I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

2012/24

GOODNESS-OF-FIT TESTS FOR A SEMIPARAMETRIC MODEL UNDER RANDOM DOUBLE TRUNCATION

MOREIRA, C., DE UNA-ALVAREZ, J. and I. VAN KEILEGOM

Goodness-of-fit tests for a semiparametric model under random double truncation

Carla Moreira [*] Jacobo de Uña-Álvarez [§]

Ingrid Van Keilegom [††]

July 31, 2012

Abstract

Doubly truncated data are commonly encountered in areas like medicine, astronomy, economy, among others. A semiparametric estimator of a doubly truncated random variable has been proposed by Moreira and de Uña-Álvarez (2010b). Their estimator is based on a parametric specification of the distribution function of the truncation times. This semiparametric estimator outperforms the nonparametric maximum likelihood estimator when the parametric information is correct, but might behave badly when the assumed parametric model is far off. In this paper we introduce several goodness-of-fit tests for the parametric model. The proposed tests are investigated through simulations. For illustration purposes, the tests are also applied to data on the induction time to AIDS for blood transfusion patients.

[*]Department of Statistics and OR, University of Vigo; and Centro de Matemática, University of Minho. E-mail address: carlamgmm@gmail.com. Research supported by research grants MTM2008-03129 and MTM2011-23204 (FEDER support included) of the Spanish Ministerio de Ciencia e Innovación and SFRH/BPD/68328/2010 grant of Portuguese Fundação Ciência e Tecnologia.

[§]Department of Statistics and OR, University of Vigo. E-mail address: jacobo@uvigo.es. Research supported by research grants MTM2008-03129 and MTM2011-23204 (FEDER support included) of the Spanish Ministerio de Ciencia e Innovación and by grant PGIDIT07PXIB300191PR of the Xunta de Galicia.

[††]Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain. E-mail address: ingrid.vankeilegom@uclouvain.be. Research supported by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy), by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, and by the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16-039 of the "Communauté française de Belgique" (granted by the "Académie universitaire Louvain").

1 Introduction

Truncated data play an important role in the statistical analysis of survival times as well as in other fields like astronomy or economy. Generally speaking, random truncation occurs when one is only able to observe the variable of interest when its value falls within a certain (random) set. This happens for example when analyzing time from HIV infection to AIDS, when the recruited times correspond to individuals who developed AIDS prior to some specific date (right truncation). Another typical truncation setup can be found in medicine, where the survival time of prevalent cases must be larger than the time from diagnosis to the cross-section date to enter the sample, resulting in random left truncation. In all these situations, estimation is based on sampling information coming from a conditional distribution, which is different from the distribution of the population of interest, and suitable corrections of ordinary estimators are needed. This problem goes back to Turnbull (1976).

Among the various existing problems with random truncation, the literature has mainly focused on the left truncation model or, more generally, on one-sided truncation setups. See for example Woodroofe (1985) and Stute (1993). Left truncation and right censoring was considered by many authors, including Tsai et al. (1987), Wang (1991) and Zhou and Yip (1999), among many others. However, in some scenarios two-sided (rather than one-sided) truncation occurs. As an example, consider a situation in which one only observes the survival times of individuals with terminating event falling between two specific dates, t_0 and t_1. In this case, the observed survival times X are those satisfying $U \leq X \leq V$, where U (resp. V) is the time elapsed between the initial event and t_0 (resp. t_1). This double truncation phenomenon was recognized in applications with AIDS data (Bilker and Wang, 1996), astronomic data (Efron and Petrosian, 1999), and cancer data (Moreira and de Uña-Álvarez, 2010a).

The literature on two-sided or double random truncation is much more scarce, probably due to the technical complications in the computation of estimators and in the derivation of statistical properties. Efron and Petrosian (1999) introduced the nonparametric maximum likelihood estimator (NPMLE) of the distribution function (df) under double truncation, while Shen (2010a) formally established the uniform strong consistency and the weak convergence of the NPMLE. Bootstrap methods to approximate the finite sample distribution of the NPMLE with doubly truncated data were explored in Moreira and de Uña-Álvarez (2010a). An R package to compute the NPMLE of a doubly truncated df was presented in Moreira et al. (2010).

In some scenarios, it may be convenient to introduce some parametric information regarding the truncation variables. For example, in epidemiological applications, the distribution of

the truncation times is related to the so-called incidence rate of a disease, so information on the behavior of this rate (such as stationarity) may lead to specific models for the truncation times. See Wang (1989) and Asgharian et al. (2002) for further discussions. This semiparametric approach, in which the distribution of the truncation times is assumed to belong to a given parametric family, was investigated in Moreira and de Uña-Álvarez (2010b), see also Shen (2010b). Interestingly, these authors showed that the semiparametric estimator may outperform the NPMLE in the sense of the mean squared error (MSE). However, it was also pointed out that missspecification of the parametric family may introduce a systematic estimation bias, and hence goodness-of-fit methods for the parametric model are needed in practice. This is the problem we address in this paper.

The rest of the paper is organized as follows. In Section 2 we introduce the NPMLE, the semiparametric estimator, and six different test statistics for the goodness-of-fit problem considered above. Also, a bootstrap algorithm is proposed to approximate the null distribution of the tests in practice. In Section 3 the finite sample performance of the proposed tests is investigated through simulations. A real data illustration is given in Section 4, while Section 5 reports the main conclusions of our investigation.

