University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2002 *Paper* 114

Bivariate Current Status Data

Mark J. van der Laan* Nicholas P. Jewell†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, jewell@uclink.berkeley.edu

Bivariate Current Status Data

Mark J. van der Laan and Nicholas P. Jewell

Abstract

In many applications, it is often of interest to estimate a bivariate distribution of two survival random variables. Complete observation of such random variables is often incomplete. If one only observes whether or not each of the individual survival times exceeds a common observed monitoring time C, then the data structure is referred to as bivariate current status data (Wang and Ding, 2000). For such data, we show that the identifiable part of the joint distribution is represented by three univariate cumulative distribution functions, namely the two marginal cumulative distribution functions, and the bivariate cumulative distribution function evaluated on the diagonal. The EM algorithm can be used to compute the full nonparametric maximum likelihood estimator of these three univariate cumulative distribution functions; however, we show that smooth functionals of these univariate cumulative cdfs can be efficiently estimated with easy to compute nonparametric maximum likelihood estimators (NPMLE), based on reduced data consisting of univariate current status observations. We use these univariate current status NPMLEs to obtain both a test of independence of the two survival random variables, and a test of goodness of fit for the copula model used in Wang & Ding (2000). Finally, we extend the data structure by allowing the presence of covariates, possibly time-dependent processes that are observed until the monitoring time C. We show that applying the locally efficient estimator, developed in van der Laan and Robins (1998), to the reduced univariate current status data yields locally efficient estimators.

1 Introduction

Consider a study in which interest focuses on the bivariate distribution F of two positive random variables (T_1, T_2) which cannot be directly measured. Rather, for each individual, we observe, at a random monitoring, or censoring, time, C, whether T_j exceeds C or not for each $j = 1, 2$. That is, on each subject, we observe:

$$(C, \Delta_1 \equiv I(T_1 \leq C), \Delta_2 \equiv I(T_2 \leq C)).$$

Following Wang & Ding (2000), we call this data structure bivariate current status data since it generalizes the well-known current status data structure $(C, I(T \leq C))$ for a univariate survival time T. It is assumed here that C is independent of $\vec{T} = (T_1, T_2)$, although this is weakened later in the paper when covariates are present. The density of this observed data structure, conditional on the monitoring time $C = c$, is:

$$\begin{aligned} p(\delta_1, \delta_2 | C = c) &= F_3(c)^{\vec{\delta}=(1,1)}(1 + F_3 - F_1 - F_2)(c)^{\vec{\delta}=(0,0)} \\ &\quad \times (F_1 - F_3)(c)^{\vec{\delta}=(1,0)}(F_2 - F_3)^{\vec{\delta}=(0,1)}, \end{aligned}$$

where $F_1(t) = P(T_1 \leq t)$, $F_2(t) = P(T_2 \leq t)$ and $F_3(t) = P(T_1 \leq t, T_2 \leq t)$. It follows that only the three univariate cdf's F_1, F_2 and F_3 are identifiable. Although the complete bivariate distribution, F, is not identified, the dependence measure $F_3 - F_1 F_2$ is identifiable from the data, so that assessment of independence of T_1 and T_2 is possible.

Here, we consider estimation of F_1, F_2, F_3 and smooth functionals of these marginal cumulative distribution functions of the type $\mu_j(r) = \int r(s)\{1 - F_j(s)\}ds$ for a given function r, $j = 1, 2, 3$. Note that, with $R(x) = \int_0^x r(t)dt$, by integration by parts, we have:

$$\int_0^\tau r(t)(1 - F_j(t))dt = R(t)\ (1 - F_j(t))|_0^\tau + \int_0^\tau R(t)dF_j(t),$$

for any τ (including $\tau = \infty$). Hence if $\lim_{t \to \tau}(1 - F_j(t))R(t)$ is zero or known, then an estimate of $\int_0^\tau r(1 - F_j)dt$ provides us with an estimate of $\int_0^\tau RdF_j$. In particular, with this condition, if $r(t) = 1$ and the support of F_j is in $[0, \tau]$, then $\mu_j(r) = ET_j$, and, if $r(t) = kt^{k-1}$, then $\mu_j(r) = ET_j^k$, $j = 1, 2, 3$. Moreover, by setting $r(t) = K(\{t - t_0\}/h)/h$ for some kernel K and bandwidth h, an estimator of $\mu_j(r)$ provides a smooth estimator of $S_j(t_0) = 1 - F_j(t_0)$.

