Smoothing Survival Functions using an Empirical Saddlepoint Approximation Based Method with Doubly Censored Data

Manjari Dissanayake, Adao Alexandre Trindade
Texas Tech University

Abstract

A commonly used non-parametric procedure to estimate survival functions is through the Kaplan-Meier (KM) estimator. The distribution the KM estimator delivers is a discrete one with approximations to the distribution only at the observation times given. The proposed method is a non-parametric method to produce smooth KM survival functions using an empirical saddlepoint approximation. The resulting distribution is constructed by inverting the moment generating function (MGF) for the KM estimated discrete Simulation studies are conducted to demonstrate the performance of the method among competing parametric method and the semi-parametric spline-based method.

Simulation Study Design

Simulate interval censored data from 4 distributions

- Model 1: $Weibull(1,10)$
- Model 2: $Weibull(1/3,10)$
- Model 3: $Weibull(2,\sqrt{(2/3)})$
- Model 4: $LogLogistic(2,1)$

Sample size (n) = 10, 25, 50, 75, 100

Percentage of censoring (per_cen) = 10%, 30%, 60%

Use the three methods to obtain the PDF:

- Parametric (par)

The underlying distribution should be known.

Then use maximum likelihood estimation to find unknown parameters.

- Semi-parametric (spar)

Fit cubic splines for knots at $t_1, t_2, \dots, t_m$

- Empirical Saddlepoint Approximation (spa)

Compare the integrated squared errors between the true density and calculated PDF's.

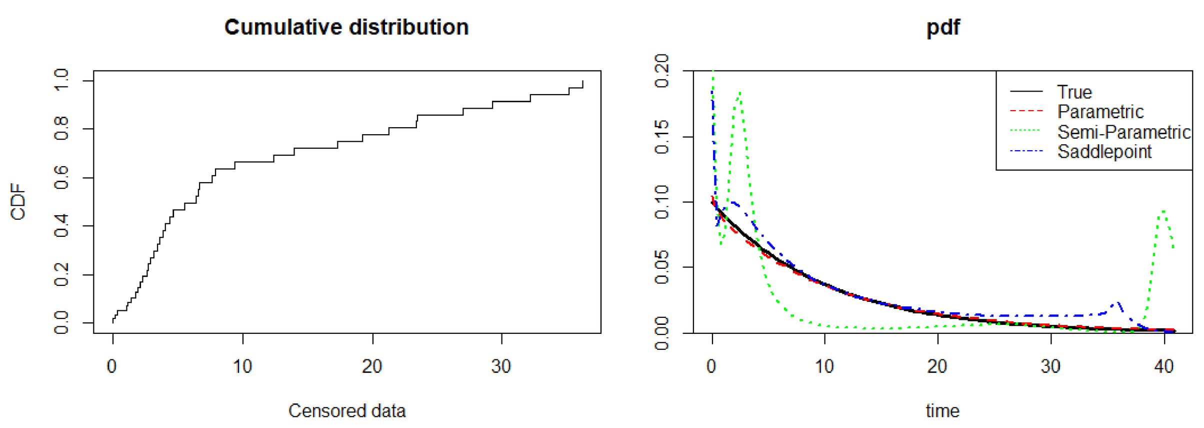


Figure 1: A discrete CDF (right) and PDF's obtained by above methods (left).

Introduction and Motivation

Data: Failure time data

- Times to certain events (failure or survival event).

Eg: Death, The failure of a mechanical component of a machine, or learning something.

- Occurrence of the event: failure.

The variable of interest : Survival variable.

Doubly-censored data (Interval-censored data)

- Failure time (T) is observed as an interval.

$T \in (L, R], \quad L \leq R$

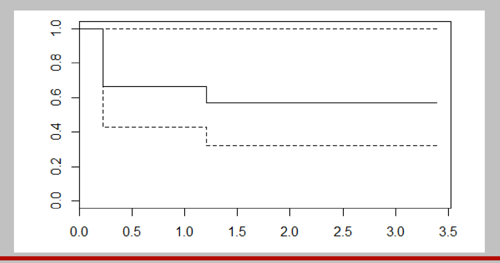
Eg: Times to breast retraction of 6 patients.

(45,], (25,37], (37,], (4,11], (17,25], (6,10]

Survival function (Not smooth)

- Defined as the probability that failure time T exceeds time t.

$S(t) = P(T > t), \quad 0 < t < \infty$



Can we get a good smooth probability function using empirical saddle point approximation method for the interval censored data?

