Published in final edited form as:

Lifetime Data Anal. 2014 January ; 20(1): 23–37. doi:10.1007/s10985-013-9263-7.

Cross-Ratio Estimation for Bivariate Failure Times with Left Truncation

Tianle Hu,

Eli Lilly and Company, Indianapolis, IN 46285

Xihong Lin, and

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

Bin Nan

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109

Abstract

The cross-ratio is an important local measure that characterizes the dependence between bivariate failure times. To estimate the cross-ratio in follow-up studies where delayed entry is present, estimation procedures need to account for left truncation. Ignoring left truncation yields biased estimates of the cross-ratio. We extend the method of Hu et al. (2011) by modifying the risk sets and relevant indicators to handle left-truncated bivariate failure times, which yields the cross-ratio estimate with desirable asymptotic properties that can be shown by the same techniques used in Hu et al. (2011). Numerical studies are conducted.

Keywords

Bivariate survival; Cross-ratio; Left truncation; Pseudo-partial likelihood; Right censoring

1 Introduction

One common focus in the analysis of correlated random variables is the correlation structure. Non-parametric dependence measures such as Kendall's τ and Spearman's correlation coefficient can be used to characterize the dependence. But for the purpose of interpretation, measures specifically developed for bivariate survival data are preferred. The most prominent one of these is the cross-ratio. Formulated as the ratio of two conditional hazard functions, the cross-ratio measures the relative hazard of one time component conditional on another time component at some time point and beyond (Kalbfleisch and Prentice 2002, chap. 10).

Clayton (1978) considered a constant cross-ratio that yields an explicit closed-form bivariate survival function, known as the Clayton copula model. Reparameterizing the Clayton model, Oakes (1982, 1986, 1989) formulated the bivariate failure time distribution in the frailty model framework, where a common latent frailty variable induces correlation between survival times. For inference on a constant cross-ratio when the marginal distributions are completely unspecified, many parameter estimation methods have been developed. In the absence of censoring, Clayton (1978) proposed a pseudo likelihood approach by combining a series of conditional probabilities; Oakes (1982, 1986) proposed an estimator based on Kendall's τ . In the presence of censoring, Shih and Louis (1995) investigated the two-stage

semiparametric estimation procedure, where Kaplan-Meier estimators were used to estimate the marginal survival functions; Glidden (2000) extended the approach of Shih and Louis (1995) to allow the marginal hazard for failure times to follow a stratified Cox model; Glidden and Self (1999) discussed an approximate EM-algorithm for the maximum likelihood estimation with marginal Cox specifications; and Prentice and Hsu (1997) adopted an estimating equation approach for the estimation of the marginal hazard ratio and the cross-ratio.

The Clayton copula belongs to a larger class known as the Archimedean copulas. For the Archimedean copulas, Oakes (1989) showed that the cross-ratio function depends on time only through some function of the joint survival probability. Other common copulas which do not belong to the Archimedean class include the Gaussian copula. For instance, Li and Lin (2006) and Li, Prentice, and Lin (2008) characterized the dependence of bivariate survival data through the correlation coefficient of normally transformed bivariate survival times. Copula methods, Archimedean or not, provide a convenient way of characterizing the cross-ratio function. The choice of a specific copula and its robustness in terms of the cross-ratio estimation would be of interest and worth further investigation.

To estimate the cross-ratio as a function of time explicitly, Nan, Lin, Lisabeth, and Harlow (2006) partitioned the sample space of the bivariate survival time into rectangular regions and assumed that the cross-ratio is constant in each rectangular region. A Clayton-type model was established for the joint survival function, and a two-stage likelihood based method similar to Shih and Louis (1995) was used to estimate the piecewise constant cross-ratio. In the context of competing risks, Bandeen-Roche and Ning (2008) proposed a non-parametric method for estimating the piecewise constant time-varying cause-specific cross-ratio using the binned survival data. Motivated by Cox's partial likelihood idea, Hu, Nan, Lin and Robins (2011) proposed an pseudo-partial likelihood method, where they approximated the logarithm of an arbitrary cross-ratio function by finite dimensional polynomial functions of bivariate survival times.

Related to the cross-ratio, a class of weighted average of local dependence measures have been proposed (Fan, Hsu and Prentice 2000; Fan, Prentice and Hsu 2000). Fan and Prentice (2002) later generalized these dependence measures to accommodate regression effects on the marginal hazard functions. Hsu, Prentice, Zhao and Fan (1999) considered an estimating equation approach for nonparametric estimation of the cross-ratio in matched case-control studies. Hsu and Prentice (1996) also considered the correlation between cumulative hazard variates as a measure of dependence between pairs of failure time variates from a different perspective. Gorfine, Hsu and Prentice (2003) considered the cross-ratio estimation in the Clayton-Oakes model in the presence of measurement error in the covariate by introducing a bias reduction technique for the bivariate survival function in copula models while assuming an additive measurement error model and replicated measurements for the covariates.