1.1 Motivating Examples

Many examples of univariate current status data yield related bivariate current status structures. Thus, for example, serial-sacrifice carcinogenicity experiments of a single occult non-lethal tumor provide simple examples of current status data. Such experiments, with the possibility of occult non-lethal tumors at two different sites (e.g. liver and brain), then yield bivariate current status data.

Similarly, HIV transmission studies of the partners of HIV-infected index cases often produce current status information on the time, or number of contacts, between infection of the index case and the partner, since the latter event usually does not lead to clinically observed symptoms (Jewell and Shiboski, 1990). In this case, there may be additional cross-sectional information available at the monitoring time such as the disease status of the index case. If T_1 is the time to infection of the partner, and T_2 is the time to diagnosis of AIDS for the index case, both measured from the date of infection of the index case, the random variable (T_1, T_2) is bivariate current status data, assuming that only whether or not the index case has been diagnosed with AIDS is measured at the monitoring time C. Here, association between T_1 and T_2 may suggest greater infectivity when the index case suffers from rapidly progressing HIV disease.

A quite different example arises in twin pair studies in genetics where observed phenotypes are the ages at onset of a specific disease. For conditions such as Alzheimer's disease, the exact age of onset is usually imprecise even when a definitive diagnosis is available. If T_j is the age of onset for the jth twin, then in such cases only bivariate current status information is observed for (T_1, T_2), where the monitoring time C is here the common age of the twins at observation. Interest may focus on the strength of association between T_1 and T_2 for both non-identical and identical twins.

In all of these examples, it may be possible to measure time-independent and time-dependent covariate processes up till time C, in addition to (C, Δ_1, Δ_2). Denote such processes by $\bar{L}(C) = \{L(s) : s \leq C\}$, where L may be of high dimension. For example, in

the carcinogenicity example noted above, suppose that study mice are randomly allocated to dose groups, a fixed covariate. In addition, daily measurements of the weight of each mouse, a time-dependent covariate, are taken prior to sacrifice. Let $L(u)$ represent the measurements taken at time u, including the weight at time u and dose. We only observe the covariate process up to time C: $\overline{L}(C) = \{L(u) : 0 < u < C\}$. Thus, for each mouse, $Y = (C, \Delta_1 = I(T_1 \leq C), \Delta_2 = I(T_2 \leq C), \overline{L}(C))$ is observed. Accommodating the effects of covariates is not only of interest in terms of their relationship to (T_1, T_2), but also allows for the possibility of choosing monitoring times that depend on the observed covariate processes, as we note in Section 1.2.

1.2 Outline

In Section 2, we consider the nonparametric maximum likelihood estimators (NPMLE) of (F_1, F_2, F_3), computed via the EM algorithm. We note that easy to compute estimators of F_j, $j = 1, 2, 3$, are available. With $\Delta_3 = \Delta_1\Delta_2$, $F_j(t)$ can be represented in terms of a monotonic regression of Δ_j on C since $F_j(t) = E(\Delta_j \mid C = t)$, for $j = 1, 2, 3$. This suggest the estimator $F_{jn}(t)$ of F_j that minimizes $\sum_{i=1}^{n}(\Delta_{ji} - F_j(C_i))^2$ over all distribution functions F_j. The solution of this problem can be computed using the rapid pool-adjacent-violators algorithm (PAVA, see Barlow *et al.* 1972). This estimator happens to correspond with the NPMLE based on the reduced data (C, Δ_j). From Groeneboom & Wellner (1992), it follows that these reduced data NPMLE's converge, under appropriate conditions, at rate $n^{-1/3}$, to known asymptotic distributions. In spite of the simplicity of these three reduced data NPMLE's relative to the full NPMLE, it is shown, in §3, that, at most data generating distributions, the reduced data NPMLE's yield efficient estimators of smooth functionals of (F_1, F_2, F_3). For estimation of smooth functionals of F_j, we thus recommend these simple estimators instead of the more complex full NPMLE. We doubt whether the full NPMLE of F_j has better finite sample performance than the simple estimators, F_{jn}, $j = 1, 2, 3$.

