{G} . One could, in this case, estimate \bar{G} by using the Kaplan-Meier estimator based on n observations of $(Z(t), 1 - \Delta(t))$, where now $V^*(t)$ plays the role of the censoring variable for C . However, this does not precisely correspond to the Hu and Tsiatis estimator. To see how their estimator is an IPCW estimator, let $\bar{F}_{v^*(t)}(x) \equiv pr(V^*(t) \geq x)$ and $\bar{F}_{z(t)}(x) \equiv pr(Z(t) \geq x)$. By the independence between C and $(T, V^*(t))$,

$$\frac{1}{\bar{G}(x)} = \frac{\bar{F}_{v^*(t)}(x)}{\bar{F}_{z(t)}(x)}.$$

This suggests the following estimator of $F(t)$:

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^n \frac{\bar{F}_{v^*(t), KM}(Z_i(t))}{\bar{F}_{z(t), n}(Z_i(t))} I(T_i \leq t) \Delta_i(t), \quad (6)$$

where $\bar{F}_{v^*(t), KM}$ is the Kaplan-Meier estimator based on $Z_i(t) = V_i^*(t) \wedge C_i$, $\Delta_i(t)$ and $\bar{F}_{z(t), n}(s)$ is simply the empirical proportion of subjects with $Z_j(t)$ larger or equal than s . Hu & Tsiatis (1996) make their estimator depend on a constant $C(t)$, but setting $C(t) = \infty$ in their representation makes it equal to (6). Thus, their constant $C(t)$ appears to be artificial and unnecessary. Finally, one can generalise the Hu and Tsiatis estimator to account for dependent censoring by replacing the marginal estimate of the distribution of C by a conditional estimate of C given the covariates relevant to T . This general approach is discussed further in the next section.

2.2 A simple inverse probability of censoring weighted estimator

In this section, we define an initial estimator for our one-step estimator that weights the observed $I(T_i \leq t)$ by the correct probability of censoring. We will call this estimator the ‘simple’ estimator. As above, we exploit the following key identity to construct the simple estimator, given (4),

$$E \left\{ \frac{I(T \leq t) \Delta}{\bar{G}(\bar{T} | X)} \right\} = F(t), \quad (7)$$

where $\bar{G}(c | X) = pr(C \geq c | X)$ denotes the conditional survival function of C , given X . This identity follows directly from

$$E(\Delta | X) = pr(C \geq V(T) | X) = \bar{G}(V(T) | X),$$

which shows that the conditional expectation given X of the left-hand side of (7) equals $I(T \leq t)$. This suggest the following ad hoc estimator of $F(t)$:

$$F_n^0(t) = \frac{1}{n} \sum_{i=1}^n \frac{I(T_i \leq t) \Delta_i}{\bar{G}_n(\bar{T}_i | X_i)}, \quad (8)$$

where \bar{G}_n is an estimator of \bar{G} assuming the given model (2). Note that, by the coarsening at random assumption (1), $\bar{G}(\tilde{T} | X)$ is only a function of $Y = (\tilde{T} = C \wedge V(T), \Delta = I(V(T) \leq C), \bar{X}(\tilde{T}))$, so that $F_n^0(t)$ indeed only depends on Y_1, \dots, Y_n . If one assumes the Cox proportional hazard model for G , then one can use standard software to obtain the maximum partial likelihood estimator of the baseline hazard and the regression coefficients. In particular, if C is conditionally independent of (T, V_1) , given W , then one only has to include time-dependent and/or time-independent covariates extracted from W . If one assumes that C is completely independent of X as in Hu & Tsiatis (1996), then one can consistently estimate \bar{G} with the Kaplan-Meier estimator based on the n observations of $(\tilde{T} = C \wedge V(T), 1 - \Delta)$, where $V(T)$ now plays the role of the censoring variable for C . In the case that $G(\cdot | X) = G(\cdot)$ and \bar{G}_n is the Kaplan-Meier estimator, simulations in §4 show that this estimator (8) is competitive with the estimator proposed in Hu & Tsiatis (1996). Crucial to common types of disease survey data, the simple estimator of $F(t)$ only requires observing $(\tilde{T} = C \wedge V(T), \Delta, \Delta T)$, and does thus not require observing the whole delay process V_1 . Thus, we can apply the simple estimator to a dataset which only keeps track of the delays in reporting death, whereas the Hu & Tsiatis (1996) estimator cannot be applied to this setting.