This article concerns the cross-ratio estimation for bivariate survival data in a general setting. As in univariate survival analysis, observations in bivariate survival applications are often subject to delayed entry, also known as left truncation. In such studies, only pairs in which both failure events occur after the corresponding left truncation times are observed. Consider a bivariate lefttruncated, right-censored data set (Ino et al. 2001) comprising survival time pairs for 50 brain tumor patients. Time from diagnosis to initiation of radiation therapy and time from diagnosis to tumor progression are left-truncated by time from diagnosis to chemotherapy and right-censored by time of last follow-up. To be precise, let T_1 be the time from diagnosis to initiation of radiation therapy, T_2 be the time from diagnosis to tumor progression and U be the time from diagnosis to chemotherapy. The failure times (T_1, T_2) are observed only if $T_1 > U$ and $T_2 > U$. Martin and Betensky (2005)

used this data set to study the quasi-independence of (T_1, T_2) and U . In this paper we shift the focus to the dependence between T_1 and T_2 , characterized by the cross-ratio in the presence of truncation time U as well as right censoring.

Like left truncation, right truncation could also arise in AIDS incubation cohort study of HIV-positive subjects without AIDS. One well known such dataset is Transfusion Related AIDS (TR AIDS) data (Lagakos et al. 1988; Kalbfleisch and Lawless 1989; Gürlér 1996). With reverse-time transformation, the right-truncated data can be analyzed as left-truncated data.

We propose a pseudo-partial likelihood estimator for bivariate survival data subject to both left truncation and right censoring. The method is an extension of the estimation procedure developed by Hu et al. (2011) which approximates an arbitrary cross-ratio function on a log scale using a set of basis functions. Each pair of continuous failure times (T_1, T_2) are subject to left truncation and right censoring respectively by a pair of truncation times (U_1, U_2) and censoring times (C_1, C_2) . We estimate the dependence between T_1 and T_2 based on observed data of the form $(U_1, U_2, X_1, X_2, \Delta_1, \Delta_2)$ where $X_1 = \min(T_1, C_1)$, $X_2 = \min(T_2, C_2)$, $\Delta_1 = I(T_1 < C_1)$, and $\Delta_2 = I(T_2 < C_2)$.

This article is organized as follows. In Section 2, we propose the pseudo-likelihood approach for left-truncated bivariate survival data. In Section 3, we provide asymptotic results based on key results previously reported in Hu et al. (2011). In section 4, we conduct simulation studies by generating data from the Frank family and the piecewise cross-ratio model. We also randomly generate left truncation times to the Australian Twin Study data set to illustrate the validity of the proposed approach. We analyze two real data sets in Section 5, and provide a brief discussion in Section 6. Appendix contains a detailed asymptotic expression of the estimating function resulted from the pseudo-partial likelihood.

2 Cross-ratio estimation for left-truncated data

Let (T_1, T_2) be a pair of absolutely continuous failure times. The cross-ratio function of T_1 and T_2 (Clayton 1978; Oakes 1982, 1986, 1989) is defined as

$$\theta(t_1, t_2) = \frac{\lambda_2(t_2|T_1=t_1)}{\lambda_2(t_2|T_1>t_1)} = \frac{\lambda_1(t_1|T_2=t_2)}{\lambda_1(t_1|T_2>t_2)}, \quad (1)$$

where λ_1 and λ_2 are the conditional hazard functions of T_1 and T_2 , respectively. The problem here is to estimate the cross-ratio function $\theta(t_1, t_2)$ based on n independent and identically distributed (i.i.d.) random draws from the conditional distribution of $(U_1, U_2, X_1, X_2, \Delta_1, \Delta_2)$ given that $(X_1 > U_1, X_2 > U_2)$. In other words, we only observe $(U_1, U_2, X_1, X_2, \Delta_1, \Delta_2)$ if $(X_1 > U_1, X_2 > U_2)$. The estimator we propose makes use of the connection between the cross-ratio definition in (1) and the Cox model partial likelihood for the two-group regression problem.

Following Hu et al. (2011), the partial likelihood function is constructed in the following manner. On one time axis the occurrence of an event at some time point or beyond is treated as a binary covariate, whereas the other time component is treated as the survival outcome variable. We then modify the risk sets and relevant indicators to handle left truncations just like in the Cox model. Parameter estimates are obtained by maximizing the constructed pseudo-partial likelihood function. Such construction does not need any model for either the joint or the marginal survival function, and is robust against model mis-specification.

Specifically, treating $\{j : T_{1j} = t_1\}$ and $\{j : T_{1j} > t_1\}$ as the “exposure” and “non-exposure” groups and T_{2j} as the survival outcome, the cross-ratio $\theta(t_1, t_2)$ can be viewed as the hazard

ratio of T_2 between these two groups. To further account for the biased sampling induced by left truncation, we propose redefining the at-risk process by taking into account the truncation information. Assume there are no ties among observed times for each of the two time components. Mimicking the partial likelihood idea and accounting for left truncation through modified risk sets and the relevant indicators, we can construct an objective function as follows:

$$L_n = \prod_{i=1}^n L_i^{(1)} L_i^{(2)}, \quad (2)$$

with

$$L_i^{(1)} = \prod_{j=1}^n \left[\frac{\theta(X_{1i}, X_{2j})^{I(X_{1j}=X_{1i} \geq U_{1j})}}{N(X_{1i}, X_{2j}) + (\theta(X_{1i}, X_{2j}) - 1)I(U_{2i} \leq X_{2j} \leq X_{2i})} \right]^{I(X_{1j} \geq X_{1i} \geq U_{1j})\Delta_{2j}\Delta_{1i}},$$

