Analysis of incomplete data due to double truncation

ISCB 2009

Carla Moreira and Jacobo de Uña-Álvarez

Department of Statistics and OR
University of Vigo
Spain

30th Annual Conference of the International Society for Clinical Biostatistics

August 23rd - 27th, 2009 - Prague - Czech Republic

Outline

Motivation examples

- Astronomy
- Economy
- Epidemiology
- Survival Analysis

Motivation examples

- Astronomy
- Economy
- Epidemiology
- Survival Analysis

Related with Epidemiology and/or Survival Analysis:

- Time from HIV infection to diagnosis of AIDS (Bilker and Wang, 1996)

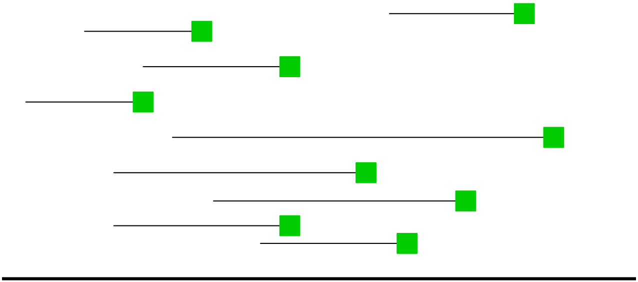
Motivation examples

- Astronomy
- Economy
- Epidemiology
- Survival Analysis

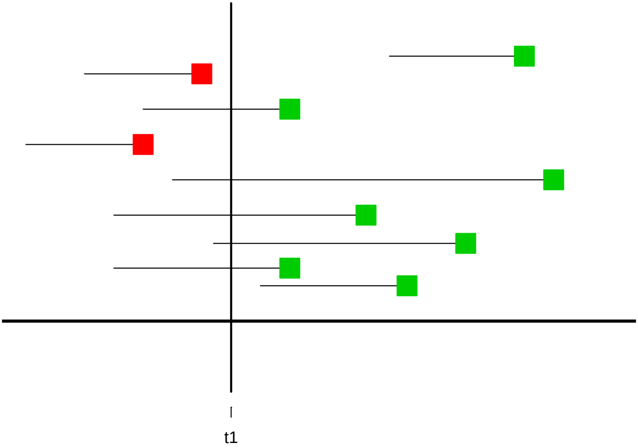
Related with Epidemiology and/or Survival Analysis:

- Time from HIV infection to diagnosis of AIDS (Bilker and Wang, 1996)
- Time from birth to diagnosis in childhood cancer (Moreira and De Uña-Álvarez, 2007)

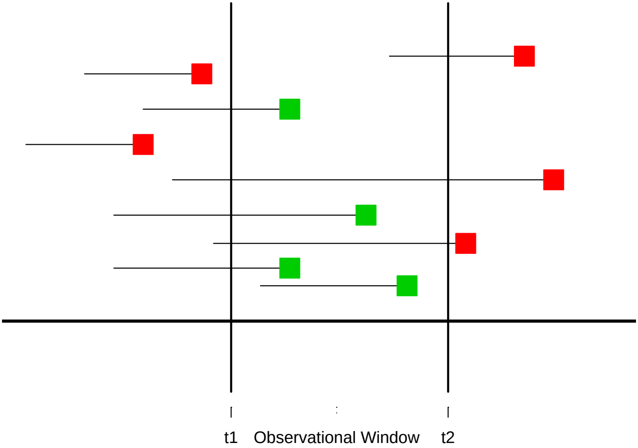
Truncation Scheme



Truncation Scheme



Truncation Scheme



Truncation Scheme

- Let X^* be the ultimate time of interest with df F
- (U^*, V^*) the pair of truncation times, with joint df K
- We observe (U^*, X^*, V^*) if and only if $U^* \leq X^* \leq V^*$
- Let $(U_i, X_i, V_i), i = 1, ..., n$ be the observed data.