Of course, if censoring depends on T through the observed covariates or through the observed ascertainment process according to a Cox proportional hazards model, then the independent censoring model of Hu & Tsiatis (1996) fails to hold and their estimator will be inconsistent. However, using G estimated by Cox regression, both the simple estimator and the generalisation of the Hu and Tsiatis estimator will still be consistent. Finally, if there is no reporting delay ($V(T) = T$) and no relevant covariate, then the ICPW estimators reduce to the Kaplan-Meier estimator of $F(t)$ based on the n independent and identically distributed observations (\tilde{T}, Δ)

3 THE LOCALLY EFFICIENT ONE-STEP ESTIMATOR

In this section, we construct a locally efficient one-step estimator by adding to the estimator in $F_n^0(t)$ in (8) an estimate of the empirical mean of the efficient influence function. Our representation of the efficient influence function has two pieces. The first is the influence function of $F_n^0(t)$ using the known G :

$$IC_0(Y | G, F(t)) \equiv \frac{I(T \leq t)\Delta}{\bar{G}(\tilde{T} | X)} - F(t).$$

The second piece is a projection of IC_0 on the tangent space, which is a function IC_{nu}^* of Y :

$$IC_{nu}^*(Y | F_X, G) = - \int F(t | \bar{X}(u), \tilde{T} > u) \frac{dM(u)}{G(u | X)}, \quad (9)$$

where

$$dM(u) \equiv I(C \in du, \Delta = 0) - \Lambda_C(du | X)I(\tilde{T} > u)$$

and $F(t | \bar{X}(u), \tilde{T} > u)$ is the conditional probability that $T \leq t$, given $\bar{X}(u)$ and $\tilde{T} > u$. It is important to emphasise that, for any function $H(u, \bar{L}(u))$, the stochastic integral

$$\int H(u, \bar{X}(u))dM(u) = H(C, \bar{X}(C))(1 - \Delta) - \int_0^{\tilde{T}} H(u, \bar{X}(u))\Lambda_C(du | X)$$

is a function of the observed data Y because $\lambda_C(u | X)$ depends on X only through $\bar{X}(u)$.

In a technical report by the authors it is shown that the efficient influence curve for estimation of $F(t)$ is

$$IC^*(Y | F_X, G, F(t)) \equiv IC_0(Y | G, F(t)) - IC_{nu}^*(Y | F_X, G).$$

Let $IC_{nu}^*(\cdot | F_{X,n}, G_n)$ be an estimator of $IC_{nu}^*(\cdot | F_X, G)$ obtained by substitution of estimators of $F(t | \bar{X}(u), \tilde{T} > u)$ and G , where $G_n = G_{\eta_n}$ suppresses, in the notation, the dependence on the parameter η . In the next subsection we propose an estimator of $F(t | \bar{X}(u), \tilde{T} > u)$. Note that IC_{nu}^* depends on G also through the measure $dM(u)$. One can now estimate $F(t)$ with the one step estimator

$$F_n^1(t) = F_n^0(t) + \frac{1}{n} \sum_{i=1}^n \left\{ IC_0(Y_i | G_n, F_n^0(t)) - IC_{nu}^*(Y_i | F_X^n, G_n) \right\}, \quad (10)$$

where F_n^0 is the IPCW-estimator defined in (8). Let $Pf \equiv \int f dP$ for a probability measure P and measurable function f . Let P_n be the empirical distribution function so that $P_n f = 1/n \sum_{i=1}^n f(Y_i)$. Note that $P_n IC_0(\cdot | G_n, F_n^0(t)) = 0$ and therefore one can delete the IC_0 -term in (10). We chose to retain the IC_0 -term in order to show that $F_n^1(t)$ is just the classical one-step estimator as defined in Bickel, et al. (1993, p. 395).