$$L_i^{(2)} = \prod_{j=1}^n \left[\frac{\theta(X_{1j}, X_{2i})^{I(X_{2j}=X_{2i} \geq U_{2j})}}{N(X_{1j}, X_{2i}) + (\theta(X_{1j}, X_{2i}) - 1)I(U_{1i} \leq X_{1j} \leq X_{1i})} \right]^{I(X_{2j} \geq X_{2i} \geq U_{2j})\Delta_{1j}\Delta_{2i}},$$

where $N(t_1, t_2) = \sum_{k=1}^n I(X_{1k} \geq t_1 \geq U_{1k}, X_{2k} \geq t_2 \geq U_{2k})$ is the size of risk set at (t_1, t_2) , and $L_i^{(2)}$ is defined as $L_i^{(1)}$ with the roles of T_1 and T_2 switched. The estimator obtained by maximizing (2) is then called the pseudo-partial likelihood estimator. A referee pointed out that this approach, in absence of left truncation and right censoring, resembles a similar idea considered by Clayton (1978) for a constant θ . However, as pointed out by Oakes (1982), the asymptotic theory provided in Clayton (1978) does not hold. Hu et al. (2011) provided the asymptotic theory for a general setup that includes Clayton's estimator as a special example.

To proceed, we consider a parametric model for the cross-ratio

$$\beta(t_1, t_2; \gamma) = \log \theta(t_1, t_2)$$

with a Euclidean parameter γ and denote $\dot{\beta}_\gamma(t_1, t_2; \gamma) = \dot{\beta}(t_1, t_2; \gamma) / \gamma$. Taking the logarithm of the objective function L_n and differentiating it with respect to γ we obtain the following estimating function for γ :

$$U_n(\gamma) = \partial \log L_n(\gamma) / \partial \gamma = U_n^{(1)}(\gamma) - U_n^{(2)}(\gamma) + U_n^{(3)}(\gamma) - U_n^{(4)}(\gamma),$$

where

$$U_n^{(1)}(\gamma) = U_n^{(3)}(\gamma) = \frac{1}{n} \sum_{i=1}^n \Delta_{1i} \Delta_{2i} \dot{\beta}_\gamma(X_{1i}, X_{2i}; \gamma),$$

and

$$U_n^{(2)}(\gamma) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \frac{\Delta_{1i} \Delta_{2j} I(X_{1j} \geq X_{1i} \geq U_{1j}) I(U_{2i} \leq X_{2j} \leq X_{2i})}{N(X_{1i}, X_{2j}) - I(U_{2i} \leq X_{2j} \leq X_{2i})(1 - e^{\beta(X_{1i}, X_{2j}; \gamma)})} \times e^{\beta(X_{1i}, X_{2j}; \gamma)} \dot{\beta}_\gamma(X_{1i}, X_{2j}, \gamma),$$

$$U_n^{(4)}(\gamma) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \frac{\Delta_{1j} \Delta_{2i} I(X_{2j} \geq X_{2i} \geq U_{2j}) I(U_{1i} \leq X_{1j} \leq X_{1i})}{N(X_{1j}, X_{2i}) - I(U_{1i} \leq X_{1j} \leq X_{1i})(1 - e^{\beta(X_{1j}, X_{2i}; \gamma)})} \times e^{\beta(X_{1j}, X_{2i}; \gamma)} \dot{\beta}_\gamma(X_{1j}, X_{2i}, \gamma).$$

Then an estimator $\hat{\gamma}_n$ can be obtained by solving the equation $U_n(\gamma) = 0$ using the Newton-Raphson algorithm.

3 Asymptotic properties

We consider a parametric model with a finite number of basis functions for $\beta(t_1, t_2; \gamma)$. In particular, we assume

$$\beta(t_1, t_2; \gamma) = z(t_1, t_2)' \gamma, \quad (3)$$

where γ is the vector of coefficients. It is easily seen that $\beta_\gamma(t_1, t_2, \gamma) = z(t_1, t_2)$ and is free of γ

We find that a quadratic or cubic polynomial model $z(t_1, t_2)' \gamma = \sum_{k,l} \gamma_{kl} t_1^k t_2^l$ in (3) often yields satisfactory estimates for a smooth cross-ratio functions. Other parametric models, e.g. the piecewise constant model of Nan et al. (2006), can also be considered. In this section, we provide asymptotic results for the estimation of γ in (3). In addition to the mutual independence of the pairs of times (T_1, T_2) , (C_1, C_2) and (U_1, U_2) , we also consider the following regularity conditions:

C1 The failure times are truncated at (τ_1, τ_2) , $0 < \tau_1, \tau_2 < \infty$, such that $\Pr(T_1 > \tau_1 > U_1, C_1 > \tau_1 > U_1, T_2 > \tau_2 > U_2, C_2 > \tau_2 > U_2) > 0$.

C2 The parameter space Γ is a compact set and the true value γ_0 is an interior point of Γ .

C3 The matrix $E\{\Delta_1 \Delta_2 z(X_1, X_2)^{\otimes 2} | X_1 \geq U_1, X_2 \geq U_2\}$ is positive definite. Here $z^{\otimes 2} = zz'$.