2 The test statistics

Let X^* be the random variable of ultimate interest, with df F. We assume that X^* is subject to double truncation by the random pair (U^*, V^*) with joint df K, where U^* and V^* ($U^* \leq V^*$) are the left and right truncation variables respectively. This means that the triplet (U^*, X^*, V^*) is observed if and only if $U^* \leq X^* \leq V^*$, while no information is available when $X^* < U^*$ or $X^* > V^*$. Assume that X^* is independent of (U^*, V^*). Let (U_i, X_i, V_i), $i = 1, \ldots, n$, denote the available data, drawn independently from the conditional distribution of (U^*, X^*, V^*) given $U^* \leq X^* \leq V^*$. Let $\alpha = P(U^* \leq X^* \leq V^*)$ be the probability of no-truncation. For any df W denote the left and right endpoints of its support by $a_W = \inf\{t : W(t) > 0\}$ and $b_W = \inf\{t : W(t) = 1\}$, respectively. Let $K_1(u) = K(u, \infty)$ and $K_2(v) = K(\infty, v)$ be the marginal df's of U^* and V^*, respectively. (Woodroofe, 1985) showed that F and K are both identifiable when $a_{K_1} \leq a_F \leq a_{K_2}$ and $b_{K_1} \leq b_F \leq b_{K_2}$, which we will assume from now on.

We know that the NPMLE of (F, K) is a discrete distribution supported by the set of observed data (Turnbull, 1976). Let $\varphi = (\varphi_1, \ldots, \varphi_n)$ be the vector of mass probabilities corresponding to the weighted empirical df $\sum_{i=1}^n \varphi_i I_{[X_i \leq x]}$. Similarly, let $\psi = (\psi_1, \ldots, \psi_n)$ be the vector of weights corresponding to the estimator $\sum_{i=1}^n \psi_i I_{[U_i \leq u, V_i \leq v]}$ of the joint distribution of (U, V). Under the assumption of independence between X^* and (U^*, V^*), the

full likelihood, $\mathcal{L}(\varphi, \psi)$, can be decomposed into a product of the conditional likelihood of the X_i's given the (U_i, V_i)'s, say $\mathcal{L}_1(\varphi)$, and the marginal likelihood of the (U_i, V_i)'s, say $\mathcal{L}_2(\varphi, \psi)$:

$$\mathcal{L}(\varphi, \psi) = \prod_{j=1}^{n} \frac{\varphi_j}{\Phi_j} \times \prod_{j=1}^{n} \frac{\Phi_j \psi_j}{\sum_{i=1}^{n} \Phi_i \psi_i} = \mathcal{L}_1(\varphi) \times \mathcal{L}_2(\varphi, \psi), \tag{2.1}$$

where $\Phi_i = \sum_{m=1}^{n} \varphi_m J_{im}$, and where $J_{im} = I_{[U_i \le X_m \le V_i]}$. The conditional NPMLE of F (Efron and Petrosian, 1999) is defined as

$$F_n(x) = \sum_{i=1}^{n} \hat{\varphi}_i I_{[X_i \le x]},$$

where $\hat{\varphi} = (\hat{\varphi}_1, \ldots, \hat{\varphi}_n)$ is the maximizer of $\mathcal{L}_1(\varphi)$ in equation (2.1).

Shen (2010a) proved that the conditional NPMLE F_n maximizes indeed the full likelihood, which can also be written as the product

$$\mathcal{L}(\varphi, \psi) = \prod_{j=1}^{n} \frac{\psi_j}{\Psi_j} \times \prod_{j=1}^{n} \frac{\Psi_j \varphi_j}{\sum_{i=1}^{n} \Psi_i \varphi_i} = \mathcal{L}_1(\psi) \times \mathcal{L}_2(\psi, \varphi),$$

where $\Psi_i = \sum_{m=1}^{n} \psi_m I_{[U_m \le X_i \le V_m]} = \sum_{m=1}^{n} \psi_m J_{mi}$, for $i = 1 \ldots, n$. Here, $\mathcal{L}_1(\psi)$ denotes the conditional likelihood of the (U_i, V_i)'s given the X_i's, and $\mathcal{L}_2(\psi, \varphi)$ refers to the marginal likelihood of the X_i's. Let $\hat{\psi} = (\hat{\psi}_1, ..., \hat{\psi}_n)$ be the maximizer of $\mathcal{L}_1(\psi)$. Then, $K_n(u, v) = \sum_{i=1}^{n} \hat{\psi}_i I_{[U_i \le u, V_i \le v]}$ is the NPMLE of K (Shen, 2010a).

The NPMLE of F also admits the representation

$$F_n(x) = \alpha_n \int_{-\infty}^{x} \frac{dF_n^*(t)}{G_n(t)} \equiv \int_{-\infty}^{x} \frac{dF_n^*(t)}{G_n(t)} \Big/ \int_{-\infty}^{\infty} \frac{dF_n^*(t)}{G_n(t)},$$

where F_n^* is the ordinary empirical df of the X_i's, i.e. $F_n^*(t) = n^{-1} \sum_{i=1}^{n} I_{[X_i \le t]}$,

$$G_n(t) = \int_{\{u \le t \le v\}} K_n(du, dv)$$

is a nonparametric estimator of the conditional probability of sampling a lifetime $X^* = t$, i.e. $G(t) = P(U^* \le t \le V^*)$, and $\alpha_n = [\int_{-\infty}^{\infty} G_n(t) dF_n^*(t)]^{-1}$ is an estimator of α. Shen (2010a) established the uniform strong consistency and the weak convergence of F_n.

