Published in final edited form as:

Lifetime Data Anal. 2014 July ; 20(3): 335–354. doi:10.1007/s10985-013-9287-z.

Computationally simple estimation and improved efficiency for special cases of double truncation

Matthew D. Austin,

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

David K. Simon, and

Department of Neurology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA 02115, USA

Rebecca A. Betensky

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

Matthew D. Austin: austmatt@gmail.com; David K. Simon: dsimon1@bidmc.harvard.edu; Rebecca A. Betensky: betensky@hsph.harvard.edu

Abstract

Doubly truncated survival data arise when event times are observed only if they occur within subject specific intervals of times. Existing iterative estimation procedures for doubly truncated data are computationally intensive (Turnbull 38:290–295, 1976; Efron and Petrosian 94:824–825, 1999; Shen 62:835–853, 2010a). These procedures assume that the event time is independent of the truncation times, in the sample space that conforms to their requisite ordering. This type of independence is referred to as *quasi-independence*. In this paper we identify and consider two special cases of quasi-independence: complete quasi-independence and complete truncation dependence. For the case of complete quasi-independence, we derive the nonparametric maximum likelihood estimator in closed-form. For the case of complete truncation dependence, we derive a closed-form nonparametric estimator that requires some external information, and a semi-parametric maximum likelihood estimator that achieves improved efficiency relative to the standard nonparametric maximum likelihood estimator, in the absence of external information. We demonstrate the consistency and potentially improved efficiency of the estimators in simulation studies, and illustrate their use in application to studies of AIDS incubation and Parkinson's disease age of onset.

Keywords

Piecewise exponential; Quasi-independence; Semi-parametric; Survival analysis

1 Introduction

Doubly truncated survival data arise when event times are observed only if they occur within subject specific intervals of times. The lower and upper limits of these intervals are termed the left and right truncation times, respectively. Existing methods for doubly truncated data assume that the event time is independent of the truncation times in the sample space that conforms to their requisite ordering, i.e., *quasi-independence* (Tsai 1990). In this paper, we address doubly truncated data that satisfy more restrictive independence conditions. An AIDS incubation cohort study from Lagakos et al. (1988) recruited subjects who were diagnosed with AIDS prior to 1986 and who became HIV infected through a blood transfusion. Age of AIDS diagnosis, the event time of interest, is thus doubly truncated by age at transfusion and age at 1986. For the youngest members of this cohort, not only is the event time independent of the truncation times, but the three times are mutually independent in their observable region. At the opposite extreme, a Parkinson's disease (PD) clinical trial recruited subjects who were within 5 years of their diagnosis. Age at Parkinson's diagnosis and its relationship to certain candidate genotypes were of interest. Within this trial, age at diagnosis is doubly truncated by age at study entry minus 5 years on the left, and age at study entry on the right. This exemplifies complete, functional dependence between the truncation times.

For general double truncation, Turnbull (1976) provided a self-consistency algorithm to obtain a consistent estimator for the distribution function of the failure time of interest. Efron and Petrosian (1999) and Shen (2010a) also proposed nonparametric estimators for the failure time distribution in the presence of double truncation. The procedure from Efron and Petrosian (1999) is an iterative procedure that begins with the product-limit estimator of the distribution function of the failure time, based on the failures and the left truncation times only. They found their algorithm to converge faster than that of Turnbull (1976). Like Turnbull's estimator, their estimator suffers from poor performance in the presence of small and minimally overlapping risk sets. Shen (2010a) proved that the conditional NPMLE of Efron and Petrosian (1999) is also the unconditional NPMLE. Shen (2010a) also derived an inverse probability weighted estimator that starts with an initial estimate of the failure time distribution, and then iterates between calculation of the two inverse probability weighted estimators for the truncation and failure time distribution until convergence. This formulation allowed Shen (2010a) to apply general results for consistency and weak convergence provided by van der Laan (1996) to this context of double truncation. Moreira and Uña-Álvarez (2010a) found this estimator to have a similar convergence rate as that of Efron and Petrosian (1999). Unlike the estimators of Turnbull (1976) and Efron and Petrosian (1999), that proposed by Shen (2010a) exhibits adequate performance in the presence of sparse risk sets. None of these three estimators can be expressed in closed form and all are computationally intensive.

All three of these estimators make the assumption of *quasi-independence* (Tsai 1990) of the failure time and the truncation times. Here, we consider two special cases of quasi-independence that are motivated by real examples, such as those described above. Under complete quasi-independence, the two truncation times are additionally independent of each other in the observable region. Under complete truncation dependence, one truncation time

is a deterministic function of the other. This function is known, as it arises from an explicit sampling design.

Complete quasi-independence can occur in studies that have long accrual periods, followed by fixed calendar time sampling of subjects who have experienced onset of disease. If disease risk is analyzed on the age scale, the age of disease onset is left-truncated by age at study entry, and it is right-truncated by age at substudy sampling. Given the long accrual period, it is plausible that the ages at entry and sampling (i.e., left and right truncation times) are independent of each other over the observable region, in addition to being quasi-independent of the age of onset of the disease of interest. The complete truncation dependence case, along with that of general double truncation, was considered in the nonparametric setting by Moreira and Uña-Álvarez (2010a). Their interest in this case arose from a study of survival following childhood cancer. Children from North Portugal diagnosed with cancer between January 1, 1999 and December 31, 2003, were recruited into the study. That is, the right truncation time (i.e., the child's age at the end of recruitment) is exactly 5 years greater than the left truncation time (i.e., the child's age at the beginning of recruitment). These authors established uniform consistency of the inverse probability weighted estimator for the failure time adapted to this setting, which Shen (2010a) did not. They also provided simulation results to suggest weak convergence. Moreira and Uña-Álvarez (2010b) and Shen (2010b) developed general asymptotic results for a semi-parametric estimator for the complete truncation dependence case (in addition to the more general case), including variance estimators.

A test of quasi-independence in the general double truncation setting was proposed by Martin and Betensky (2005). The special cases of quasi-independence must be shown to hold prior to the use of our new, proposed estimators; we derive novel tests for this purpose. For the complete quasi-independence case, our novel contributions are a three degree-of-freedom test for complete quasi-independence, a closed-form nonparametric maximum likelihood estimator for the failure time distribution and a closed-form variance estimator. These offer considerable computational improvement over currently available iterative estimators. We also provide a novel estimator for the truncation probability in the population. For the complete truncation dependence case, we derive a one degree-of-freedom test for quasi-independence in this setting. We derive a closed-form nonparametric estimator for use when an external estimate for the distribution of the right truncation time is available. This is computationally advantageous as compared to iterative algorithms, and also allows for simpler asymptotic analysis. We also derive a flexible semi-parametric estimator for use when an external estimate is not available and for improved efficiency over the Turnbull estimator. Our estimator has increased flexibility and robustness over previously derived estimators, which were shown to be inconsistent if the parametric truncation distribution is misspecified. It benefits from the availability of convergence results and a variance estimator derived by Moreira and Uña-Álvarez (2010b) and Shen (2010b).

