Universidade de Vigo

Bootstrapping the NPMLE for doubly truncated data

Carla Moreira and Jacobo de Uña Álvarez

Report 09/01

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://eioweb.uvigo.es/
E-mail: depc05@uvigo.es

Universidade de Vigo

Bootstrapping the NPMLE for doubly truncated data

Carla Moreira and Jacobo de Uña Álvarez

Report 09/01

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita:
Universidade de Vigo
Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

Bootstrapping the NPMLE for doubly truncated data

First author: Carla Moreira

Second author: Jacobo de Uña-Álvarez*

Department of Statistics and OR, University of Vigo, Vigo, Spain

*Corresponding author. Address: Departamento de Estadística e I.O., Facultad de CC. Económicas y Empresariales, Universidad de Vigo, Campus Lagoas-Marcosende, 36310 Vigo, Spain. E-mail: jacobo@uvigo.es, Phone: (+34) 986812492 Fax: (+34) 986812401

Doubly truncated data appear in a number of applications, including astronomy and survival analysis. In this paper we review the existing methods to compute the NPMLE under double truncation, which has no explicit form and must be approximated numerically. We introduce the bootstrap as a method to estimate the finite sample distribution of the NPMLE under double truncation. The performance of the bootstrap is investigated in a simulation study. The nonstandard case in which the right and left truncation times determine each other is covered. As an illustration, nonparametric estimation and inference on the birth process and the age at diagnosis for childhood cancer in North Portugal is considered.

Keywords: Double truncation; nonparametric maximum likelihood; bootstrap

1 Introduction

Censored and truncated data appear in a number of applications, including astronomy and survival analysis. Turnbull (1976) introduced a substitute for the ordinary empirical distribution function (*df*) when the data are incomplete due to grouping, censoring and/or truncation. Later, statistical methods for more specific problems involving truncation were investigated. Woodroofe (1985) considered the nonparametric maximum likelihood estimator (NPMLE) of a *df* under left-truncation, and he derived its asymptotic properties. This estimator was further investigated by Stute (1993), who gave an almost sure representation of the NPMLE as a sum of iid random variables plus a negligible remainder. The presence of censoring from the right in the left-truncation model was considered by Tsai et al. (1987) and Zhou and Yip (1999), among others. These methods can be properly adapted to deal with right-truncated data. In sum, we can say that the problem of one-side truncation is quite solved and well-understood nowadays.

However, literature about double truncation is more scarce. The situation of double truncation will arise in practice whenever both small and large values of the variable of interest are less probably observed, due to the presence of some random bounds which may vary from individual to individual. The NPMLE of a *df* observed under doubly truncation was studied by Efron and Petrosian (1999). The asymptotic properties of this estimator were formally established by Shen (2008). It seems a bit surprising that these two papers are (for the best of our knowledge) almost the only contributions in literature devoted to nonparametric statistics for the double truncation phenomenon. A possible reason is that the non-explicit form of the NPMLE greatly complicates its analysis. Several numerical methods for approximating the NPMLE have been proposed (Efron and Petrosian, 1999; Shen, 2008), but their convergence properties have not been described in detail. Besides, the issue of truncation is ignored many times by practitioners, who do not care about it even when (one-side or double) truncation may dramatically deteriorate the observational procedure and introduce a systematic bias in estimation. All these facts could explain the lack of research and interest on this topic during the last years.

In this paper, we are manly concerned with the revision of the existing numerical algorithms and technical results on the NPMLE for doubly truncated data (Section 2). The main contribution of our work is the investigation of the bootstrap as a method to approximate the finite sample distribution of the NPMLE (Section 3). This is a relevant problem, since the asymptotic distribution of this estimator is complicated and (so far) it has not been used in the development of practical inference methods. In Section 4, we consider a real data application in which double truncation naturally arises; this Section 4 serves for the purpose of illustration of the NPMLE and their bootstrap approximations. Finally, Section 5 reports the main conclusions of our study.

2 The NPMLE revisited

Let X^* be the time of ultimate interest, with df F, and let (U^*, V^*) be the pair of truncation times, with (joint) df K, so one is only able to observe (U^*, V^*, X^*) when $U^* \leq X^* \leq V^*$ (otherwise, nothing is observed). Let $(U_i, V_i, X_i), i = 1,...,n$ be the observed data. Under the assumption of independence between X^* and (U^*, V^*), the full likelihood is given by

$$L_n(f,k) = \prod_{j=1}^{n} \frac{f_j k_j}{\sum_{i=1}^{n} F_i k_i},$$

where $f = (f_1, f_2, ..., f_n)$ and $k = (k_1, k_2, ..., k_n)$ are distributions putting probability f_i on X_i and k_i on (U_i, V_i) respectively, and where F_i is defined through $F_i = \sum_{m=1}^{n} f_m J_{im}$, where $J_{im} = I_{[U_i \leq X_m \leq V_i]} = 1$ if $U_i \leq X_m \leq V_i$ and equal to zero otherwise.