Instead of estimating K nonparametrically, we can also assume that K belongs to a parametric family of df's $\{K_\theta\}_{\theta \in \Theta}$, where θ is a vector of parameters and Θ is a compact, finite dimensional parameter space. In that case, $G(t)$ is parametrized as

$$G_\theta(t) = \int_{\{u \le t \le v\}} K_\theta(du, dv).$$

The parameter θ can be estimated by the maximizer $\hat{\theta}$ of the conditional likelihood of the (U_i, V_i)'s given the X_i's, i.e. $\hat{\theta} = \text{argmax}_{\theta\in\Theta}\mathcal{L}_1^*(\theta)$, where

$$\mathcal{L}_1^*(\theta) \equiv \mathcal{L}_1^*(\psi) = \prod_{i=1}^{n} \frac{k_\theta(U_i, V_i)}{G_\theta(X_i)},$$

and where $k_\theta(u, v) = \frac{\partial^2}{\partial u \partial v} P(U^* \leq u, V^* \leq v) = K_\theta(du, dv)$ is the joint density of (U^*, V^*) (assumed to exist).

Once θ is estimated, a semiparametric estimator of F is obtained:

$$F_{\widehat{\theta}}(x) = \alpha_{\widehat{\theta}} \int_{-\infty}^{x} \frac{dF_n^*(t)}{G_{\widehat{\theta}}(t)},$$

where $\alpha_{\widehat{\theta}} = [\int_{-\infty}^{\infty} G_{\widehat{\theta}}(t) dF_n^*(t)]^{-1}$. Moreira and de Uña-Álvarez (2010b) established the asymptotic normality of both $\hat{\theta}$ and $F_{\widehat{\theta}}$. They also showed via simulations that $F_{\widehat{\theta}}$ may perform much more efficiently than the NPMLE. As a drawback, the semiparametric estimator requires preliminary specification of a parametric family, which may eventually introduce a bias component when it is far away from reality (Moreira and de Uña-Álvarez, 2010b). It is therefore necessary to have a procedure at hand that allows to test the appropriateness of the parametric family, i.e. we need to develop a goodness-of-fit test for the null hypothesis $H_0 : K = K_{\theta_0}$ for some $\theta_0 \in \Theta$.

To measure the distance between the null hypothesis and the data, several approaches are possible. Firstly, since the null refers to the joint distribution of (U^*, V^*), an initial approach is to introduce a test statistic which measures the distance between K_n and $K_{\widehat{\theta}}$. However, as long as the focus is the estimation of F, one may argue that, rather than the distribution K_θ, the relevant issue is the closeness of $F_{\hat{\theta}}$ to F_n. This motivates the use of testing methods which evaluate the difference $F_n(x) - F_{\widehat{\theta}}(x)$ along all possible x-values. Thirdly, it should be noted that F_n (respectively $F_{\widehat{\theta}}$) is a weighted average of the indicators $I_{[X_i \leq x]}$, where the weights are proportional to $G_n(X_i)^{-1}$ (respectively $G_{\widehat{\theta}}(X_i)^{-1}$). Therefore, it also makes sense to look at the difference between G_n and $G_{\widehat{\theta}}$. According to this, the tests proposed below are based on the differences $G_n(x) - G_{\widehat{\theta}}(x)$, $F_n(x) - F_{\widehat{\theta}}(x)$ and $K_n(u, v) - K_{\widehat{\theta}}(u, v)$. Both Kolmogorov-Smirnov type and Cramér-von Mises type test statistics are considered. More precisely, define:

$$
\begin{aligned}
D_n(G) &= \sup_{x\in\mathbb{R}} |G_n(x) - G_{\widehat{\theta}}(x)| && (2.2)\\
D_n(F) &= \sup_{x\in\mathbb{R}} |F_n(x) - F_{\widehat{\theta}}(x)| && (2.3)\\
D_n(K) &= \sup_{u,v\in\mathbb{R}} |K_n(u,v) - K_{\widehat{\theta}}(u,v)| && (2.4)\\
W_n^2(G) &= \int (G_n(x) - G_{\widehat{\theta}}(x))^2 F_n(dx) && (2.5)\\
W_n^2(F) &= \int (F_n(x) - F_{\widehat{\theta}}(x))^2 F_n(dx) && (2.6)\\
W_n^2(K) &= \int\int (K_n(u,v) - K_{\widehat{\theta}}(u,v))^2 dK_n(u,v). && (2.7)
\end{aligned}
$$

These test statistics may be grouped into three different classes: G-based tests (equations (2.2) and (2.5)), F-based tests (equations (2.3) and (2.6)), and K-based tests (equations (2.4) and (2.7)). Note that the squared differences in (2.5) and (2.6) are weighted by the NPMLE of Efron and Petrosian (1999), which means that more weight is given to X_i's that have a small probability of being observed.

