

Inference with Bivariate Truncated Data

Christopher M. Quale and Mark J. van der Laan
Dept. of Biostatistics, University of California at Berkeley

May 15, 2001

Abstract

In this paper we build on previous work for estimation of the bivariate distribution of the time variables T_1 and T_2 when they are observable only on the condition that one of the time variables, say T_1 , is greater than (left-truncation) or less than (right truncation) some observed time variable C_1 . In this paper, we introduce several results based on the Influence Curve (which we derive in this paper) of the NPMLE of the distribution F of (T_1, T_2) developed by van der Laan (van der Laan, 1996). Specifically we will: prove that the NPMLE is asymptotically equivalent to an estimator developed by Gürler (Gürler, 1997), derive the asymptotic distribution of the NPMLE based on its Influence Curve, present tests to determine the amount of dependence between T_1 and T_2 , present the results of simulation studies that compare the NPMLE and Gürler's estimator and evaluate the performance of both the above mentioned tests and confidence intervals of F based on the asymptotic distribution of the NPMLE, and finally we will apply the methods in a data analysis in which we also point out practical issues that arise in the implementation of the estimator.

Keywords: Bivariate Truncation, Non- Parametric Maximum Likelihood, Influence Curves.

1 Introduction

The TR-AIDS (Transfusion related AIDS) dataset (Wang, 1989, Gürler, 1996) recorded age at transfusion and time from transfusion to AIDS. However, only those subjects who had received a diagnosis of AIDS prior to July 1986, the end of the study, were included in the study. Thus, if we define T_1 as time from transfusion to AIDS, T_2 as age at transfusion and C_1 as time from transfusion to July 1986, we are only able to observe (T_1, T_2, C_1) if $T_1 \leq C_1$.

This dataset is a special case of randomly right truncated bivariate survival data. Left (right) univariate truncation of a time variable T occurs when a subject is entered into a study conditional on $T \geq C$ ($T \leq C$) for some other time variable C . The bivariate left (right) truncation case corresponds to the situation where we observe (T_1, T_2) conditional on $T_1 \geq C_1$ and $T_2 \geq C_2$ ($T_1 \leq C_1$ and $T_2 \leq C_2$). In this paper we will focus on bivariate survival data which is subject to truncation on only T_1 , i.e. $C_2 = 0$ ($C_2 = \infty$).

Van der Laan developed the Non-Parametric Maximum Likelihood Estimator (NPMLE) for the survivor function of (T_1, T_2) when both of the time variables are subject to either left or right truncation by some observed time variables (C_1, C_2) (van der Laan, 1996). Van der Laan showed that under certain regularity conditions, $\sqrt{n}(S_n - S)$ converges weakly to a Gaussian process and that S_n is a consistent, asymptotically linear and efficient estimator of S (see van der Laan, 1996 for details).

For the case we focus on, in which there is truncation on only one variable, van der Laan showed that the NPMLE is explicit. Gürler also developed several estimators for this data structure (Gürler, 1997). We will prove that her best performing estimator (based on her simulation results) is asymptotically equivalent to the NPMLE. We present simulation studies comparing Gürler's best estimator with the NPMLE, in which the NPMLE performed slightly better than Gürler's estimator under various simulated data structures.

In section 3 we will develop methods for doing inference with the NPMLE, based on the assertion that the NPMLE is an asymptotically linear estimator of S . Van der Laan (1996) did not derive the influence curve for the NPMLE, We will show how to construct confidence intervals for $S(t_1, t_2)$ and will present results from a simulation study evaluating the performance of these confidence intervals. In addition, we will propose two tests which investigate the dependence between the random variables T_1 and T_2 . The first tests the independence of the events $T_1 \geq t_1$ and $T_2 \geq t_2$. We will derive this test by comparing the full estimator of the bivariate survivor function of (T_1, T_2)

with an estimator based on the product of the marginal survivor functions $S(\cdot, 0)$ and $S(0, \cdot)$ of T_1 and T_2 , respectively. Using the properties of the estimator S_n we will show how to get a distribution of our proposed test statistic under the null hypothesis of independence of the events $T_1 \geq t_1$ and $T_2 \geq t_2$. The second test extends the first by providing a global test of the null hypothesis of independence of the events $T_1 \geq t_{1j}$ and $T_2 \geq t_{2j}$ over a set of points $(t_1, t_2) = ((t_{11}, t_{21}), (t_{12}, t_{22}), \dots, (t_{1J}, t_{2J}))$ versus dependence of $T_1 \geq t_{1j}$ and $T_2 \geq t_{2j}$ for at least one $j \in \{1, 2, \dots, J\}$.

Finally, we will apply the proposed methods for the uncensored case to the TR AIDS dataset mentioned above. This will give us an opportunity to utilize the inferential methods we develop, including the pointwise and global test of the independence of sets of events.

We note here that the methods we develop for truncated data are easily applied to either right or left truncated data. Namely, if we solve the problem of estimation for left truncated data, the results apply immediately to right truncated data. Right truncated data may be seen as a “mirror image” of left truncated data. Suppose we have developed an estimator S_n^T for the case in which we observe (T_1, T_2, C_1) conditional on $T_1 \geq C_1$, where S^T is the survivor function of (T_1, T_2) . If we are presented with the following right truncated data: (U_1, U_2, D_1) conditional on $U_1 \leq D_1$, we apply the methods for left truncated data to the following: $(-U_1, -U_2, -D_1)$ conditional on $-U_1 \geq -D_1$, which is simply the specification for left truncated data. For $u_1 \geq 0$ and $u_2 \geq 0$ the estimate $S_n^U(-u_1, -u_2) = \hat{P}(-U_1 \geq -u_1, -U_2 \geq -u_2)$ corresponds to an estimate F_n^U of $F^U(u_1, u_2)$, and all results obtained for left truncation will apply to the estimator F_n^U . For ease of presentation, we will show the results for left truncation with the understanding that the conclusions are the same for right truncation.

