An R package for analyzing truncated data

Carla Moreira[1*], Jacobo de Uña-Álvarez[1] and Rosa M. Crujeiras[2]

[1] Department of Statistics and OR, University of Vigo
[2] Department of Statistics and OR, University of Santiago de Compostela

*carlamgmm@gmail.com

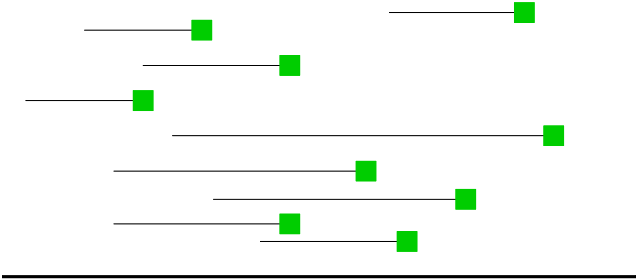
Outline

Motivation examples

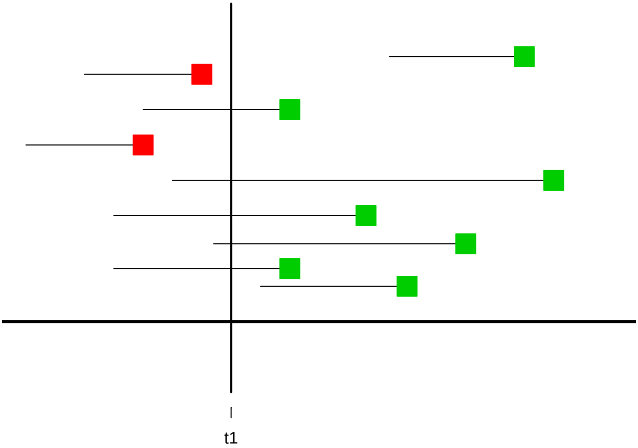
- Astronomy
- Epidemiology
- Economy
- Survival Analysis

In these cases, we must apply specialized statistical models and methods due the need to accommodate the event of losses in the sample, such as grouping, censoring or truncation.

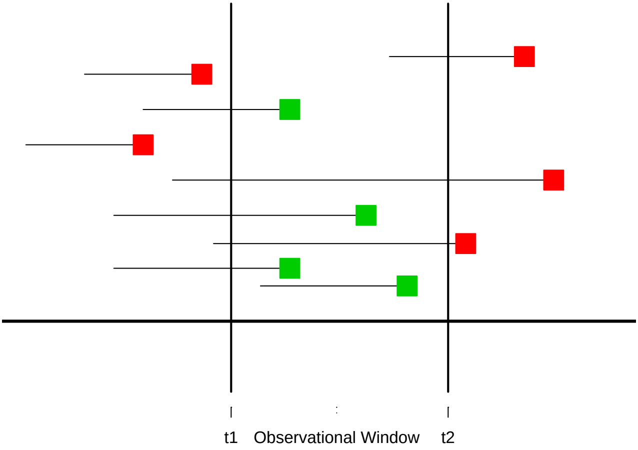
Truncation Scheme



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 estimators

The condicional $NPMLE$ 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.

DTDA-package

- `efron.petrosian(X,...)`
- `lynden(X,...)`
- `shen(X,...)`

DTDA-package

- `efron.petrosian(X,...)`
- `lynden(X,...)`
- `shen(X,...)`
- 3 examples data sets with $X \sim$ `Unif(0,1)`and:

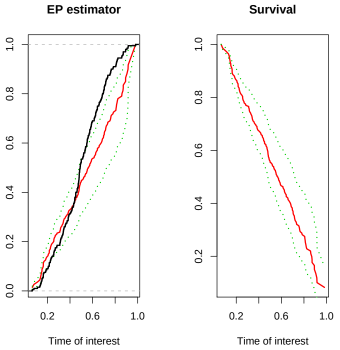
  Ex.1 $U \sim$ `Unif(0,0.5)`, $V \sim$ `Unif(0.5,1)`
  Ex.2 $U \sim$ `Unif(0,0.25)`, $V \sim$ `Unif(0.75,1)`
  Ex.3 $U \sim$ `Unif(0,0.67)`, $V \sim$ `Unif(0.33,1)`