In practice, the null distribution of the six proposed test statistics must be estimated. For this we introduce a semiparametric bootstrap resampling plan, which makes use of the information contained under the parametric null model. Fix B and let T be any of the six test statistics considered above for $H_0 : K \in \{K_\theta\}_{\theta\in\Theta}$. The bootstrap resampling plan is as follows. All data are generated independently of each other (between and within resamples):

1. For $b = 1, \ldots, B$:

   For $i = 1, \ldots, n$:

   Step 1 Draw (U_i^b, V_i^b) from $K_{\hat{\theta}}$.

   Step 2 Draw independently X_i^b from F_n.

   Step 3 If $U_i^b \le X_i^b \le V_i^b$ is violated, go back to Step 1. Otherwise, keep (U_i^b, X_i^b, V_i^b).

   Let T^{*b} be the test statistic obtained from the bootstrap resample (U_i^b, X_i^b, V_i^b), $i = 1, \ldots, n$.

2. Compute the critical value C_α, which is the $100(1-\alpha)\%$ percentile of $T^{*1}, \ldots, T^{*B}$.

3. If the realization of the test statistic is greater than or equal to C_α, then reject H_0.

The simulation results in the following section are based on this resampling plan.

3 Simulation study

In this section we study the finite sample behavior of the test statistics defined in (2.2)–(2.7) through simulations. We consider two different situations of double truncation, Case 1 and Case 2. In Case 1, U^*, V^* and X^* are mutually independent. In Case 2, we simulate U^* independently of X^* and then we define $V^* = U^* + \tau$ for some fixed constant $\tau > 0$. Case 2 follows the spirit of the AIDS Blood Transfusion Data, which we will consider in Section 4. The recruited patients in that data set are those with terminating events falling between two specific dates. Under the null hypothesis the data are generated as follows. For Case 1 we draw X^* from a uniform $U(0,1)$, while U^* and V^* are drawn independently from uniform distributions with respective supports (0, c) and (d, 1), where c and d are chosen as follows: $c = d = 0.5$ (Model 1.1), and $c = 0.25, d = 0.75$ (Model 1.2). Truncation occurs whenever the condition $U^* \leq X^* \leq V^*$ is violated. For Case 2, we take $\tau = 0.25$ and $U^* \sim U(0, 0.75)$, $X^* \sim U(0,1)$ (Model 2.1) and $U^* \sim U(0,1)$, $X^* \sim 0.75 Beta(3/4, 1) + 0.25$ (Model 2.2). The function G of each simulated model is shown in Figure 1. The depicted functions indicate that small and large values of the variable of interest (X^*) are observed with a relatively small probability in the first three models, while there is no observational bias in Model 2.2 (because G remains constant on the support of X^*). In these simulated models, the percentage of truncation is 50% (Model 1.1), 25% (Model 1.2) and 75% (Model 2.1 and Model 2.2).

The parametric family of distributions $\{K_\theta\}_{\theta\in\Theta}$ is defined as follows. We always consider a $Beta(\theta_1, 1)$ distribution for U^* in Case 2, and the independent product of this distribution and a $Beta(1, \theta_2)$ for V^* in Case 1. Since under H_0 the support of the data (generated as indicated above) may be different from the interval $[0, 1]$, we adapted the support of these $Beta$ parametric models to the corresponding supporting intervals. In this way, Models 1.1, 1.2, 2.1, and 2.2 belong to the null hypothesis.

Under the alternative hypothesis, we generate U^* from a $Beta(1, a)$ with $a \neq 1$ (note that $a = 1$ corresponds to the null hypothesis). The values $a = 1/10, 1/5, 1/2, 2$ and 4 are considered. The alternative shapes of the G function are shown in Figure 1. From this Figure we see that, for Case 1, the deviation from the null is only important when x is small, while G is changed all along its support for Case 2.

The simulations are carried out for samples of size $n = 50$, 100 and 250. The results are based on 1000 Monte Carlo trials and for each of them 300 bootstrap replications are taken. In Tables 1 to 4 (corresponding respectively to Models 1.1, 1.2, 2.1 and 2.2) the rejection percentages of the six test statistics introduced in Section 2 are given for $\alpha = 0.05$. As expected, the power increases with the sample size and with the distance between the null

and the alternative (given by $|a-1|$).

When investigating the rejection proportions under the null ($a = 1$), we see that, generally speaking, all the test statistics respect the nominal level well. As an exception, we note that the statistics based on the comparison between the two estimators of F are a bit anti-conservative, particularly for small sample sizes. For example, for Models 1.1 and 2.1, the level in the simulations is always above 10% in the case $n = 50$. We also investigate the rejection proportions under the different considered alternatives. For this we consider separately Case 1 and Case 2, which show somewhat different results.