2 Tests for special cases of quasi-independence

Let X denote the failure time of interest. Sampling into the study requires that $X \in (L, R)$, where L denotes the left truncation time and R the right truncation time. Quasi-independence in this setting means independence between the failure time and the truncation times in the observable region, denoted as $X \perp\!\!\!\perp (L, R) \mid_Q L < X < R$. Martin and Betensky (2005) proposed a U -statistic to test for quasi-independence. Their two degree-of-freedom test statistic is based on an estimate of a bivariate conditional Kendall's tau,

$$\tau_c = \begin{pmatrix} E[\text{sign}\{(L_1 - L_2)(X_1 - X_2)\} \mid \Omega_{12}] \\ E[\text{sign}\{(R_1 - R_2)(X_1 - X_2)\} \mid \Omega_{12}] \end{pmatrix}.$$

where $\Omega_{12} = I\{\max(L_1, L_2) \leq \min(X_1, X_2), \max(X_1, X_2) \leq \min(R_1, R_2)\}$ indicates the requisite comparability criteria for the two sets of observations.

2.1 Complete quasi-independence

The first special case of quasi-independence imposes independence between L and R , in addition to the standard independence between X and (L, R) , in the observable region. To test this assumption, we extend the test proposed by Martin and Betensky (2005) through use of a trivariate conditional Kendall's tau,

$$\tau_c = (\tau_{LX} \tau_{LR} \tau_{XR})^T = \begin{pmatrix} E[\text{sign}\{(L_1 - L_2)(X_1 - X_2)\} \mid \Omega_{12}] \\ E[\text{sign}\{(L_1 - L_2)(R_1 - R_2)\} \mid \Omega_{12}] \\ E[\text{sign}\{(R_1 - R_2)(X_1 - X_2)\} \mid \Omega_{12}] \end{pmatrix}.$$

A consistent estimator of this parameter is

$$\hat{\tau}_c = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=1}^n \begin{bmatrix} \text{sign}\{(L_i - L_j)(X_i - X_j)\} \\ \text{sign}\{(L_i - L_j)(R_i - R_j)\} \\ \text{sign}\{(R_i - R_j)(X_i - X_j)\} \end{bmatrix} I(\Omega_{ij}) = \mathbf{U}_c \begin{pmatrix} n \\ 2 \end{pmatrix} / M,$$

where $\mathbf{U}_c = (U_{LX} \ U_{LR} \ U_{XR})^T$ is a vector of U -statistics and $M = \sum_{i=1}^{n-1} \sum_{j=i}^n I(\Omega_{ij})$ is the total number of comparable pairs. The expected values of $M / \begin{pmatrix} n \\ 2 \end{pmatrix}$, U_{LX} , U_{LR} , and U_{XR} are $pr(\Omega_{12}) = \mu$, $\tau_{LX}\mu$, $\tau_{LR}\mu$, and $\tau_{XR}\mu$, respectively. Applying standard results for U -statistics (Randles and Wolfe 1979), it follows that under H_0 , $n^{1/2}(\mathbf{U}_c - \tau_c \mu)$ is asymptotically $N(\mathbf{0}, 4\boldsymbol{\Sigma})$, where sample elements of the symmetric matrix $\boldsymbol{\Sigma}$ are of the form

$$\begin{aligned} \Sigma(1, 1) &= E[\text{sign}\{(L_1 - L_2)(X_1 - X_2)\} \times \text{sign}\{(L_1 - L_3)(X_1 - X_3)\} I(\Omega_{12} \cap \Omega_{13})] - (\tau_{LX}\mu)^2 \\ \Sigma(3, 2) &= E[\text{sign}\{(L_1 - L_2)(R_1 - R_2)\} \times \text{sign}\{(X_1 - X_3)(R_1 - R_3)\} I(\Omega_{12} \cap \Omega_{13})] - (\tau_{LR}\tau_{XR}\mu^2). \end{aligned}$$

Similarly, $n^{1/2}(\hat{\tau}_c - \tau_c)$ is asymptotically $N(0, 4\mathbf{\Sigma}\mu^2)$, and

$n\hat{\tau}_c^T \sum_c^{-1} \hat{\tau}_c \hat{\mu}^2 / 4 = n\mathbf{U}_c^T \sum_c^{-1} \mathbf{U}_c / 4$ is asymptotically distributed as χ_3^2 . The matrix, $\mathbf{\Sigma}$, is consistently estimated by averaging its elements over all observed triples.

This test must be conducted and “accepted” prior to applying our new estimator for the distribution of X . If complete quasi-independence is rejected, we would proceed to test general quasi-independence using the test proposed by Martin and Betensky (2005). If that test does not reject, we would then calculate the more computationally burdensome estimator of Turnbull (1976).

2.2 Complete truncation dependence

The case of complete truncation dependence involves functional dependence between L and R , i.e., $L = H(R)$, for some function $H(\cdot)$ over the support of R . A simple, but actual example is $H(R) = R - c$. That is, there may be a fixed interval between the left and right truncation times. To establish that the data satisfy this sampling model, we must “accept” a test of $H_0: X \perp\!\!\!\perp R \mid_Q H(R) < X < R$. The parameter of interest for this purpose is the conditional Kendall’s tau,

$$\tau_c = E \left[\text{sign}\{(X_1 - X_2)(R_1 - R_2)\} \mid \Omega'_{12} \right],$$

where $\Omega'_{12} = I\{\max(H(R_1), H(R_2)) \leq \min(X_1, X_2); \max(X_1, X_2) \leq \min(R_1, R_2)\}$. A consistent estimator of this parameter is

$$\hat{\tau}_c = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i}^n \text{sign}\{(X_i - X_j)(R_i - R_j)\} I(\Omega'_{ij}) = U_c \binom{n}{2} / M,$$

where $M = \sum_{i=1}^{n-1} \sum_{j=i}^n I(\Omega'_{ij})$ is the total number of comparable pairs. The expected values of $M / \binom{n}{2}$ and U_c are $pr(\Omega'_{12}) = \mu$ and $\tau_c \mu$, respectively. Under H_0 , $n^{1/2}(U_c - \tau_c \mu)$ is asymptotically $N(0, 4\sigma^2)$, a one degree-of-freedom test, where

$$\sigma^2 = E \left[\text{sign}\{(X_1 - X_2)(R_1 - R_2) \times (X_1 - X_3)(R_1 - R_3)\} I(\Omega'_{12} \cap \Omega'_{13}) \right] - (\tau_c \mu)^2.$$

We note that the actual transformation, $H(\cdot)$, need not be known for this test. We note also that the test for quasi-independence under general double truncation proposed by Martin and Betensky (2005) reduces to this one degree of freedom test when there is *monotone* functional dependence between L and R , even without knowledge of the relationship.