Results

- Model 1: $Weibull(1,10)$

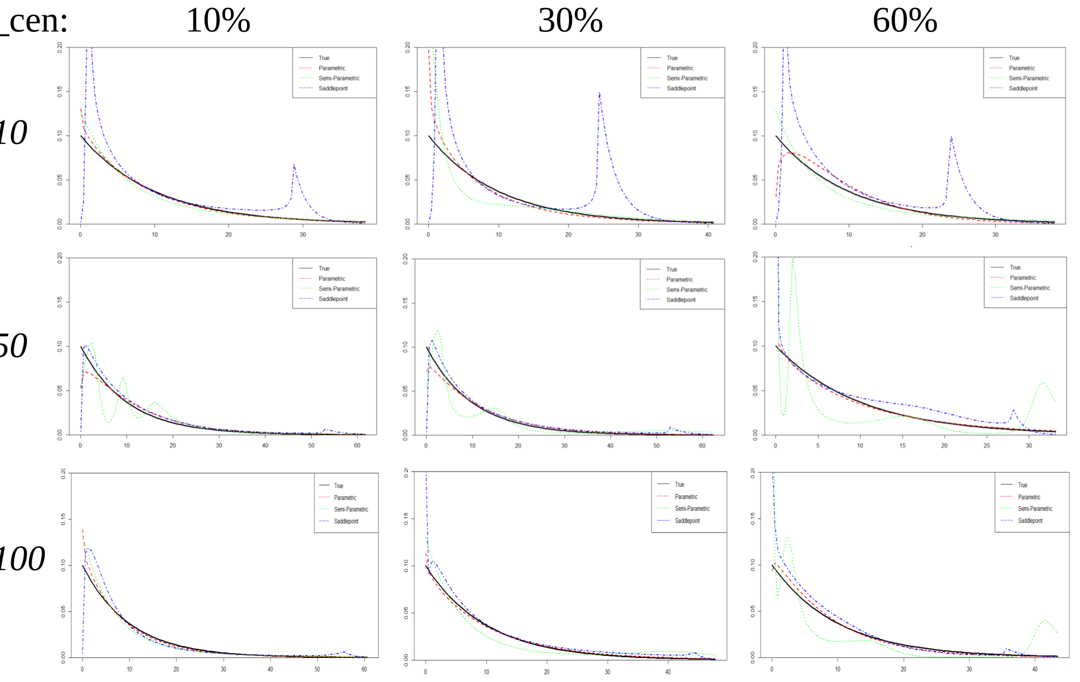


Figure 2: PDF's obtained for $Weibull(1,10)$ by above methods for *n=10, 50, 100* and per_cen 10, 30, and 60.

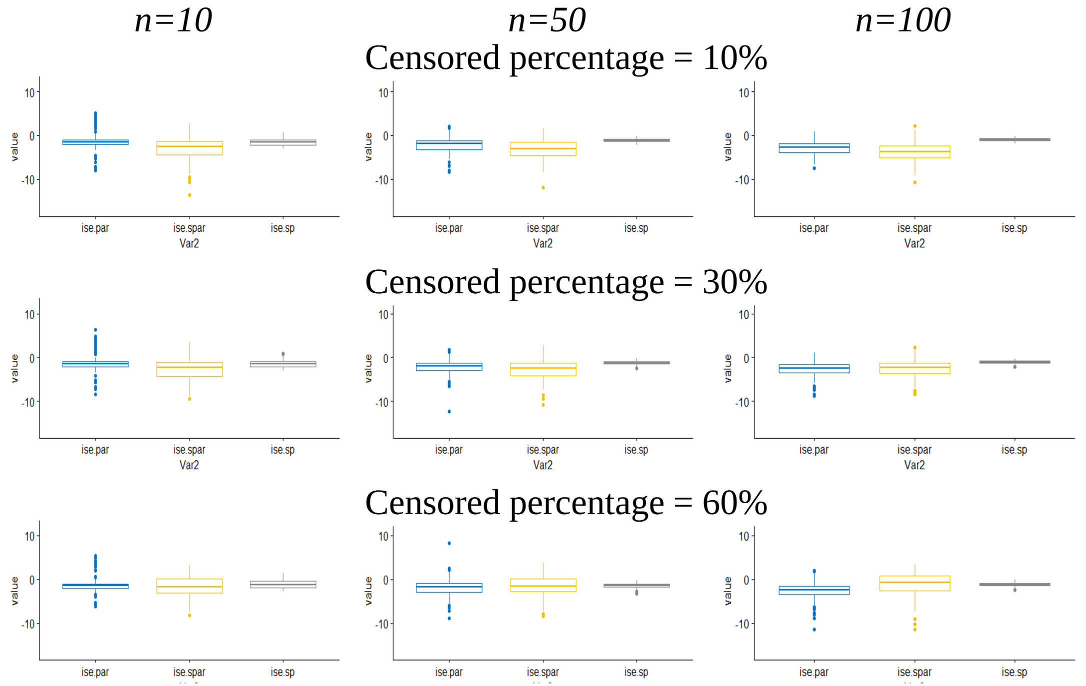


Figure 3: Boxplots of ISE obtained for $Weibull(1,10)$ by above methods for *n=10, 50, 100* and per_cen 10, 30, and 60.