Theorem 1 Under Conditions C1–C3, the solution of $U_n(\gamma) = 0$, denoted by $\hat{\gamma}_n$, is a consistent estimator of γ_0 . Furthermore, $n^{1/2}(\hat{\gamma}_n - \gamma_0)$ converges in distribution to a normal random variable with mean zero and variance $\mathbf{I}(\gamma_0)^{-1} \Sigma(\gamma_0) \mathbf{I}(\gamma_0)^{-1}$, where

$$\mathbf{I}(\gamma_0) = 2E\left\{\Delta_1 \Delta_2 z(X_1, X_2)^{\otimes 2} | X_1 \geq U_1, X_2 \geq U_2\right\}$$

and $\Sigma(\gamma_0)$ is the asymptotic variance of $U_n(\gamma_0)$, which is provided in the Appendix.

The proofs of consistency and asymptotic normality in Theorem 1 are essentially no different to the proofs of Theorems 1 and 2 in Hu et al. (2011), except for the additional care given to the risk sets with appropriately adjusted indicator functions for left truncation. Details are therefor omitted. The asymptotic representation of $U_n(\gamma_0)$ leads to an estimator of $\Sigma(\gamma_0)$.

4 Simulations

4.1 Frank family

First, we consider the Frank family with unit exponential marginal distributions as in Hu et al (2011). The corresponding cross-ratio function is

$$\frac{(\alpha - 1)\log(\alpha)\alpha^{2-e^{-t_1}-e^{-t_2}}}{(\alpha^{1-e^{-t_1}} - \alpha) \times (\alpha^{1-e^{-t_2}} - \alpha)} \left[-1 + e^{-t_1} + e^{-t_2} + \log_{\alpha} \left\{ 1 + \frac{(\alpha^{1-e^{-t_1}} - 1)(\alpha^{1-e^{-t_2}} - 1)}{\alpha - 1} \right\} \right].$$

We choose $\alpha = 0.0023$ and generate truncation times (U_1, U_2) and censoring times (C_1, C_2) from a Uniform $(0, 1)$ and a Uniform $(0, 2.3)$ distribution respectively. The estimated cross-ratio is obtained using the quadratic polynomial model by maximizing the pseudo-partial likelihood function (2) with respect to coefficients γ . Note that our estimator only provides an approximation of the cross-ratio in the Frank family. Results are averaged over 1000 simulation runs, each with a sample of 300 observations, of which approximately 50% are censored. Approximately 35% of the data are truncated, and therefore not included in the observed data.

The top panel of Figure 1 compares our estimator with the naive estimator that ignores left truncation. It can be seen that our estimator corrects the bias. The bottom panel gives the cross-ratio as a function of one time component while fixing the other time component between $t = 0.25$ and $t = 1.50$ with a 0.25 time unit increment. Confidence bands for β are calculated based on the empirical variance of $\hat{\gamma}$. By exponentiating, we obtain the empirical confidence bands for θ . Figure 1 shows that the proposed method estimates the true cross-ratio of the Frank family very well, despite the fact that the working model is only an approximation of the true $\theta(t_1, t_2)$. From Figure 1 we also observe increased variability for small t_1/t_2 and reduced variability for large t_1/t_2 compared with results in Hu et al (2011) where a cubic polynomial log cross-ratio was applied with a sample size of 400. This is not unexpected because an important effect of left truncation is that the sizes of the risk sets do not necessarily decrease in time due to late entrants. So it may no longer be true that cross-ratio functions for large (t_1, t_2) are estimated less precisely than those for small (t_1, t_2) .

To check the performance of the model based variance estimator, we choose nine points based on the quartiles of the marginal failure time distributions. The empirical variances based on 1000 replications are compared with the average of the model based variance estimates. The results given in Table 1 show that the model based variance estimator works well.

4.2 Piecewise constant model

We also simulate data using the algorithm in Nan et al. (2006). First we assume that the cross-ratio $\theta(t_1, t_2)$ is piecewise constant over four intervals: $\theta(t_1, t_2) = .9$ when $t_1 \in [0, .25)$, $\theta(t_1, t_2) = 2.0$ when $t_1 \in [.25, .5)$, $\theta(t_1, t_2) = 4.0$ when $t_1 \in [.5, .75)$, and $\theta(t_1, t_2) = 1.5$ when $t_1 > .75$. Additionally, we generate truncation times U_1 and U_2 from Uniform(0,1). The results based on sample sizes of 400 and 800 summarized in Table 2 show that our estimators work well. The proposed variance model based variance estimator again works well since the coverage probabilities are all close to the nominal coverage probability, 95%.

4.3 Australian Twin Study

For the purpose of illustration, we also analyze the Australian Twin Study data set (Duffy et al. 1990; Hu et al. 2011). The original data set is not left truncated, but we artificially generate bivariate truncation times to illustrate how our proposed method can be applied to correct bias. Moreover since we have results based on complete data which we can compare the new results with, we can assess the performance of the proposed method.