2 An overview of estimation with left truncated data

First we will look at the univariate case. Let $T \sim F$ and $C \sim G$. Denote (T', C') as a draw from the conditional distribution of (T, C) given $T \geq C$. If we observe T'_i and C'_i for $i = 1, \dots, n$ then the NPMLE for the survivor function $S(t) = 1 - F(t)$ is given by the product limit estimator. This

estimator is given by:

$$S_{pl}(t) = \prod_{(0,t]} (1 - \Lambda_n(ds))$$

where

$$\Lambda_n(ds) = \frac{\sum_{i=1}^n I(T'_i \in ds)}{\sum_{i=1}^n I(T'_i \geq s, C'_i \geq s)}$$

is an estimate of the hazard probability $\Lambda(ds) = P(T \in ds \mid T \geq s)$.

The asymptotic properties of this estimator have been studied by Woodroffe (Woodroffe, 1985), Wang, (Wang, Jewell and Tsai, 1986), Gill (Gill and Keiding, 1990), van der Vaart (van der Vaart 1991) and Bickel (Bickel et al, 1993). These authors have shown that the product limit estimator is an asymptotically linear and efficient estimator of S , where asymptotic linearity means:

$$\sqrt{n}(S_{pl}(t) - S(t)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n IC_{pl}(T'_i, C'_i \mid F, G, t) + o_p(1) \quad (1)$$

where $IC_{pl}(T'_i, C'_i \mid S, t)$ is the influence curve of the estimator. The influence curve of the product limit estimator is given by (see Bickel et al 1993):

$$IC_{pl}(T', C' \mid F, G, t) = -S(t) \left\{ \frac{I(T' \leq t)}{(G^e - F^e)(T')} - \alpha \int_0^{t \wedge T'} \frac{F(du)}{S^2(u)G(u)} \alpha \int_0^{t \wedge C'} \frac{F(du)}{S^2(u)G(u)} \right\} \quad (2)$$

where $\alpha \equiv P(T \geq C)$ and

$$(G^e - F^e)(t) = P(T \geq t, C \leq t \mid T \geq C)$$

Suppose we observe truncated right censored data. More precisely, suppose we observe $(Y = T \wedge C^*, C, \Delta = I(T \leq C^*))$ conditional on $Y \geq C$, where C^* is a censoring variable and T is the time variable of interest. Then the product limit estimate of S for this case requires one to estimate $\Lambda(ds)$ by:

$$\Lambda_n(ds) = \frac{\sum_{i=1}^n I(Y'_i \in ds, \Delta_i = 1)}{\sum_{i=1}^n I(Y'_i \geq s, C'_i \geq s)}$$

where, as above, (Y'_i, C'_i) denote draws from the conditional distribution of Y given that $Y \geq C$. The influence curve given above may also be extended to this case.

For truncated bivariate survival data in which we observe (T_1, T_2, C_1) conditional on $T_1 \geq C_1$, Gürler (Gürler, 1997) has developed several estimators, and the general NPMLE applied to this special case solves this problem explicitly. Again, we will use the convention that (T'_1, T'_2, C'_1) represents a draw from the conditional distribution of (T_1, T_2, C_1) given $T_1 \geq C_1$.

Let $(T_1, T_2) \sim F$ and $C_1 \sim G$, let S be the corresponding survivor function of (T_1, T_2) , and define S_{pl} to be the product limit estimate of the distribution of T_1 . Van der Laan (van der Laan 1996) showed that the NPMLE S_n had the following form:

$$\begin{aligned} S_n(t_1, t_2) &= \int_{t_1}^{\infty} \int_{t_2}^{\infty} \frac{F_n^e(ds_1, ds_2)}{\int_0^{s_1} \frac{G_n^e(dc_1)}{S_{pl}(c_1)}} \\ &= \sum_{i=1}^n \frac{I(T'_{1i} \geq t_1, T'_{2i} \geq t_2)}{\sum_{j=1}^n \frac{I(C'_{1j} \leq T'_{1i})}{S_{pl}(C'_{1j})}} \end{aligned} \quad (3)$$

where $F_n^e(ds_1, ds_2)$ and $G_n^e(ds_1)$ are the empirical estimates of the probabilities $P(T_1 \in ds_1, T_2 \in ds_2 \mid T_1 \geq C_1)$ and $P(C_1 \in ds_1 \mid T_1 \geq C_1)$, respectively. Specifically,

$$\begin{aligned} F_n^e(ds_1, ds_2) &= \frac{1}{n} \sum_{i=1}^n I(T'_{1i} \in ds_1, T'_{2i} \in ds_2) \\ G_n^e(ds_1) &= \frac{1}{n} \sum_{i=1}^n I(C'_{1i} \in ds_1) \end{aligned}$$

Gürler also developed several estimators for the bivariate distribution under univariate truncation. The estimator presented below represents the one that performed best in her simulation studies (Gürler, 1997).

One way to view Gürler's estimator of the distribution of (T_1, T_2) is to recognize that

$$\begin{aligned} F^e(ds_1, ds_2) &= \frac{P(T_1 \in ds_1, T_2 \in ds_2 \mid T_1 \geq C_1)}{P(T_1 \in ds_1, T_2 \in ds_2, T_1 \geq C_1)} \\ &= \alpha \end{aligned}$$

Therefore

$$F^e(ds_1, ds_2) = \frac{G(s_1)F(ds_1, ds_2)}{\alpha}$$

where $\alpha = P(T_1 \geq C_1)$. We now have:

$$F(ds_1, ds_2) = \frac{\alpha F^e(ds_1, ds_2)}{G(s_1)}$$

We observe that

$$\begin{aligned} \frac{G(s_1)S_1(s_1)}{\alpha} &= \frac{P(C_1 \leq s_1, T_1 \geq s_1)}{P(T_1 \geq C_1)} = P(C_1 \leq s_1, T_1 \geq s_1 \mid T_1 \geq C_1) \\ &= \int_{s_1}^{\infty} \int_0^{s_1} \tilde{F}^e(dc_1, dt_1) \end{aligned}$$

where $\tilde{F}^e(dc_1, dt_1) \equiv P(C_1 \in dc_1, T_1 \in dt_1 \mid T_1 \geq C_1)$. Using the product limit estimator as defined above, Gürler's representation may be expressed in the following way:

$$\begin{aligned} S_n^{Gur}(t_1, t_2) &= \int_{t_1}^{\infty} \int_{t_2}^{\infty} \frac{S_{pl}(s_1, 0)F_n^e(ds_1, ds_2)}{\int_0^{s_1} \int_{s_1}^{\infty} \tilde{F}_n^e(dc_1, dt_1)} \\ &= \sum_{i=1}^n \frac{S_{pl}(T'_{1i})I(T'_{1i} \geq t_1, T'_{2i} \geq t_2)}{\sum_{j=0}^n I(C'_{1j} \leq T'_{1i}, T'_{1j} \geq T'_{1i})} \end{aligned}$$