3 NPMLE for complete quasi-independence

Under the complete quasi-independence sampling model, the likelihood of the data is

$$L = \prod_{i=1}^n \frac{dF_X(x_i) F_L(x_i) S_R(x_i)}{\int F_L(u) S_R(u) dF_X(u)} \frac{dF_L(l_i) dF_R(r_i)}{F_L(x_i) S_R(x_i)},$$

where $S_Z(\cdot)$ denotes the survivorship function for the random variable Z and $F_Z(\cdot) = 1 - S_Z(\cdot)$. We seek to find the MLE for $F_X(\cdot)$. A similar problem for left truncated data was addressed in the nonparametric setting by Turnbull (1976) and Vardi (1985), and in the semi-parametric setting by Wang (1989). In the context of left truncated data, Vardi (1985) showed that the first component of our full likelihood,

$$\prod_{i=1}^n \frac{dF_X(x_i) F_L(x_i) S_R(x_i)}{\int F_L(u) S_R(u) dF_X(u)},$$

can be treated as a multinomial likelihood and that the NPMLE for $F_X(\cdot)$, $\hat{F}_X(\cdot)$, satisfies

$$\frac{d\hat{F}_X(x_i) F_L(x_i) S_R(x_i)}{\int F_L(u) S_R(u) d\hat{F}_X(u)} = \frac{1}{n} \sum_{j=1}^n I(x_j = x_i) \text{ for all } x_i.$$

Thus, analogous to the solution provided by Vardi (1985), the NPMLE for F_X is

$$\hat{F}_X(u) = \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{\hat{S}(x_i)} \right\}^{-1} \frac{1}{n} \sum_{i=1}^n \frac{I(X_i \leq u)}{\hat{S}(x_i)},$$

where $\hat{S}(u) = \hat{S}_R(u) \hat{S}_L^*(u^*)$, $F_L(u) = \hat{S}_L^*(u^*)$, $u^* = \tau - u$, and $L^* = \tau - L$, where τ is the upper limit of the support of X . The leftmost term in the expression of $\hat{F}_X(u)$ is an estimator of the truncation probability, $\text{pr}(L < X < R)$. Thus, operationally, we first calculate the usual Lynden-Bell nonparametric maximum likelihood estimates for F_R and F_L from the conditional likelihood components $\prod_{i=1}^n dF_L(l_i)/F_L(x_i)$ and $\prod_{i=1}^n dF_R(r_i)/S_R(x_i)$, respectively, and then form products of $\hat{S}_R$ and $\hat{F}_L$ to use as weights for the inverse probability weighted estimator of $F_X(\cdot)$. Analogous to all estimators in the presence of truncation, our estimator is limited to the conditional distribution of X in its observed interval, which extends from the minimum of the support of L to the maximum of the support of R .

In the supplementary material, we establish the weak convergence of $n^{1/2}\{\hat{F}_X(u; \hat{S}) - F_X(u)\}$ to a zero mean Gaussian process, with covariance function

$$\sum(s, t) = \int \int \frac{\Lambda(u, v)}{S(u)S(v)} \{F_X(s) - I(u \leq s)\} \{F_X(t) - I(v \leq t)\} dF_X(u) dF_X(v) + \Lambda^*(s, t),$$

where $s = t, v = u, \beta = \int dF_X(y)S(y)^{-2}, \gamma = \text{pr}(L < X < R)$,

$$\begin{aligned} \Lambda(u, v) &= S(u)S(v) \left\{ \int_0^{\min(u, v)} \frac{dK_R(y)}{Q_R^2(y)} + \int_0^{\min(u^*, v^*)} \frac{dK_{L^*}(y^*)}{Q_{L^*}^2(y^*)} \right\}, \\ \Lambda^*(u, v) &= \beta \gamma^2 \left\{ \frac{1 - F_X(u)}{\beta} \int_0^v \frac{dF_X(y)}{S(y)^2} + F_X(u) F_X(v) - \frac{F_X(v)}{\beta} \int_0^u \frac{dF_X(y)}{S(y)^2} \right\}, \\ K_Y(u) &= \text{pr}(Y \leq u | Y > X) \text{ and } Q_Y(u) = \text{pr}(X \leq u \leq Y | Y > X). \end{aligned}$$

4 Estimators for complete truncation dependence

Under the complete truncation dependence sampling model, the likelihood of the data is

$$L = \prod_{i=1}^n \frac{dF_X(x_i) [F_R\{H^{-1}(x_i)\} - F_R(x_i)]}{\int [F_R\{H^{-1}(u)\} - F_R(u)] dF_X(u)} \frac{dF_R(r_i)}{[F_R\{H^{-1}(x_i)\} - F_R(x_i)]}.$$

As in Vardi (1985) and Wang (1989), the NPMLE of $F_X(\cdot), F_X(\hat{\cdot})$, satisfies

$$\frac{d\hat{F}_X(x_i) [F_R\{H^{-1}(x_i)\} - F_R(x_i)]}{\int [F_R\{H^{-1}(u)\} - F_R(u)] d\hat{F}_X(u)} = \frac{1}{n} \sum_{j=1}^n I(x_j = x_i),$$

for all x and is given by

$$\hat{F}_X(u) = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{[F_R\{H^{-1}(x_i)\} - F_R(x_i)]} \right)^{-1} \frac{1}{n} \sum_{i=1}^n \frac{I(X_i \leq u)}{[F_R\{H^{-1}(x_i)\} - F_R(x_i)]}. \quad (1)$$

If F_R is available from external sources, then plugging it in to (1) yields a closed form estimator for F_X , as in the complete quasi-independence case. If an external estimate is not available, it must be estimated from the data. Unfortunately, inspection of the likelihood reveals that the conditional likelihood component of the full likelihood that is available for estimation of F_R is not of the simple form of that for a singly truncated random variable, as it was in the complete quasi-independence case, and thus a simple, closed form estimator for F_R is not available to plug in for calculation of F_X . Instead, we can similarly express $F_R(\hat{\cdot})$ as

$$\hat{F}_R(u) = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{[F_X(r_i) - F_X\{H(r_i)\}] } \right)^{-1} \frac{1}{n} \sum_{i=1}^n \frac{I(X_i \leq u)}{[F_X(r_i) - F_X\{H(r_i)\}]}, \quad (2)$$

and iteratively update our estimates of $F_X(\cdot)$ and $F_R(\cdot)$ through alternating between (1) and (2). A similar iterative estimation procedure was proposed by Shen (2010a).

