Constrained estimation for binary and survival data

Jeremy M. G. Taylor
Yong Seok Park
John D. Kalbfleisch
Biostatistics, University of Michigan

May, 2010

Outline

- Motivation
- Two Binomial probabilities, $p_1 \leq p_2$
- Survival functions, $S_1(t) \geq S_2(t)$

Challenges

- What estimator to use?
- General Approaches
  - Restricted MLE
    - Isotonic regression
    - Pooled adjacent violators algorithm
  - Bayesian: Impose restriction through prior distribution
- Inference: Confidence intervals

Motivation

- New cancer treatment. Drug 3 levels, $d_1 < d_2 < d_3$
- Possible toxic side effects
  - $p_j = P(Toxicity|d_j)$
  - Know $p_1 \leq p_2 \leq p_3$
  - Utilize this information in the analysis
- Data
  - $Y_1 \sim Binomial(n_1, p_1)$
  - $Y_2 \sim Binomial(n_2, p_2)$
  - $Y_3 \sim Binomial(n_3, p_3)$
- Want $\hat{p}_1 \leq \hat{p}_2 \leq \hat{p}_3$
- Why
  - Gain efficiency, e.g. $n_1 = 15, n_2 = 3, n_3 = 14$
  - Consistent with truth

Restricted MLE for two binomial probabilities

- $Y_j \sim Binomial(n_j, p_j)$
- $p_1 \leq p_2$
- restricted MLE is given by
  - $\hat{p}_{1n} = \min\{d_1/n_1, (d_1 + d_2)/(n_1 + n_2)\}$
  - $\hat{p}_{2n} = \max\{d_2/n_2, (d_1 + d_2)/(n_1 + n_2)\}$.
- Construction of confidence intervals is difficult if p_1 is close to p_2
- Inference is difficult near or on boundary of parameter space

Simulation results: Biases and Efficiency

Table: Restricted MLE and the unrestricted MLE ($n_1 = 50, n_2 = 100$).

	Restricted MLE	
	p_1	p_2
	bias	
$p_1 = 0.5, p_2 = 0.5$	-0.024	0.010
$p_1 = 0.5, p_2 = 0.52$	-0.017	0.009
$p_1 = 0.5, p_2 = 0.7$	0.001	0.001
$p_1 = 0.5, p_2 = 0.9$	-0.001	-0.001
Efficiency: Var(Restricted)/Var(Unrestricted)		
$p_1 = 0.5, p_2 = 0.5$	0.562	0.784
$p_1 = 0.5, p_2 = 0.52$	0.620	0.818
$p_1 = 0.5, p_2 = 0.7$	0.993	0.996
$p_1 = 0.5, p_2 = 0.9$	1	1

- **Theorem 0: CLT.** Suppose that $p_1 < p_2$. Then $\sqrt{n_j}(\hat{p}_{jn} - p_j) \to_d N(0, p_j(1-p_j))$.
- **Theorem 1.** Suppose that $p_1 = p_2$, $\lim_{n\to\infty} n_2/n_1 = c$, and $0 < c < \infty$. Then

$$\sqrt{n_1}(\hat{p}_{1n} - p_1) \to_d \min\left\{W_1, \frac{1}{1+c}W_1 + \frac{\sqrt{c}}{1+c}W_2\right\},$$

  and

$$\sqrt{n_2}(\hat{p}_{2n} - p_2) \to_d \max\left\{W_2, \frac{\sqrt{c}}{1+c}W_1 + \frac{c}{1+c}W_2\right\},$$

  as $n \to \infty$, where W_1 and W_2 are independent and identically distributed as $N(0, p_1(1-p_1))$.
- Asymptotic results not useful or accurate for small n