3 Influence curves and inference in the bivariate case

Consider the case of bivariate truncation mentioned above, namely when we observe (T_1, T_2, C_1) conditional on $T_1 \geq C_1$, where $(T_1, T_2) \sim F$ and $C_1 \sim G$. In order to obtain estimates of the variance of the NPMLE S_n we will show that S_n is an asymptotically linear estimator of S . Using the definition given in (1), S_n will be asymptotic linear at $t = (t_1, t_2)$ if

$$\sqrt{n}(S_n(t) - S(t)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n IC(T'_{1i}, T'_{2i}, C'_{1i} \mid F, G, t) + o_p(1)$$

For a proof of the asymptotic linearity and consistency of S_n , we refer the reader to the work of van der Laan (van der Laan 1996), who showed the result for the general case (i.e. truncation of both T_1 and T_2 as well as uniform consistency over the rectangle $(0, \tau]$, where $\tau = (\tau_1, \tau_2)$). The method to prove the asymptotic linearity of both S_n and S_n^{Gur} at a given point $t = (t_1, t_2)$ is a direct application of the functional delta method (see Gill, 1995).

Although van der Laan proved the asymptotic linearity of the NPMLE S_n , he did not provide the functional form of the influence curve. Our approach of obtaining the influence curve is as follows. First we notice that S_n as in (3) may be represented as a functional of F_n^e , G_n^e and S_{pl} , which we will define to be $\Psi(F_n^e, G_n^e, S_{pl})(t)$. We approximate $\Psi(F_n^e, G_n^e, S_{pl})(t) - \Psi(F^e, G^e, S_1)(t)$ by a linear functional of $F_n^e - F^e$, $G_n^e - G^e$, and $S_{pl} - S_1$, which will give us the desired influence curve. Similarly, we may express Gürler's estimator $S_n^{Gur}(t)$ as a functional of F_n^e , \tilde{F}_n^e and S_{pl} , defined by $\Omega(F_n^e, \tilde{F}_n^e, S_{pl})(t)$. Then we approximate $\Omega(F_n^e, \tilde{F}_n^e, S_{pl})(t) - \Omega(F^e, \tilde{F}^e, S_1)(t)$ by a linear functional of $F_n^e - F^e$, $\tilde{F}_n^e - \tilde{F}^e$ and $S_{pl} - S_1$. The following theorem gives the influence curve of S_n obtained through this method and states that this is also the influence curve of S_n^{Gur} which implies that the two estimators are asymptotically equivalent.

Theorem 1 *Under the assumptions given in Appendix A, at the point $t = (t_1, t_2)$, $S_n(t)$ and $S_n^{Gur}(t)$ are asymptotically linear, efficient and consistent estimators of the survivor function S at t . The influence curve of both $S_n^{Gur}(t)$ and $S_n(t)$ is given by:*

$$\begin{aligned}
& IC(T'_1, T'_2, C'_1 \mid F, G, t) \\
&= \int_{t_1}^{\infty} \int_{t_2}^{\infty} \frac{I(T'_1 \in ds_1, T'_2 \in ds_2) - F^e(ds_1, ds_2)}{H(G^e, S)(s_1)} \\
&\quad - \int_{t_1}^{\infty} \int_{t_2}^{\infty} \frac{F^e(ds_1, ds_2)}{H^2(G^e, S)(s_1)} \left\{ \frac{I(C'_1 \leq s_1)}{S(C'_1, 0)} - \int_0^{s_1} \frac{G^e(dc_1)}{S(c_1, 0)} \right\} \\
&\quad + \int_{t_1}^{\infty} \int_{t_2}^{\infty} \frac{F^e(ds_1, ds_2)}{H^2(G^e, S)(s_1)} \int_0^{s_1} \frac{IC_{pl}(T', C' \mid F, G, c_1)}{S(c_1, 0)} G^e(dc_1)
\end{aligned} \tag{4}$$

where $IC_{pl}(T', C' \mid F, G, c_1)$ is given by (2) and $H(G^e, S)(s_1) \equiv \int_0^{s_1} \frac{G^e(dc_1)}{S(c_1, 0)}$

We refer the reader to the work of Quale (Quale and van der Laan, 1998) for the details of the derivation of this influence curve for both the NPMLE and for Gürler's estimator.

From (1), we see that once we have an expression for the influence curve, not only do we have a method to estimate the variance of $S_n - S$, we are able to apply the central limit theorem to get the asymptotic distribution of $S_n - S$ (which also follows from the fact that $\sqrt{n}(S_n - S)$ converges to a Gaussian process). This will allow us to construct confidence intervals about our estimates of $S(t)$ in the following manner:

- Get an estimate $S_n(t)$ of $S(t)$ from (3).
- Plug in estimates F_n^e , G_n^e , S_{pl} and \widehat{IC}_{pl} and the data $Y_i \equiv (T'_{1i}, T'_{2i}, C'_{1i})$ into (4) and obtain an estimate $\widehat{IC}_i(t_1, t_2) \equiv \widehat{IC}(Y_i | F_n, G_n, t)$ for $i = 1, \dots, n$
- Then, we may estimate the variance of $S_n(t_1, t_2)$ by $\frac{1}{n}$ times the sample variance of $\widehat{IC}(Y | F_n, G_n, t)$, namely

$$\widehat{\sigma}_{S_n(t_1, t_2)}^2 = \frac{1}{n} \left\{ \frac{1}{n} \sum_{i=1}^n (\widehat{IC}_i(t_1, t_2) - \overline{IC}(t_1, t_2))^2 \right\}$$

where $\overline{IC}(t_1, t_2) \equiv \frac{1}{n} \sum_{i=1}^n \widehat{IC}_i(t_1, t_2)$.

- Thus the $(1-\alpha)$ confidence interval for $S(t_1, t_2)$ is given by:

$$S_n(t_1, t_2) \pm z_{1-\frac{\alpha}{2}} \widehat{\sigma}_{S_n(t_1, t_2)}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$ quantile of the standard normal distribution.