Given that nonparametric estimation is computationally intensive if an estimate of F_R is not available, consideration of a semi-parametric estimator for potential gains in efficiency does not materially increase the necessary computational burden. Thus, we propose a semi-parametric estimator, extending that proposed by Wang (1989) for left truncated data. We assume a finite piecewise exponential distribution for R , with hazard function defined as $h_R(t; \theta) = \theta_k$ for $v_{k-1} < t < v_k$, for $k = 1, \dots, K$. For K finite, on the order of 10 or less, this semi-parametric model enjoys the flexibility of a nonparametric approach, with the gained efficiency of a parametric modeling approach (Lindsey and Ryan 1998).

The likelihood is now parameterized by θ as

$$L = \prod_{i=1}^n \frac{dF_X(x_i) [F_R\{H^{-1}(x_i); \theta\} - F_R(x_i; \theta)]}{\int [F_R\{H^{-1}(u); \theta\} - F_R(u; \theta)] dF_X(u)} \frac{dF_R(r_i; \theta)}{[F_R\{H^{-1}(x_i); \theta\} - F_R(x_i; \theta)]},$$

and the semi-parametric maximum likelihood estimator for $F_X(\cdot)$ is given by

$$\hat{F}_X(u; \hat{\theta}_c) = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{[F_R\{H^{-1}(x_i); \hat{\theta}_c\} - F_R(x_i; \hat{\theta}_c)]} \right)^{-1} \times \frac{1}{n} \sum_{i=1}^n \frac{I(X_i \leq u)}{[F_R\{H^{-1}(x_i); \hat{\theta}_c\} - F_R(x_i; \hat{\theta}_c)]},$$

where $\hat{\theta}_c$ is the vector of hazards that maximizes the conditional likelihood,

$$\prod_{i=1}^n dF_R(r_i; \theta) / [F_R\{H^{-1}(x_i); \theta\} - F_R(x_i; \theta)]. \quad (3)$$

Note that (3) is the likelihood for a doubly truncated random variable, which R is as it is bounded below by X and above by $H^{-1}(X)$. Weak convergence of the process $n^{1/2}\{F_X(\cdot; \hat{\theta}_c) - F_X(\cdot)\}$ was established by Moreira and Uña-Álvarez (2010b) and Shen (2010b) for finite dimensional parametric models, and is applicable to our estimator with finite K . These authors also derived an analytical variance estimator.

There are several choices to make when fitting a piecewise exponential distribution: the number of pieces, the locations of the changepoints in the hazard function, and the numerical maximization algorithm. To address these issues we propose iterative implementation of an informal Kolmogorov–Smirnov (KS) test for goodness-of-fit for the estimated piecewise exponential distribution. While the nominal null distribution of the KS test does not hold exactly for truncated data (Koziol and Byar 1975) and while the target piecewise exponential is not fully specified (i.e., its parameters are estimated), we nonetheless use the nominal p values as measures of goodness-of-fit within each iteration of our algorithm (and acknowledge that they cannot be interpreted as actual p values). Moreira et al. (2012) proposed a bootstrap based goodness-of-fit test for a semi-parametric estimator

of a doubly truncated random variable, which could be used for more accurate p value estimation within our algorithm. However, this test substantially increases the computational burden of our estimator as it requires estimation of the general NPMLE and as a bootstrap test would be required within each iteration of our algorithm. Our simulations indicate that our p values are sufficiently good measures of goodness-of-fit. Our algorithm entails pre-selection of α , a p value threshold for goodness-of-fit. This will typically be some value larger than 0.05, possibly in the range 0.10–0.30. There is a trade-off involved in the choice of α ; larger values will lead to lower bias and higher variance than smaller values. Our algorithm consists of the following steps:

- Step 1: Initialize $K = 1$ (K is the number of pieces in the piecewise exponential distribution for R).
- Step 2: Estimate $\hat{\theta}_c$ using (3).
- Step 3: Calculate p , the approximate KS p value for the test of equality of the nonparametric Turnbull estimator of R , treated as doubly truncated (by X and $X + c$, as in (3)), to a piecewise exponential distribution with the estimated hazard from Step 2.
- Step 4: If $p < \alpha$, increment K by 1 and return to Step 2. If $p \geq \alpha$, stop.

Regarding the locations of the pieces, a simple approach is to use percentiles of the crude empirical distribution of R . For example, for $K = 2$ we fit one hazard from the minimum through the median of R , and one from the median to the maximum. An alternative approach is to properly estimate the distribution of R given that it is left-truncated by X , using the Turnbull estimator, and then to pre-select pieces visually to correspond to bends in the estimated cumulative hazard function. The order of entry of these pieces to be used by our proposed KS algorithm would also be pre-specified to increasingly capture the important features of the distribution of R . We took the first approach in our simulation studies as it was simple to automate. We took the second approach in our analyses of the AIDS and Parkinson's studies.

5 Simulation studies

5.1 Complete quasi-independence

We examined the consistency of the NPMLE derived in Sect. 3 for the case of complete quasi-independence for overall probabilities of observation of $\text{pr}(L < X < R)$ ranging from 0.02 to 0.64. We obtained these probabilities by varying the degree of the single-sided truncation via the probabilities $\text{pr}(L < X)$ and $\text{pr}(X < R)$; we did this in order to learn about the impact of an imbalance in truncation relative to X , as well as that of the overall magnitude of truncation.

We generated the data as follows: $L \sim N(\mu_L, \sigma_L^2)$, $X \sim \text{Gamma}(25, .5)$, and $R \sim N(\mu_R, \sigma_R^2)$. We varied the values of μ_L , σ_L , μ_R , and σ_R to achieve the desired truncation probabilities. We generated the components of an observation, (l, x, r) , independently from their respective distributions, and retained the observation only if it satisfied $l < x < r$. We repeated this process until we obtained 200 or 400 observations of (l, x, r) . This data

generation process satisfies all of the conditions of complete quasi-independence. We computed the NPMLE for X derived in Sect. 3 and repeated this process 500 times (5,000 times for the variance estimators).