As noted by Shen (2008), this likelihood can be written as a product of the conditional likelihood of the X_i's given the (U_i, V_i)'s, say $L_1(f)$, and the marginal likelihood of the (U_i, V_i)'s, say $L_2(f,k)$:

$$L_n(f,k) = \prod_{j=1}^{n} \frac{f_j}{F_j} \times \prod_{j=1}^{n} \frac{F_j k_j}{\sum_{i=1}^{n} F_i k_i} \equiv L_1(f) \times L_2(f,k).$$

The conditional NPMLE of *F* (Efron and Petrosian, 1999) is defined as the maximizer of $L_1(f)$. This criterion leads to an estimator $\hat{f}$ satisfying

$$\frac{1}{\hat{f}_j} = \sum_{i=1}^{n} J_{ij} \frac{1}{\hat{F}_i}, \ (j = 1,...,n) \tag{1}$$

where $\hat{F}_i = \sum_{m=1}^{n} \hat{f}_m J_{im}$. Equation (1) was used by Efron and Petrosian (1999) to introduce the following EM algorithm to compute $\hat{f}$:

Step *EP1*. Compute the initial estimate $\hat{F}_{(0)}$ corresponding to $\hat{f}_{(0)} = (1/n,...,1/n)$;

Step *EP2*. Apply (1) to get an improved estimator $\hat{f}_{(1)}$ and compute the $\hat{F}_{(1)}$ pertaining to $\hat{f}_{(1)}$;

Step *EP3*. Repeat Step EP2 until convergence criterion is reached.

Efron and Petrosian (1999) also suggested a different algorithm based on a modified Lynden-Bell's (1971) method for one-sided truncation. As claimed by these authors, this second method converges faster than EP1-EP3; however, in our experience, we have found that it may converge to a wrong solution in some situations, so in principle we do not recommend using it in practice.

By interchanging the roles of the X_i's and the (U_i,V_i)'s, the full likelihood can be written as the product

$$L_n(f,k) = \prod_{j=1}^{n} \frac{k_j}{K_j} \times \prod_{j=1}^{n} \frac{K_j f_j}{\sum_{i=1}^{n} K_i f_i} \equiv L_1(k) \times L_2(k,f)$$

where $K_i = \sum_{m=1}^{n} k_m I_{[U_m \le X_i \le V_m]} = \sum_{m=1}^{n} k_m J_{mi}$, and where $L_1(k)$ is the conditional likelihood of the (U_i,V_i)'s and $L_2(k,f)$ is the marginal likelihood of the X_i's. Maximization of $L_1(k)$ leads to a $\hat{k}$ such that

$$\frac{1}{\hat{k}_j} = \sum_{i=1}^{n} J_{ji} \frac{1}{\hat{K}_i}, \ (j = 1,...,n) \qquad (2)$$

where $\hat{K}_i = \sum_{m=1}^{n} \hat{k}_m J_{mi}$. Shen (2008) proved that the solutions to (1) and (2) are (not only the conditional NPMLEs but also) the unconditional NPMLEs of F and K respectively. Besides, he showed that both estimators can be obtained in a simultaneous way by solving the following system of equations:

$$\hat{f}_j = \left[\sum_{i=1}^{n} \frac{1}{\hat{K}_i} \right]^{-1} \frac{1}{\hat{K}_j}, \ \text{j=1,…,n} \qquad (3)$$

$$\hat{k}_j = \left[\sum_{i=1}^{n} \frac{1}{\hat{F}_i} \right]^{-1} \frac{1}{\hat{F}_j}, \quad \text{j=1,…,n} \tag{4}$$

This system of equations suggests the following algorithm to compute the NPMLE (Shen, 2008):

Step S1. Compute the initial estimate $\hat{F}_{(0)}$ corresponding to $\hat{f}_{(0)} = (1/n,...,1/n)$;

Step S2. Apply (4) to get the first step estimator of k, $\hat{k}_{(1)}$, and compute the $\hat{K}_{(1)}$ pertaining to $\hat{k}_{(1)}$;

Step S3. Apply (3) to get the first step estimator of f, $\hat{f}_{(1)}$, and its corresponding $\hat{F}_{(1)}$;

Step S4. Repeat Steps S2 and S3 until convergence criterion is reached.

In practice, algorithms EP1-EP3 and S1-S4 will give the same solution $\hat{f}$. By implementing and running both methods, we have confirmed that they converge in a number of iterations of the same order. Interestingly, the latter method provides the NPMLEs of both F and K in a simultaneous way. In the simulations below, we use EP1-EP3 with stopping rule that the maximum distance among the f_i's computed in two successive steps is below $1/(10\times n)$.

Shen (2008) includes an asymptotic analysis of the NPMLE of F. Specifically, the uniform consistency and the weak convergence of the estimator are established. The asymptotic distribution of $\hat{F}$ is complicated, so these results do not provide practical answers to inference problems under doubly truncation. In our next Section we propose the bootstrap as a method to approximate the finite sample distribution of $\hat{F}$.

3 Bootstrap approximation

In this Section we consider the simple bootstrap (Gross and Lai, 1996) as a method to approximate the distribution of the NPMLE of F in finite samples. We also investigate via simulations the performance of the bootstrap when computing confidence limits for F.

Let (U_{ib}, V_{ib}, X_{ib}), $i = 1,...,n$, be a bootstrap resample taken from the initial data by putting weight $1/n$ at each of the observations $(U_i, V_i, X_i), i = 1,...,n$. Repeat this procedure a large number B of times so we have $b = 1,..., B$ bootstrap resamples. Put $\hat{F}_b$ for the estimator $\hat{F}$ computed from the b th bootstrap resample, $b = 1,..., B$. Then, the values $\hat{F}_1(t),..., \hat{F}_b(t)$ can be used to empirically approximate the finite sample distribution of $\hat{F}(t)$ for a given t.