We will use these methods in our data analyses in section 5.

3.1 Tests of independence

In this section we propose two tests which will determine the amount of dependence between our random variables T_1 and T_2 . The first test we describe looks at a particular point (t_1, t_2) and tests the null hypothesis that the events $T_1 \geq t_1$ and $T_2 \geq t_2$ are independent events. The second test is an extension of the first; namely, we select a vector of points $(t_1, t_2) = ((t_{11}, t_{21}), (t_{12}, t_{22}), \dots, (t_{1J}, t_{2J}))$ and conduct a test of the null hypothesis that the events $T_1 \geq t_{1j}$ and $T_2 \geq t_{2j}$ are independent for all

$j \in \{1, 2, \dots, J\}$ versus the alternative that $T_1 \geq t_{1j}$ and $T_2 \geq t_{2j}$ are independent for at least one $j \in \{1, 2, \dots, J\}$.

The two tests have their own strengths in the context of bivariate analysis. The first test gives us the ability to examine local areas of dependence, i.e. it can tell us the regions over which T_1 and T_2 are highly dependent. The second gives us a global α level test of dependence of T_1 and T_2 since $T_1 \geq t_{1j}$ and $T_2 \geq t_{2j}$ being dependent for at least one j will give us that T_1 and T_2 are dependent random variables.

3.1.1 Pointwise test of independence between T_1 and T_2

Let $S_n(t_1, t_2)$ represent the NPMLE of the survivor function $S(t_1, t_2)$. Given two time variables (T_1, T_2) , it is of great interest to examine the question of whether or not the event $T_1 \geq t_1$ is independent of the event $T_2 \geq t_2$. The test statistic we propose to examine this relationship is based on an examination of the difference between the estimator of S based on no assumption of independence, and the estimator of S based on an assumption of independence. Namely:

$$D_n(t_1, t_2) \equiv S_n(t_1, t_2) - S_{1n}(t_1, 0)S_{2n}(0, t_2) \quad (5)$$

will be used as a measure of the dependence between the events $T_1 \geq t_1$ and $T_2 \geq t_2$. In this case, $S_{1n}(t_1, 0)$ and $S_{2n}(0, t_2)$ represent the estimated marginal distribution functions of T_1 and T_2 , respectively, again using the NPMLE as the estimator. If these two events are indeed independent, then we would expect (5) to be close to zero. However, if there were dependence between the two events, then we would expect there to be some deviation from zero.

Theorem 2 *Given $t = (t_1, t_2)$, if the events $T_1 \geq t_1$ and $T_2 \geq t_2$ are independent, then $\sqrt{n}D_n(t)$ is asymptotically normal with mean zero and*

$$\text{var}(\sqrt{n}D_n(t)) = \text{var}(IC_{test}(T'_{1i}, T'_{2i}, C'_{1i} \mid F, G, t))$$

where

$$\begin{aligned} IC_{test}(T'_{1i}, T'_{2i}, C'_{1i} \mid F, G, t) &= IC(T'_{1i}, T'_{2i}, C'_{1i} \mid F, G, t) - S(0, t_2)IC_1(T'_{1i}, T'_{2i}, C'_{1i} \mid F, G, t_1) \\ &\quad - S(t_1, 0)IC_2(T'_{1i}, T'_{2i}, C'_{1i} \mid F, G, t_2) \end{aligned}$$

where IC_1 and IC_2 are such that

$$\begin{aligned}\sqrt{n}(S_n(t_1, 0) - S(t_1, 0)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n IC_1(T'_{1i}, T'_{2i}, C'_{1i} \mid F, G, t_1) + o_p(1) \\ \sqrt{n}(S_n(0, t_2) - S(0, t_2)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n IC_2(T'_{1i}, T'_{2i}, C'_{1i} \mid F, G, t_2) + o_p(1)\end{aligned}$$

Furthermore, since the test is based on the efficient influence curves IC, IC_1, IC_2 , it is an efficient test.

The method to prove this result is similar to the proof of Theorem 1. First one must show the asymptotic linearity of $D_n - (S(t_1, t_2) - S(t_1, 0)S(0, t_2))$ by expressing it as a functional linear in $S_n(t_1, t_2) - S(t_1, t_2)$, $S_n(t_1, 0) - S(t_1, 0)$ and $S_n(0, t_2) - S(0, t_2)$. This can be accomplished using the telescoping identity $a_n b_n - ab = (a_n - a)b_n + a(b_n - b)$. We already have that $S_n(t_1, t_2) - S(t_1, t_2)$ is asymptotically linear, and we can show the same for the marginals at the points t_1 and t_2 . Further details on the derivation of the asymptotic distribution may be found in the appendix. We will demonstrate an application of this test in our analysis of the TR AIDS dataset.

3.1.2 Summary test of independence

Suppose that we would like a summary test of dependence over a set of points $(t_1, t_2) = \{(t_{1j}, t_{2j}) \mid j = 1, 2, \dots, J\}$. Specifically, define the null hypothesis H_0 as follows: The events $T_1 \geq t_{1j}$ and $T_2 \geq t_{2j}$ are independent for all $j \in \{1, 2, \dots, J\}$. The alternative hypothesis H_a is simply that for some $j \in \{1, 2, \dots, J\}$ the events $T_1 \geq t_{1j}$ and $T_2 \geq t_{2j}$ are dependent, which would imply, of course, that the random variables T_1 and T_2 are dependent random variables.