- Model 4: $LogLogistic(2,1)$

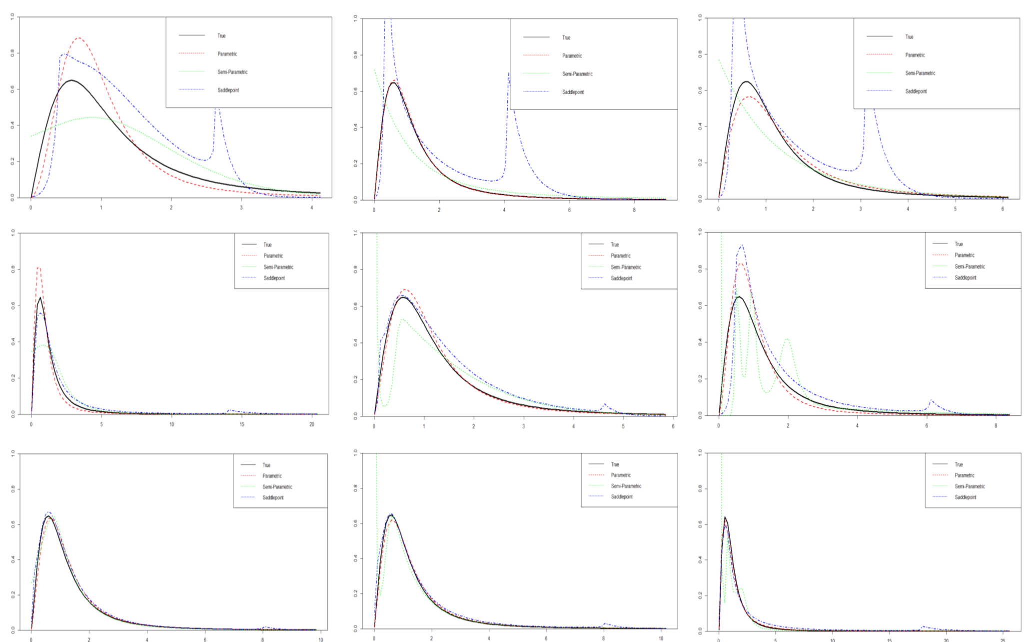


Figure 4: PDF's obtained for $LogLogistic(2,1)$ by above methods for *n=10, 50, 100* and per_cen 10, 30, 60.