Table 1 displays the true (conditional) values of F_X for a pre-specified set of quantiles of X , along with the average bias of our new NPMLE and that of the Turnbull estimator, the average analytical variance (see supplementary material for the derivation) of our NPMLE, its empirical variance for comparison, and the ratio of the Turnbull mean squared error to that of our NPMLE. We used the self-consistency algorithm originally proposed by Turnbull (1976) for calculation of the Turnbull estimator. The top set of scenarios in the table are for heavy truncation, as defined by $\{\text{pr}(X < R), \text{pr}(L < X), \text{pr}(L < X < R)\} = (0.33, 0.33, 0.02), (0.33, 0.66, 0.10), (0.66, 0.33, 0.10), (0.66, 0.66, 0.32)$ and the bottom set of four scenarios are for more moderate truncation, as defined by $\{\text{pr}(X < R), \text{pr}(L < X), \text{pr}(L < X < R)\} = (0.50, 0.50, 0.10), (0.50, 0.80, 0.32), (0.80, 0.50, 0.32), (0.80, 0.80, 0.64)$. The sample size is 200 in each scenario.

The new and Turnbull NPMLE's are similar in bias. This is not surprising, given the results of Mandel (2007), who showed that the analogues of these estimators (i.e. an inverse probability weighted estimator and a self-consistency estimator) in the case of one-sided truncation without censoring are asymptotically equivalent. The Turnbull NPMLE appears to have slightly lower MSE than the new NPMLE in several, though not all, of the scenarios considered. While we might expect a decrease in variance in the proposed NPMLE relative to the Turnbull due to its use of additional model information, this is not the case due to the need to estimate F_L and S_R , separately, for estimation of F_X , and thus effectively twice the number of parameters as the Turnbull estimator. The advantage of our new NPMLE is its ease of practical implementation and its facilitation of asymptotic analysis, due to its closed form expression. We also note that our new analytical variance formula provides estimates that are quite close to the simulation based variances.

As seen in the top half of Table 1, the NPMLE mostly has small bias even with the heavy overall probabilities of observation of 0.02–0.32. The concentration of the bias varies with the degree of left versus right sided truncation; with heavier left truncation (i.e., smaller $\text{pr}(L < X)$) there is greater bias for small $F(X)$ and with heavier right truncation (i.e., smaller $\text{pr}(X < R)$) there is greater bias for large $F(X)$. Symmetric left and right truncation appears to diminish the bias at the endpoints. The bottom half of Table 1 displays similar, though less severe, patterns of behavior due to the more moderate levels of truncation imposed in those simulations.

Table 2 lists the simulation results for a sample size of 400, for the heaviest truncation case of Table 1. As expected, the increase in sample size leads to improved bias and variance.

5.2 Complete truncation dependence

We examined the consistency of the semi-parametric maximum likelihood estimator derived in Sect. 4 for the case of complete truncation dependence for overall truncation probabilities of $\text{pr}(L < X < R) = 0.25$ and 0.50. We obtained these probabilities by varying the degree of

the single-sided truncation via the probabilities $\text{pr}(L < X)$ and $\text{pr}(X < R)$; we did this in order to learn about the impact of an imbalance in truncation, as well as the degree of truncation.

We generated the data as follows: $R \sim N(\mu_R, \sigma_R^2)$, $X \sim \text{Gamma}(25, .5)$, and $L = R - c$. We varied the parameters, μ_R and σ_R , in order to achieve the desired truncation probabilities. We generated the components of an observation, (x, r) , from their respective distributions and retained the observation only if it satisfied $r - c < x < r$. We repeated this process until we obtained 200 or 400 observations of $(r - c, x, r)$. This data generation process satisfies all of the conditions of complete truncation dependence. We calculated the SPMLE for X derived in Sect. 4, using p value thresholds of $\alpha = 0.10$ and 0.25 , and repeated this process 500 times.

Table 3 displays our results for truncation scenarios defined by $\{\text{pr}(R - c < X), \text{pr}(X < R), \text{pr}(R - c < X < R)\} = \{(0.50, 0.50, 0.25), (0.50, 0.75, 0.25), (0.75, 0.50, 0.25), (0.75, 0.75, 0.50)\}$ for $c = 15$, $N = 200$, $\alpha = 0.10$. Table 4 examines the effect of increased sample size ($N = 400$) and p value threshold ($\alpha = 0.05, 0.1$ and 0.25). Both tables list the true (conditional) F_X for a pre-specified set of quantiles of X , along with the average bias, average analytical variance, and the mean squared error ratio based on the analytical variances of the SPMLE and NPMLE. Table 3 additionally lists the bias and empirical variance of the Turnbull estimator.

In all but one scenario of Tables 3 and 4, the Turnbull MSE is 1–5 times larger than that of the SPMLE for $F(X) = 0.33$. The lower MSE of the Turnbull NPMLE for smaller values of $F(X)$ is due to the higher variability of the SPMLE at the left endpoint of the support of X . This demonstrates the increased efficiency of the SPMLE. It is seen in Table 3 that the analytical variances of the SPMLE coincide reasonably well with the empirical percentile estimates, indicating the quality of the analytic approximations of Moreira and Uña-Álvarez (2010b) and Shen (2010b) when applied to the piecewise exponential truncation distribution. We also see that the semi-parametric maximum likelihood estimator has small bias, even with the heavy truncation represented in this table. As expected, there is slightly higher variance associated with the higher p value thresholds (Table 4), due to their correspondence with more hazard function pieces in the SPMLE. We also examined K , the number of pieces for the piecewise exponential estimate for R associated with the three p value thresholds used in Table 4. As expected, the average K is increasing with the threshold; it was 2.8 for the threshold of 0.05, 3.0 for the threshold of 0.1, and 3.6 for the threshold of 0.25. We note that the analytical variances listed in Table 3 do not account for estimation of K , and thus their close approximation of the empirical variances is promising.