- **Theorem 2.** Suppose that $p_2 = p_1 + \Delta/\sqrt{n_1}$, $\lim_{n\to\infty} n_2/n_1 = c$, and $0 < c < \infty$. We have, when $p_1 = p_2$,

$$\sqrt{n_1}(\hat{p}_{1n} - p_1) \to_d \min\left(W_1, \frac{1}{1+c}W_1 + \frac{\sqrt{c}}{1+c}W_2 + \frac{c}{1+c}\Delta\right),$$

  and

$$\sqrt{n_2}(\hat{p}_{2n} - p_2) \to_d \max\left(W_2, \frac{\sqrt{c}}{1+c}W_1 + \frac{c}{1+c}W_2 - \frac{\sqrt{c}}{1+c}\Delta\right),$$

  as $n \to \infty$, where W_1 and W_2 are independent with distribution $N(0, p_1(1-p_1))$.
- Confidence intervals don't have good coverage rates

Bootstrap Confidence Intervals

- Group 1, n_1 observations, (0,1,1,0,1,.....,0)
- Group 2, n_2 observations, (1,1,0,0,0,.....,1)
- Resample within groups
- Bootstrap percentile confidence intervals
    - Coverage rates good at moderate sample sizes
    - Coverage rates OK at small sample sizes

Table: Simulation: Coverage rates of 95% confidence intervals

$n_1 = 50$, $n_2 = 100$					
$p_1 = 0.5$		$p_2 = 0.5$	$p_2 = 0.52$	$p_2 = 0.7$	$p_2 = 0.9$
Restricted MLE	p_1	0.94	0.93	0.90	0.93
Theorem 2	p_2	0.94	0.94	0.96	1.00
percentile bootstrap CI	p_1	0.94	0.95	0.96	0.96
based on restricted MLE	p_2	0.95	0.95	0.96	0.96
$p_1 = 0.8$		$p_2 = 0.8$	$p_2 = 0.82$	$p_2 = 0.85$	$p_2 = 0.9$
Restricted MLE	p_1	0.96	0.92	0.88	0.86
Theorem 2	p_2	0.95	0.96	0.96	0.97
percentile bootstrap CI	p_1	0.95	0.96	0.97	0.95
based on restricted MLE	p_2	0.94	0.94	0.95	0.95

Estimation of Survival Functions

Stochastic Ordering

Survival Function

$S(t) = Pr(T > t)$

Definition of Stochastic Ordering:

$T_1 \leq_{st} T_2$ if $Pr(T_1 > t) \leq Pr(T_2 > t)$ for $t \in R$

One-sample Case: Estimation of $S_1(t)$

- Bounded Below: $S_1(t) \geq S_2(t)$, where $S_2(t)$ is known;
- Bounded Above: $S_1(t) \leq S_2(t)$, where $S_2(t)$ is known.

Two-sample Case:

- $S_1(t) \geq S_2(t)$, $S_1(t)$ and $S_2(t)$ are unknown.

Motivation - Survival Analysis in Cancer Study

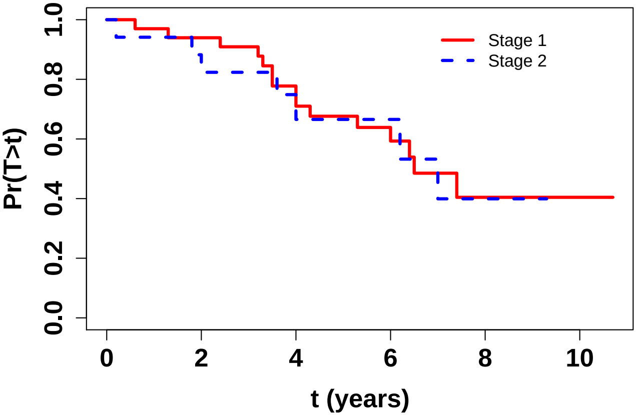


Figure: Kaplan-Meier plots of larynx cancer patients(Kardaun, 1983)