In our simulations below, 95% confidence limits based on the bootstrap are computed. That is, the 2.5% upper and 2.5% lower values of the $\hat{F}_b(t)$'s are removed to construct the confidence interval. We focus on the attained coverages to see if the simple bootstrap behaves consistently. As specific values of t, we consider the nine deciles of the simulated F .

Our first simulated model is as follows: U^*, V^* and X^* are mutually independent; X^* is drawn from a Uniform$(0,1)$ model, while U^*, V^* are drawn from Uniform distributions with respective supports $(0, a)$ and $(b,1)$, where a and b are chosen to get the following percentages of truncation: 25% $(a = 0.25, b = 0.75)$, 50% $(a = b = 0.5)$, and 67% $(a = 0.67, b = 0.33)$. The truncation occurs when $U^* \leq X^* \leq V^*$ is violated.

We repeat the drawing until forming samples of a given size; sample sizes of $n = 50$, $n = 100$, $n = 150$ and $n = 250$ were considered. The number of bootstrap resamples was taken to be $B = 500$ (partial results obtained for $B = 1000$ were very similar to those reported here). We performed 500 trials for each situation. Results are displayed in Tables 1 to 3.

Table 1. Coverages of the 95% bootstrap confidence intervals for the NPMLE of F along 500 trials for several sample sizes n. $X^* \sim U(0,1), U^* \sim U(0,0.25)$, $V^* \sim U(0.75,1)$ were simulated as mutually independent (percentage of truncation PT=25%). Means and standard deviations of the interval lengths are also reported.

PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI	PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI
25%	50	1	0.804	0.2456243	0.04416079	25%	150	1	0.844	0.1842230	0.016439369
		2	0.864	0.3351561	0.02757534			2	0.910	0.2180977	0.010801898
		3	0.914	0.3576116	0.01663728			3	0.918	0.2222340	0.008241057
		4	0.936	0.3655469	0.01159498			4	0.940	0.2257118	0.006925665
		5	0.946	0.3703399	0.01098973			5	0.926	0.2285656	0.007081697
		6	0.952	0.3715230	0.01414495			6	0.934	0.2314368	0.008704853
		7	0.930	0.3696328	0.02163229			7	0.932	0.2338441	0.011991991
		8	0.916	0.3555943	0.03315225			8	0.902	0.2336715	0.016148443
		9	0.766	0.2770630	0.06085439			9	0.868	0.2122450	0.023700052
25%	100	1	0.840	0.2101497	0.025889812	25%	250	1	0.866	0.1635029	0.009018960
		2	0.902	0.2627010	0.016048790			2	0.938	0.1785669	0.005931400
		3	0.934	0.2703862	0.011099727			3	0.938	0.1787695	0.004462368
		4	0.940	0.2745797	0.008857473			4	0.956	0.1782646	0.003442428
		5	0.948	0.2776489	0.008755461			5	0.968	0.1777068	0.003238050
		6	0.938	0.2804290	0.011428271			6	0.948	0.1772098	0.003723815
		7	0.918	0.2819764	0.016347936			7	0.934	0.1771925	0.004974283
		8	0.906	0.2797369	0.023488212			8	0.914	0.1762222	0.006726882
		9	0.822	0.2432781	0.038462042			9	0.878	0.1622074	0.009823084

Table 2. Coverages of the 95% bootstrap confidence intervals for the NPMLE of F along 500 trials for several sample sizes n. $X^* \sim U(0,1), U^* \sim U(0,0.5)$, $V^* \sim U(0.5,1)$were simulated as mutually independent (percentage of truncation PT=50%). Means and standard deviations of the interval lengths are also reported.

PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI	PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI
50%	50	1	0.800	0.2893772	0.06326777	50%	150	1	0.820	0.2039402	0.024850033
		2	0.884	0.3896346	0.04333030			2	0.876	0.2526884	0.015606256
		3	0.906	0.4252954	0.02672420			3	0.920	0.2613354	0.010427008
		4	0.934	0.4320490	0.01799294			4	0.940	0.2644056	0.008244577
		5	0.946	0.4320227	0.01488952			5	0.956	0.2664698	0.007781465
		6	0.924	0.4317530	0.01807301			6	0.954	0.2678976	0.009054365
		7	0.904	0.4268223	0.02659023			7	0.938	0.2684207	0.012415691
		8	0.874	0.3932083	0.04345826			8	0.914	0.2614831	0.018086663
		9	0.642	0.2870272	0.08478940			9	0.838	0.2232881	0.028837889
50%	100	1	0.842	0.2517481	0.04761928	50%	250	1	0.862	0.1903549	0.020159429
		2	0.894	0.3191751	0.03044891			2	0.904	0.2144114	0.013712314
		3	0.924	0.3302741	0.02022693			3	0.928	0.2180754	0.009590434
		4	0.936	0.3325236	0.01452287			4	0.950	0.2175388	0.007245497
		5	0.948	0.3325854	0.01219506			5	0.950	0.2168241	0.006348315
		6	0.942	0.3327959	0.01366703			6	0.938	0.2170453	0.006665531
		7	0.930	0.3310709	0.01869327			7	0.936	0.2162816	0.008297049
		8	0.890	0.3174481	0.02830693			8	0.908	0.2111539	0.011569802
		9	0.814	0.2611456	0.04801988			9	0.862	0.1899454	0.017602704

Table 3. Coverages of the 95% bootstrap confidence intervals for the NPMLE of F along 500 trials for several sample sizes n. $X^* \sim U(0,1)$, $U^* \sim U(0,0.67)$, $V^* \sim U(0.33,1)$ were simulated as mutually independent (percentage of truncation PT=25%). Means and standard deviations of the interval lengths are also reported.

PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI	PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI
67%	50	1	0.830	0.3122912	0.06753057	67%	150	1	0.862	0.2195450	0.02850742
		2	0.876	0.4066955	0.04630946			2	0.892	0.2653109	0.01873450
		3	0.916	0.4449330	0.02914732			3	0.900	0.2785490	0.01330000
		4	0.930	0.4575003	0.01909018			4	0.908	0.2859858	0.01068658
		5	0.912	0.4600851	0.01596511			5	0.918	0.2897380	0.01090830
		6	0.906	0.4554927	0.01919461			6	0.916	0.2899473	0.01350828
		7	0.896	0.4435533	0.02792070			7	0.910	0.2853926	0.01853920
		8	0.864	0.4038760	0.04657200			8	0.900	0.2765086	0.02726752
		9	0.664	0.3001341	0.08849003			9	0.852	0.2419259	0.04136601
67%	100	1	0.852	0.2604485	0.04700933	67%	250	1	0.860	0.1790041	0.013800119
		2	0.896	0.3237959	0.03136434			2	0.922	0.2060986	0.009266317
		3	0.922	0.3374700	0.02089404			3	0.944	0.2154767	0.006891761
		4	0.932	0.3429401	0.01447854			4	0.950	0.2204567	0.005964958
		5	0.934	0.3428345	0.01151395			5	0.944	0.2238385	0.006184628
		6	0.934	0.3393893	0.01204575			6	0.918	0.2237231	0.007700082
		7	0.934	0.3299615	0.01574352			7	0.926	0.2212286	0.010562632
		8	0.886	0.3107754	0.02353785			8	0.914	0.2150350	0.014762260
		9	0.838	0.2557840	0.04073803			9	0.878	0.1939044	0.020935150

In these Tables 1 to 3 we see that the bootstrap improves its performance with an increasing sample size. The proportion of truncated data plays some role too, and it can be seen that higher truncation rates lead in general to poorer coverages. However, this is not always the case, a fact that should not be taken as extremely surprising since the final sample sizes are the same regardless the proportion of truncation. Another issue that follows from the reported results is that the bootstrap coverages tend to underestimate the nominal 95% at both tails of the distribution; however, the performance of the method between percentiles 30% and 70% seems to be quite satisfactory.

In Tables 1 to 3 we report the mean and the standard deviation of the length of the bootstrap confidence interval along the 500 replicates. The relatively large values of the standard deviations at both tails of F suggest that the double truncation provokes a serious damage of the sampling information at these extreme points, where the variance of $\hat{F}(t)$ is not properly estimated.

We are also concerned with the situation in which U^* and V^* determine each other via $U^*=V^*-\delta$, where δ is a known, positive constant which represents in practice the width of an observational window (see our Section 4). Because of this, we simulated a second model in the following way. X^* is drawn from a Uniform$(0,15)$ distribution and, independently, U^* is drawn from a Uniform$(-5,15)$ distribution; then, we compute $V^*=U^*+5$. This simulated example is interesting because of two reasons. First, it reproduces a situation similar to the practical setup that will be explored in Section 4. Second, it represents a case in which the truncated distribution (that is, the distribution of X^* conditionally on $U^*\leq X^*\leq V^*$) coincides with that of interest; in other words, there is no observational bias, in the sense that the truncation does not change the sampling probabilities of each X_i. Nevertheless, 37.5% of the observations are truncated. Simulation results (again along 500 Monte Carlo trials) are displayed in Table 4. In Table 4 we see that the bootstrap coverages are quite close to the nominal 95% even for moderate sample sizes.

Table 4. Coverages of the 95% bootstrap confidence intervals for the NPMLE of F along 500 trials for several sample sizes n. $X^*\sim U(0,15)$, $U^*\sim U(-5,15)$ were independently simulated, $V^*=U^*+5$ (percentage of truncation PT=37.5%). Means and standard deviations of the interval lengths are also reported.

PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI	PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI
37,5%	50	1	0.926	0.2963838	0.033498505	37,5%	150	1	0.920	0.1334691	0.002107902
		2	0.958	0.4118613	0.027939769			2	0.934	0.1908358	0.002776801
		3	0.960	0.4750338	0.018368003			3	0.952	0.2212942	0.002710846
		4	0.964	0.5057550	0.012083352			4	0.946	0.2348739	0.002339495
		5	0.964	0.5172609	0.009944428			5	0.966	0.2371179	0.002180414
		6	0.954	0.5091094	0.013326961			6	0.946	0.2320010	0.002235822
		7	0.956	0.4786123	0.020850551			7	0.942	0.2189926	0.002442784
		8	0.960	0.4222021	0.031123942			8	0.940	0.1885122	0.002545796
		9	0.938	0.3078716	0.038154186			9	0.940	0.1339126	0.001946357
37,5%	100	1	0.946	0.1779742	0.005417392	37,5%	250	1	0.952	0.09770567	0.0005203785
		2	0.966	0.2519635	0.005733617			2	0.960	0.13559469	0.0006667569
		3	0.972	0.2940953	0.004752728			3	0.958	0.15420616	0.0007077924
		4	0.970	0.3149343	0.003803556			4	0.958	0.16224800	0.0006776715
		5	0.970	0.3203841	0.003489434			5	0.956	0.16473085	0.0007322845
		6	0.966	0.3147412	0.003953728			6	0.962	0.16221405	0.0008096485
		7	0.952	0.2922541	0.005136242			7	0.948	0.15352497	0.0008784952
		8	0.934	0.2513246	0.006331729			8	0.936	0.13427046	0.0009156716
		9	0.946	0.1768550	0.005384688			9	0.944	0.09728982	0.0006555822

