PERGAMON

Computers & Operations Research 29 (2002) 423–431

computers & operations research

www.elsevier.com/locate/dsw

Computing the nonparametric estimator of the survivor function when all observations are either left- or right-censored with tied observation times

Lawrence M. Leemis

Department of Mathematics, The College of William & Mary, P.O. Box 8795, Williamsburg, VA 23187-8795, USA

Abstract

A data set of missiles tested at various times consists entirely of left- and right-censored observations. We present an algorithm for computing the nonparametric maximum likelihood estimator of the survivor function. When there are a significant number of tied observations, the algorithm saves significant computation time over a direct implementation of the survivor function estimate given in Andersen and Rønn, (Biometrics 1995; 51:323–9).

Scope and purpose

The algorithm presented here is of use to a modeler interested in computing the nonparametric maximum likelihood estimator of the survivor function for a data set that consists solely of left- and right-censored observations, which also contains tied observation times. This estimator is of use to a modeler in analyzing a data set of nonnegative response times, as in the case of a reliability engineer modeling component survival times or a biostatistician modeling patient survival times.

1. Introduction

The Kaplan–Meier product-limit estimator is commonly used for survivor function estimation for a data set containing right-censored observations [1]. We consider the problem of estimating the survivor function for a data set consisting solely of left- and right-censored observations. Computing the nonparametric maximum likelihood estimate (NPMLE) of the survivor function in this case is not trivial, although no iterative methods are required. This paper gives an algorithm

E-mail address: leemis@math.wm.edu (L.M. Leemis).

PII: S0305-0548(00)00077-0

for computing the NPMLE of the survivor function $S(t)$ using an estimator given in Andersen and Rønn [2]. The algorithm is applied to a data set of missile tests.

Following Andersen and Rønn's notation, let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be the vector of i.i.d. failure times with survivor function S and let $\mathbf{T} = (T_1, T_2, \dots, T_n)$ be the vector of observation times, independent of $\mathbf{X}$. The i.i.d. pairs (T_i, D_i) are observed for $i = 1, 2, \dots, n$, where the indicator function $D_i = I(X_i \leqslant T_i)$ is one when the observation is left-censored (i.e., $X_i \leqslant T_i$), and zero when the observation is right-censored (i.e., $X_i > T_i$).

One could envision several scenarios where an algorithm of this type might be used. Two such scenarios follow.

- An educational psychologist is interested in the age at which a child is able to master a particular task. The psychologist tests n children and records the age of child i, T_i, and whether or not child i was able to master the task D_i. All observations are either left censored ($D_i = 1$ if the child successfully completes the task) or right censored ($D_i = 0$ if the child fails to complete the task).
- An automobile manufacturer may routinely require their dealerships to inspect a particular part (e.g., the muffler) on every vehicle that enters their service facility. The dealership inspects n vehicles and records the odometer reading T_i, along with whether the muffler is defective ($D_i = 1$) or passes inspection ($D_i = 0$).

Andersen and Rønn [2] indicate that the NPMLE for the survivor function is the solution to an isotonic regression problem. Assuming that the observation times are ordered so that $T_1 \leqslant T_2 \leqslant \cdots \leqslant T_n$, define

$$H_i = \sum_{j=1}^{i} D_j$$

for $i = 0, 1, \dots, n$, and consider the step function associated with plotting i vs. H_i. The greatest convex minorant is the piecewise linear function that falls below the step function and connects the bottoms of the steps under the constraint that the greatest convex minorant must be a convex function. $1 - \hat{S}(T_i)$ is the left-continuous derivative at i of the piecewise-linear function, and $\hat{S}(t) = \hat{S}(T_i)$ for $t \in (T_i, T_{i+1})$, where $T_0 = 0$ and $T_{n+1} = +\infty$.

Some additional notation will be helpful when developing an algorithm for computing the NPMLE of the survivor function when ties in the observation times are present. Let k be the number of distinct observation times. For $j = 1, 2, \dots, k$, let Y_j be the observation time, n_j be the number of items tested at time Y_j, and L_j be the number of left-censored observations at time Y_j. It is assumed that the left-censored observations are placed before the right-censored observations in computing the greatest convex minorant for reasons discussed in Section 2.

Table 1 contains a left- and right-censored data set consisting of $n = 2534$ missiles tested at $k = 55$ distinct observation times ranging from $Y_1 = 2$ to $Y_{55} = 60$ months. Each missile tested either fails the test, contributing a left-censored observation, or passes the test, contributing a right-censored observation. A total of $\sum_{j=1}^{55} L_j = 171$ missiles failed the test, each contributing a left-censored observation.