Applying our method with a cubic polynomial model for the log cross-ratio, we compare the average results from 100 truncated data sets with 20% truncation rate with the complete data

analysis for monozygotic twins and dizygotic twins separately (Figure 2). Fixing one time component at 10, 20 and 30, it is clear that our proposed method successfully recovers the cross-ratio estimated from the complete data.

5 Data Application

5.1 Anaplastic Oligodendroglioma Study

We first illustrate the the proposed method with the left-truncated, right-censored anaplastic oligodendroglioma data set (Ino et al. 2001). Malignant gliomas are the most common type of primary brain tumor, with 12000 new cases diagnosed each year in the United States. Surgical resection and radiation therapy have been the mainstays of treatment. Clearly, more precise identification of treatment sensitive tumors and potential long-term survivors at diagnosis would afford greater flexibility in initial treatment decisions for such patients. Improved identification of resistant tumors might justify prescribing novel treatments at diagnosis for patients with a poor prognosis. With a rational basis for delivering or withholding radiotherapy, effective treatments can be appropriately prescribed, and ineffective and costly treatments specifically avoided.

The data set contains 50 patients treated at the London Regional Cancer Center with histologically confirmed anaplastic oligodendrogliomas in whom chemotherapy was used as an integral part of an overall patient management strategy from diagnosis. Thirty-four of these patients received radiation therapy after chemotherapy (as consolidation or at recurrence), 5 received radiation therapy concurrent with chemotherapy, and 11 were not irradiated. Median follow-up time from diagnosis was 107 months (minimum, 7 months).

Time from diagnosis to initiation of radiation therapy T_1 and time from diagnosis to tumor progression T_2 are left-truncated by time from diagnosis to chemotherapy and right-censored by time from diagnosis to the end of study. Since the number of observations in the study is small, we start with quadratic polynomial basis functions for $\beta(t_1, t_2; \gamma)$, i.e.,

$$\beta(t_1, t_2; \gamma) = \gamma_0 + \gamma_1 t_1 + \gamma_2 t_2 + \gamma_3 t_1^2 + \gamma_4 t_2^2 + \gamma_5 t_1 t_2$$

and use backward elimination for variable selection with $p = 0.15$ as the cut-off. The final selected model is $\hat{\theta}(t_1, t_2) = \exp(2.86 - 1.72 \cdot 10^{-1} \cdot t_2 + 2.10 \cdot 10^{-3} \cdot t_2^2)$ (Figure 3). It is interesting to see that the dependence between T_1 and T_2 is a function of T_2 only and this dependence is positive for $T_2 < 25$ months and negative, though not statistically significant, after 25 months. A careful examination of the data shows that majority of the subjects received radiation therapy within 25 months and about 40% of the subjects had recurrence within 25 months as well. Because radiation therapy is typically given right after recurrence, this early positive dependence is not unexpected. However, our analysis hints at a potential subgroup of patients with delayed recurrences who could benefit from early radiation therapy, although caution needs to be exercised over the interpretation of the results given that the pointwise 95% confidence band contains $\theta = 1$. Due to the small sample size of this trial and the fact that we are estimating a parametric surface, further study that is powered to detect such pattern is warranted. Meanwhile, exploratory biomarker analysis may help elucidate if such subgroup of patients that could benefit from earlier radiation therapy really exists.

5.2 Transfusion related AIDS

Like left truncation, right truncation could also arise in an AIDS incubation cohort study of HIV-positive subjects without AIDS. One well known such data set is the Transfusion

Related AIDS (TR AIDS) data (Lagakos et al. 1988; Kalbfleisch and Lawless 1989; Gürlér 1996). Of patients who were thought to be infected with HIV by blood or blood-product transfusion, the data record include the patient's age T_1 , the date of HIV infection T , and the date of AIDS diagnosis S . The observation period is terminated at U , which is same for all patients. Only those individuals for whom the incubation time (the time from HIV infection to the manifestation of AIDS) $T_2 = S - T$ is less than $U_2 = U - T$ can be observed, where T_2 is right truncated by U_2 . The primary interest here is in the association between incubation time and age at infection. One would suspect that disease progression is dependent on age at HIV infection. The established inferential procedures for a right truncated sample focus on the reverse-time transformation. Specifically, let $\tau = \max(U_2)$ be the largest observed time in the truncated sample. The transformed variable $T_2^* = \tau - T_2$ is left truncated by $U_2^* = \tau - U_2$. Methods developed in the context of left truncation can be readily extended to right truncation with reverse-time transformation. For the TR AIDS data, where T_2 is of primary interest, Lagakos et al. (1988) studied the distribution function on the reverse time axis. Kalbfleisch and Lawless (1989) discussed the Cox regression analysis on the reverse-time hazard.

The data set includes 295 observations of 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 age at transfusion, T_2 as time from transfusion to AIDS, and U_2 as time from transfusion to July 1986, we are only able to observe (T_1, T_2, U_2) if $T_2 \leq U_2$. This is a special case of randomly right truncated bivariate survival data where $U_1 = \infty$.

We discarded one questionable observation in this data set. The observation we discarded had a T_2 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. In order to apply the method developed for left truncated data, we reverse the time axis of T_2 by subtracting it from the maximum U_2 . Specifically, let $\tau = \max(U_2)$ be the largest observed time in the truncated sample. The transformed variable T_2^* is left truncated by $U_2^* = \tau - U_2$ (Figure 4).