Define the test statistic $D_n(t_{1j}, t_{2j})$ as in section 3.1.1 above. Let $\tilde{D}_n(t_1, t_2)$ be defined as the vector $(D_n(t_{1j}, t_{2j}), \dots, D_n(t_{1J}, t_{2J}))$. An easy extension of Theorem 2 gives us that under the null hypothesis H_0 defined above, $\sqrt{n}\tilde{D}_n(t_1, t_2)$ is asymptotically multivariate normal with mean vector 0 and $J \times J$ covariance matrix Σ . Note that $\Sigma_{j,j'} = cov(IC(t_{1j}, t_{2j}), IC(t_{1j'}, t_{2j'}))$ where

$$IC(t_{1j}, t_{2j}) = IC_{test}(T'_1, T'_2, C'_1 \mid F, G, t_{1j}, t_{2j})$$

In order to carry out the test, we will estimate $\Sigma_{j,j'}$ by the sample covariance based on estimates of the influence curve:

$$\widehat{\Sigma}_{j,j'} = \frac{1}{n} \sum_{i=1}^n (\widehat{IC}_i(t_{1j}, t_{2j}) - \overline{IC}(t_{1j}, t_{2j})) (\widehat{IC}_i(t_{1j'}, t_{2j'}) - \overline{IC}(t_{1j'}, t_{2j'}))$$

where

$$\widehat{IC}_i(t_{1j}, t_{2j}) = \widehat{IC}_{test}(T'_{1i}, T'_{2i}, C'_{1i} \mid F, G, t_{1j}, t_{2j})$$

and

$$\overline{IC}(t_{1j}, t_{2j}) = \frac{1}{n} \sum_{i=1}^n \widehat{IC}_i(t_{1j}, t_{2j})$$

Under the same conditions as needed for the asymptotic linearity of the estimator $S_n(\cdot, \cdot)$, $\widehat{\Sigma}$ converges in probability to Σ . Therefore, the standardized test statistic $\widetilde{M}_n(t_1, t_2)$, defined as $\widetilde{M}_n(t_1, t_2) \equiv \sqrt{n} \widehat{\Sigma}^{-\frac{1}{2}} \widetilde{D}_n(t_1, t_2)$ will be distributed as asymptotically normal with $J \times J$ identity covariance matrix I and mean vector 0 under the null hypothesis H_0 .

Consider the squared euclidean norm of the test statistic $\widetilde{M}_n(t_1, t_2)$:

$$\|\widetilde{M}_n(t_1, t_2)\|^2 = \sum_{j=1}^J M_n(t_{1j}, t_{2j})^2$$

Since, under H_0 , $\widetilde{M}_n(t_1, t_2)$ is a vector of asymptotically independent, normal 0,1 random variables, $\|\widetilde{M}_n(t_1, t_2)\|^2$ will be asymptotically distributed as a chi square random variable with J degrees of freedom. Note that the test has power against deviations from independence at one point (t_{1j}, t_{2j}) since $\|\widetilde{M}_n(t_1, t_2)\|^2 \geq c$ occurs if $|M_n(t_{1j}, t_{2j})| \geq \sqrt{c}$ for at least one $j \in \{1 \dots J\}$. Thus in order to conduct a level α test of the null hypothesis H_0 that T_1 and T_2 are independent over the set of points (t_1, t_2) , we need to determine a critical point c such that:

$$\alpha = P(\|\widetilde{M}_n(t_1, t_2)\|^2 \geq c \mid H_0)$$

Clearly c is equal to the $1 - \alpha$ quantile of the chi-square distribution with J degrees of freedom.

If it is necessary to assign more or less influence to $M_n(t_{1j}, t_{2j})$, we can weight each of these by w_j to create a weighted global test statistic. This weighted test statistic is then $\sum_{j=1}^J w_j M_n(t_{1j}, t_{2j})^2$. In order to get a critical point for the weighted test, one could follow the following procedure:

- Simulate 1000 copies of the random variable $B_i = \sum_{j=1}^J w_j X_j$, where X_j is generated from a χ_1^2 distribution.
- Determine the empirical $1 - \alpha$ quantile \hat{c} of the generated vector.

We will implement this test in both the simulation study and then use it in our analysis of the TR-AIDS dataset.

4 Simulation Results

Here we present the results for three simulation studies concerning bivariate survival data when one of the variables is left truncated: one comparing the NPMLE with Gürler's representation, another looking at the performance of estimated confidence intervals using the NPMLE, and one evaluating the performance of the tests of independence. The criteria for the comparison of the two estimators was mean squared error (MSE) at a 6x6 grid of values of (t_1, t_2) . The criteria for the confidence interval simulations was the empirical coverage probabilities of estimated 0.95 confidence intervals at a 6x6 grid of values of (t_1, t_2) over 625 iterations. The criteria in the test simulations was the empirical rejection probability of rejection for 0.05 level tests, under a model where the data were generated such that T_1 and T_2 were independent.

The amount of truncation in the simulations was determined by the quantity $\alpha \equiv P(T \geq C)$. In the confidence interval and comparison simulations, two levels of truncation were studied, corresponding to $\alpha = 0.66$ and $\alpha = 0.33$. For the test simulation, the levels of truncation studied corresponded to $\alpha = 0.75$ and $\alpha = 0.50$. Since the sample size of the simulated datasets is actually determined by the level of truncation, the sample size is random. All simulations were run so that there were approximately 300 observations per iteration.

The simulated data had the following structure:

$$\begin{aligned} T_{1i} &= \gamma R_{1i} + (1 - \gamma) R_{3i} \\ T_{2i} &= \gamma R_{2i} + (1 - \gamma) R_{3i} \\ C_{1i} &= R_{4i} \end{aligned}$$

where R_{ji} for $j = 1, 2, 3$ were independent mean 10 exponential random variables, R_{4i} was an exponential random variable (independent of R_{1i}, R_{2i} , and R_{3i}) whose mean was tuned to vary the amount of truncation (namely, the

larger the mean of C , the more truncation) and γ , a quantity that controlled the amount of dependence between T_1 and T_2 , was set at 0.5 for the confidence interval and comparison simulations and $\gamma = 1$ for the test simulations. Data would be generated as above, and an individually generated observation would be included in the simulated dataset if $T_{i1} \geq C_1 i$.

The results for the comparison simulation in Table 1 indicate that the estimators do perform similarly with respect to MSE, with some differences in the tails of the distribution. We notice that the NPMLE performs slightly better than Gürler's representation for most time points, except in the right hand tail under low truncation, where Gürler's estimator performs better. Other simulations varying the amount of dependence between T_1 and T_2 (not shown here) mimic the pattern we see in Table 1.

The confidence interval simulation results in Table 2 indicate that for both high and low truncation, the coverage probabilities remained close to the ideal 0.95. Other simulations (not shown here) varying the amount of dependence between T_1 and T_2 , showed similar good performance of the confidence intervals.