5.3 Kendall's tau simulations

In Sect. 2 we proposed two Kendall's tau tests for each of the special cases of double truncation. The tests differ in what they assume and what they test regarding quasi-independence. The first test is for complete quasi-independence. In our simulations, we assessed the power of this test to detect dependence of (L, R) , in the setting of quasi-independence, i.e., $X \perp\!\!\!\perp (L, R) \mid_Q L < X < R$. We chose to isolate this particular component of potential quasi-dependence as this is the novel aspect of our tau test. In practice, one might

first test for complete quasi-independence using our proposed test; if it does not reject, then use of our proposed estimator is appropriate. If it does reject, then one would proceed to the test of Martin and Betensky (2005) for focused quasi-independence with respect to X . If that test does not reject, then the Turnbull estimator may be used. The second tau test that we proposed is for quasi-independence of X relative to (L, R) under an assumed functional relationship between L and R . This relationship does not need to be known for the execution of this tau test. We expect that in the case of complete truncation dependence, the dependence model is known to the investigator, as it arises from structural features of the sampling. Thus, to assess this test, we simulated quasi-dependence of (X, L) , with specified complete truncation dependence ($R = L + c$).

To assess our test for complete quasi-independence case, we generated observations as $X \sim \text{Gamma}(25, .5)$ and $(L, R) \sim$ bivariate normal with mean vector $\mu = (\mu_L, \mu_R) = (40, 60)$, variances $\sigma_L^2 = \sigma_R^2 = 3$, and a range of correlations from 0.05 to 0.75. Note that quasi-independence with respect to X is maintained in this setting. To assess our test for complete truncation dependence, we generated $(X, R) \sim$ bivariate normal with mean vector $\mu = (\mu_X, \mu_R) = (50, 49)$, variances $\sigma_X^2 = \sigma_R^2 = 3$, and a range of correlations from 0.05 to 0.75. We set $L = R - 9$. In this scenario, quasi-independence is violated, while the complete truncation dependence model is upheld; this is realistic as we assume it to be known. We assess the power of our Kendall's tau test to identify the violation of quasi-independence in this setting.

Table 5 lists the results of these simulations. In all cases, we generated 500 data sets of $N = 200$ observations such that $L < X < R$. Using 0.05-level significance tests, we have high power (89 %) to detect violations of complete truncation independence due to moderately low correlation (0.25) of L and R . Under complete truncation dependence, we have about 56 % power to detect a violation of quasi-independence using our proposed tau test when the correlation between L and X is 0.75. This power increases substantially with increased sample size.

6 Applications

6.1 AIDS incubation study

A registry established by the US Centers for Disease Control and Prevention (CDC) includes subjects who were infected with HIV through a blood transfusion, and who were diagnosed with AIDS prior to July 1, 1986. Thus, the age of AIDS onset is doubly truncated by age at contaminated blood transfusion on the left and age at July 1, 1986 on the right. Several datasets from this study have been used by various authors; we use the dataset given by Lagakos et al. (1988), with truncation events as defined above. We also note that these data have typically been analyzed as right truncated, with time of infection as the time origin. Bilker and Wang (1996) also treated the data as double truncated, though using a specific calendar date as the time origin; this creates complete truncation dependence. Here, we analyze the data using birth as the time origin, which creates double truncation that potentially satisfies complete quasi-independence. Lagakos et al. (1988) and Wang (1989) analyzed these data separately for three strata of age in years at HIV infection: (0–4], [5,59],

and [60,85]. This was motivated by the strong association between immunocompetence and age. We begin by testing for complete quasi-independence using the test from Sect. 2.1; the p values for the three age strata are 0.28, 0.002, and 0.0005. Therefore, we are justified in using our NPMLE for complete quasi-independence for the youngest age group of children, but not for the older age strata. This finding might be due to more intensive medical monitoring for the youngest children. It also might be due to the considerably smaller age range for the youngest age group than for the others, in which quasi-independence seems more plausible. The rejection of quasi-independence in the older age groups is consistent with observations of correlation between ages at HIV infection and incubation time (Rosenberg et al. 1994).

Figure 1 displays the NPMLE for the age of AIDS diagnosis for the (0–4] stratum, along with the estimate from Turnbull (1976) and the empirical cumulative distribution function. We include the latter because it is commonly used in this setting when truncation is not recognized. As expected, the NPMLE and the Turnbull are quite different from the empirical cumulative distribution function, and their sparseness illustrates the limited information in this subgroup. The estimated probability of observation, $\hat{\text{pr}}(L < X < R)$, (derived in the supplementary material) is 0.42, confirming that there is substantial truncation in this age group. The use of the age scale in this example makes it somewhat artificial, as age at AIDS might not be of scientific interest. Nonetheless, this illustrates the use of the complete quasi-independence test and a substantial discrepancy between estimates that do and do not account for truncation.

6.2 Parkinson's disease study

Parkinson's disease (PD) is a degenerative disorder of the central nervous system. Symptoms include tremor, rigidity or stiffness of the limbs and trunk, slowness of movement, and impaired balance and coordination. A multi-center phase 3 study (NCT00449865) sponsored by the U.S. National Institute of Neurological Disorders and Stroke (NINDS) was undertaken to determine if the nutritional supplement, creatine, slows the progression of PD over time (The NINDS NET-PD Investigators 2006). The study closed enrollment with a total of 1741 subjects who were randomized to receive either creatine or placebo. Subjects were eligible for the study only if they were diagnosed with PD within 5 years prior to their potential study entry. Genotyping of a few candidate genes was done on a subset of subjects for whom DNA samples were available. One ancillary question was whether a particular genotype is associated with age of onset of PD. This question could be addressed by analyzing only baseline data, and thus did not require waiting until the study had been completed. In fact, the data set that we obtained did not contain the genetic information, and it included only the 397 subjects who had accrued to the study by the end of 2007. Thus, we focus on how to estimate age of onset given the truncated sampling.

Letting X denote age at PD diagnosis and R denote age at study entry, the sampling for this study requires that $R - 5 < X < R$. Note that in the general framework that we presented in Sect. 2.2, this amounts to $H(R) = R - 5$. If we had genotype information available, we would then seek to estimate the distributions of X separately for each genotype, and ultimately to compare the distributions. Since we do not yet have the genotype information, we