The results for smooth functionals are exploited, in §3.1–3.3, to construct (i) simple to compute tests of independence of T_1 and T_2, and (ii) a goodness of fit test for the semipara-

metric copula model for the bivariate distribution function F, assumed by Wang & Ding (2000).

Finally, in §4, we briefly describe locally efficient estimators for the extended data structure $Y = (C, \Delta_1, \Delta_2, \bar{L}(C))$ that includes observation of covariate processes up to time C. The assumption of independence between $\vec{T}$ and C is now assumed *conditional* on the observed $\bar{L}(C)$. This therefore allows dependence between the monitoring time C and the T_j's that arises solely through $\bar{L}(C)$. To illustrate the importance of this extension, consider a mouse tumorigenicity experiment designed to estimate the distributions of time to development of liver adenoma and time to development of brain tumor, and dependence between these two onset times. Suppose that $L(u)$ includes weight at time u, and that for each individual $Y = (C, \Delta_1 = I(T_1 \leq C), \Delta_2 = I(T_2 \leq C), \overline{L}(C))$ is observed. A reasonable monitoring scheme is to increase the 'hazard' of monitoring shortly after a mouse begins to lose weight, since if the sacrifice time is closer to the time of tumor onset then more efficient estimation is possible. This monitoring scheme introduces dependence between C and $\vec{T}$ and estimators that ignore this dependence will be biased. Collecting information on a surrogate process, and allowing the monitoring time to depend on it, is a superior design to experiments that require independent censoring, and thus can be used to improve estimation.

1.3 Previous Work and Comparison with our Results

Previous work and examples of univariate current status data can be found in Diamond, *et al.* (1986), Jewell & Shiboski (1990), Diamond & McDonald (1991), Keiding (1991), Sun & Kalbfleisch (1993), among others. In its nonparametric setting, it is also known as interval censoring, case I (Groeneboom & Wellner, 1992).

For a single random variable T, the NPMLE of the distribution function, F, of T, based on current status data, is the pool-adjacent-violators estimator for the monotone regression $F(t) = E(\Delta \mid C = t)$ of Barlow *et al.* (1972), where $\Delta = I(T \leq C)$ is the current status indicator at time C. The asymptotic distribution of this estimator has been analyzed by Groeneboom & Wellner (1992), and efficiency of the NPMLE of smooth functionals of F

(such as its mean and variance) has been proved by Groeneboom & Wellner (1992), van de Geer (1994), and Huang & Wellner (1995). Estimation of regression coefficients, associated with fixed covariates, when survival time is subject to current status observation, has been considered by several authors including Rabinowitz, Tsiatis & Aragon (1995), and Huang (1996).

The addition of time-dependent covariates to the data structure is considered in van der Laan & Robins (1998). They develop locally efficient estimators of smooth functionals of the distribution of T. By incorporating information on the process, $\overline{L}(C)$, their estimators are guaranteed both (i) to be more efficient than the NPMLE that ignores data on $\overline{L}(C)$, and (ii) to remain consistent and asymptotically normal, whatever the joint distribution of $(T, \overline{L})$. The NPMLE that incorporates data on $\overline{L}(C)$ fails to attain these goals, when L has high dimension, because of the curse of dimensionality (Robins & Ritov, 1997).

Wang & Ding (2000) were the first to consider bivariate current status data. To avoid identifiability issues, they assumed a semiparametric copula model for the bivariate distribution, parametrizing the complete bivariate distribution by its marginals and a single real valued parameter α. They proposed estimation of the marginals by the reduced data estimators, F_{jn}, substitution of these estimators into the likelihood, and then maximization of the plug-in-likelihood w.r.t. α. As a consequence, their estimate of dependence will be biased if the true bivariate distribution is not adequately described by the copula model. As a result, this paper provides an important extension of their work since it directly estimates what is identifiable from the data. In particular, we provide a goodness-of-fit test for the copula model. The extension of these ideas to incorporate covariate processes, with application of the developed locally efficient estimators, is also of considerable value as noted.