As a technical remark, we mention that our second simulated model is not covered by the theory in Shen (2008). This is because the density of $(U*,V*)$ does not exist, and hence his conditions do not apply. However, we believe that the asymptotic properties of the NPMLE hold (when properly rewritten) even in this situation. Of course, the role of the joint density of $(U*,V*)$ in the Theorems in Shen (2008) must be played by one of the marginal densities in such a case. The simulations reported in Table 4 also suggest that the consistency of the NPMLE does not rely on the existence of the joint density of the truncation times.

Finally, in Tables 5 to 7 we report the results attained by the bootstrap in a third simulated scenario. Here, following Shen (2008), $X*$ was simulated according to a Weibull model with shape parameter $\delta_f = 4$, while $U*$ and $V*$ were simulated as Exponential random variables with scale parameters δ_g and δ_q chosen to give the following proportions of truncation: 24% $(\delta_g = 0.25, \delta_q = 4)$, 61% $(\delta_g = 0.25, \delta_q = 1)$, and 77% $(\delta_g = 1, \delta_q = 1)$; the variables $X*$, $U*$ and $V*$ being mutually independent. The figures in these Tables 5 to 7 are similar to (with coverages at the tails even better than) those reported for the other simulated scenarios.

Table 5. Coverages of the 95% bootstrap confidence intervals for the NPMLE of F along 500 trials for several sample sizes n. $X^* \sim Weibull(4)$, $U^* \sim Exp(0.25)$ and $V^* \sim Exp(4)$ were independently simulated, (percentage of truncation PT=24%). Means and standard deviations of the interval lengths are also reported.

PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI
24%	50	1	0.934	0.1975237	0.005823337
		2	0.932	0.2967048	0.005574116
		3	0.926	0.3733609	0.005871491
		4	0.932	0.4281354	0.006704173
		5	0.926	0.4664573	0.008870403
		6	0.922	0.4887661	0.015113794
		7	0.908	0.4782586	0.025869852
		8	0.896	0.4167960	0.039955903
		9	0.574	0.2746207	0.004540717
24%	100	1	0.940	0.1353184	0.0009537175
		2	0.940	0.2089121	0.0011641410
		3	0.936	0.2656524	0.0015496776
		4	0.926	0.3084118	0.0021468187
		5	0.938	0.3395079	0.0031899226
		6	0.926	0.3566491	0.0052257673
		7	0.920	0.3554470	0.0091432918
		8	0.920	0.3294152	0.0158875977
		9	0.804	0.2457709	0.0307463406
24%	150	1	0.944	0.1085990	0.0004499294
		2	0.942	0.1670520	0.0005694156
		3	0.946	0.2127983	0.0006119813
		4	0.942	0.2481035	0.0006804801
		5	0.942	0.2724610	0.0008568508
		6	0.924	0.2859906	0.0015149589
		7	0.922	0.2866114	0.0025428711
		8	0.920	0.2681944	0.0046969563
		9	0.854	0.2069035	0.0110171636
24%	250	1	0.970	0.08447693	0.0001588800
		2	0.966	0.13018770	0.0002033821
		3	0.946	0.16561479	0.0002415360
		4	0.952	0.19272548	0.0003168499
		5	0.954	0.21344269	0.0004521632
		6	0.946	0.22542940	0.0007007448
		7	0.936	0.22731152	0.0012425856
		8	0.930	0.21610019	0.0021260028
		9	0.910	0.17743371	0.0042369239

Table 6. Coverages of the 95% bootstrap confidence intervals for the NPMLE of F along 500 trials for several sample sizes n. $X^* \sim Weibull(4)$, $U^* \sim Exp(0.25)$ and $V^* \sim Exp(1)$ were independently simulated, (percentage of truncation PT=61%). Means and standard deviations of the interval lengths are also reported.

PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI
61%	50	1	0.908	0.1856790	0.0047243227
		2	0.916	0.2526272	0.0026333367
		3	0.928	0.2883946	0.0013351945
		4	0.920	0.3054700	0.0007121448
		5	0.916	0.3096677	0.0004312397
		6	0.914	0.3016381	0.0005555537
		7	0.938	0.2815981	0.0009679642
		8	0.936	0.2435847	0.0016900318
		9	0.922	0.1785869	0.0025686772
61%	100	1	0.920	0.1398757	0.0011737566
		2	0.944	0.1834375	0.0006288367
		3	0.946	0.2072275	0.0003725595
		4	0.952	0.2190578	0.0002032854
		5	0.952	0.2219913	0.0001554677
		6	0.944	0.2163113	0.0001703713
		7	0.952	0.2011812	0.0002595534
		8	0.930	0.1730449	0.0004353625
		9	0.932	0.1266531	0.0006219371
61%	150	1	0.936	0.1162993	5.245997e-4
		2	0.934	0.1510687	3.108046e-4
		3	0.934	0.1700682	1.883559e-4
		4	0.934	0.1800177	1.079951e-4
		5	0.936	0.1818839	8.505465e-5
		6	0.944	0.1768257	1.004030e-4
		7	0.948	0.1642230	1.418396e-4
		8	0.940	0.1425388	2.123472e-4
		9	0.932	0.1063069	2.631238e-4
61%	250	1	0.938	0.08991838	2.016862e-4
		2	0.934	0.11725864	1.319403e-4
		3	0.926	0.13182968	8.252832e-5
		4	0.944	0.13981962	5.914922e-5
		5	0.936	0.14163626	5.663164e-5
		6	0.950	0.13792121	4.646030e-5
		7	0.948	0.12857017	5.987313e-5
		8	0.952	0.11096217	8.142365e-5
		9	0.916	0.08232521	1.060810e-4