demonstrate our methods on the study population as a whole. Given the sampling based on a 5 year window, the complete truncation dependence model is most appropriate, and we apply our test from Sect. 2.2 to establish its validity. We obtain a p value of 0.186, and so it is plausible that $X \perp\!\!\!\perp Q(R - 5, R) \mid R - 5 < X < R$. Figure 2 displays two semi-parametric maximum likelihood estimates, the Turnbull estimate, and the empirical cumulative distribution function. For illustration, we display the semi-parametric estimate based on five hazard pieces (with associated KS p value of 0.06) and that based on eight hazard pieces (with KS p value of 0.38). As expected, more pieces are associated with better goodness-of-fit (less bias), and thus larger KS p value. Also as expected, the empirical cumulative distribution function is heavily biased; the estimated median age of onset based on the empirical cumulative distribution function is 60.6, versus 64.9 for the semi-parametric estimate with 8 partitions and 64 for the estimate from Turnbull (1976). We note that the empirical CDF that ignores the truncation lies above the estimator that accounts for the double truncation. This suggests that in some overall sense, smaller ages at PD diagnosis are oversampled in this population, relative to larger ages at PD diagnosis. In the case of double truncation, the nature of the bias of an unadjusted estimator, such as the empirical CDF, is difficult to predict, as it is driven by the relative overlap of the sampling intervals, as well as the placements of the variable of interest within those intervals. The estimated probability of observation is 0.44, which confirms that there was substantial truncation. This indicates the importance of accounting for sampling, even in the context of a randomized clinical trial. The only scenario in which the sampling can be ignored for this sampling model is when R is uniformly distributed (Moreira and Uña-Álvarez 2010b).

Parenthetically, an alternative estimation strategy is to separately estimate the distribution of age at study entry from a PD clinical trial that did not have any requirement on age of PD onset relative to age at study entry and then to plug this in to the inverse probability weighted equation for F_X in Sect. 4 to obtain a closed-form nonparametric estimator. This would obviate the need for a semi-parametric estimator, and it would greatly expedite computation for the real question of interest as soon as the valuable genetic information is available: the association between genotype and age at PD onset.

7 Discussion

Doubly truncated data arise due to various retrospective sampling designs. We have derived estimators for the distribution of the variable of interest in the presence of two special cases of double truncation. In future work, we will further study these sampling models and develop methods for multi-group comparisons, such as age of PD onset within genotypes, and for regression analysis that are applicable when they hold. Hopefully, we will also discover computational efficiency for these analyses in the case of complete quasi-independence. Ideally, standard software will be adaptable to this setting. This may be true, as well, in the case of complete truncation dependence if an external estimate is available for F_R . We will extend the results of Mandel and Betensky (2007) to test whether R follows a uniform distribution in the complete truncation dependence case. If it does, then the sampling can be ignored and standard methods can be applied.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Bilker WB, Wang MC. A semiparametric extension of the Mann-Whitney test for randomly truncated data. *Biometrics*. 1996; 50:10–20. [PubMed: 8934583]
- Efron B, Petrosian V. Nonparametric methods for doubly truncated data. *J Am Stat Assoc*. 1999; 94(447):824–825.
- Koziol J, Byar D. Percentage points of the asymptotic distributions of one and two sample K-S statistics for truncated or censored data. *Technometrics*. 1975; 17(4):507–510.
- Lagakos SW, Barraj LM, DeGruttola V, Hu FB. Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika*. 1988; 75(3):515.
- Lindsey J, Ryan L. Methods for interval-censored data. *Stat Med*. 1998; 17(2):219–238. [PubMed: 9483730]
- Mandel M. Censoring and truncation-highlighting the differences. *Am Stat*. 2007; 61(4):321–324.
- Mandel M, Betensky RA. Simultaneous confidence intervals based on the percentile bootstrap approach. *Comput Stat Data Anal*. 2007; 52(4):2158–2165. [PubMed: 19137059]
- Martin EC, Betensky RA. Testing quasi-independence of failure and truncation times via conditional Kendall's tau. *J Am Stat Assoc*. 2005; 100(470):484–492.
- Moreira C, de Uña-Álvarez J. Bootstrapping the NPMLE for doubly truncated data. *J Nonparametr Stat*. 2010a; 22(5):567–583.
- Moreira C, de Uña-Álvarez J. A semiparametric estimator of survival for doubly truncated data. *Stat Med*. 2010b; 29:3147–3159. [PubMed: 21170909]
- Moreira, C.; de Uña-Álvarez, J.; Van Keilegom, I. Goodness-of-fit tests for a semiparametric model under random double truncation. 2012. Unpublishedmanuscript
- Randles, R.; Wolfe, D. Introduction to the theory of nonparametric statistics. Wiley; New York: 1979.
- Rosenberg P, JIG, RJB. Effect of age at seroconversion on the natural AIDS incubation distribution. Multicenter Hemophilia Cohort Study and the International Registry of Seroconverters. *AIDS*. 1994; 8(6):803–810. [PubMed: 8086140]
- Shen P. Nonparametric analysis of doubly truncated data. *Ann Inst Stat Math*. 2010a; 62:835–853.
- Shen P. Semiparametric analysis of doubly truncated data. *Commun Stat*. 2010b; 39(17):3178–3190.
- The NINDS NET-PD Investigators. A randomized, double-blind, futility clinical trial of creatine and minocycline in early Parkinson disease. *Neurology*. 2006; 66(5):664–671. [PubMed: 16481597]
- Tsai WY. Testing the assumption of independence of truncation time and failure time. *Biometrika*. 1990; 77(1):169.
- Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J R Stat Soc Ser B (Methodol)*. 1976; 38(3):290–295.
- van der Laan MJ. Nonparametric estimation of the bivariate survival function with truncated data. *J Multivar Anal*. 1996; 58:107–131.
- Vardi Y. Empirical distributions in selection bias models. *Ann Stat*. 1985; 13(1):178–203.
- Wang MC. A semiparametric model for randomly truncated data. *J Am Stat Assoc*. 1989; 84(407):742–748.

Fig. 1.
Age at AIDS diagnosis: age < 4

Fig. 2.
Age at PD diagnosis

Table 1
Simulation results (500 repetitions) for the complete quasi-independence NPMLE, in comparison to the Turnbull NPMLE