Under the assumption of independence between X^* and (U^*, V^*):

The full likelihood is given by:

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

Truncation Scheme

Where:

- $f = (f_1, f_2, ..., f_n)$
- $k = (k_1, k_2, ..., k_n)$
- $F_i = \sum_{m=1}^{n} f_m J_{i_m}$

and

$$J_{i_m} = I_{[U_i \leq X_m \leq V_i]} = 1 \quad \text{if} \quad U_i \leq X_m \leq V_i,$$

or zero otherwise.

As noted by Shen (2008):

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

Efron-Petrosian estimator

The conditional NPMLE of F (Efron-Petrosian, 1999) is defined as the maximizer of $L_1(f)$.

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

where $\hat{F}_i = \sum_{m=1}^{n} \hat{f}_m J_{im}$.

This equation was used by Efron and Petrosian (1999) to introduce the EM algorithm to compute $\hat{f}$.

EM algorithm from Efron and Petrosian (1999)

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

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

EP3. Repeat Step EP2 until convergence criterion is reached.

Shen estimator

Interchanging the roles of X's and (U_i, V_i):

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

where

$$K_i = \sum_{m=1}^{n} k_m I_{[U_m \leq X_i \leq V_m]} = \sum_{m=1}^{n} k_m J_{im}$$

and maximizing $L_1(k)$:

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

with $\hat{K}_i = \sum_{m=1}^{n} \hat{k}_m J_{im}$.

Shen Estimator

Shen (2008) showed that the solutions are the unconditional NPMLE of F

and K, respectively, and both estimators can be obtained by:

$$\hat{f}_j = \left[\sum_{i=1}^{n} \frac{1}{\hat{K}_j} \right]^{-1} \frac{1}{\hat{K}_j}, \quad j = 1, ..., n$$

$$\hat{k}_j = \left[\sum_{i=1}^{n} \frac{1}{\hat{F}_j} \right]^{-1} \frac{1}{\hat{F}_j}, \quad j = 1, ..., n$$

EM algorithm from Shen (2008)

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

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

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

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

Simple bootstrap procedure

- From the original data, we take a bootstrap resample (U_{ib}, V_{ib}, X_{ib}), $i = 1, ..., n$ 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
- Put $\hat{F}_b$ for the estimator $\hat{F}$ computed from the b^{th} bootstrap resample, $b = 1, ..., B$
- The values of $\hat{F}_1(t), ..., \hat{F}_b(t)$ can be used to empirically approximate the finite sample distribution of $\hat{F}(t)$ for a given t

Simulated model

- X^* is independent of (U^*, V^*) but $U^* = V^* - \delta$
- $X^* \sim Unif(0, 15)$, $U^* \sim Unif(-5, 15)$ and $V^* = U^* + 5$

Simulated model

PT	n	Deciles	Coverage	Mean Length CI	Length sd. CI
		1	0.926	0.3019516	0.033585412
		2	0.951	0.4139273	0.027835971
		3	0.958	0.4704103	0.018342575
		4	0.971	0.4981912	0.011472745
37,5%	50	5	0.957	0.5042559	0.008720808
		6	0.960	0.4942161	0.010959723
		7	0.955	0.4624988	0.017726775
		8	0.940	0.3994099	0.026072178
		9	0.917	0.2852907	0.032080445
		1	0.950	0.09643203	0.0005328409
		2	0.941	0.13397596	0.0007621275
		3	0.950	0.15348897	0.0007692665
		4	0.950	0.16222925	0.0006908011
37,5%	250	5	0.956	0.16459729	0.0006598123
		6	0.959	0.16209428	0.0006474786
		7	0.958	0.15367155	0.0006482495
		8	0.951	0.13382406	0.0006319017
		9	0.975	0.09619246	0.0004300545

Table: Coverages of the 95% bootstrap confidence intervals for the NPMLE of F along 1000 trials for sample sizes 50 and 250. $X^* \sim Unif(0, 15)$, $U^* \sim Unif(-5, 15)$ were independently simulated and $V^* = U^* + 5$. Means and standard deviations of the interval lengths are also reported. Simple bootstrap method was considered.