Motivation - Constrained Estimator

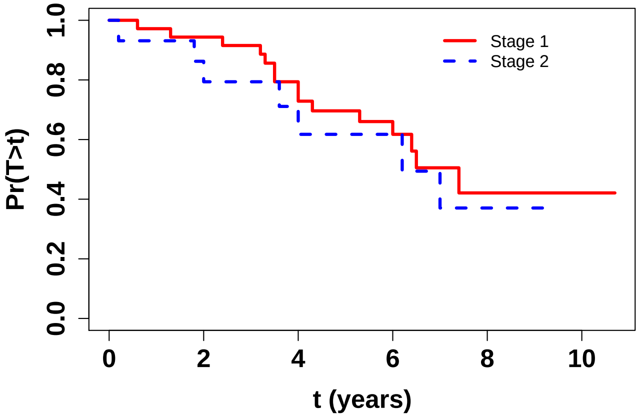


Figure: Constrained NPMLE

Motivation - Cont.

Wide Range of Applications.

- biomedical research;
- engineering sciences;
- economics;
- software reliability.

Estimators from separate samples may not satisfy constraint

- random variation;
- small sample size;

Constrained Estimator

- Potential to gain efficiency
- Realistic estimate

Literature

C-NPMLE: Two-sample case without censoring.

Brunk et al. (1966).

C-NPMLE: One- & two-sample with right censoring.

Dykstra (1982) - $\langle$ Correct in Bounded Below Case $\rangle$
Some possible outcomes were not properly handled.

May not be C-NPMLE in bounded above and two-sample case.

Alternative: One-sample case.

Puri and Singh (1992); Rojo and Ma (1996).

Alternative: Two-sample case.

Lo (1987) - **swapping estimates if violated**;
Rojo (2004) - **averaging estimates if violated**;
Park et al (2010) - **pointwise constrained MLE**.

One sample, No constraints

NPMLE: Kaplan-Meier estimator.

Product limit estimator.

Distribution is discrete. Jumps at the event times.

- $h_i = \log[S(t_i)/S(t_{i-1})]$
- Discrete hazard $= 1 - \exp(h_i)$
- $S(t_i) = \exp[\sum_{j=1}^{i} h_j]$
- d_i = number of events at time t_i
- n_i = number of at risk at time t_i

The NPMLE of $S(\cdot)$ is given by

$$\hat{h}_i = \begin{cases} \log(1 - \frac{d_i}{n_i}) & d_i > 0 \\ 0 & d_i = 0 \end{cases}$$

One-sample Bounded Above

Problem

Data

(Y_{1i}, Δ_{1i}), $i = 1, \cdots, n$;
$\Delta_{1i} = 1$ if event occurred or
$\Delta_{1i} = 0$ if right censored

Goal

Estimate $S_1(t)$ under $S_1(t) \leq S_2(t)$.

Likelihood

$L = \prod_{i=1}^{n} [S_1(Y_{1i}-) - S_1(Y_{1i})]^{\Delta_{1i}} S_1(Y_{1i})^{1-\Delta_{1i}}$

Discrete Case:
$L = \prod_{j=1}^{m} [S_1(a_{j-1}) - S_1(a_j)]^{d_{1i}} S_1(a_j)^{c_{1i}}$

Definitions

C-NPMLE

Constrained Nonparametric MLE: nonparametric estimator that maximizes the likelihood amongst those that satisfy the constraint. C-NPMLE may not be unique.

MC-NPMLE

Maximum C-NPMLE, which is C-NPMLE that maximizes the estimate of the survivor function in the class of all C-NPMLE.