Applying the proposed method to the transformed data, we fit the following model and obtain an estimate of the cross-ratio for age and reversed incubation time:

$$\beta(t_1, t_2^*; \gamma) = \gamma_0 + \gamma_1 t_1 + \gamma_2 t_2^* + \gamma_3 t_1^2 + \gamma_4 t_2^{*2} + \gamma_5 t_1 t_2^*.$$

Intuitively, the dependence between T_1 and T_2 is in the opposite direction of the dependence of T_1 and T_2^* . To ease interpretation, we further take the reciprocal of the estimated cross-ratio so that, with respect to T_1 , the result can be interpreted as the dependence between T_1 and T_2 . Based on the estimated cross-ratio surface, we see that age and disease incubation are highly associated at a younger age. This implies that for children, disease incubation is much shorter than for adults. Such dependence diminishes quickly for older patients. This finding is consistent with previous findings based on discrete age groups (children (1–4), adults (5–59) and elderly (60+)) determined according to immuno competence (Lagakos et al. 1988; Kalbfleisch and Lawless 1989; Gürlér 1996).

6 Discussion

Most approaches of estimating the cross-ratio as a measure of dependence among correlated failure times have been based on the assumption that the cross-ratio function is a constant or piece-wise constant over time. With such an assumption, it is possible to come up with a likelihood based estimation method for the cross-ratio. However, it is not clear how flexible

those approaches are when left truncation arises, which is very common in observational studies. In this paper, we show how a pseudo-partial likelihood approach motivated by the Cox model can be applied to doubly truncated data or bivariate left-truncated, right-censored data, that cannot be easily handled by existing methods for estimating the cross-ratio. The use of risk sets as in the Cox partial likelihood function allows simple modification to accommodate left truncation.

We analyzed TR AIDS data which is right truncated. A right-truncated variable can be converted to a left-truncated variable if one reverses the time axis. Using a model developed for left truncation with the reverse-time technique, we estimated the cross-ratio for age and reversed incubation time. However, interpretation of such a result is awkward, and therefore, inference on the regular forward-time cross-ratio for right truncated data warrants further investigation.

van der Laan (1996) described a study collecting bivariate survival times of HIV positive patients where T_1 is the incubation time (lag time between human immunodeficiency virus infection and AIDS) and T_2 is the survival time (lag time between HIV infection and death). T_2 is left-truncated if HIV infection occurs before the start of the study. In that study, one might be interested in the dependence between T_1 and T_2 . The cross-ratio is well suited for characterizing such a dependence. However, one needs to be cautious if T_2 can censor T_1 .

Additionally, developing novel methods for estimating the covariate effect on the cross-ratio in the presence of both left truncation and right censoring is of interest. The idea of constructing a similar pseudo-partial likelihood seems to apply, but extra care needs to be taken. This problem is under investigation and will be pursued elsewhere

Acknowledgments

Lin's research is supported in part by NCI grants R37CA76404 and P01-CA134294. Nan's research is supported in part by NSF grant DMS-1007590 and NIH grant R01-AG036802. The authors would like to thank Dr. David Duffy for the Australian Twin Study data, Dr. Emily Martin and Dr. Rebecca Betensky for the anaplastic oligodendroglioma data, and the Editor and two anonymous referees for their constructive comments.

Appendix

We use similar notation as in Hu et al. (2011). Let $(U_1^*, U_2^*, X_1^*, X_2^*, \Delta_1^*, \Delta_2^*)$ be an independent copy of $(U_1, U_2, X_1, X_2, \Delta_1, \Delta_2)$. Let $S(t_1, t_2) = \Pr(X_1 \geq t_1, U_1, X_2 \geq t_2, U_2 | X_1 \geq U_1, X_2 \geq U_2)$. Following van der Vaart and Wellner (1996), we use $\mathbb{P}_n$ and $\mathbb{Q}_n$ to denote the empirical measures of n independent copies of $(U_1^*, U_2^*, X_1^*, X_2^*, \Delta_1^*, \Delta_2^*)$ and $(U_1, U_2, X_1, X_2, \Delta_1, \Delta_2)$ that follow the distributions P and Q , respectively. In fact, $\mathbb{P}_n \equiv \mathbb{Q}_n$ and $P \equiv Q$ because the two samples are identical. The purpose of creating such notation is to keep track the double summations in $U_n(\gamma)$. Let

$$\tilde{g}(\Delta_2, U_1, U_2, X_1, X_2, \Delta_1^*, U_1^*, U_2^*, X_1^*, X_2^*; \gamma) = \Delta_1^* \Delta_2 z(X_1^*, X_2) \frac{I(X_1 \geq X_1^* \geq U_1) I(U_2^* \leq X_2 \leq X_2^*) e^{\beta(X_1^*, X_2; \gamma)}}{S(X_1^*, X_2)},$$

and define

$$\begin{aligned} \tilde{h}_Q(\delta_1, u_1, u_2, x_1, x_2; \gamma) \\ &= Q \tilde{g}(\Delta_2, U_1, U_2, X_1, X_2, \delta_1, u_1, u_2, x_1, x_2; \gamma), \tilde{h}_P(\delta_2, u_1, u_2, x_1, x_2; \gamma) \\ &= P \tilde{g}(\delta_2, u_1, u_2, x_1, x_2, \Delta_1^*, U_1^*, U_2^*, X_1^*, X_2^*; \gamma). \end{aligned}$$