efron.petrosian illustration under double truncation

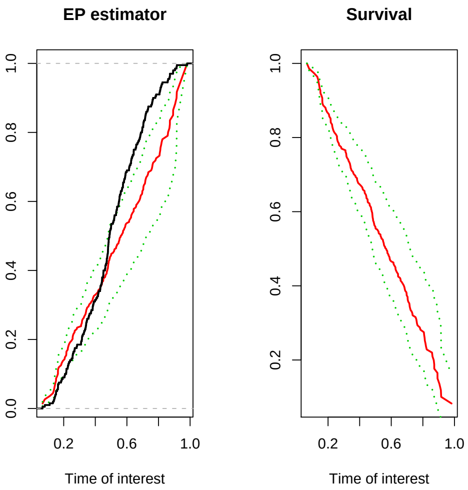
EX.1-50% of truncation

efron.petrosian(X,U,V,...)

>iter
>f
>FF
>S
>Sob
>upperF
>lowerF
>upperS
>lowerS



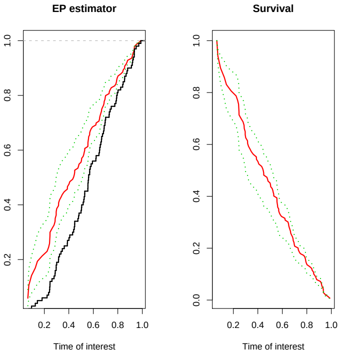
efron.petrosian illustration under double truncation



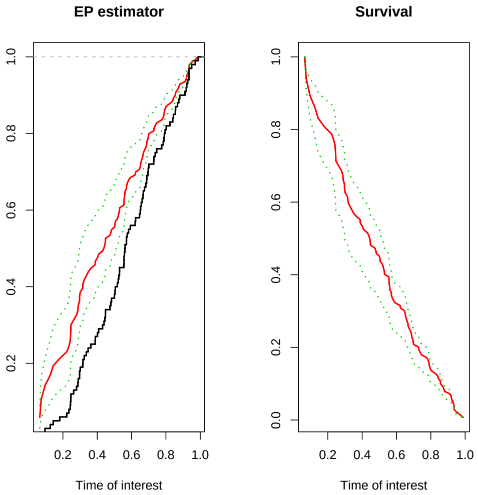
efron.petrosian illustration under left truncation

EX.1

```
efron.petrosian(X,U,...)
```



efron.petrosian illustration under left truncation

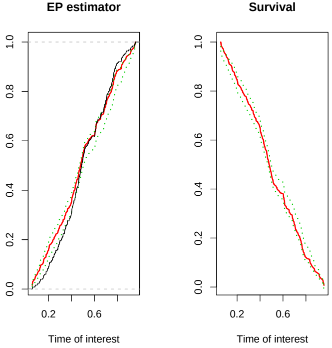


lynden illustration under double truncation

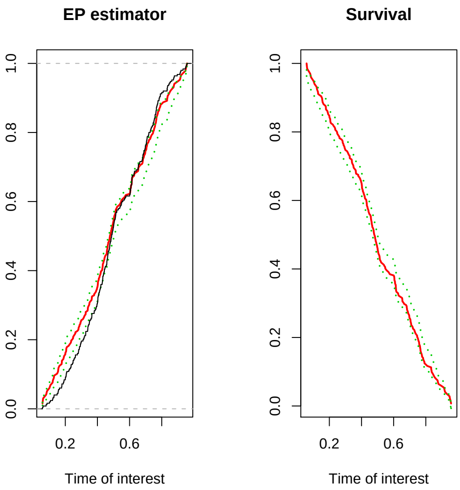
EX.2-25% of truncation

`lynden(X,U,V,...)`

- >iter
- >NJ
- >f
- >FF
- >h
- >S
- >Sob
- >upperF
- >lowerF
- >upperS
- >lowerS



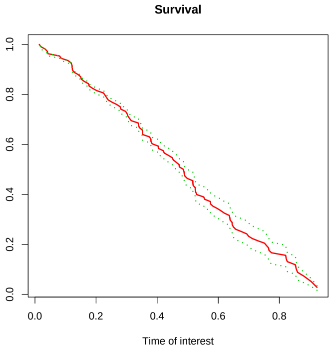
lynden illustration under double truncation