Table 7. Coverages of the 95% bootstrap confidence intervals for the NPMLE of F along 500 trials for several sample sizes n. $X^* \sim Weibull(4)$, $U^* \sim Exp(1)$ and $V^* \sim Exp(1)$ were independently simulated, (percentage of truncation PT=77%). Means and standard deviations of the interval lengths are also reported.

PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI	PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI
77%	50	1	0.926	0.1748611	0.0035380332	77%	150	1	0.938	0.1096601	3.887405e-4
		2	0.926	0.2424130	0.0021247580			2	0.938	0.1443336	2.412177e-4
		3	0.926	0.2801101	0.0011204420			3	0.952	0.1640779	1.520827e-4
		4	0.936	0.2987272	0.0005444046			4	0.950	0.1754095	8.998838e-5
		5	0.912	0.3047008	0.0004058266			5	0.948	0.1793936	7.433879e-5
		6	0.934	0.2999724	0.0005604185			6	0.948	0.1760500	9.045534e-5
		7	0.946	0.2812444	0.0001155338			7	0.946	0.1649955	1.147859e-4
		8	0.932	0.2440442	0.0018353241			8	0.960	0.1447237	1.711888e-4
		9	0.894	0.1809869	0.0033074950			9	0.928	0.1089519	2.730569e-4
77%	100	1	0.920	0.1332339	0.0007762434	77%	250	1	0.932	0.0851074	1.336451e-4
		2	0.948	0.1743842	0.0004721978			2	0.946	0.1116196	8.727653e-5
		3	0.952	0.2000963	0.0002667188			3	0.962	0.1270808	5.752648e-5
		4	0.946	0.2135039	0.0001683143			4	0.952	0.1355112	4.617004e-5
		5	0.932	0.2179024	0.0001391024			5	0.962	0.1385305	3.769955e-5
		6	0.932	0.2136059	0.0001428158			6	0.946	0.1355736	4.332742e-5
		7	0.938	0.2005204	0.0002367707			7	0.958	0.1268817	5.517355e-5
		8	0.966	0.1762387	0.0003677982			8	0.938	0.1112504	8.089469e-5
		9	0.936	0.1325477	0.0006467370			9	0.922	0.0839049	1.184319e-4

4 Real data illustration

The childhood cancer data includes all the cases diagnosed in North Portugal between January 1st 1999 and December 31st 2003, on children aged below 15 years old, with a follow-up until April 30th 2006. The available statistical information is contained in the following variables: birth date; date of death; censoring status (value 1 if death is observed or 0 otherwise); source of diagnosis (institution at which the diagnosis took place); residence (including parishes, small towns and districts); sex; age at diagnosis (in years); date of the first symptom; date of first examination; date of diagnosis; and type of cancer (leukaemia, lymphoma, central nervous system cancers, neuroblastoma, retinoblastoma, renal cancers, hepatic tumours, bone tumours, soft tissues tumours, germ cell tumours, melanomas and others epithelial tumours; according to paediatric classification tumours whose based according the International Childhood Cancer Classification, 3 rd Edition.

The data correspond to 409 children diagnosed from cancer, 180 female and 229 male, the birth date varying between May 13th 1984 and July 2nd 2003. In the five years of recruitment (between January 1st 1999 and December 31st 2003), the yearly number of cases diagnosed ranged from 63 (2002) to 90 (2003). The most precocious diagnosis corresponded to a 6 days old baby, and the latest diagnostic case verified corresponded to an adolescent with almost 15 years old. The more frequent diagnostics are the precocious: 50% of the cases correspond to children below six years old, and 75% of the cases correspond to children below ten years old. The mean age of diagnosis was 6.5 (in years). We concentrate on the 393 cases (220 males, 173 females) which report complete information about the progress of the disease (see Moreira and de Uña-Álvarez, 2007, for details about these data and estimation of cancer survivorship).

Let X^* be the age (in years) at diagnosis and let U^* be the age of the individual at January 1st 1999. Note that (U^*, X^*) is observed only when $U^* \leq X^* \leq U^*+5$. Hence, the distribution of X^* is doubly truncated by (U^*, V^*) where $V^* = U^*+5$. This implies that (in principle) ordinary methods for estimating the distribution of the age at diagnosis should not be applied. In Figure 1 we depict the NPMLE of the *df* of X^* (computed from the algorithm EP1-EP3 in Section 2) along with the 95% pointwise confidence band based on the simple bootstrap. This estimator indicates that in most of the cases the diagnosis occurs at early ages. For comparison purposes, the ordinary empirical *df* is also reported. It turns out that both functions almost coincide along their support. This suggests that the double truncation issue does not induce an observational bias in the diagnosis age. This should not be taken as a surprising fact; indeed, when U^* is uniformly distributed, it is easily seen that the distribution of X^* conditionally on $U^* \leq X^* \leq U^*+5$ is the same as that of X^*.