		KS (D_n)			CM (W_n^2)		
a	n	G	F	K	G	F	K
	50	0.983	0.791	0.982	0.728	0.765	0.977
1/10	100	0.995	0.914	0.994	0.931	0.906	0.993
	250	0.997	0.988	0.996	0.987	0.987	0.996
	50	0.977	0.642	0.965	0.754	0.604	0.905
1/5	100	0.992	0.844	0.992	0.966	0.828	0.992
	250	0.997	0.959	0.997	0.993	0.961	0.997
	50	0.367	0.278	0.301	0.329	0.251	0.240
1/2	100	0.731	0.378	0.668	0.649	0.333	0.591
	250	0.979	0.681	0.973	0.968	0.634	0.964
	50	0.047	0.108	0.046	0.056	0.102	0.037
1	100	0.056	0.096	0.055	0.066	0.089	0.047
	250	0.052	0.062	0.056	0.057	0.058	0.046
	50	0.336	0.143	0.318	0.230	0.150	0.195
2	100	0.700	0.226	0.659	0.672	0.226	0.546
	250	0.983	0.596	0.978	0.993	0.604	0.966
	50	0.922	0.272	0.897	0.837	0.311	0.637
4	100	0.988	0.594	0.988	0.992	0.607	0.979
	250	0.994	0.829	0.994	0.997	0.852	0.994

Table 1: Power under $H_1 : K \notin \{K_\theta\}_{\theta\in\Theta}$ along 1000 trials for Model 1.1 with $\alpha = 0.05$.

For Models 1.1 and 1.2 (Case 1, Tables 1 and 2), the test statistics based on the comparison between the nonparametric and the parametric estimator of G are the most powerful. Indeed, $D_n(G)$ gives the largest rejection proportions in almost all cases (and, when it is not the largest, it is improved by $W_n^2(G)$ by only a small amount). On the contrary, the statistics measuring the departure between the nonparametric and the semiparametric estimator of F

		KS (D_n)			CM (W_n^2)		
a	n	G	F	K	G	F	K
	50	0.994	0.558	0.994	0.521	0.557	0.994
1/10	100	0.997	0.755	0.997	0.787	0.738	0.996
	250	0.998	0.954	0.998	0.980	0.944	0.998
	50	0.991	0.521	0.991	0.699	0.502	0.987
1/5	100	0.997	0.697	0.997	0.941	0.652	0.997
	250	0.999	0.921	0.999	0.993	0.896	0.999
	50	0.635	0.253	0.534	0.445	0.221	0.416
1/2	100	0.951	0.366	0.904	0.827	0.318	0.847
	250	1.000	0.642	1.000	0.999	0.567	1.000
	50	0.047	0.010	0.054	0.045	0.106	0.047
1	100	0.046	0.081	0.044	0.048	0.067	0.043
	250	0.056	0.073	0.060	0.062	0.063	0.049
	50	0.658	0.134	0.631	0.483	0.127	0.515
2	100	0.944	0.252	0.929	0.917	0.235	0.847
	250	0.998	0.569	0.998	0.997	0.558	0.997
	50	0.988	0.310	0.986	0.954	0.309	0.938
4	100	0.995	0.566	0.994	0.995	0.570	0.994
	250	0.999	0.802	0.999	1.000	0.816	0.999

Table 2: Power under $H_1 : K \notin \{K_\theta\}_{\theta\in\Theta}$ along 1000 trials for Model 1.2 with $\alpha = 0.05$.

perform poorly, giving the smallest powers. This indicates that, in Case 1, the alternatives are more easily detected when looking at the function G. The test statistics based on K ($D_n(K)$ and $W_n^2(K)$) are competitive for the largest considered sample size ($n = 250$), but for $n = 50$ and $n = 100$ they are often less able to reject the null than the G-based tests. This is interesting, since the G-based tests are only looking at a portion of the joint distribution of (U^*, V^*).

Case 2 (Tables 3 and 4) is more difficult to summarize. The more visible change with respect to Models 1.1 and 1.2 is that the F-based test statistics are the more powerful ones in special situations. This is true, for example, for Model 2.1 and the closest alternatives $a = 1/2$ and $a = 2$ (at least with moderate sample sizes), and for Model 2.2 for $a = 2$ and $a = 4$. In these situations, the F-based tests may have about five times the power of the G-based or the K-based tests. On the contrary, for Model 2.1, the test statistic leading to the largest power when considering more distant alternatives is often $W_n^2(K)$, while for

		KS (D_n)			CM (W_n^2)		
a	n	G	F	K	G	F	K
	50	0.477	0.393	0.427	0.497	0.581	0.525
1/10	100	0.906	0.703	0.888	0.692	0.800	0.905
	250	0.979	0.910	0.977	0.818	0.926	0.978
	50	0.399	0.254	0.320	0.309	0.406	0.416
1/5	100	0.843	0.574	0.775	0.674	0.698	0.774
	250	0.980	0.883	0.979	0.922	0.928	0.979
	50	0.075	0.133	0.054	0.052	0.222	0.078
1/2	100	0.318	0.226	0.202	0.249	0.271	0.238
	250	0.794	0.664	0.575	0.679	0.682	0.589
	50	0.013	0.114	0.013	0.018	0.119	0.016
1	100	0.019	0.068	0.027	0.045	0.072	0.028
	250	0.039	0.074	0.054	0.057	0.066	0.057
	50	0.036	0.245	0.041	0.064	0.117	0.031
2	100	0.147	0.291	0.197	0.263	0.230	0.223
	250	0.566	0.592	0.683	0.666	0.547	0.723
	50	0.112	0.287	0.092	0.246	0.121	0.139
4	100	0.773	0.424	0.744	0.831	0.271	0.777
	250	0.973	0.827	0.970	0.965	0.707	0.968

Table 3: Power under $H_1 : K \notin \{K_\theta\}_{\theta\in\Theta}$ along 1000 trials for Model 2.1 with $\alpha = 0.05$.