2 The Nonparametric Maximum Likelihood Estimator.

The general EM algorithm can be used to compute the NPMLEs of F_1, F_2 and F_3. We note that the masses of the NPMLEs can only be determined up to the intervals $C_{r-1} \leq t < C_r$,

for $r = 1, \ldots, n+1$, where $C_0 = 0$ and $C_{n+1} = \infty$. To describe the 'full' data for the EM algorithm, consider the grid of $(n+1)^2$ rectangles in the positive quadrant where the rs^{th} rectangle is $R_{rs} = \{(t_1, t_2) : C_{r-1} \leq t_1 < C_r, C_{s-1} \leq t_2 < C_s\}$. The full data is then n_{rs}, the number of observations in R_{rs}, and the unknown parameters are p_{rs}, the mass given to R_{rs} by the joint distribution of (T_1, T_2), for $r, s = 1, \ldots, n+1$.

Given current estimates, F^k of F, the E step requires computation of $E(n_{rs}|(C, \Delta_1, \Delta_2)_i, i = 1, \ldots, n, F^k)$ for each r, s. Since n_{rs} is simply the sum of indicators that reflect whether an observation belongs to R_{rs} or not, this computation is straightforward; for example when $(C, \Delta_1 = \Delta_2)_i = (C_i, 1, 1)$, we assign the mass of the single observation to each R_{rs} in $\{(t_1, t_2) : 0 \leq t_1, t_2 < C_i\}$ according to the relative mass that F^k gives to R_{rs}, conditional on being in $\{(t_1, t_2) : 0 \leq t_1, t_2 < C_i\}$. Given the updated estimates $\hat{n}_{rs}$ thus calculated, the estimate of p_{rs} is then just $\hat{n}_{rs}/n$.

Note that each F^k estimates more than what is identifiable from the data. Thus, at each stage of the algorithm, only $F_i^k, i = 1, 2, 3$, should be evaluated for convergence assessment. At convergence, we again only consider the estimates of $F_i, i = 1, 2, 3$, derived from the limit of F^k. As is typical with the EM algorithm, care must be used in selecting an appropriate starting value; in particular, if the starting value for F puts no mass on a given R_{rs}, then no subsequent iterative estimates of F will place mass there either. In such cases, the algorithm will not necessarily converge to the NPMLE. Note however that, since the likelihood function is strictly concave, the EM algorithm will converge to the global maximum so long as the choice of support points of the starting value is sufficiently rich; if, for example, it places positive mass on all rectangles R_{rs} (van der Laan, 1996).

3 Efficient Estimation of Smooth Functionals.

For estimation of F_j, $j = 1, 2, 3$, we have shown that one can use the full NPMLE or the much simpler reduced data NPMLE's F_{jn}. In this section, we show that for the purpose of estimation of smooth functionals of F_j, $j = 1, 2, 3$, there is no loss in efficiency with the

estimators F_{jn} at many data generating distributions. This result is made specific by the following theorem. We assume that the random monitoring time C follows the distribution G (with associated density function g).

Theorem 1 *Let $j \in \{1,2,3\}$ be given. Consider the nonparametric model for $Y = (C, \Delta_1, \Delta_2)$, where C is independent of $\vec{T}$ and the distribution F_j is unspecified. We observe n i.i.d. observations of Y. Let $\mu_j = \int (1 - F_j)(u) r(u) du$ for a given function r. Consider the estimator $\mu_{jn} = \int (1 - F_{jn})(u) r(u) du$, where F_{jn} is the isotonic regression estimator of $F_j(c) = E(\Delta_j \mid C = c)$. Then, μ_{jn} is regular and asymptotically linear at any (F_j, G) for which F_j is continuous with density $f_j > 0$ on $[0, M_j]$ and zero elsewhere ($M_j < \infty$), $r/g(x) < M < \infty$ for $x \in [0, M_j]$.*

The influence curve of μ_{jn} is given by:

$$IC(C, \Delta_j \mid F_j, g, r) = \frac{r(C)}{g(C)} \left[F_j(C)(1 - \Delta_j) - \{1 - F_j(C)\}\Delta_j \right]. \tag{1}$$

The variance of this influence curve is given by:

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

Finally, at any $P_{F,G}$ satisfying the conditions of Lemma 1 (Appendix), we have that μ_{jn} is an asymptotically efficient estimator of μ_j.

Since each F_{jn} is just the NPMLE for simple univariate current status data, the regularity and asymptotic linearity of μ_{jn} follows from the results of Huang & Wellner (1995). Lemma 1 shows that for the full data distribution, $P_{F,G}$, the tangent space is the entire space $L_0^2(P_{F,G})$, which implies that any regular and asymptotic linear estimator is asymptotically efficient (Bickel et al., 1993).