$N = 200$	$F(X)$	$\text{pr}(X < R) = 0.33$				$\text{pr}(X < R) = 0.66$			
		Bias		Variance ^a		MSE ratio ^b		Bias	
		New	Turnbull	Analytical	Sim	Turnbull	Sim	New	Turnbull
0.33	$\text{pr}(L < X)$	0.04	-0.05	-0.04	0.000	0.50	0.000	-0.04	-0.02
		0.15	-0.12	-0.11	0.008	0.94	0.009	-0.11	-0.09
		0.33	-0.11	-0.11	0.016	1.09	0.015	-0.09	-0.06
		0.54	-0.08	-0.07	0.012	1.09	0.013	-0.06	-0.03
		0.73	-0.04	-0.04	0.008	0.80	0.010	-0.01	-0.02
0.66		0.87	-0.02	-0.02	0.007	0.88	0.007	0.01	0.01
		0.04	-0.01	-0.01	0.003	1.24	0.003	-0.02	-0.01
		0.15	0.01	0.01	0.005	0.91	0.005	-0.01	0.01
		0.33	0.04	0.04	0.008	0.82	0.009	0.00	0.05
		0.54	0.07	0.08	0.014	0.80	0.015	0.02	0.02
0.80		0.73	0.10	0.09	0.020	0.92	0.020	0.03	0.03
		0.87	0.10	0.08	0.010	0.76	0.011	0.04	0.02
		$\text{pr}(X < R) = 0.50$							
		0.04	-0.04	-0.02	0.004	0.82	0.005	-0.04	-0.03
		0.15	-0.03	-0.01	0.010	1.15	0.009	-0.06	-0.05
		0.33	-0.01	0.01	0.011	0.58	0.012	-0.04	-0.05
		0.54	0.02	0.02	0.011	0.69	0.010	-0.02	-0.02
		0.73	0.04	0.04	0.011	0.66	0.013	-0.01	-0.01
		0.87	0.05	0.05	0.011	0.82	0.015	0.00	0.00
		0.04	0.00	-0.01	0.002	1.91	0.001	0.00	0.00
		0.15	0.01	0.01	0.003	1.10	0.004	0.00	0.00
		0.33	0.03	0.04	0.006	0.79	0.006	0.01	0.01
		0.54	0.05	0.08	0.009	1.23	0.010	0.01	0.01
		0.73	0.07	0.09	0.014	1.12	0.013	0.01	0.01
		0.87	0.09	0.08	0.013	0.89	0.013	0.02	0.00

^aVariances are based on 5000 repetitions: "analytical" is novel (see supplementary material); "sim" is empirical

Simulation results (500 repetitions) for the complete quasi-independence NPMLE, in comparison to the Turnbull NPMLE: $N = 400$ and heavy truncation

Table 2

$N = 400$ $\text{pr}(L < X) = \text{pr}(X < R) = 0.33$					
F(X)	Bias		Variance ^a		MSE ratio ^b
	New	Turnbull	Analytical	Simulation	
0.04	-0.04	-0.03	0.000	0.000	0.61
0.15	-0.10	-0.10	0.008	0.007	1.06
0.33	-0.10	-0.11	0.015	0.013	1.15
0.54	-0.07	-0.06	0.010	0.010	0.94
0.73	-0.04	-0.04	0.008	0.009	0.81
0.87	-0.02	-0.02	0.005	0.006	0.92

^aVariances are based on 5,000 repetitions: “analytical” is novel (see supplementary materials); “simulation” is empirical

^bRatio of Turnbull MSE divided by new NPMLE MSE

Table 3

Simulation results (500 repetitions) for the complete truncation dependence SPMLE, in comparison to the Turnbull NPMLE: p value threshold=0.1

N = 200	F(X)	pr(X < R) = 0.75					pr(X < R) = 0.50					
		Bias		Variance		MSE ratio ^b	Bias		Variance		MSE ratio ^b	
		New	Turnbull	Analytical ^a	Sim		New	Turnbull	Analytical ^a	Sim		
pr(L < X)	0.75	0.04	-0.01	-0.01	0.000	0.000	1.34	0.01	0.02	0.000	0.000	0.67
		0.15	-0.01	0.01	0.000	0.000	1.94	0.02	0.03	0.002	0.003	0.45
		0.33	0.00	0.03	0.002	0.002	2.97	0.02	0.03	0.007	0.007	2.33
		0.54	0.03	0.02	0.003	0.004	3.13	0.01	0.09	0.014	0.013	3.23
		0.73	0.04	0.03	0.005	0.005	4.31	-0.09	-0.08	0.039	0.028	1.61
0.5		0.87	0.03	0.02	0.006	0.006	4.89	-0.01	-0.01	0.038	0.036	1.57
		0.04	-0.04	0.02	0.000	0.000	0.36	0.03	0.02	0.000	0.000	0.71
		0.15	-0.08	0.03	0.001	0.002	0.43	0.02	0.03	0.002	0.002	0.54
		0.33	-0.07	0.03	0.005	0.005	1.24	-0.08	0.03	0.007	0.007	1.88
		0.54	-0.07	0.07	0.015	0.015	1.85	-0.11	-0.09	0.015	0.014	2.21
		0.73	-0.03	-0.04	0.027	0.028	1.75	-0.08	-0.07	0.027	0.026	2.34
		0.87	0.03	-0.01	0.037	0.044	1.07	-0.04	-0.02	0.032	0.033	1.45

^a Analytical variances are equivalent to those derived by Moreira and Uña-Álvarez (2010b) and Shen (2010b)

^b Ratio of Turnbull MSE divided by SPMLE MSE

Simulation results (500 repetitions) for the complete truncation dependence SPMLE, in comparison to the Turnbull NPMLE: $N = 400$, p value thresholds = 0.05, 0.1 and 0.25, moderate truncation

F(X)	$N = 400$ $\text{pr}(L < X) = 0.75, \text{pr}(X < R) = 0.50$									
	p-value threshold = 0.05			p-value threshold = 0.10			p-value threshold = 0.25			
	Bias	Var ^a	MSE ratio ^b	Bias	Var ^a	MSE ratio ^b	Bias	Var ^a	MSE ratio ^b	
0.04	0.014	0.000	0.59	0.010	0.000	0.90	0.009	0.000	0.19	
0.15	0.014	0.002	0.42	0.012	0.002	0.35	0.016	0.002	0.39	
0.33	0.018	0.006	2.63	0.013	0.006	3.13	0.014	0.008	2.16	
0.54	0.011	0.010	3.43	0.011	0.010	2.96	0.011	0.012	2.73	
0.73	-0.075	0.032	1.38	-0.063	0.031	1.72	-0.069	0.050	0.97	
0.87	-0.017	0.025	1.66	-0.009	0.026	1.62	-0.010	0.042	1.05	

^a Analytical variances are equivalent to those derived by Moreira and Uña-Álvarez (2010b) and Shen (2010b)

^b Ratio of Turnbull MSE divided by SPMLE MSE

Table 5

Power of 0.05-level Kendall's tau tests

Correlation ^a	Complete quasi- independence	Complete truncation dependence
0.05	0.06	0.06
0.10	0.21	0.07
0.25	0.89	0.21
0.50	1.00	0.43
0.75	1.00	0.56

^a $Corr(L, R)$, for test for complete quasi-independence; $Corr(X, R)$, for test for quasi-independence under complete truncation dependence