The one-step estimator $F_n^1(t)$ depends on estimates of G and $F(t | \bar{X}(u), \tilde{T} > u)$. Theorem 2 in the Appendix can be used as a template to prove the local efficiency result for the one-step estimator $F_n^1(t)$. In §4 we will apply this theorem to a specific one-step estimator for the marginal data structure of Hu & Tsiatis (1996). Generally speaking, Theorem 2 shows that, if G_n is estimated consistently, then the estimator $F_n^1(t)$ is asymptotically linear, and, if IC_{nu}^* , i.e. $F(t | \bar{X}(u), \tilde{T} > u)$, is also estimated consistently, then $F_n^1(t)$ is even asymptotically

efficient. The protection against inconsistent estimation of this conditional probability comes from the fact that $\int H(u, \bar{X}(u))dM(u)$ has conditional mean zero, given X , for any function H , because $E(dM(u) | X, C > u) = 0$.

3.1 Estimation of IC_{nu}^* using the conditional expectation representation

The idea of representing the conditional probability $F(t | \bar{X}(u), \tilde{T} > u)$ as a regression of a random variable $O_G(Y)$ on observed covariates is due to Robins (1993) and Robins & Rotnitzky (1992). It has a powerful application in estimating $F(t | \bar{X}(u), \tilde{T} > u)$. First,

$$F(t | \bar{X}(u), \tilde{T} > u) = E \left(I(T \leq t) \Delta \frac{\bar{G}(u | X)}{\bar{G}(\tilde{T} | X)} | \bar{X}(u), \tilde{T} > u \right). \quad (11)$$

Assume $\lambda_C(c | x)$ satisfies (3). We can estimate α_0 with the partial likelihood equations only involving α and the corresponding estimator of the baseline hazard, which itself is a simple function of this partial likelihood estimator and the data (Andersen, et al., 1993). This yields an estimator of the martingale M and $\bar{G}(u | X)$. Note also that $d\hat{M}$ is only nonzero at the observed censoring times. If we use the representation (11), then it remains to estimate the conditional expectation of a random variable

$$O_G \equiv \frac{I(T \leq t) \Delta \bar{G}(u | X)}{\bar{G}(\tilde{T} | X)},$$

given $\bar{X}(u), \tilde{T} > u$ at a u corresponding with an observed censoring variable. Given an estimate G_n of G and for a given t , O_{G_n} is an observed random variable. Consequently, for every u corresponding to an observed C_i , one can carry out a parametric or nonparametric regression estimation of O_{G_n} on one or a number of relevant, for T , summary measures Z_1, \dots, Z_k of $\bar{X}(u)$, only using the observations with $\tilde{T}_i > u$.

If one assumes that V_1 is independent of C and T , then one only has to select covariates from $\bar{W}(V_1(u))$. As particular summary measure of $\bar{W}(V_1(u))$ one can take a prediction of T only based on $\bar{W}(V_1(u))$. In many practical situations such a covariate is known to have a monotone effect on the distribution of T so that monotonic regression will be most appropriate.

3.2 Truncation

In this section, we consider an alteration of the data that increases the small sample performance of the locally efficient estimator. Consider $F_n^1(t)$ for a given t . Assume that $V(T) \leq T + C_0$ for some $C_0 < \infty$ with probability tending to 1. From the representation

(9) it follows trivially that the efficient influence curve $IC^*(Y | F_X, G, F(t))$ is statistically dependent on $Y = (\tilde{T}, \Delta, \bar{X}(\tilde{T}))$ only through $\bar{X}(t + C_0)$; this is a direct consequence of the fact that, if $u > t + C_0$, then $P(T \leq t | \bar{X}(u), \tilde{T} > u) = 0$. In other words, information on the subject collected after $t + C_0$ is no longer relevant for the estimation of $F(t)$.

The regression method represents $P(T \leq t | \bar{X}(u), \tilde{T} > u)$ as a regression of observed O_G on observed covariates. However, note that, even when $u > t + C_0$, there is no guarantee that the estimator of this regression equals zero, as it should. This creates unnecessary variability of the estimator since the estimator $F_n^1(t)$ will now use \bar{G}_n in the tail of the distribution. Therefore, before carrying out the regression estimation procedure, those Y_i with $\tilde{T}_i > t + C_0$ should be replaced by $(\tilde{T}_i = t + C_0, \Delta_i = 1, \bar{X}_i(t + C_0))$. As information after $t + C_0$ is of no use for estimating $F(t)$, truncating the data in this way does not lead to a loss in efficiency. If C_0 is not known, then one just uses a reasonable C_0 . An incorrectly small C_0 , such that $pr\{V(T) > T + C_0\} > 0$, will lead to a loss in efficiency, but the estimator still remains consistent and asymptotically linear, and typically results in a more robust estimator. On the other hand, if one chooses C_0 very large, then the regression estimate will be unnecessarily unstable.