3.1 Test for independence.

We now apply the results of the previous section to obtain a test of independence of T_1 and T_2. Let $M_3 < \infty$ be the end point of the assumed compact support of F_3. For a given function $w(\cdot)$ satisfying $\int_0^{M_3} w(s) ds = 1$ and $w(s) = 0$ for $s \geq M_3$, define $\mu_I(w) =$

$\int\{F_1F_2 - F_3\}(s)w(s)ds$. Note that, if T_1, T_2 are independent, $\mu_I(w) = 0$ for any such w. Let $\mu_{I,n}(w) = \int\{F_{1n}F_{2n} - F_{3n}\}(s)w(s)ds$ be the plug-in estimate of $\mu_I(w)$. We have the following result which is a corollary of Theorem 1.

Theorem 2 *Assume that F_j is continuous with density $f_j > 0$ on $[0, M_j]$ and zero elsewhere ($M_j < \infty$), $w(x)/g(x) < M < \infty$ for $x \in [0, M_j]$, $j = 1, 2, 3$. Then, assuming the conditions of Lemma 1 (Appendix), $\mu_{I,n}(w)$ is an asymptotically efficient estimator of $\mu_I(w)$ with influence curve*

$$IC_I(Y \mid F, g, w) = -IC(Y \mid F_1, g, r = F_2 w) - IC(Y \mid F_2, g, r = F_1 w) + IC(Y \mid F_3, g, r = w), \quad (2)$$

where the three influence curves on the right-hand side are defined in Theorem 1.

The calculation of the influence curve (2) follows from the results of Theorem 1, the definition of $\mu_{I,n}(w)$, and a standard telescoping algebraic argument as for product differentiation.

Let σ^2 be the variance of $IC_I(Y)$ and let $\hat{\sigma}^2 = (1/n)\sum_{i=1}^n \widehat{IC}^2(Y_i)$ be the estimate of σ^2 obtained by plugging in estimates of F_j, g, $j = 1, 2, 3$, into the formula (2) for $IC_I(Y)$. Any reasonable density estimate of g suffices, although, as for univariate current status data, it may be helful to base the bin or bandwith on the support points for F_{1n}, F_{2n} and F_{3n}, all of which are a subset of $c_1, \ldots, c_n$. Now, the statistic

$$D_n(w) = \frac{\mu_{I,n}(w)}{\hat{\sigma}}.$$

can be used to test independence of T_1, T_2. If T_1 is independent of T_2, then, under the conditions of Theorem 2, $D_n(w)$ is asymptotically normally distributed with mean zero and variance one.

A global test of independence is based on taking w to be constant over the support of F_3. For more refined, or local, tests of independence, we take w to be some form of kernel function around any chosen point in $[0, M_3]$. Computing the test statistic for a collection of such w's, weighting different regions in $[0, M_3]$, allows us, in principal, to determine areas of $[0, M_3]$ where violations of independence occur. A combined J degree of freedom test

of independence, based on a set of J weight functions, can also be constructed by using a multivariate version of $D_n(w)$ with each element corresponding to a different weight function.

3.2 Goodness of Fit Test of a Copula Model

Consider the copula model for the joint survival function, described in Wang and Ding (2000):

$$S(t_1, t_2) = C_\alpha(S_1(t_1), S_2(t_2)), \tag{3}$$

where $C_\alpha(\cdot,\cdot) : [0,1]^2 \to [0,1]$ indexes a parametric family of survival functions on the unit square, with fixed marginals. The function C determines the local dependence structure and $\alpha \in \mathbb{R}$ is a global association parameter related to Kendall's tau, denoted as τ, via the following equation:

$$\tau = 4\int_0^1\int_0^1 C_\alpha(u,v)dudv - 1.$$

Ding and Wang (2000) construct a 'pseudomaximum likelihood' estimator of α, α_n, that converges at $\sqrt{n}$ rate; specifically, their estimate is the maximum liklihood estimate of α, assuming the copula model and substituting the reduced data NPMLEs of F_1 and F_2; the latter immediately yield corresponding estimates of S_1 and S_2, similarly labeled S_{1n} and S_{2n}). The estimates of α and the joint survival function depend on the copula model being correct so that it is of value to have a goodness-of-fit test of this assumption available.