The plot of i vs. H_i in Fig. 1 for the missile data shows that the convex minorant has seven pieces, and hence there will be seven discontinuous drops in the survivor function estimate. The survivor

Table 1
Missile test data

j	Y_j	n_j	L_j	j	Y_j	n_j	L_j
1	2	1	0	29	33	64	9
2	4	5	0	30	34	38	10
3	6	14	1	31	35	77	6
4	7	6	0	32	36	81	5
5	8	3	0	33	37	67	6
6	10	12	2	34	38	26	6
7	11	29	1	35	39	81	4
8	12	18	1	36	40	61	4
9	13	41	2	37	41	65	9
10	14	10	0	38	42	57	6
11	15	6	0	39	43	93	1
12	16	6	0	40	44	35	1
13	17	21	0	41	45	36	4
14	18	27	5	42	46	44	5
15	19	29	2	43	47	12	0
16	20	15	3	44	48	20	4
17	21	33	1	45	49	15	3
18	22	24	3	46	50	43	6
19	23	36	0	47	51	31	1
20	24	95	8	48	52	10	3
21	25	80	6	49	53	16	1
22	26	22	13	50	54	10	1
23	27	80	5	51	55	12	1
24	28	77	3	52	56	1	0
25	29	13	5	53	57	2	1
26	30	28	6	54	58	4	0
27	31	88	6	55	60	3	0
28	32	11	1				

function estimate, including the risers associated with the seven steps, is shown in Fig. 2, and the values of $\hat{S}(t)$ are given in Table 2. The estimate is not plotted beyond the last downward step.

2. Computing $\hat{S}(t)$

Tied observation times imply that not all of the bottoms of the steps need to be considered when computing the greatest convex minorant. The assumption that left-censored observations be placed before right censored observations when computing the greatest convex minorant impacts the NPMLE of $S(t)$. As shown in Fig. 3 for $n_j = 5$ and $L_j = 3$, the convex piecewise linear function between the two steps differs if the left-censored values are placed first (e.g., the figure on the left) or last (e.g., the figure on the right). The solid dots represent the i vs. H_i values and the greatest convex minorant is shown by the dashed line connecting the bottoms of the steps. Thus only the last step associated with a time value Y_j where $L_j > 0$ need be considered in the plot of i vs. H_i.

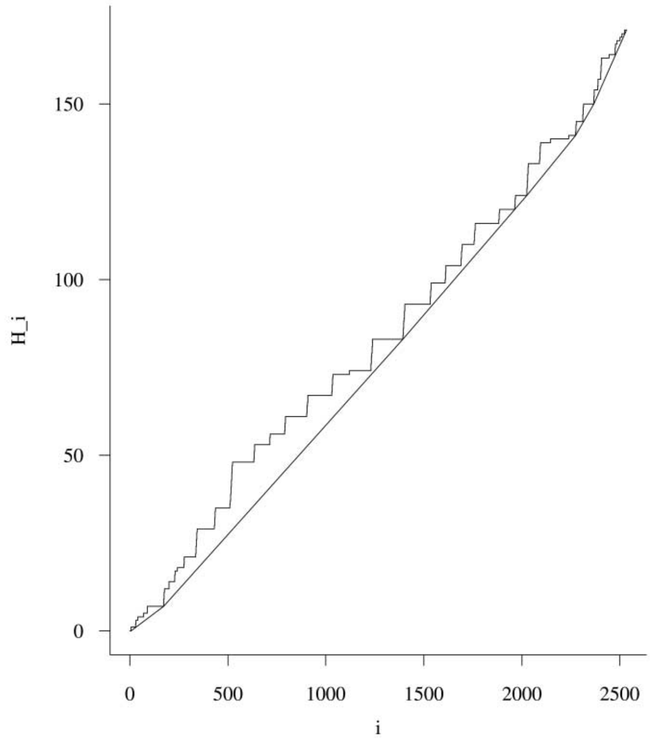

Fig. 1. The step function and greatest convex minorant for the missile data.

Table 2
Survivor function estimates

Interval	$\hat{S}(t)$
[0, 17)	1.0000
[17, 33)	0.9578
[33, 39)	0.9378
[39, 40)	0.9351
[40, 44)	0.9344
[44, 47)	0.9320
[47, 60)	0.9022
[60, ∞)	0.8743

Furthermore, if $N_1, N_2, \ldots, N_k$ are the cumulative number of items on test and $M_1, M_2, \ldots, M_k$ are the cumulative number of left censorings, then the potential corner points of the greatest convex minorant occur at the points (N_j, M_j), for $j = 1, 2, \ldots, k$. Special accommodations must be made for $j = 1$.