For the lower level of truncation, the test simulations performed well, however the results in Table 3 demonstrate the need for caution in finite sample implementation. The empirical rejection probabilities for the global test at $\alpha = 0.75$ and $\alpha = 0.5$ were 0.05 and 0.17, respectively. The results for the pointwise test can be found in Table 3. The pointwise tests for $\alpha = 0.75$ were all close to the ideal 0.05, For $\alpha = 0.5$. the pointwise tests were elevated toward the lower tail of the distribution of (T_1, T_2) , but performed better away from the lower quantiles. The poor performance of the pointwise test at the lower end of the distribution and of the global test is most likely due to the sparseness of data at smaller values of T_1 that occurs at the higher level of truncation. Simulation studies (not shown here) indicate that increasing the sample size stabilized the tests at higher levels of truncation.

From our simulation results, for high levels of truncation and smaller sample sizes, the global test should be interpreted very carefully. However, in high truncation / small sample size situations, the pointwise tests away from the lower end of the distribution (for left truncation) can be considered to be reliable.

It should be noted that the implementation of these methods for right truncated data is completely analogous to the implementation for left truncated data, as mentioned in the introduction. Simulation studies (not shown here) showed analogous results for right truncated data as those presented

above.

5 Data Analysis

In this section our goal is to give a brief example of the implementation of our methods for bivariate truncated data using the TR AIDS data. Recall that in this dataset, we observe a patient’s time from blood transfusion to AIDS only if this time is less than the time from transfusion to July 1986, the end of the study. In this dataset, then, we observe 293 i.i.d copies of (T_1, T_2, C_1) conditional on $T_1 \leq C_1$, (thus an observation may be denoted (T'_1, T'_2, C'_1)) where $T_1 =$ time from transfusion to AIDS, $T_2 =$ age at transfusion and $C_1 =$ time from transfusion to July 1986.

Since this data is right truncated, we apply the “mirroring” argument mentioned in the introduction to adapt the NPMLE S_n for this type of data. As a result, we obtain an estimate of the distribution function F , but, as mentioned before, the inferential results obtained for left truncation are applicable.

It should be noted that we discarded two questionable observations in this dataset. The first observation we discarded had a T'_1 value of zero, which means that this subject most likely contracted AIDS from a source other than the blood transfusion and is thus not of interest here. The other observation was one whose T'_1 value (89 months) was at the tail of the distribution, very close to the value of the observed truncation variable C'_1 (90 months). This point represents a violation of the assumptions stated in the appendix (specifically assumption 2). Recall that we are using the mirroring argument in our estimation for this dataset. The analog of the second assumption (for a point (t_1, t_2) we assume that $\int_0^{t_1} \int_0^{t_2} \frac{dF(s_1, s_2)}{G(s_1, s_2)} < \infty$) for the mirroring argument is that for a point (t_1, t_2) , we assume that $\int_{t_2}^{\infty} \int_{t_2}^{\infty} \frac{dF(s_1, s_2)}{G(s_1, s_2)} < \infty$. Essentially, violation of this assumption leads to an unbounded influence curve. A good diagnostic tool to guard against such violations is a plot of the actual values of the influence curve (y-axis) by subject (x-axis) for various values of (t_1, t_2) . Figure 1 is such a plot for the TR AIDS data. The points denoted by the “X” are those for the aforementioned subject with $T'_1 = 89$ and $C'_1 = 90$. We see that the values of the influence curves for this subject are significantly smaller compared with the other observations (especially for larger values of (t_1, t_2)). Later, we will see the effect that this kind of observation can have on the variance estimation. Thus, unless noted otherwise, the results that

we will present will be without these observations.

We first present two figures showing the error that would be made if the truncation on the time variable T_1 is ignored and F is estimated simply by the empirical distribution function. Figure 2 shows an estimate of the marginal distribution of age at transfusion along with the corresponding empirical distribution estimate. Although we see that the curves are quite similar, the empirical estimate is larger than the NPMLE for younger ages, which may reflect some dependence between age at transfusion and time from transfusion to AIDS at these younger ages. This dependence is reflected in the test results below. Figure 3 indicates that ignoring the truncation leads to severe overestimation of the marginal distribution of time from transfusion to AIDS, since this estimate does not account for the fact that we are less likely to see large values due to the right truncation of the time variable T_1 .

Figure 4 shows the estimated bivariate distribution of time from transfusion to AIDS and age at transfusion. In Figure 5 we see the estimated marginal distribution of time from transfusion to AIDS with associated 95% confidence intervals derived as in section 3. Figure 6 shows the estimator with the outlying observation mentioned above included, and we see that the variance estimates are inflated.

From a subject matter point of view, Figure 7 is probably the most interesting. Separate curves were fit according to the following age groups: children(1-4), adults (5-59) and elderly (60+). The age groups were determined according to immuno competence (Gürler, 1996). This plot demonstrates the wide differences in disease incubation profiles for the different age groups. While disease incubation does not seem to differ much between the adults and the elderly, we see that the disease incubation for children is much faster, as would be suspected. This is a scenario where the tests proposed in section 3.1.1 is extremely useful.

In Table 4 we present the result for the test of the independence of the events $T_1 \leq t_1$ and $T_2 \leq t_2$ for various values of t_1 and t_2 (keep in mind that t_1 is in units of months and t_2 is in units of years). The P-value for the global test was less than 0.000001, which may indicate that the time from transfusion to AIDS and age at transfusion are dependent random variables. The pointwise tests at the upper end of the distribution of (T_1, T_2) should be interpreted very carefully, as should the results of the global test. However, the tests closer to the lower end of the distribution of (T_1, T_2) will be more reliable (as they are away from the region of truncation) and can be more confidently interpreted as real departures from independence. We will thus

concentrate our analysis of the test results on the portions of Table 4 on the area from t_1 from 4 to 50 and t_2 from 4 to 50. The results in the table reflect what we see in Figure 7. We see that the following pairs of events appear to be dependent: $T_2 \leq 4$ years and $T_1 \leq 10$ months, $T_2 \leq 4$ years and $T_1 \leq 30$ months, $T_2 \leq 4$ years and $T_1 \leq 50$ months, $T_2 \leq 10$ years and $T_1 \leq 10$ months, $T_2 \leq 10$ years and $T_1 \leq 30$ months, $T_2 \leq 10$ years and $T_1 \leq 50$ months and $T_2 \leq 30$ years and $T_1 \leq 10$ months and $T_2 \leq 30$ years. Namely we see that age and disease incubation appear to be related for those of young age.