Note that, in general, $S(t,t) = 1 + F_3(t) - F_1(t) - F_2(t)$ which can be estimated non-parametrically by $S_n(t,t) = 1 + F_{3n}(t) - F_{1n}(t) - F_{2n}(t)$. On the other hand, using the copula model, we can also estimate $S(t,t)$ by $S_{n,cop}(t,t) = C_{\alpha_n}(S_{1n}(t_1), S_{2n}(t_2))$ One can then assess the goodness of fit of the copula model using the test statistic $\mu_{fit,n}(w) = \int\{S_{n,cop}(t,t) - S_n(t,t)\}w(t)dt$ for a given weight function w.

If the true data generating distribution follows the assumed copula model, then, under regularity conditions $\int\{S_{n,cop}(t,t) - S_n(t,t)\}w(t)dt$ is asymptotically linear with a certain influence curve $IC_{cop}(Y)$. With the regularity conditions of Theorem 1, we also have that $\int\{S_n(t,t) - S(t,t)\}w(t)dt$ is asymptotically linear with influence curve $IC_{NP}(Y) = -IC(Y \mid$

$F_3, g, w) + IC(Y \mid F_1, g, w) + IC(Y \mid F_2, g, w)$, where the latter influence curves are defined in (1). Thus, if the copula model is correct, then $\mu_{fit,n}(w)$ is asymptotically linear with influence curve $IC_{cop}(Y) - IC_{NP}(Y)$. If $\hat{\sigma}$ is the plug-in empirical estimate of the standard deviation of $IC_{cop}(Y) - IC_{NP}(Y)$, then the test statistic $\mu_{fit,n}(w)/\hat{\sigma}$ is asymptotically standard normal if the copula model is correct. As for the proposed tests of independence, this test statistic can be computed for a collection of w's, weighting different regions in $[0, M_3]$, assuming sufficient data is available. In this manner, areas of $[0, M_3]$ where deviations from the copula model occur can be determined.

4 The Locally Efficient One-Step Estimator Including Covariate Processes.

We now turn to estimation for extended bivariate current status data, $Y = (C, \Delta_1, \Delta_2, \overline{L}(C))$, where $L(\cdot)$ is a vector of covariates, possibly time-dependent, as introduced in §1. As before, we can reduce the data to univariate current status data $(C, \Delta_j, \bar{L}(C))$ on T_j and apply the locally efficient one-step estimators of functionals $\mu_j(r)$ of van der Laan & Robins (1998). Under regularity conditions, van der Laan & Robins showed that these one-step estimators are locally efficient for this reduced data structure. Lemma 2 (Appendix) proves that the efficient influence curve for the parameter $\mu_j(r)$ for the complete bivariate current status data structure $(C, \vec{\Delta}, \bar{L}(C))$ equals the efficient influence curve for the parameter $\mu_j(r)$ for the reduced data structure $(C, \Delta_j, \bar{L}(C))$, $j = 1, 2, 3$, at most data generating distributions of interest. As a consequence, the one-step estimators of $\mu_j(r)$ based on the reduced data are also locally efficient for the complete bivariate current status data structure.

We make some brief comments regarding these one-step estimators, referring to van der Laan & Robins (1998) for a more detailed treatment. First we state the assumptions regarding the monitoring time C. As noted earlier, we now allow dependence between C and $\vec{T}$, but only through the observed covariates. That is, the 'hazard' of monitoring at time t, given the full, unobserved, data $X = (\vec{T}, \overline{L})$, is only a function of the observed portion of the covariate

process, $\overline{L}(t)$:

$$\lambda_C(t \mid X) = \lambda_C(t \mid \overline{L}(t)). \tag{4}$$

This implies $G(\cdot \mid X)$, the conditional distribution function of C, satisfies coarsening at random (Robins, 1993). Coarsening at random, originally formulated by Heitjan & Rubin (1991), was generalized by Jacobsen & Keiding (1995) and Gill *et al.* (1997). If no covariate process $\bar{L}$ is available, then (4) implies that C is independent of $\vec{T}$. The principal regularity condition for the estimators is that $r(\cdot)/g(\cdot \mid X) < M < \infty$ F_X-a.e. which requires that the monitoring density is positive at any point s with $r(s) > 0$.