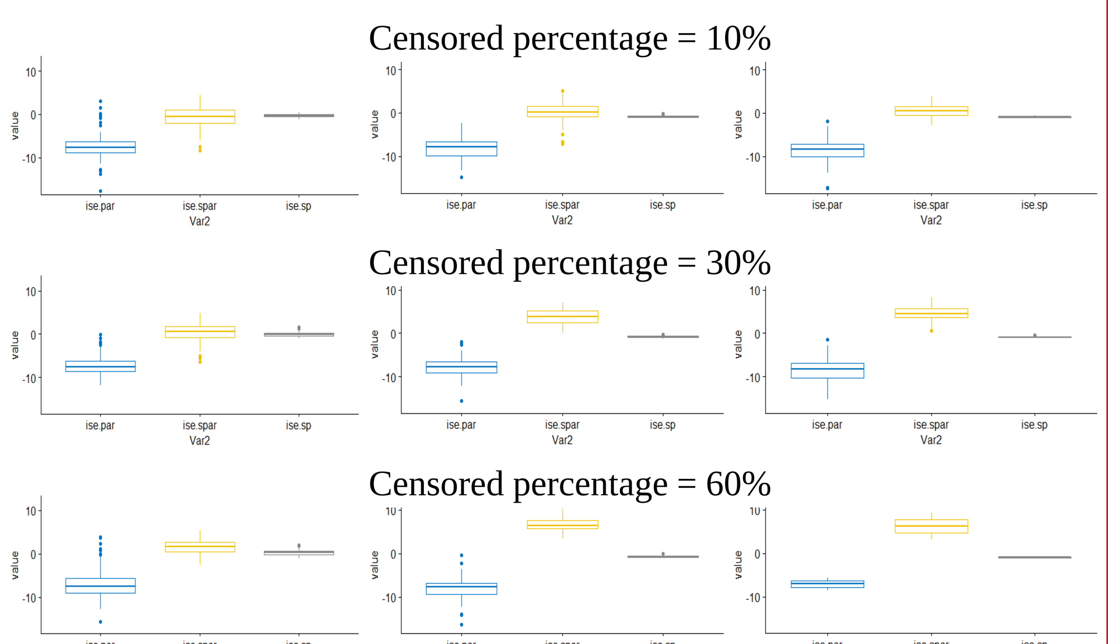


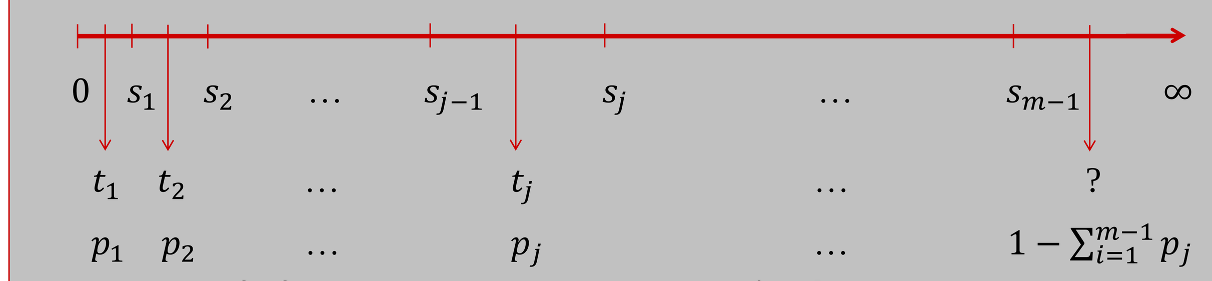
Figure 5: Boxplots of ISE obtained for $LogLogistic(2,1)$ by above methods for *n=10, 50, 100* and per_cen 10, 30, and 60.

Conclusions

- We do not need any previous assumptions, such as in parametric methods, but performs as good as parametric methods.
- We can see clearly that when sample size is small, the empirical saddlepoint based approximation works better in model fitting, than other methods. (The splines methods doesn't work for sample sizes less than 6)
- The method works really well in modeling the underlying density function when it has sharp turns. (Eg: $LogLogistic(2,1)$)

Proposed Method

- Arrange $\{0, L_i, R_i\}$ in ascending order → Name them $0, s_1, s_2, \dots, s_{m-1}$



- Use *survfit* function in R to give optimum $\hat{p}$ values using Non-parametric Maximum Likelihood Estimation (NPMLE).
- Probability Density Function (PDF): $\hat{f}(t) = \hat{f}(t_j) = \hat{p}_j$ when $t_j < t \leq t_{j+1}$
- The Cumulative Distribution Function (CDF): $\hat{F}(t) = \sum_{t_j \leq t} \hat{p}_j$
- Obtain the Moment Generating Function (MGF): $M(r) = \sum_{j=1}^{m} e^{rt_j} \hat{f}(t_j)$
- Obtain the Cumulant Generating Function (CGF): $K(r) = \ln M(r)$
- The saddlepoint CDF (Lugannani & Rice, 1980):

$$\hat{F}_S(t) = \begin{cases} \Phi(\hat{w}) + \phi(\hat{w})(\hat{w}^{-1} - \hat{u}^{-1}) & if\ r \neq 0 \\ \frac{1}{2} + \frac{K^{(3)}(0)}{6\sqrt{2\pi}K^{(2)}(0)^{\frac{3}{2}}} & if\ r = 0 \end{cases}$$

where $\hat{w} = sign(r)\sqrt{2[rt - K(r)]}$ & $\hat{u} = r\sqrt{K^{(2)}(r)}$

- Then the saddlepoint PDF:

$$\hat{f}_S(t) = \left[\frac{1}{2\pi K^{(2)}(r)}\right]^{1/2} \exp\{K(r) - rt\}$$

where $K^{(1)}(r) = t$

Future Work

- Integrated squared error comparisons for CDF's.
- Using the empirical saddlepoint based method there sometimes is a spike in the pdf obtained. Need to introduce a correction term to the model.
- The penalty term in splines method needs the number of uncensored observations. We need to come-up with a method to treat data with censoring percentage 100%.

References

1. Kaplan EL and Meier P (1958). "Nonparametric estimation from incomplete observations", Journal of American Statistical Association, vol. 53, no. 282, pp. 457–481.

2. Klein JP and Moeschberger ML (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd edition, Springer-Verlag.

3. Sun J (2006). *The Statistical Analysis of Interval-censored Failure Time Data*, Springer.