Model 2.2 the best tests for $a < 1$ are $D_n(G)$ and the K-based statistics. The fact that the F-based tests lead to the largest power in some cases of Model 2.1 could be due to their anti-conservatism. However, F-based tests preserve the level well in Model 2.2. In this latter case, differences between the nonparametric and the semiparametric estimators of F lead more frequently to reject the null model than when considering departures for G or K.

When comparing Kolmogorov-Smirnov to Cramér-von Mises type statistics, it is found that the former reject the null more frequently and, therefore, they are generally preferable. However, Case 2 models report different results, particularly when focusing on F and K-based tests. Consequently, it is worthwhile to keep both measures when testing for the parametric model.

		KS (D_n)			CM (W_n^2)		
a	n	G	F	K	G	F	K
	50	0.220	0.078	0.230	0.067	0.116	0.125
1/10	100	0.350	0.134	0.345	0.222	0.132	0.413
	250	0.870	0.224	0.853	0.751	0.206	0.830
	50	0.175	0.054	0.179	0.059	0.083	0.099
1/5	100	0.267	0.099	0.252	0.170	0.108	0.300
	250	0.778	0.192	0.751	0.617	0.189	0.770
	50	0.109	0.033	0.106	0.034	0.049	0.050
1/2	100	0.116	0.079	0.126	0.091	0.085	0.125
	250	0.299	0.119	0.303	0.217	0.111	0.347
	50	0.083	0.037	0.089	0.023	0.049	0.042
1	100	0.046	0.055	0.047	0.045	0.055	0.034
	250	0.041	0.066	0.058	0.062	0.068	0.061
	50	0.050	0.139	0.048	0.061	0.124	0.018
2	100	0.052	0.251	0.056	0.096	0.249	0.056
	250	0.072	0.397	0.180	0.303	0.406	0.264
	50	0.040	0.285	0.025	0.064	0.125	0.017
4	100	0.125	0.535	0.070	0.212	0.388	0.115
	250	0.727	0.862	0.576	0.834	0.771	0.667

Table 4: Power under $H_1 : K \notin \{K_\theta\}_{\theta\in\Theta}$ along 1000 trials for Model 2.2 with $\alpha = 0.05$.

4 Real data application

For illustration purposes, in this section we consider epidemiological data on transfusion-related Acquired Immune Deficiency Syndrome (AIDS). The AIDS Blood Transfusion Data are collected by the Centers for Disease Control (CDC), which is from a registry data base, a common source of medical data (see Kalbfleisch and Lawless, 1989; Bilker and Wang, 1996). The variable of interest (X^*) is the induction or incubation time, which is defined as the time elapsed from Human Immunodeficiency Virus (HIV) infection to the clinical manifestation of full-blown AIDS. The CDC AIDS Blood Transfusion Data are subject to double truncation. Indeed, the data were retrospectively ascertained for all transfusion-associated AIDS cases in which the diagnosis of AIDS occurred prior to the end of the study, thus leading to right truncation. Moreover, since HIV was unknown prior to 1982, any cases of transfusion-related AIDS before this time would not have been properly classified and thus would have been

missed. Thus, in addition to right truncation, the data are also truncated from the left. See Bilker and Wang (1996), Section 5.2, for further discussion.

The data include 494 cases reported to the CDC prior to January 1, 1987, and diagnosed prior to July 1, 1986. Of the 494 cases, 295 had consistent data, and the infection could be attributed to a single transfusion or short series of transfusions. Our analyses are restricted to this subset, which is entirely reported in Kalbfleisch and Lawless (1989), Table 1. The variable U^* is defined as the length of time between January 1, 1982, and the moment of HIV infection, while V^* is the time from HIV infection to the end of study (July 1, 1986). Note that the difference between V^* and its respective U^* is always 4.5 years.

We choose to work with a $Beta(\theta_1, \theta_2)$ model for U^* under the null hypothesis. The parameters θ_1 and θ_2 are estimated by maximizing the conditional likelihood of the truncation times (see Section 2 for more details). Note that in this case the pair (U^*, V^*) does not have a density, and the likelihood $\mathcal{L}_1^*(\theta)$ must be properly re-defined by replacing the density k_θ by the density of U^* in that expression, see Remark 2.1 in Moreira and de Uña-Álvarez (2010b) for further details. The estimated values of the parameters are $\widehat{\theta}_1 = 1.289$ and $\widehat{\theta}_2 = 3.119$.

Informal testing of the null hypothesis can be done by plotting the fitted parametric model versus the NPMLE of the biasing function G. Both curves are shown in Figure 2, right. The label for the horizontal axis refers to the values of U_i and V_i in the sample, which are the jump points of G_n. The semiparametric and the nonparametric estimators of the cumulative distribution function of X^* are also given in Figure 2, left. The transformation $t \to (t+4.5)/8.5$ has been applied to the time axis in both figures, so that X^* belongs to the interval $[0.5, 1.5]$, while U^* is supported on $[0, 1]$. Note that, for the test statistic $W_n^2(G)$, the differences between the curves in Figure 2, right, are only relevant on the interval $[0.5, 1.5]$, while the full support $[0, 1.5]$ must be considered for the computation of $D_n(G)$.