The one-step estimators of μ_j, $j = 1, 2, 3$, are consistent and asymptotically normal if we succeed in consistently estimating $\lambda_C(\cdot \mid X)$ at a suitable rate under the assumption (4). One such case is the experiment described in Section 1.2 where $\lambda_C(t \mid \overline{L}(t))$ is known by design because it is under the control of the investigator (so estimation of $\lambda_C(t \mid \overline{L}(t))$ is not even necessary). In general, a correctly specified semiparametric model which admits a consistent estimator for $\lambda_C(t \mid \overline{L}(t))$ can be used. van der Laan & Robins (1998) recommend modeling $\lambda_C(t \mid \overline{L}(t))$ by a time-dependent Cox proportional hazards model:

$$\lambda_C(t \mid \bar{L}(t)) = \lambda_0(t) \exp(\alpha^\top W(t)), \tag{5}$$

where $W(t)$ is a function of $\bar{L}(t)$. The model for the observed data distribution is now complete since the observed data distribution $P_{F_X,G}$ of Y is indexed by the full data distribution F_X which is left unspecified and the conditional distribution $G(\cdot \mid X)$ which needs to satisfy a semiparametric model such as (5).

Implementing the one-step estimators require an estimator of $F_j(t \mid \bar{L}(u)) = P(T_j \leq t \mid \bar{L}(u))$, $j = 1, 2, 3$, for various u's and t. By the curse of dimensionality, one needs to specify a lower dimensional working model for this conditional distribution and estimate it accordingly. The results of van der Laan & Robins (1998) show that the resulting one-step estimator is locally efficient for the data structure $(C, \Delta_j, \bar{L}(C))$ in the sense that it is asymptotically efficient for our model if the working model contains the truth, and it remains consistent and asymptotically normal otherwise.

Note that the methods of §3.1-3.2 can be generalized to the extended bivariate current status data structure, only assuming a semiparametric model (5) for $g(c \mid X)$.

5 Discussion

The EM algorithm described in §2 may be slow to converge. It would be of interest to derive an alternative algorithm for computing the NPMLE of F_1, F_2 and F_3 by extending the multivariate isotonic algorithm of Jewell & Kalbfleisch (2002).

Throughout the paper we have assumed a common monitoring time C for both T_1 and T_2. In some applications, it may be natural that the monitoring times will be different for the two survival time components. For example, this occurs in studies of age of onset for siblings who are examined at a common time but, of course, have different ages. This is a substantially more complex problem and the methods discussed here do not easily extend to cover this data structure.

APPENDIX: SATURATED TANGENT SPACE RESULTS.

We first provide a result for the marginal bivariate current status data structure.

Lemma 1 *Assume that G is absolutely continuous w.r.t. Lebesgue measure with a density g with support $[0, K]$, that F has a Lebesgue density with support $[0, M_1] \times [0, M_2]$. Then the tangent space at $P_{F,G}$ equals $L_0^2(P_{F,G})$.*

Proof. Let $A : L_0^2(F) \to L_0^2(P_{F,G})$ be defined by $A(h)(Y) = E(h(\vec{T}) \mid Y)$. Then the adjoint $A^\top : L_0^2(P_{F,G}) \to L_0^2(F)$ is given by $A^\top(v)(\vec{T}) = E(v(Y) \mid \vec{T})$. Explicitly,

$$\begin{aligned} A^\top(v)(\vec{T}) &= \int_{T_1 \vee T_2} v(c,1,1)dG(c) + \int_0^{T_1 \wedge T_2} v(c,0,0)dG(c) \\ &\quad - I(T_2 < T_1) \int_{T_2}^{T_1} v(c,0,1)dG(c) - I(T_1 < T_2) \int_{T_1}^{T_2} v(c,1,0)dG(c). \end{aligned}$$

The tangent space is given by $\overline{R(A) + L_0^2(G)}$. Since $\overline{R(A)}^\perp = N(A^\top)$ it suffices to prove that $N(A^\top) = L_0^2(G)$. Fixing $T_2 = M_2$ at the end point of its support $[0, M_2]$, taking the

derivative w.r.t T_1 at $t_1 < M_2$ yields $V(t_1, 2)g(t_1) = V(t_1, 4)g(t_1)$ for all $t_1 < M_2$. For fixed $T_2 = 0$, taking the derivative w.r.t T_1 at $t_1 > 0$ yields $V(t_1, 1)g(t_1) = V(t_1, 3)g(t_1)$. Fixing $T_1 = 0$, taking the derivative w.r.t. T_2 at $t_2 > 0$ yields $V(t_2, 1)g(t_2) = V(t_2, 4)g(t_2)$. Fixing $T_1 = M_1$ at the end point of its support $[0, M_1]$, taking the derivative w.r.t. T_2 at $t_2 < M_1$ yields $V(t_2, 2)g(t_2) = V(t_2, 3)g(t_2)$. This proves that any $V(C, \vec{\Delta}) \in N(A^{\top})$ does not depend on $\vec{\Delta}$. $\square$