4 THE LOCALLY EFFICIENT ESTIMATOR IN THE MARGINAL MODEL

In this section we discuss the special case that there is no relevant covariate process, which corresponds to the marginal data structure studied in Hu & Tsiatis (1996). First, $Y = (\tilde{T} = C \wedge T, \Delta = I(V(T) \leq C), \bar{X}(\tilde{T}))$, where $\bar{X}(\tilde{T}) = (\bar{V}_1(\tilde{T}), \bar{R}(V_1(\tilde{T})))$. If C is independent of V_1 , as assumed by Hu & Tsiatis (1996), then we can estimate $\bar{G}(u | X) = \bar{G}(u)$ with the Kaplan-Meier estimator of G based on the n identically and independently distributed observations of $(\tilde{T}, 1 - \Delta)$. Alternatively, if there is reason to expect dependence between censoring and the delay process, then one can model $\bar{G}(u | X)$ with the Cox proportional hazards model with time-dependent covariates extracted from $\bar{V}_1(u)$. Below, we focus on the independent censoring model of Hu & Tsiatis (1996) with G_n being the Kaplan-Meier estimator. The initial estimator $F_n^0(t)$ is given by

$$F_n^0(t) = \frac{1}{n} \sum_{i=1}^n \frac{I(T_i \leq t) \Delta_i}{\bar{G}_n(\tilde{T}_i)}.$$

Furthermore,

$$pr(T \leq t \mid \bar{X}(u), \tilde{T} > u) = pr(T \leq t \mid \bar{V}_1(u), \tilde{T} > u),$$

which yields a IC_{nu}^* given by

$$IC_{nu}^*(Y \mid F_X, G) = - \int P(T \leq t \mid \bar{V}_1(u), \tilde{T} > u) \frac{dM(u)}{\bar{G}(u)}.$$

This proves that the efficient influence curve $IC^*(Y \mid F_X, G, F(t))$ in the model of Hu & Tsiatis (1996) is given by

$$IC^*(Y) = \frac{I(T \leq t)\Delta}{\bar{G}(V(\tilde{T}))} - F(t) + \int pr(T \leq t \mid \bar{V}_1(u), \tilde{T} > u) \frac{dM(u)}{\bar{G}(u)}.$$

Thus the variance of this function of Y at a given F_X, G provides the optimal limiting variance. A necessary condition for construction of a fully efficient estimator, i.e. an estimator which achieves this bound at every F_X, G , is to estimate IC^* consistently at every F_X, G (Bickel, et al., 1993). This requires nonparametric estimation of the conditional distribution of T , given the time-dependent ascertainment process $\bar{V}_1(u)$. The curse of dimensionality renders this impossible.

However, according to Theorem 2 we can assume a submodel of $pr(T \leq t \mid \bar{V}_1(u), \tilde{T} > u)$ and estimate this conditional probability accordingly, while being protected against misspecification. The resulting one-step estimator will now be locally efficient at this submodel and still be consistent and asymptotically linear at every other distribution. A submodel of interest is to assume that T is independent of the ascertainment process. Then $pr(T \leq t \mid \bar{X}(u), \tilde{T} > u) = pr(T \leq t \mid \tilde{T} > u)$ so that $IC_{nu}^*(F_X^1, G)$, at an F_X^1 with T and V_1 independent, is given by

$$- \int pr(T \leq t \mid \tilde{T} > u) \frac{dM(u)}{\bar{G}(u)}.$$

For a given u , we can estimate $F(t \mid \tilde{T} > u) = pr(T \leq t \mid \tilde{T} > u)$ with the initial estimator F_n^0 based on the subsamples with $\tilde{T}_i > u$, $i = 1, \dots, n$. Alternatively, using the conditional expectation representation (11), we can estimate, for a given u , $pr(T \leq t \mid \tilde{T} > u)$ with the average of

$$Z_i = \frac{I(T_i \leq t)\Delta_i \bar{G}(u)}{G_n(\tilde{T}_i)} \text{ using the subsample with } \tilde{T}_i > u.$$