Theorem: Bounded Above

For $S_1(\cdot)$ and $S_2(\cdot)$ discrete the MC-NPMLE of $S_1(\cdot)$ is given by

$$\hat{h}_{1i} = \begin{cases} \log(1 - \dfrac{d_{1i}}{n_{1i} - \hat{k}^i}) & d_{1i} > 0 \\ \min\left[0, \ \displaystyle\sum_{j=1}^{i} h_{2j} - \sum_{j=1}^{i-1} \hat{h}_{1j}\right] & d_{1i} = 0 \end{cases}$$

and $\hat{k}^i = \min_{a \le i} \max_{b \ge i} \min(K^-(a,b), n_{1b})$, where
(Dykstra 1982: $\hat{k}^i = \min_{a \le i} \max_{b \ge i} K^-(a,b)$)
$K^-(a,b) = \max\{0, -K(a,b)\}$ and $K^-(a,b)$ is the solution of
$\sum_{j=a}^{b} \log(1 - \frac{d_{1j}}{n_{1j}+k}) - \sum_{j=a}^{b} h_{2j} = 0$.

Algorithm: Bounded Above

1. Set $i_0 = 0$, $\ell = 1$, $m' = \max(i : n_{1i} > 0)$.
2. Let $i_\ell = \min_{b > i_{\ell-1}}\{b : H(i_{\ell-1}+1, b, 0) > 0\}$. If no such i_ℓ exists, set $i_\ell = m'$ and $k_\ell = 0$ and go to step 6, otherwise go to step 3.
3. If $d_{1i_\ell} = 0$ and $H(i_{\ell-1}+1, i_\ell, -n_{1i_\ell}) \geq 0$, then set $k_\ell = n_{1i_\ell}$ and go to step 5, otherwise set $k_\ell = -K(i_{\ell-1}+1, i_\ell)$ and go to step 4.
4. Let $I = \min_{b > i_\ell}\{b : n_{1b} > k_\ell \text{ and } H(i_\ell + 1, b, -k_\ell) > 0\}$. If no such I exists, then go to step 5. Otherwise, set $i_\ell = I$ and go to step 3.
5. Let $\hat{h}_{1j} = \log[1 - d_{1j}/(n_{1j} - k_\ell)], i_{\ell-1} + 1 \leq j \leq i_\ell - 1$

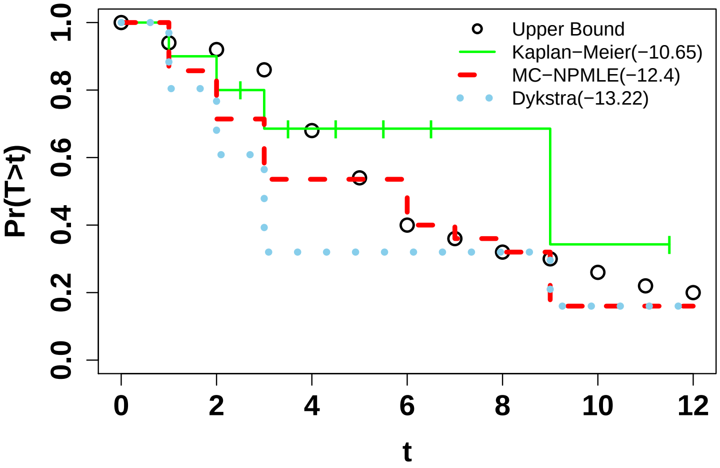
$$\hat{h}_{1i_\ell} = \begin{cases} \log[1 - d_{1i_\ell}/(n_{1i_\ell} - k_\ell)], & \text{if } k_\ell < n_{1i_\ell} \\ \sum_{j=i_{\ell-1}+1}^{i_\ell} \hat{h}_{2j} - \sum_{j=i_{\ell-1}+1}^{i_\ell - 1} \hat{h}_{1j}, & \text{if } k_\ell = n_{1i_\ell} \end{cases}$$