The algorithm is given in Appendix A and an Splus implementation is given in Appendix B, which is available from the author. Phase 1 of the algorithm consists of computing the cumulative

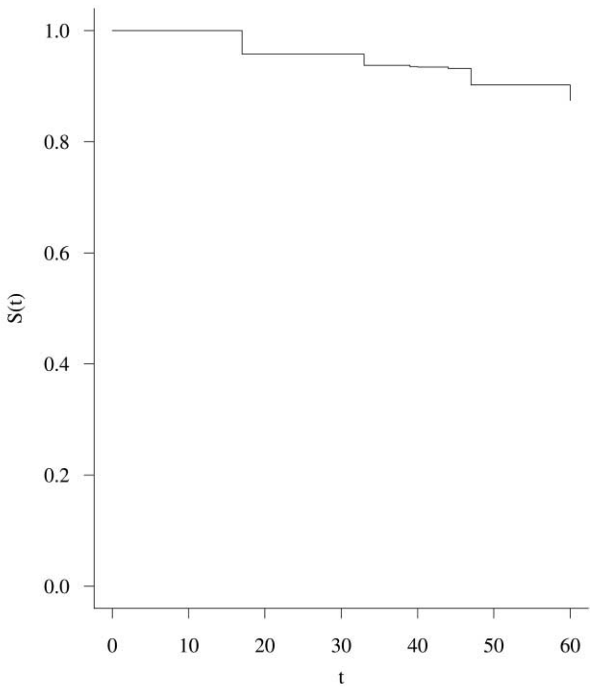

Fig. 2. NPMLE of $S(t)$ for the missile data.

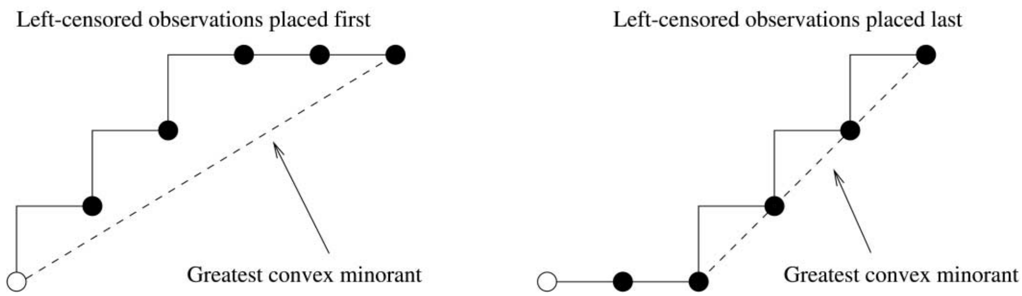

Fig. 3. The effect of the placement of left-censored data values on the greatest convex minorant. The solid points denote five $(i,\ H_i)$ pairs associated with an observation time with ties.

number of items on test and the cumulative number of left censorings. In order for the algorithm to handle the case where $L_1 > 0$, M_0 and N_0 are set to zero. In order for the algorithm to handle the case where $L_k > 0$, L_{k+1} is set to zero. Phase 2 of the algorithm computes the greatest convex minorant by checking for the minimal slope between potential corner points. The j loop index is either incremented by one (if $L_j = 0$) or incremented to the next index associated with the corner points of the greatest convex minorant (if $L_j > 0$). The index r is incremented every time the survivor function estimate takes a step downward. Since no slope can exceed one, the variable *MinSlope* is initially set to 1.01, and replaced in the search for the smallest slope. The estimated survivor function values are stored in $s_1, s_2, \ldots$, and the associated drop times in the estimated survivor function are stored in $d_1, d_2, \ldots$. Considering the computational complexity, Phase 1 of the algorithm is $O(k)$, whereas Phase 2 is order $O(k^2)$. The time savings occurs over the naive implementation of the estimator when there are numerous ties, as was the case with the missile data.

Acknowledgements

The author gratefully acknowledges the assistance of John Backes and Mike Crawford from the Yorktown Naval Weapons Station, and Per Kragh Andersen from the University of Copenhagen for helpful suggestions.

Appendix A. Algorithm

Procedure NPMLE: Compute the nonparametric survivor function estimate for a data set consisting entirely of left- and right-censored observations.

Input: For $j = 1, 2, \ldots, k$,
j: index associated with the observation times.
Y_j: observation time.
n_j: number tested at time Y_j.
L_j: number of left-censored observations at time Y_j.
Output: The nonparametric estimate of the survivor function of the time to failure given by the survivor function values $s_1, s_2, \ldots$ and corresponding drop times $d_1, d_2, \ldots$.
[*Phase* 1: Compute the cumulative number of items on test $N_1, N_2, \ldots, N_k$ and the cumulative number of left censorings $M_1, M_2, \ldots, M_k$. The artificial values $M_0 = 0$ and $N_0 = 0$ are inserted to allow the algorithm to correctly handle the boundary condition $L_1 > 0$. The artificial value $L_{k+1} = 0$ is inserted to allow the algorithm to correctly handle the boundary condition $L_k > 0$.]