Let $F_n(t \mid \tilde{T} > u)$ be the estimator of $F(t \mid \tilde{T} > u)$. The resulting one-step estimator (10) is

$$F_n^1(t) = F_n^0(t) + \frac{1}{n} \sum_{i=1}^n \int F_n(t \mid \tilde{T} > u) \frac{dM_{i,G_n}(u)}{G_n(u)},$$

where we used the notation M_{i,G_n} to stress the dependence of M on G_n and that M is a function of Y . By verifying the conditions of Theorem 2 it is shown in our technical report under weak conditions that $F_n^1(t)$ is asymptotically linear with influence curve $IC_0(Y | G, F(t)) - \int F(t | \tilde{T} > u) \frac{dM(u)}{G(u)}$. In particular, this implies that $F_n^1(t)$ is asymptotically efficient if T is independent of V_1 . This result is stated in the following theorem.

Theorem 4.1 *Let t be given and let $C_0 < \infty$ be such that $\text{pr}(V(T) < T + C_0) = 1$. Assume that $\text{pr}(V(T) > t + C_0) > \delta > 0$ and $\bar{G}(t + C_0) > \delta > 0$ for some δ . Moreover, assume that the estimator $F_n(t | \tilde{T} > u)$ is such that*

$$u \rightarrow F_n(t | \tilde{T} > u)$$

has variation smaller than a fixed $M < \infty$ with probability tending to 1. Then $F_n^1(t)$ is consistent and asymptotically linear with influence curve

$$IC(Y | F_X, G) = \frac{I(T \leq t)\Delta}{\bar{G}(\tilde{T})} - F(t) + \int F(t | \tilde{T} > u) \frac{dM(u)}{G(u)}.$$

If T is independent of the ascertainment process V_1 , then $F_n^1(t)$ is asymptotically efficient.

Theorem 1 implies under weak conditions that

$$\sqrt{n} \{F_n^1(t) - F(t)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n IC(Y_i | F_X, G) + o_P(1).$$

Thus the left-hand side is asymptotically normally distributed with mean zero and variance given by $\sigma^2 = \text{var}\{IC(Y | F, G)\}$. This variance can be estimated with

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n IC^2(Y_i | F_n, G_n),$$

where $IC(Y_i | F_n, G_n)$, $i = 1, \dots, n$, was already needed to compute $F_n^1(t)$. Now, a 0.95-asymptotic confidence interval is given by

$$F_n^1(t) \pm 1.96\hat{\sigma}/\sqrt{n}.$$

5 SIMULATION RESULTS

A simulation study was performed to examine the performance of the locally efficient estimators relative to that of the competing estimators. For each simulation we computed the ‘naive’ Kaplan-Meier estimate of survival that uses the last monitoring time before follow-up as the censoring time for those individuals who are censored. We compared the mean

squared errors of four estimators to this Kaplan-Meier estimator: the estimator proposed by Hu & Tsiatis (1996), our simple IPCW estimator (8), the locally efficient one-step estimator without a covariate and the locally efficient estimator with a covariate $W = T + 0.1e$, where $e \sim N(0, 1)$. It is correctly assumed that T is independent of the censoring time C . Thus, for the locally efficient estimator, we must estimate

$$E \left(\frac{I(T \leq t) \Delta \bar{G}(u)}{\bar{G}(\tilde{T})} \mid \bar{X}(u), \tilde{T} > u \right). \quad (12)$$

Let G_n be the Kaplan-Meier estimator of G . Recall that, for the case without covariates, for a fixed u this expectation is estimated using an average of the random variables

$$Z_i = \frac{I(T_i \leq t) \Delta_i \bar{G}_n(u)}{\bar{G}_n(\tilde{T}_i)}$$

based on the subsample with $\tilde{T}_i > u$. When implementing the locally efficient estimator with the covariate, we estimated the expectation as a smooth regression with the supersmoother as presented in a Stanford Univeristy technical report of J.H. Friedman (1984) (see also Härdle, 1993) of the random variables Z_i against the covariate W_i .