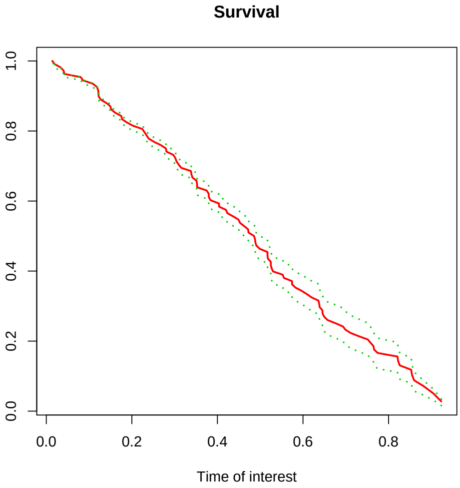
lynden illustration under right truncation

EX.2

```
lynden(X,V,...)
```



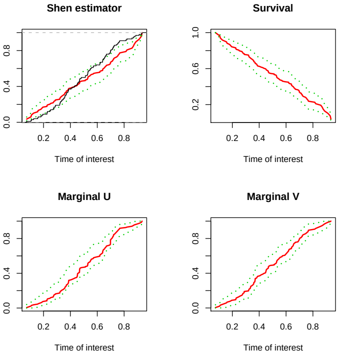
lynden illustration under right truncation



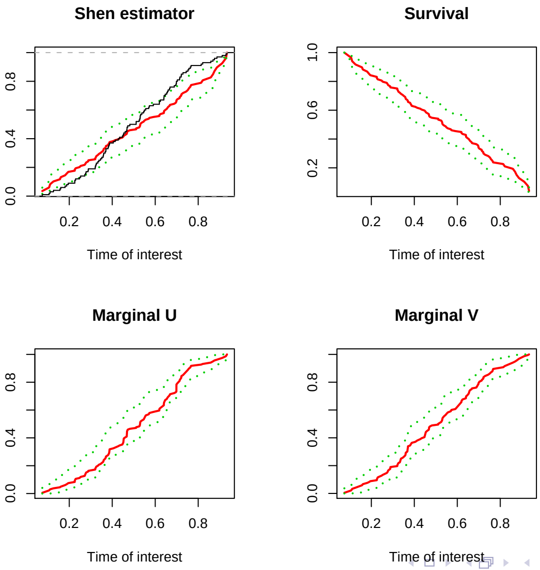
shen illustration under double truncation

EX.3-67% of truncation

```
shen(X,U,V...)
    >iter
    >f
    >FF
    >S
    >Sob
    >k
    >fU
    >fV
    >upperF
    >lowerF
    >upperS
    >lowerS
```



shen illustration under double truncation



Summary

- The DTDA package provides different algorithms for analyzing randomly truncated data, one-sided and two-sided (i.e. doubly) truncated data being allowed.

Summary

- The DTDA package provides different algorithms for analyzing randomly truncated data, one-sided and two-sided (i.e. doubly) truncated data being allowed.
- This package incorporates the functions efron.petrosian, lynden and shen, which call the iterative methods introduced by Efron and Petrosian (1999)and Shen (2008).

Summary

- The DTDA package provides different algorithms for analyzing randomly truncated data, one-sided and two-sided (i.e. doubly) truncated data being allowed.
- This package incorporates the functions efron.petrosian, lynden and shen, which call the iterative methods introduced by Efron and Petrosian (1999)and Shen (2008).
- Estimation of the lifetime and truncation times distributions is possible, together with the corresponding pointwise confidence limits based on the bootstrap.

Summary

- The DTDA package provides different algorithms for analyzing randomly truncated data, one-sided and two-sided (i.e. doubly) truncated data being allowed.
- This package incorporates the functions efron.petrosian, lynden and shen, which call the iterative methods introduced by Efron and Petrosian (1999)and Shen (2008).
- Estimation of the lifetime and truncation times distributions is possible, together with the corresponding pointwise confidence limits based on the bootstrap.
- Plots of cumulative distributions and survival functions are provided.

Summary

- The DTDA package provides different algorithms for analyzing randomly truncated data, one-sided and two-sided (i.e. doubly) truncated data.
- This package incorporates the functions efron.petrosian, lynden and shen, which call the iterative methods introduced by Efron and Petrosian (1999)and Shen (2008).
- Estimation of the lifetime and truncation times distributions is possible, together with the corresponding pointwise confidence limits based on the bootstrap.
- Plots of marginal cumulative distributions and survival functions are provided.
- There are no R packages with double truncation scheme.

Acknowledgments

- Work supported by the research Grant MTM2008-03129 and MTM2008-0310 of the Spanish Ministerio de Ciencia e Innovación
- Grant PGIDIT07PXIB300191PR of the Xunta de Galicia

References

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

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.(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.