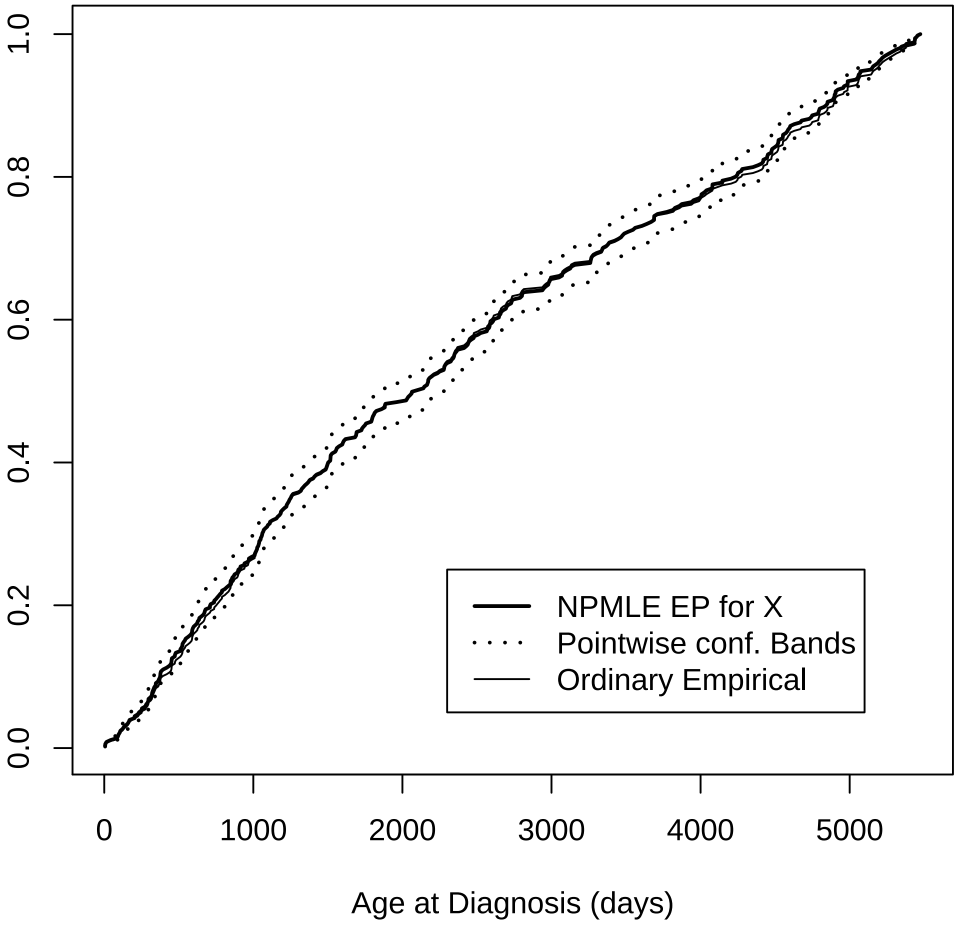

Figure 1. NPMLE of the distribution of the age at diagnosis for the childhood cancer data, and 95% pointwise confidence band based on the bootstrap. The ordinary empirical distribution of the age at diagnosis is included for comparison.

We also estimated the df of $V^* = U^* + 5$ by means of its NPMLE. For doing this, we just consider V^* as doubly truncated by the pair $\left(X^*, X^* + 5\right)$ and then we apply algorithm EP1-EP3. Note that V^* measures time from birth to the end of recruitment (December 31st 2003); hence, the distribution of V^* can be interpreted from the viewpoint of the birth process of the individuals suffering from cancer during their childhood. The resulting estimator is displayed in Figure 2. When looking at the 95% limits, we see that there is no disagreement between the NPMLE and the uniform distribution (included in Figure 2 for comparison). Although our proposed bootstrap only provides pointwise confidence limits, this suggests that the uniform assumption on the birth process could be acceptable.