We performed both the simulations that Hu & Tsiatis (1996) used to compare their estimator with the Kaplan-Meier estimator, but we report only the first of them: for the second simulation, the Kaplan-Meier estimate that uses the last available monitoring time as the censoring time for censored observations is nearly consistent and so there is little to be gained by accounting for the delay. In their second simulation, Hu & Tsiatis (1996) reported a large potential reduction in bias over the naive Kaplan-Meier. This discrepancy could have occurred if they used the follow-up time rather than the last available monitoring time as the censoring time.

In our simulation, the alive status of all individuals was recorded immediately after patients' visits to the hospital; that is, $U_j = A_j$, $j = 1, \dots, k - 1$. Moreover, the failure time T was generated from an exponential distribution with a mean of one year, independent of the follow-up time C . The entry-times of the patients were uniformly distributed over a 2 year period: $I \sim Un(0, 2)$. The time of analysis is at 2 years and the censoring time is simply the follow-up time $C = 2 - I$. The visit times were generated from a Poisson process with mean inter-arrival time of six months and death was reported immediately; note that the data analyst does not know this, but only knows that T is larger than the last monitoring time. Since T is independent of the ascertainment process, our proposed locally efficient estimator is asymptotically efficient in this simulation. We used a sample size of 100 and performed

625 replicates.

The results suggest that the simple and the locally efficient estimator without covariates have similar performance relative to the biased Kaplan-Meier estimator. The locally efficient estimator with covariate $T + 0.1e$ gains efficiency over the competing estimators. However, as greater error is added to T , the performance converges to that of the other estimators. More significantly, the increase in efficiency occurs even though the procedure for estimating the conditional expectation (12) ignores the fact that $\bar{X}(u) = T + 0.1e$. If the covariate is T itself, i.e. $e = 0$, and we estimate the expectation which now equals $I(T \leq t)$ consistently, then the one-step estimator is asymptotically equivalent to the empirical distribution based on T_1, \dots, T_n . The locally efficient estimator works best when there are strong predictive covariates and the expectation is fitted well, but even with a misspecified fit the estimator is consistent and asymptotically normal. Thus, one can gain the most efficiency by having strong covariates and fitting the correct parametric model, but even when neither circumstance exists, one still rarely loses by trying. This is confirmed in a more extensive simulation study of locally efficient estimation in right-censored data models to be reported elsewhere.

ACKNOWLEDGEMENT

This research was supported by a FIRST award from the National Institute of General Medical Sciences, National Institute of Health. We greatly appreciate the helpful comments of the reviewer, particularly for the insight that the Hu and Tsiatis estimator can be represented as an IPCW estimator.

APPENDIX

Recall the notation $Pf = \int f(x)dP(x)$.

Theorem 5.1 *Let t be given. Let $C_0 < \infty$ be such that $V(T) \leq T + C_0$ with probability 1. Consider the one-step estimator $F_n^1(t)$. We assume the following:*

- (i) $\bar{F}(t + C_0) > \delta$ and $\bar{G}(t + C_0 | X) > \delta F_X$ almost everywhere for some $\delta > 0$;
- (ii) $IC^*(\cdot | F_{X,n}, G_n, F_n^0(t))$ falls in a $P_{F_X, G}$ -Donsker class with probability tending to 1;
- (iii) for some distribution F_X^1 we have

$$\|IC^*(\cdot | F_{X,n}, G_n, F_n^0(t)) - IC^*(\cdot | F_X^1, G, F(t))\|_{P_{F_X, G}} \rightarrow 0$$

in probability;

(iv) if for any G_1 we define

$$\Phi(G_1) = P_{F_X, G} \{ IC^*(\cdot | F_X^1, G_1, F(t)) \},$$

then

$$\begin{aligned} & P_{F_X, G} \{ IC^*(\cdot | F_X, n, G_n, F_n^0(t)) - IC^*(\cdot | F_X, n, G, F_n^0(t)) \} \\ &= \Phi(G_n) - \Phi(G) + o_P(1/\sqrt{n}); \end{aligned}$$

(v) $\Phi(G_n)$ is an asymptotically efficient estimator of $\Phi(G)$ for a model containing the true G with tangent space $T_2(P_{F_X, G})$.