We used the tests proposed in Section 2 to verify whether the parametric model assumed for the truncation variables is appropriate. The p-values of the tests (based on $B = 1000$ bootstrap resamples) are presented in Table 5. The table shows that the $Beta$ model is suitable for U^* or that, at least, there is no statistical evidence to reject it. In this case, the test statistic $W_n^2(G)$ is based on the distance between G_n and $G_{\hat{\theta}}$ on the interval $[0.5, 1.5]$ (see the curves displayed in Figure 2, right panel), while the test statistic $D_n(G)$ looks at the maximal separation along the interval $[0, 1.5]$. None of the methods rejected the $Beta$ model for U^*.

Test	KS (D_n)			CM (W_n^2)		
	G	F	K	G	F	K
p-value	0.844	0.214	0.438	0.335	0.179	0.466

Table 5: P-values for the AIDS Blood Transfusion Data computed from $B = 1000$ bootstrap resamples.

5 Conclusions

When analyzing doubly truncated data, more efficient estimation may be obtained under a semiparametric truncation model. However, in order to avoid systematic estimation bias, the semiparametric model must be tested. In this paper we proposed several Kolmogorov-Smirnov and Cramér-von Mises type test statistics for this problem. The proposed test statistics measure the distance between the nonparametric and the semiparametric estimator of several important curves, namely the cumulative df of the variable of interest, the joint df of the pair of truncation times, and the bias function G.

The more natural test is the one reporting the distance between K_n and $K_{\widehat{\theta}}$. When the parametric null model is false, this test should be able to reject H_0 when the sample size increases. However, our simulations showed that more power can be obtained in practice when focusing on special portions of the function K. More explicitly, the distance between the nonparametric and the parametric bias function (G_n and $G_{\widehat{\theta}}$) will often lead to a more optimal test. It was also demonstrated that a comparison between the nonparametric and the semiparametric estimator of the cumulative df of the lifetime of interest is quite powerful in certain scenarios. It should be noted, however, that F-based tests are sometimes anti-conservative and hence, in general, the application of G or K-based tests is recommended. A real data illustration has been provided.

References

Asgharian, M., C. M'Lan, and D. Wolfson (2002). Length-biased sampling with right-censoring: an unconditional approach. *Journal of the American Statistical Association 97*, 201–209.

Bilker, W. B. and M.-C. Wang (1996). A semiparametric extension of the Mann-Whitney test for randomly truncated data. *Biometrics 52*, 10–20.

Efron, B. and V. Petrosian (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association 94*, 824–834.

Kalbfleisch, J. D. and J. F. Lawless (1989). Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association 84*, 360–372.

Moreira, C. and J. de Uña-Álvarez (2010a). Bootstrappping the NPMLE for doubly truncated data. *Journal of Nonparametric Statistics 22*, 567–583.

Moreira, C. and J. de Uña-Álvarez (2010b). A semiparametric estimator of survival for doubly truncated data. *Statistics in Medicine 29*, 3147–3159.

Moreira, C., J. de Uña-Álvarez, and R. Crujeiras (2010). DTDA: an R package to analyze randomly truncated data. *Journal of Statistical Software 37*, 1–20.

Shen, P. (2010a). Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics 62*, 835–853.

Shen, P. (2010b). Semiparametric analysis of doubly truncated data. *Communications in Statistics – Theory and Methods 39*, 3178–3190.

Stute, W. (1993). Almost sure representations of the product-limit estimator for truncated data. *The Annals of Statistics 21*, 146–156.

Tsai, W., N. Jewell, and M. Wang (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika 74*, 883–886.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society – Series B 38*, 290–295.

Wang, M.-C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association 84*, 742–748.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association 86*, 130–143.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *The Annals of Statistics 13*, 163–177.

Zhou, Y. and P. S. F. Yip (1999). A strong representation of the product-limit estimator for left truncated and right censored data. *Journal of Multivariate Analysis 69*, 261–280.