6. If $i_\ell = m'$, stop. Otherwise, set $\ell = \ell + 1$ and go to step 2.

Proof that Algorithm gives MC-NPMLE

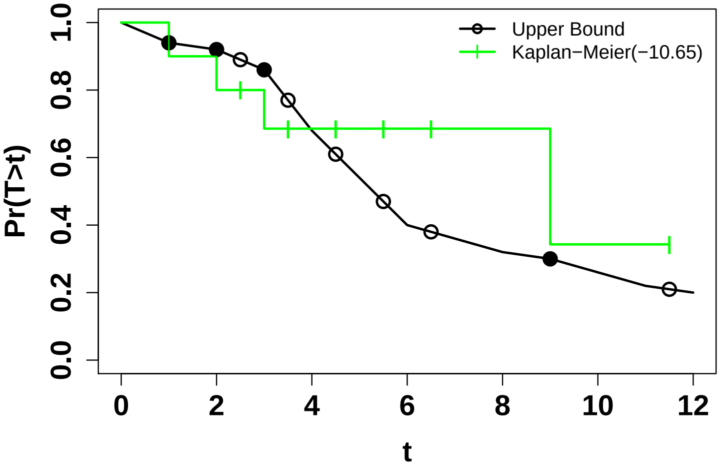
- Constrained optimization problem
- Maximize likelihood subject to some constraints
    - Max $log(L(h_1, ..., h_k)$
    - s.t. $S_1(t) \geq S_2(t)$, $h_j \leq 0$
- Kuhn-Tucker conditions
- Lagrange multipliers

Example: Bounded Above



One-sample Bounded Above with Continuous Constraint

Example


Upper Bound
Kaplan−Meier(−10.65)
Pr(T>t)
t
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12

Naïve Method

"limit approaching"

Use the limit of a discrete function to approach a continuous one;

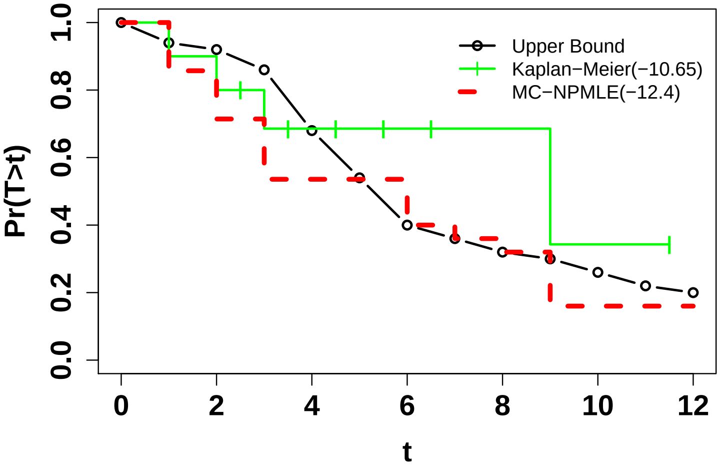
For example

Choose R evenly spaced times between 0 and $\max(Y_{1i})$ as potential death times and obtain the limiting estimate of $\hat{S}_1(t)$ with discrete method as R goes to infinity;

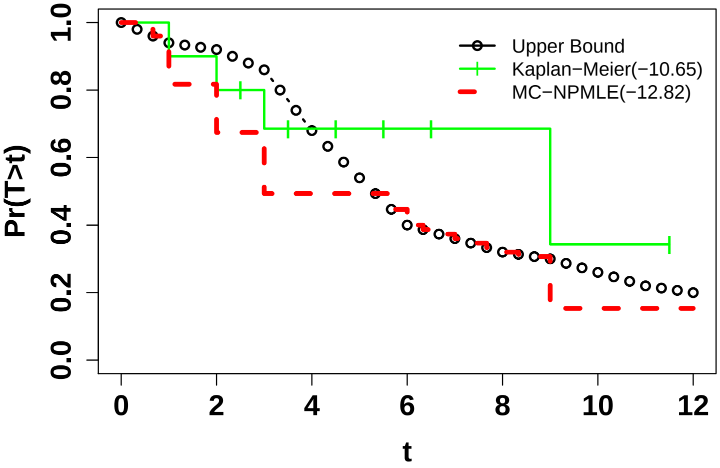
Drawback

Computationally intensive.