This lemma can be immediately generalized to the following result for the extended bivariate current status data structure.

Lemma 2 *Assume that $G(\cdot \mid X)$ is absolutely continuous w.r.t. the Lebesgue measure with a density $g(\cdot \mid X)$ with support $[0, K]$, that $F(\cdot \mid L)$ has a Lebesgue density with support $[0, K_{1,L}] \times [0, K_{2,L}]$. Then the tangent space at $P_{F_X,G}$ equals $L_0^2(P_{F_X,G})$.*

The implication of this result is that, under the conditions of Lemma 2, any regular asymptotically linear estimator of F_X is asymptotically efficient.

References

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972) *Statistical Inference under Order Restrictions*, Wiley, New York.

Bickel, P.J., Klaassen, A.J., Ritov, Y. & Wellner, J.A. (1993), *Efficient and adaptive inference in semi-parametric models*, Johns Hopkins University Press, Baltimore.

Diamond, I.D. McDonald, J.W. & 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. & 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.

van de Geer, S. (1994), *Asymptotic normality in mixture models,* preprint University of Leiden, the Netherlands.

Gill, R.D., van der Laan, M.J. & Robins, J.M. (1997), Coarsening at Random: Characterizations, Conjectures and Counter-examples, *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, 255-295, Springer, D.Y. Lin and T.R. Fleming (eds).

Groeneboom, P. & Wellner, J.A. (1992), *Information bounds and nonparametric maximum likelihood estimation*, Birkhäuser Verlag.

Heitjan, D.F. & Rubin, D.B. (1991), Ignorability and coarse data, *Annals of Statististics* **19**, 2244–2253.

Huang, J. (1996), Efficient estimation for the proportional hazards model with interval censoring, *Annals of Statististics*, **24**, 540–568.

Huang, J. & Wellner, J.A. (1995), Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I, *Statistica Neerlandica* **49**, 153–163.

Jacobsen, M. & Keiding, N. (1995) , Coarsening at random in general sample spaces and random censoring in continuous time, *Annals of Statistics*, **23**, 774–786.

Jewell, N.P. & Kalbfleisch, J.D. (2002), Maximum Likelihood Estimation of Ordered Multinomial Parameters, to appear.

Jewell, N.P. & Shiboski, S.C. (1990), Statistical analysis of HIV infectivity based on partner studies, *Biometrics*, **46**, 1133-1150.

Keiding, N. (1991) Age-specific incidence and prevalence (with discussion), *Journal of the Royal Statistical Society Ser. A*, **154**, 371–412.

van der Laan, M.J (1996), *Efficient and Inefficient Estimation in Semiparametric Models*, CWI Tract No. 114, Amsterdam: Centrum voor Wiskunde on Informatica.

van der Laan, M.J. & Robins, J.M. (1998), Locally Efficient Estimation with Current Status Data and Time-Dependent Covariates. *Journal of the American Statistical Association*, **93**, 693–701.

Rabinowitz, D., Tsiatis, A. & Aragon, A. (1995), Regression with interval-censored data, *Biometrika*, **82**, 501–513.

Robins, J.M. (1993), Information recovery and bias adjustment in proportional hazards re-

gression analysis of randomized trials using surrogate markers, *Proceedings of the Biopharmaceutical Section, American Statistical Association*, 22–33.

Robins, J. M. & Ritov, Y. (1997), Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models, *Statistics in Medicine*, **16**, 285–319.

Sun, J. & Kalbfleisch, D. (1993), The analysis of current status data on point processes, *Journal of the American Statistical Association*, **88**, 1449–1454.

Wang, W. & Ding, A.A. (2000), On assessing the association for bivariate current status data, *Biometrika*, **87**, 879–893.