A Appendix - Assumptions necessary for asymptotic linearity and consistency of S_n

The assumptions that we must make in order to claim asymptotic linearity of S_n at the point (t_1, t_2) :

1. Let $t = (t_1, t_2)$ be such that $S(t) > \delta > 0$ and assume that S has a finite number of jumps and is continuous everywhere else.
2. Assume that $\int_0^t \frac{dF}{G} < \infty$ and $G \ll F$ (G is absolutely continuous with respect to F)
3. $\frac{G(x)dG(x)}{S(x)dF(x)} < M < \infty$ on $[0, t]$
4. $G(\{0\}) > 0$

We refer the reader to the work of van der Laan (van der Laan 1996) for details on the implications and implementation of these assumptions in the proof of asymptotic linearity.

B Appendix - Details on the derivation of the distribution of the test statistic for the independence test

Recall the definition of the test statistic $D_n(t_1, t_2)$ given by (5). We have, therefore, the following:

$$\begin{aligned}
D_n(t_1, t_2) - (S(t_1, t_2) - S(t_1, 0)S(0, t_2)) &= \\
&= S_n(t_1, t_2) - S(t_1, t_2) - (S_n(t_1, 0)S_n(0, t_2) - S(t_1, 0)S(0, t_2))
\end{aligned}$$

Note that under the hypothesis of independence of the events $T_1 \geq t_1$ and $T_2 \geq t_2$, $S(t_1, t_2) = S(t_1)S(t_2)$

Recall from section 3 that we may express $S_n(t_1, t_2) - S(t_1, t_2)$ as a sum of n i.i.d random variables (namely the influence curves at the points $(T'_{1i}, T'_{2i}, C'_{1i})$ for $i = 1, \dots, n$). We must now do the same for $S_n(t_1, 0)S_n(0, t_2) - S(t_1, 0)S(0, t_2)$. Using the aforementioned telescoping identity $a_n b_n - ab = (a_n - a)b_n + a(b_n - b)$, we get the following result:

$$\begin{aligned}
S_n(t_1, 0)S_n(0, t_2) - S(t_1, 0)S(0, t_2) &= \\
&= (S_n(t_1, 0) - S(t_1, 0))S_n(0, t_2) + (S_n(0, t_2) - S(0, t_2))S(t_1, 0) \\
&\approx (S_n(t_1, 0) - S(t_1, 0))S(0, t_2) + (S_n(0, t_2) - S(0, t_2))S(t_1, 0)
\end{aligned}$$

Now we use the fact that given points t_1 and t_2 , we can show the asymptotic linearity of $S_n(t_1, 0) - S(t_1, 0)$ and $S_n(0, t_2) - S(0, t_2)$. We will suppress the notation on the influence curves with respect to $T'_{1i}, T'_{2i}, C'_{1i}$ and F, G , but recognize that the summation is over the observations $T'_{1i}, T'_{2i}, C'_{1i}$, and that the influence curves are functionals of F and G . Therefore:

$$\begin{aligned}
D_n(t_1, t_2) - (S(t_1, t_2) - S(t_1, 0)S(0, t_2)) &= \\
&= S_n(t_1, t_2) - S(t_1, t_2) - \{S(0, t_2)(S_n(t_1, 0) - S(t_1, 0)) \\
&\quad + S(t_1, 0)(S_n(0, t_2) - S(0, t_2))\} \tag{6} \\
&\approx \frac{1}{n} \sum_{i=1}^n IC_i(t_1, t_2) - S(0, t_2) \frac{1}{n} \sum_{i=1}^n IC_{1,i}(t_1) - S(t_1, 0) \frac{1}{n} \sum_{i=1}^n IC_{2,i}(t_2) \\
&= \frac{1}{n} \sum_{i=1}^n IC_{test,i}(t_1, t_2)
\end{aligned}$$

where IC_i is the influence curve of S_n , $IC_{1,i}$ is the influence curve of $S_n(\cdot, 0)$ and $IC_{2,i}$ is the influence curve of $S_n(0, \cdot)$, and the influence curves are all evaluated at the observation $(T'_{1i}, T'_{2i}, C'_{1i})$. This gives us the stated influence curve for D_n given in Theorem 2.

As noted above, if the hypothesis that $T_1 \geq t_1$ and $T_2 \geq t_2$ are independent events is true, then $S(t_1, t_2) - S(t_1, 0)S(0, t_2) = 0$, which gives us that $E(D_n) = 0$. The central limit theorem applied to (6) gives us that under the independence hypothesis, as $n \rightarrow \infty$ $\sqrt{n}D_n(t_1, t_2)$ converges in distribution to a mean zero normal random variable with variance equal to $\text{var}(IC_{test}(T'_{1i}, T'_{2i}, C'_{1i} | F, G, t_1, t_2))$.

References

- [1] P.J. Bickel, C.A.J Klassen, Y. Ritov and J.A. Wellner, *Efficient and adaptive inference in semiparametric models*, Johns Hopkins University Press, Baltimore, 1993.
- [2] Gill, R.D., van der Laan, M.J., and Wellner, J.A., "Inefficient estimators of the bivariate survival function for three models," *Ann. Inst. H Poincaré Probab. Statist.* **31** pp. 545-597, 1995.
- [3] R.D Gill and N. Keiding, "Random truncation models and Markov processes," *Anns. Statist.* **18** pp. 582-602, 1990..
- [4] Ü. Gürler, "Bivariate Estimation with Right Truncated Data," *J. Am. Stat. Assoc.* **91** pp. 1152-1165, 1996.
- [5] Ü. Gürler, "Bivariate distribution and hazard functions when a component is randomly truncated," *J. Multivariate Anal.* **60** pp. 20-47, 1997.
- [6] N.A. Hessel, B.A. Koblin, et al., "Progression of Human Immunodeficiency Virus type 1 (HIV-1) infection among homosexual men in Hepatitis B vaccine trial cohorts in Amsterdam, New York City, and San Francisco, 1978-1991," *Am. J. Epidemiol* **139** pp. 1077-86, 1994.
- [7] C.M. Quale and M.J. van der Laan, "Inference with bivariate truncated data," *Tech. Rep #69, Group in Biost, UC Berkeley*, 1998 .
- [8] M.J. van der Laan, "Nonparametric estimation of the bivariate survivor function under random truncation," *J. Multivariate Anal.* **58** pp. 107-131, 1996..
- [9] A.W. van der Vaart, "On differentiable functionals," *Ann. Statist.* **19** pp. 178-204, 1991.
- [10] M.C. Wang, N.P. Jewell and W.Y. Tsai, "Asymptotic properties of the product limit estimate under random truncation," *Ann. Statist.* **14** pp. 1597-1605, 1986.
- [11] M.C. Wang, "A semiparametric model for randomly truncated data," *Anns. Statist.* **14** pp. 742-748, 1989.