Also define

$$f(\delta_1, u_1, u_2, x_1, x_2, \delta_2^*, u_1^*, u_2^*, x_1^*, x_2^*) = \frac{\delta_1^* \delta_2 I(x_1 \geq x_1^* \geq u_1) I(u_2^* \leq x_2 \leq x_2^*) \theta(x_1^*, x_2; \gamma_0)}{S(x_1^*, x_2)^2}.$$

Denote $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - P)$, we have

$$\begin{aligned} n^{1/2} \mathbb{U}_n(\gamma_0) &= 2\mathbb{G}_n \left\{ \Delta_1 \Delta_2 z(X_1, X_2) - \tilde{h}_Q(\Delta_1, U_1, U_2, X_1, X_2; \gamma_0) + \tilde{h}_p(\Delta_2, U_1, U_2, X_1, X_2; \gamma_0) + \iint I(X_1 \geq x_1^* \geq U_1, X_2 \geq x_2 \geq U_2 \right. \\ &\quad \left. + o_p(1) \rightarrow_d N(0, \Sigma(\gamma_0)) \right\}. \end{aligned}$$

References

1. Bandeen-Roche K, Ning J. Non-parametric estimation of bivariate failure time associations in the presence of a competing risk. *Biometrika*. 2008; 95:221–232. [PubMed: 20305739]
2. Clayton DG. A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika*. 1978; 65:141–151.
3. Cox DR. Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society B*. 1972; 34:187–220.
4. Duffy DL, Martin NG, Matthews JD. Appendectomy in Australian Twins. *American Journal of Human Genetics*. 1990; 47:590–592. [PubMed: 2248678]
5. Fan J, Hsu L, Prentice RL. Dependence Estimation over a Finite Bivariate Failure Time Region. *Lifetime Data Analysis*. 2000; 6:343–355. [PubMed: 11190604]
6. Fan J, Prentice RL. Covariate-Adjusted Dependence Estimation on a Finite Bivariate Failure Time Region. *Statistica Sinica*. 2002; 12:689–705.
7. Fan J, Prentice RL, Hsu L. A Class of Weighted Dependence Measures for Bivariate Failure Time Data. *J. R. Statist. Soc. B*. 2000; 62:181–190.
8. Glidden DV. A Two-Stage Estimator of the Dependence Parameter for the Clayton-Oakes Model. *Lifetime Data Anal.* 2000; 6:141–156. [PubMed: 10851839]
9. Glidden DV, Self SG. Semiparametric Likelihood Estimation in the Clayton–Oakes Failure Time Model. *Scand. J. Statist.* 1999; 26:363–372.
10. Gorfine M, Hsu L, Prentice RL. Estimation of Dependence between Paired Correlated Failure Times in the Presence of Covariate Measurement Error. *Journal of the Royal Statistical Society. Series B*. 2003; 65(3):643–661.
11. Gürlér Ü. Bivariate Estimation With Right Truncated Data. *J. Am. Statist. Assoc.* 1996; 91:1152–1165.
12. Gürlér Ü. Bivariate Distribution and Hazard Functions When a Component Is Randomly Truncated. *J. Multivariate Anal.* 1997; 60:20–47.
13. Hsu L, Prentice RL. On Assessing the Strength of Dependency between Failure Time Variates. *Biometrika*. 1996; 83:491–506.
14. Hsu L, Prentice RL, Zhao LP, Fan JJ. On dependence estimation using correlated failure time data from case-control family studies. *Biometrika*. 1999; 86:743–753.
15. Hu T, Nan B, Lin X, Robins J. Time-Dependent Cross-Ratio Estimation for Bivariate Failure Times. *Biometrika*. 2011; 98(2):341–354. [PubMed: 22822258]
16. Ino Y, Betensky RA, Zlatescu MC, Sasaki H, Macdonald DR, Stemmer-Rachamimov AO, Ramsay DA, Cairncross JG, Louis DN. Molecular Subtypes of Anaplastic Oligodendroglioma: Implications for Patient Management at Diagnosis. *Clinical Cancer Research*. 2001; 7:839–845. [PubMed: 11309331]