Then $F_n^1(t)$ is asymptotically linear with influence curve given by

$$IC \equiv \Pi(IC^*(\cdot | F_X^1, G, F(t)) | T_2^\perp(P_{F_X, G})).$$

In particular, if $IC_{nu}^*(F_X^1, G) = IC_{nu}^*(F_X, G)$, then $F_n^1(t)$ is asymptotically efficient.

For the case that $G(c | x)$ is modelled with Cox proportional hazards the explicit form of $\Pi(\cdot | T_2)$ is given in Robins (1996). For the case that $G(c | x) = G(c)$, we have that

$$IC = \int E \left\{ F(t | \bar{X}(u), \tilde{T} > u) - F_1(t | \bar{X}(u), \tilde{T} > u) | \tilde{T} = u, \Delta = 0 \right\} \frac{dM(u)}{\bar{G}(u)}.$$

REFERENCES

- ANDERSON, P.K., BORGAN, O., GILL, R.D. & KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.
- BICKEL, P.J., KLAASSEN, A.J., RITOV, Y. & WELLNER, J.A. (1993). *Efficient and Adaptive Inference in Semi-parametric Models*. Baltimore: Johns Hopkins University.
- GILL, R.D., VAN DER LAAN, M.J. & ROBINS, J.M. (1997). Coarsening at random: Characterizations, Conjectures and Counter-Examples. In *Proceedings of the First Seattle Symposium in Biostatistics 1995*, Ed. D.Y. Lin and T.R. Fleming, pp. 255–294. New York: Springer Lecture Notes in Statistics.
- HÄRDLE, W. (1993). *Applied Nonparametric Regression*. Econometric Society Monographs, Cambridge University Press.
- HASTIE, T.J. & TIBSHIRANI, R.J. (1990). *Generalized Additive Models*. London: Chapman and Hall.
- HEITJAN, D.F. & RUBIN, D.B. (1991). Ignorability and coarse data. *Ann. of Statist.* **19**, 2244–53.

- HU, P.H. & TSIATIS, A.A. (1996). Estimating the survival function when ascertainment of vital status is subject to delay. *Biometrika* **83**, 371–80.
- JACOBSEN, M. & KEIDING, N. (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Statist.* **23**, 774–86.
- RITOV, Y. & WELLNER, J.A. (1988). Censoring, martingales, and the Cox model. *Contemp. Math.* **80**, 191–219. Providence, R.I.: American Mathematical Society.
- ROBINS, J.M. & ROTNITZKY, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers In *Aids Epidemiology: Methodological Issues*, Ed. N.P. Jewell, Dietz, K. and V.T. Farewell, pp. 297-331. Boston: Birkhäuser.
- ROBINS, J.M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. *Pro. Biopharm. Sec., Am. Stat. Assoc.*, pp. 24–33.
- ROBINS, J.M. (1996). Locally efficient median regression with random censoring and surrogate markers. In *Lifetime Data: Models in Reliability and Survival Analysis*, Ed. N.P. Jewell, A.C. Kember, Ting Lee, M. and Whitmore, G.A., pp. 263-74. Boston: Kluwer.
- VAN DER VAART, A.W. & WELLNER, J.A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer Verlag.

Table 1: *Relative Mean Squared Error for estimation of $S(t)$ relative to the Kaplan-Meier estimator, $n = 100$, based on 625 replicates. The covariate W is $T + 0.1e$, where $e \sim N(0, 1)$.*

t	$S(t)$	KM/Hu,Tsiatis	KM/Simple	KM/1-step	KM/1-step, W
0.1	0.90	1.8	1.8	1.8	1.8
0.2	0.80	2.2	2.2	2.2	2.3
0.4	0.70	2.6	2.6	2.6	2.8
0.5	0.60	2.9	2.9	2.9	3.3
0.7	0.50	2.9	2.9	3.0	3.4
0.9	0.40	3.1	3.1	3.1	4.0
1.2	0.30	3.4	3.4	3.5	4.5
1.6	0.20	3.2	3.2	3.3	4.0