- [12] M. Woodroffe, “Estimating a distribution function with truncated data,” *Ann. Statist.* **13** pp. 163-177, 1985. Correction: *Ann Statist* **15** pp. 883, 1987.

Table 1: Ratio of $\frac{MSE_{S_n(t_1, t_2)}}{MSE_{S_n^{Gur}(t_1, t_2)}}$ (Low truncation, High truncation) where $T_{1i} = 0.50R_{1i} + 0.50R_{3i}$, $T_{1i} = 0.50R_{2i} + 0.50R_{3i}$ for R_{ji} , $j = 1, 2, 3$ independent mean 10 exponential random variables and C_{1i} an independent exponential random variable with mean varied so that $n \approx 300$ for the two levels of truncation, based on 625 replicates

	$t_2 = 0$	$t_2 = 4$	$t_2 = 8$	$t_2 = 10$	$t_2 = 12$	$t_2 = 18$
$t_1 = 0$	NA,NA	0.89,0.80	0.99,0.94	1.03,0.96	1.00,0.95	1.01,0.98
$t_1 = 4$	0.94,0.87	0.94,0.87	0.98,0.94	1.01,0.99	1.00,0.98	1.00,0.99
$t_1 = 8$	0.93,0.92	0.92,0.93	0.96,0.96	1.01,0.99	1.00,1.00	1.02,1.00
$t_1 = 10$	0.94,0.92	0.93,0.91	0.97,0.95	1.01,0.97	1.00,0.98	1.02,1.00
$t_1 = 12$	0.96,0.92	0.95,0.91	0.98,0.92	1.01,0.94	1.00,0.97	1.03,0.99
$t_1 = 18$	0.96,0.87	0.93,0.88	0.98,0.93	1.01,0.94	1.00,0.96	1.09,0.99

Table 2: Empirical coverage probabilities (High truncation, Low truncation) for estimated confidence intervals of $S(t_1, t_2)$, where $T_{1i} = 0.50R_{1i} + 0.50R_{3i}$, $T_{1i} = 0.50R_{2i} + 0.50R_{3i}$ for R_{ji} , $j = 1, 2, 3$ independent mean 10 exponential random variables and C_{1i} an independent exponential random variable with mean varied so that $n \approx 300$ for the two levels of truncation, based on 625 replicates

	$t_2 = 0$	$t_2 = 4$	$t_2 = 8$	$t_2 = 10$	$t_2 = 12$	$t_2 = 18$
$t_1 = 0$	1.00,1.00	0.93,0.91	0.95,0.94	0.96,0.93	0.95,0.94	0.93,0.96
$t_1 = 4$	0.92,0.94	0.94,0.94	0.94,0.94	0.95,0.93	0.95,0.93	0.93,0.94
$t_1 = 8$	0.95,0.94	0.96,0.93	0.95,0.94	0.95,0.94	0.94,0.95	0.94,0.95
$t_1 = 10$	0.95,0.94	0.94,0.94	0.95,0.93	0.95,0.95	0.94,0.94	0.94,0.94
$t_1 = 12$	0.95,0.94	0.95,0.93	0.95,0.93	0.95,0.93	0.94,0.94	0.92,0.94
$t_1 = 18$	0.94,0.93	0.95,0.94	0.95,0.94	0.95,0.95	0.94,0.94	0.92,0.94

Table 3: Empirical rejection probabilities ($P(T \geq C) = 0.50, 0.75$) for the test of independence of the events $T_1 \geq t_1$ and $T_2 \geq t_2$ where T_1 and T_2 are independent mean 10 exponential random variables. The values of T_1 and T_2 correspond to the 0.25, 0.50 and 0.75 quantiles of the distribution of a mean 10 exponential random variable.

	$F_{0.25}$	$F_{0.50}$	$F_{0.75}$
$F_{0.25}$	0.08,0.06	0.08,0.05	0.08,0.07
$F_{0.50}$	0.07,0.06	0.06,0.06	0.06,0.07
$F_{0.75}$	0.08,0.06	0.07,0.06	0.05,0.04

Table 4: P-Values for test of independence of the events $T_1 \leq t_1$ and $T_2 \leq t_2$ for the TR-AIDS dataset. The P-Value for the global test was < 0.00001 .

	$t_2 = 4$	$t_2 = 10$	$t_2 = 30$	$t_2 = 50$	$t_2 = 70$	$t_2 = 80$
$t_1 = 4$	0.21	0.22	0.18	0.03	0.09	0.25
$t_1 = 10$	0.00	0.01	0.02	0.00	0.85	0.20
$t_1 = 30$	0.00	0.00	0.43	0.35	0.95	0.26
$t_1 = 50$	0.00	0.05	0.82	0.49	0.68	0.07
$t_1 = 70$	0.01	0.02	0.00	0.48	0.34	0.48
$t_1 = 80$	0.28	0.28	0.26	0.26	0.26	0.84

Figure 1: Plot of influence curve by subject for various values of (t_1, t_2) (the “X” denotes the possible outlier)

Figure 2: Comparison of the empirical estimate (no account of truncation) and NPMLE of the distribution of age at transfusion

Figure 3: Comparison of the empirical estimate (no account of truncation) and NPMLE of the distribution of Time from Transfusion to AIDS

Figure 4: Estimated bivariate distribution of time from transfusion to AIDS and age at transfusion

Figure 5: Estimated marginal distribution of time from transfusion to AIDS with pointwise 95% Confidence Intervals (outlier removed)

Figure 6: Estimated marginal distribution of time from transfusion to AIDS with pointwise 95% Confidence Intervals (outlier included)

Figure 7: Time from transfusion to AIDS, by age group