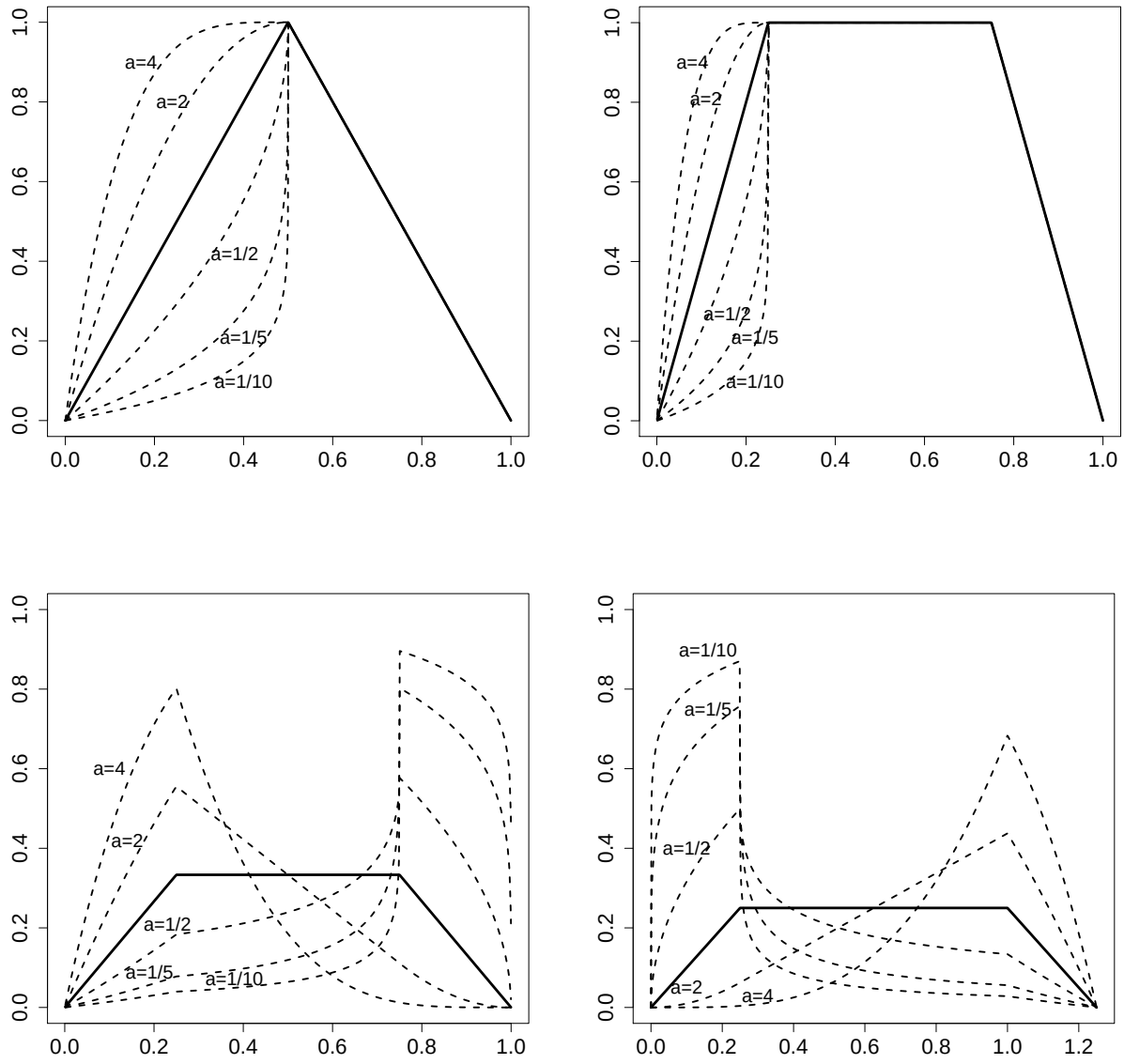

Figure 1: Shapes of the G function for the simulated models under the null hypothesis (solid line, $a = 1$) and under the considered alternative hypotheses (dashed lines, $a = 1/10, 1/5, 1/2, 2$ and 4): Model 1.1 (top-left), Model 1.2 (top-right), Model 2.1 (bottom-left), and Model 2.2 (bottom-right).

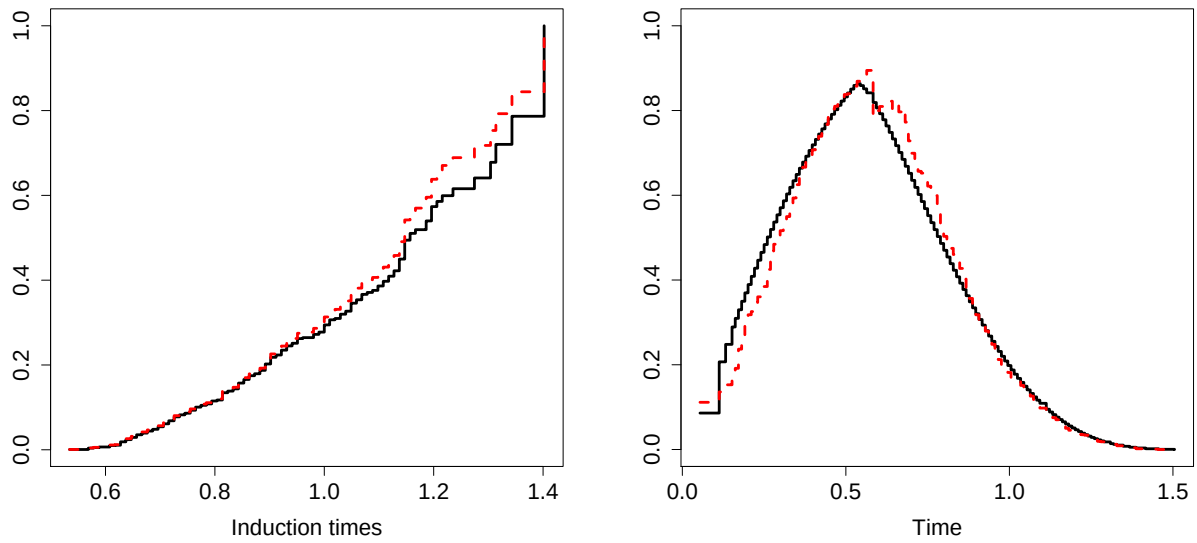

Figure 2: AIDS Blood Transfusion Data: The NPMLE and the semiparametric estimator of F (left), and the nonparametric and the parametric estimator of G (right). The solid curve is the (semi)parametric estimator, the dashed curve is the nonparametric estimator.