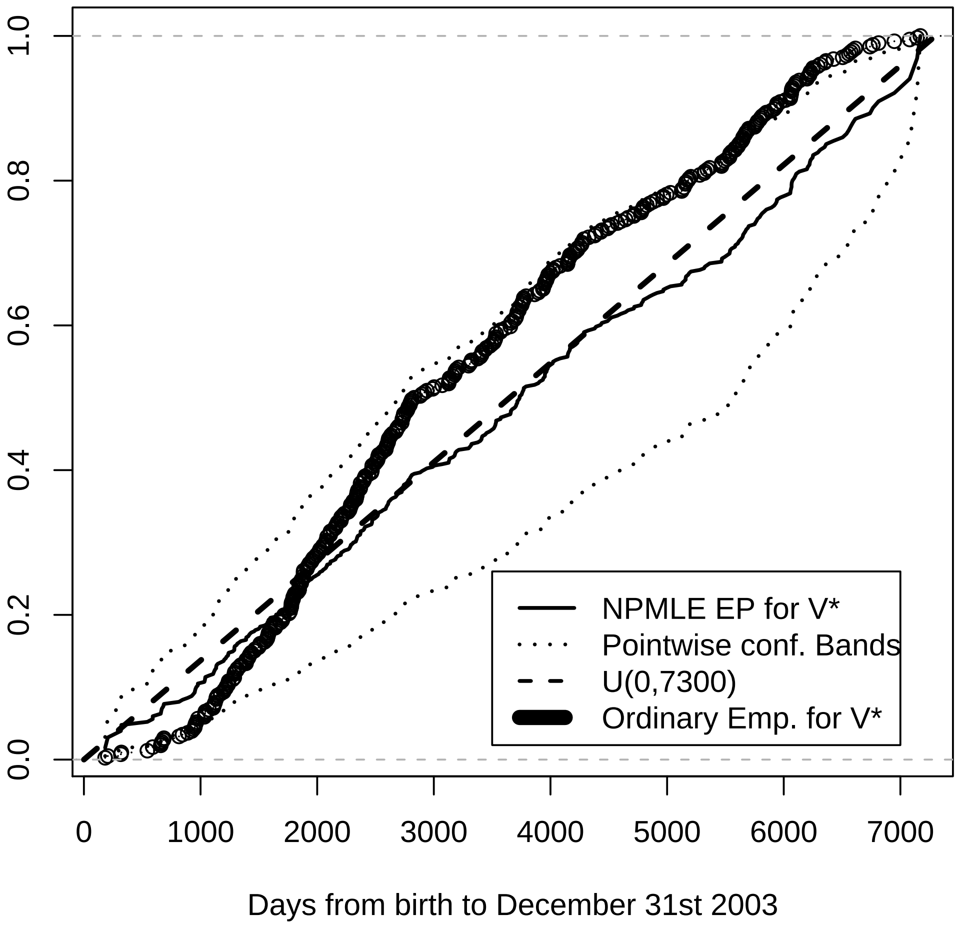

Figure 2. NPMLE of the distribution of (time from birth to December 31st 2003) for the childhood cancer data, and 95% pointwise confidence band based on the bootstrap. The uniform distribution and the ordinary empirical of V* are included for comparison.

Figure 2 also reports the ordinary empirical *df* of V^* computed from the observed right-truncation limits. Unlike for X^* (Figure 1), we see that the impact of the observational bias is remarkable in this case. Comparison of the NPMLE and the empirical *df* indicates that, due to the double truncation, small values (and also extremely large values) of V^* are less probably observed. As a consequence, if the truncation were ignored, one would underestimate the birth rate of individuals (eventually suffering from cancer along their childhood) between 2001 and 2003 (which correspond to about the first 1000 days in Figure 2).

5 Main conclusions

In this paper the NPMLE for doubly truncated data has been revisited. Existing algorithms for the numerical approximation of the NPMLE (which has no explicit form) have been reviewed. Both the estimation of the doubly truncated distribution and of the (joint) distribution of the truncation times were considered. As a recommendation to practitioners, we suggest using the first algorithm in Efron and Petrosian (1999) or the alternative method in Shen (2008) for the computation of the NPMLE. These methods may converge slowly to the NPMLE in some instances; some simulations (not reported here) indicate that the choosing of the initial estimate $\hat{f}_{(0)}$ in Step EP1 may influence the number of iteration until reaching convergence. In practice, if some information about the observational bias is available, this should be used when making a decision about this initial solution, in order to get a faster convergence of the algorithm.

Since the asymptotic distribution of the NPMLE (Shen, 2008) is complicated, the bootstrap has been introduced as a method to approximate the sampling distribution of the NPMLE. The behaviour of the simple bootstrap was tested in a simulation study, in which the coverages of the confidence intervals based on the bootstrap were computed. For a simple size of 250 or even less, it has been found that the bootstrap coverages are close to nominal at least between the 30% and 70% percentiles of the true distribution. Some problems were found at both tails, in accordance with the loss of information provoked by the double truncation. Both the situation of truncation times with and without joint density were covered.

We have applied the proposed methods to explore the age of diagnosis and the birth process for childhood cancer in North Portugal. Point estimates and pointwise confidence bands based on the bootstrap were displayed for illustration purposes. We have seen that ignoring the double truncation issue may introduce a severe bias in estimation. All the methods were implemented in R language.

Acknowledgements: Work supported by the research Grant MTM2008-03129 of the Spanish Ministerio de Ciencia e Innovación, and by the Grant PGIDIT07PXIB300191PR of the Xunta de Galicia. We thank Rosa Crujeiras for her assistance with the programming. We also thank Dr. María José Bento from the IPO (the Portuguese Institute for Cancer at Porto) for kindly providing the childhood cancer data.

References:

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

Gross S.T. and Lai T.L. (1996) Bootstrap methods for truncated and censored data. Statistica Sinica 6, 509-530.

Lynden-Bell, D. (1971) A method of allowing for known observational selection in small samples applied to 3CR quasars. Mon. Not. R. Astr. Soc. 155, 95-118.

Moreira, C. and de Uña-Álvarez, J. (2007) Childhood cancer in North Portugal:incidence, survival and methodological issues (M. I. Gomes, D. Pestana and P. Silva Eds.) Book of Abstracts of the 56th Session of the ISI.

Shen P-S. (2008) Nonparametric analysis of doubly truncated data. Annals of the Institute of Statistical Mathematics. DOI 10.1007/s10463-008-0192-2.

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

Tsai, W.-Y., Jewell, N. P. and Wang, M.-C. (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.

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

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