Childhood cancer data description

- Includes all the cases diagnosed in Northern region of Portugal between January 1st, 1999 and December 31st, 2003;
- Follow-up until April 30th, 2006;
- Variables included: birth date, date of death, censoring status, source of diagnosis, residence, sex, age at diagnosis, date of first symptom, date of first examination, date of diagnosis and type of cancer; according to paediatric classification tumours whose based according the International Childhood Cancer Classification, 3rd Edition;

Childhood cancer data description

- Data correspond to 409 children, with age below 15 years old (180 female and 229 male);
- Birth date varying between May 13th, 1984 and July 2nd, 2003;
- In the five years of recruitment, the number of cases ranged almost uniformly (63 in 2002 to 90 in 2003);
- The more frequent diagnosis 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.

Data Formulation

- Let X^* be the age (in years) at diagnosis and U^* the age of the individual at January 1 st, 1999;
- (U^*, V^*) is observed only when $U^* \leq X^* \leq U^* + 5$;
- X^* is doubly truncated by (U^*, V^*) where $V^* = U^* + 5$;
- V^* is doubly truncated by $(X^*, X^* + 5)$.

NPMLE of the df of X^*

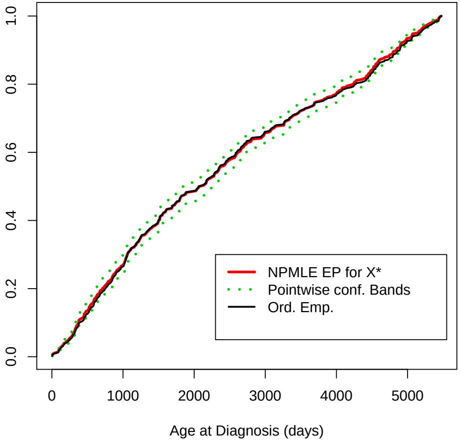


Figure: NPMLE of the distribution of the age at diagnosis for the childhood cancer data, and 95% pointwise confidence band based on the simple boostrap. The ordinary empirical distribution of the age at diagnosis is included for comparison.

NPMLE of the df of V^*

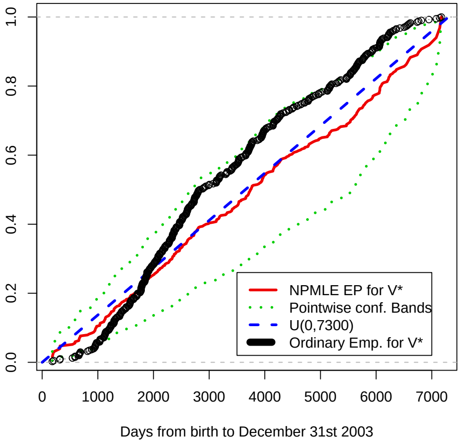


Figure: NPMLE of the distribution of NPMLE of the distribution of the time from birth to December 31st, 2003 for the childhood cancer data, and 95% pointwise confidence band based on the bootstrap. The uniform distribution and the ordinary empirical df of V^* are included for comparison.

DT vs LTRC

- Doubly truncated(DT) data are not the same as left-truncated-right-censored (LTRC) data as considered in Wang (1991) or Gross and Lai (1996).

DT vs LTRC

- Doubly truncated(DT) data are not the same as left-truncated-right-censored (LTRC) data as considered in Wang (1991) or Gross and Lai (1996).
- In LTRC setup, one would have observed those cases with $X^* > U^* + 5$, with the information on the lifetime X^* limited to $U^* + 5$ (right censored information).