17. Kalbfleisch JD, Lawless JF. Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association*. 1989; 84:360–372.
18. Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. 2nd Ed. New York: Wiley; 2002.
19. Lagakos SW, Barraj LM, De Gruttola V. Nonparametric Analysis of Truncated Survival Data With Applications to AIDS. *Biometrika*. 1988; 75:515–523.
20. Li Y, Lin X. Semiparametric Normal Transformation Models for Spatially Correlated Survival Data. *J. Am. Statist. Assoc.* 2006; 101:591–603.
21. Li Y, Prentice RL, Lin X. Semiparametric Maximum Likelihood Estimation in Normal Transformation Models for Bivariate Survival Data. *Biometrika*. 2008; 95:947–960. [PubMed: 19079778]
22. Martin EC, Betensky RA. Testing Quasi-Independence of Failure and Truncation Times via Conditional Kendall's Tau. *Journal of the American Statistical Association*. 2005; 100:484–492.
23. Nan B, Lin X, Lisabeth LD, Harlow S. Piecewise Constant Cross-Ratio Estimation for Association of Age at a Marker Event and Age at Menopause. *Journal of the American Statistical Association*. 2006; 101:65–77.
24. Oakes D. A Model for Association in Bivariate Survival Data. *Journal of the Royal Statistical Society B*. 1982; 44:414–422.
25. Oakes D. Semiparametric Inference in a Model for Association in Bivariate Survival Data. *Biometrika*. 1986; 73:353–361.
26. Oakes D. Bivariate Survival Models Induced by Frailties. *Journal of the American Statistical Association*. 1989; 84:487–493.
27. Prentice RL, Hsu L. Regression on Hazard Ratios and Cross Ratios in Multivariate Failure Time Analysis. *Biometrika*. 1997; 84:349–363.
28. Shih J, Louis TA. Inference on the Association Parameter in Copula Models for Bivariate Survival Data. *Biometrics*. 1995; 51:1384–1399. [PubMed: 8589230]
29. van der Laan MJ. Nonparametric Estimation of the Bivariate Survival Function with Truncated Data. *Journal of Multivariate Analysis*. 1996; 58:107–131.
30. van der Vaart, A.; Wellner, JA. *Weak Convergence and Empirical Processes*. New York: Springer; 1996.

Fig. 1.

Cross-ratio for the Frank family. In the top panel, the surface in gray is the true cross-ratio and the surface in black is the estimated cross-ratio. The figure on the left is the estimator accounting for truncation. The figure on the right is the naive estimator that ignores left truncation. In the bottom panel, gray lines are the true cross-ratio, the black solid lines are the estimated cross-ratio, the dotted lines are their point-wise 95% confidence bands and the black dashed lines are the naive estimates that ignore left truncation. $T_1=0.25$ to 1.5 with 0.25 increment corresponds to 22%, 39%, 53%, 63%, 71% and 78% percentiles of a unit exponential distribution.

Fig. 2.

The top left figure shows the estimated cross-ratio of monozygotic twins. The black mesh is the result of complete data and the grey mesh is based on 100 truncated data sets, each with 20% truncation rate. Fixing T_1 at 10, 20 and 30, in the three plots on the bottom left, the solid curves indicate the cross-ratio estimated from complete data and grey curves indicate the average of estimated cross-ratio based on 100 truncated data sets. On the right, the plots show the results of dizygotic twins.

Fig. 3.

Estimated cross-ratio for time from diagnosis to initiation of radiation therapy T_1 and time from diagnosis to tumor progression T_2 , both left-truncated by time from diagnosis to chemotherapy and right-censored by the end of the study. The final estimator for the cross-ratio is a function of T_2 only.

Fig. 4.

Illustration of TR AIDS data. T_1 = Age at HIV infection, T_2 = Incubation time, and U_2 = Time from HIV infection to July 1986, $\tau = \max(U_2, U_2^*)$, $U_2^* = \tau - U_2$ and $T_2^* = \tau - T_2$.

Fig. 5.
Reciprocal of estimated cross-ratio for age at transfusion T_1 and reverse transformed incubation time T_2^* .

Table 1

Comparison of empirical variance and model based variance for the Frank family. The points on both margins are the quartiles of the marginal distributions of X_1 and X_2 . The value in front of the parentheses is the average of $\hat{\beta}_n$. The first value in the parentheses is the true β . The second value is the empirical standard error, the third value is the model based standard error estimate, and the fourth value is the coverage rate of the 95% confidence interval.

n=300	X_2			
	X_1	25%	50%	75%
25%		0.99(1.01,0.17,0.17,94%)	0.82(0.82,0.20,0.19,94%)	0.68(0.63,0.28,0.28,95%)
50%		0.82(0.82,0.19,0.19,96%)	0.68(0.70,0.19,0.19,94%)	0.56(0.56,0.23,0.23,94%)
75%		0.68(0.63,0.28,0.28,95%)	0.56(0.56,0.22,0.23,96%)	0.47(0.46,0.28,0.30,97%)

Table 2

The true baseline cross-ratios are $\theta = (.9, 2.0, 4.0, 1.5)$ when t_1 is in intervals $[0, .25)$, $[.25, .5)$, $[.5, .75)$, and above $.75$. The sample sizes are 400 and 800. $\hat{\beta}_n$, point estimate average; E.SE, the empirical standard error; M.SE, the average of the model based standard error estimates; M.CP the 95% coverage probability using M.SE.

θ	n=400						n=800					
	β	$\hat{\beta}_n$	E.SE	M.SE	M.CP	$\hat{\beta}_n$	E.SE	M.SE	M.CP	$\hat{\beta}_n$	E.SE	M.SE
0.90	-0.11	-0.12	0.35	0.35	96%	-0.11	0.23	0.24	96%			
2.00	0.69	0.71	0.24	0.23	94%	0.69	0.16	0.16	96%			
4.00	1.39	1.40	0.25	0.24	95%	1.40	0.17	0.17	95%			
1.50	0.41	0.41	0.15	0.15	96%	0.41	0.11	0.11	96%			