$N_0 \leftarrow 0$
$M_0 \leftarrow 0$
for $j \leftarrow 1$ to k
  $N_j \leftarrow N_{j-1} + n_j$
  $M_j \leftarrow M_{j-1} + L_j$
endfor
$L_{k+1} \leftarrow 0$

[*Phase* 2: Compute the greatest convex minorant and associated survivor function estimate. Let **s** be a vector of the complement of the slopes of the greatest convex minorant, i.e., the **s** vector holds the survivor function estimates. Let the vector **d** hold the survivor function drop times. The index r is incremented upon each drop in the survivor function. Move from corner to corner computing slopes.]

$r \leftarrow 1$
$j \leftarrow 1$

while $\sum_{q=j}^{k+1} L_q > 0$
   if $L_j > 0$ then
      $MinSlope \leftarrow 1.01$
      for $i \leftarrow j + 1$ to $k + 1$
         if $N_{i-1} - N_{j-1} > 0$ then
            $Slope \leftarrow (M_{i-1} - M_{j-1})/(N_{i-1} - N_{j-1})$
            if $Slope \leqslant MinSlope$
               $MinSlope \leftarrow Slope$
               $p \leftarrow i$
            endif
         endif
      endfor
      $s_r \leftarrow 1 - MinSlope$
      $d_r \leftarrow Y_{p-1}$
      $r \leftarrow r + 1$
      $j \leftarrow p$
   else
      $j \leftarrow j + 1$
   endif
endwhile

Appendix B. Splus code

```
########################################
#
# Phase 1: read in the data in the form Y_i, n_i, L_i
#          compute cumulative sums
#
########################################

data ← matrix(scan("left1.d"), byrow = T, ncol = 3)
times ← data[ ,1]
numleft ← data[ ,3]
k ← length(numleft)
n ← cumsum(data[ ,2])
m ← cumsum(numleft)
numleft[k + 1] ← 0
```

```
#####################################################
#
# Phase 2: compute greatest convex minorant and survivor function
#
#####################################################


s ← rep(0, k)
d ← rep(0, k)
j ← 1
r ← 1
while (sum(numleft[j:(k + 1)]) > 0) {
  if (numleft[j] > 0) {
    minslope ← 1.01
    for (i in (j + 1):(k + 1)) {
      if (j > 1) {
        if (n[i - 1] - n[j - 1] > 0) {
          slope ← (m[i - 1] - m[j - 1]) / (n[i - 1] - n[j - 1])
        }
      }
      else
        {
        if (n[i - 1] > 0) {
          slope ← m[i - 1] / n[i - 1]
        }
      }
      if (slope <= minslope) {
        minslope ← slope
        p ← i
      }
    }
    s[r] ← 1 - minslope
    d[r] ← times[p - 1]
    r ← r + 1
    j ← p
  }
  else
  {
    j ← j + 1
  }
}
```

```
#########################################
#
# Phase 3: plot the survivor function estimator
#
#########################################

postscript(file = "left.ps", width = 6.1, height = 7.1, horizontal = F)
par(mai = c(1.0, 1.0, 0.4, 0.4))
plot(c(0, d[1]), c(1, 1), xlab = "t", ylab = "S(t)", bty = "l",
      xlim = c(0, max(times)), ylim = c(0, 1),
      sub = "", las = c(1), type = "l", font = 3)
segments(d[1], 1, d[1], s[1])
if (r > 2) {
  for (jj in 2:(r - 1)) {
     segments(d[jj - 1], s[jj - 1], d[jj], s[jj - 1])
     segments(d[jj], s[jj - 1], d[jj], s[jj])
  }
}
dev.off() # this will shut down the postscript device
```

References

[1] Andersen PK, Rønn BB. A nonparametric test for comparing two samples where all observations are either left- or right-censored. Biometrics 1995;51:323–9.

[2] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958;53:457–81.

Lawrence Leemis is professor and chair of the Mathematics Department at The College of William & Mary. He received his BS and MS degrees in Mathematics and his Ph.D. in Industrial Engineering from Purdue University. His research and teaching interests are in reliability and simulation. He has published articles in *The American Statistician*, *Communications in Statistics — Simulation and Computation*, *Computational Statistics and Data Analysis*, *IEEE Transactions on Reliability*, *IIE Transactions*, *INFORMS Journal on Computing*, *Interfaces*, *Journal of Quality Technology*, *Journal of Statistical Computation and Simulation*, *Management Science*, *Operations Research*, *Statistics and Probability Letters*, and *Technometrics*. He wrote the textbook *Reliability: Probabilistic Models and Statistical Methods*, published by Prentice-Hall.