DT vs LTRC

- Doubly truncated (DT) data are not the same as left-truncated-right-censored (LTRC) data as considered in Wang (1991) or Gross and Lai (1996).
- In LTRC setup, one would have observed those cases with $X^* > U^* + 5$, with the information on the lifetime X^* limited to $U^* + 5$ (right censored information).
- In our DT scenario, we have no information on these subjects, and hence inference procedures are expected to be less efficient than those corresponding to LTRC data.

Simulated model

- X^* is independent of (U^*, V^*) but $U^* = V^* - \delta$
- $X^* \sim Unif(0, 15)$, $U^* \sim Unif(-5, 15)$ and $V^* = U^* + 5$
- Let $(U_i, X_i, V_i), i = 1, ..., n$ be the simulated data
- Accept the pairs that verified $U_i \leq X_i$
- If $V_i < X_i, i = 1, ..., n$, the case is censored, otherwise is doubly truncated.

DT vs LTRC

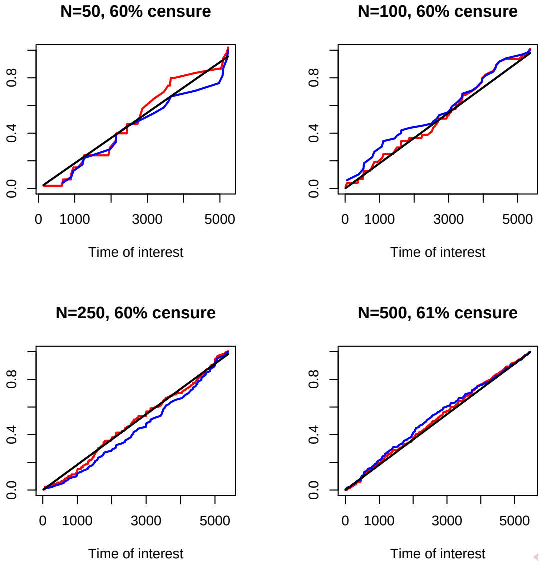


Figure 3: NPMLE of the distribution function of the time of interest for doubly truncated data (blue line), Kaplan-Meier estimator for LTRC data (red line). The interest distribution function (black line).

Summary

- NPMLE for doubly truncated data has been revisited;

Summary

- NPMLE for doubly truncated data has been revisited;
- Existing algorithms for the numerical approximation of the NPMLE has been reviewed;

Summary

- NPMLE for doubly truncated data has been revisited;
- Existing algorithms for the numerical approximation of the NPMLE has been reviewed;
- Both the estimation of the doubly truncated distribution and of the (joint) distribution of the truncation times were considered;

Summary

- NPMLE for doubly truncated data has been revisited;
- Existing algorithms for the numerical approximation of the NPMLE has been reviewed;
- Both the estimation of the doubly truncated distribution and of the (joint) distribution of the truncation times were considered;
- We suggest using the first algorithm in Efron and Petrosian (1999) or the alternative method in Shen (2008) for the computation of the NPMLE;

Summary

- 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;
- Ignoring the double truncation issue may introduce a severe bias in estimation;
- All methods were implemented in R language and included in DTDA R package.

Future Research

- Semiparametric estimator for doubly truncated data
- Regression with doubly truncated responses
- Application of the NPMLE to kernel estimation of the density and the hazard rate under double truncation

Acknowledgments

- Work supported by the research Grant MTM2008-03129 of the Spanish Ministerio de Ciencia e Innovación
- Grant PGIDIT07PXIB300191PR of the Xunta de Galicia

References

Bilker, W.B. and Wang, M.-C. (1996)
A semi-parametric extension of the Mann-Whitney test for randomly truncated data.
Biometrika, 52, 10-20.

Efron, B. 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.

References

Moreira, C. and de Uña Álvarez, J.(2007)
Childhood cancer in North Portugal: incidence, survival and methodological issues.
Book of Abstracts of the 56th Session of ISI.

Moreira, C. and de Uña Álvarez, J.(Under revision)
Bootstrapping the NPMLE for doubly truncated data.
Journal of Nonparametric Statistics.

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

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