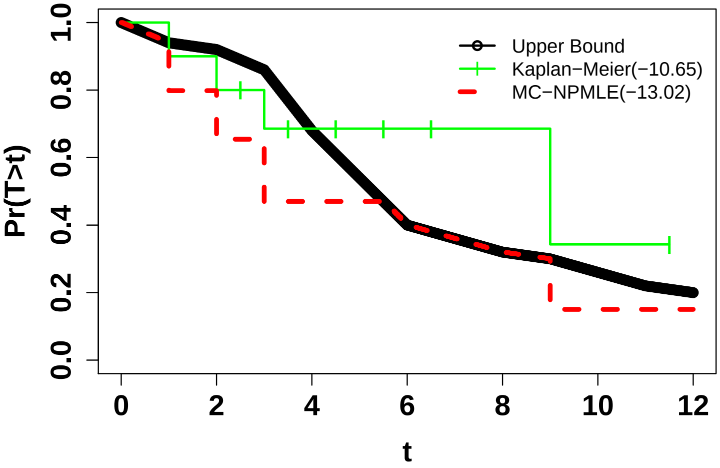
12 potential event times



36 potential event times



360 potential event times

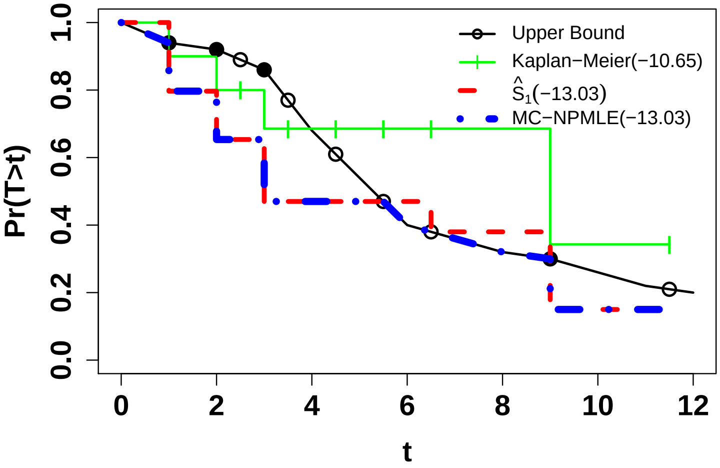


Simple algorithm

Let $C_i, i = 1, \cdots, n_c$ be all distinct observed censoring times and let X_i^- be the time just before observed death time X_i.

1. Let $X_i', i = 1, 2, \cdots, n_{tot}$ be the distinct ordered set of times from the union of X_i, X_i^- and C_i;
2. Estimate $\hat{S}_1(t)$, which is the MC-NPMLE with potential death times at $X_i', i = 1, \cdots, n_{tot}$;
3. Let $\tilde{S}_1(t) = \min(\hat{S}_1(t), S_2(t))$, which is the MC-NPMLE of $S_1(t)$ subject to $S_1(t) \leq S_2(t)$ for $t > 0$.

Simple Algorithm in Example



Two-sample Case

Problem - Two sample case

Data

(Y_{gi}, Δ_{gi}), $g = 1, 2$, $i = 1, \cdots, n_g$;
$\Delta_{gi} = 1$ if event occurred or
$\Delta_{gi} = 0$ if right censored

Goal

Estimate $S_1(t)$, $S_2(t)$ under $S_1(t) \geq S_2(t)$.

Likelihood

$L = \prod_{g=1}^{2} \prod_{i=1}^{n_g} [S_g(Y_{gi}-) - S_g(Y_{gi})]^{\Delta_{gi}} S_g(Y_{gi})^{1-\Delta_{gi}}$

Discrete Case:
$L = \prod_{g=1}^{2} \{\prod_{j=1}^{m} [S_g(a_{j-1}) - S_g(a_j)]^{d_{gi}} S_g(a_j)^{c_{gi}}\}$

Theorem for two-sample case

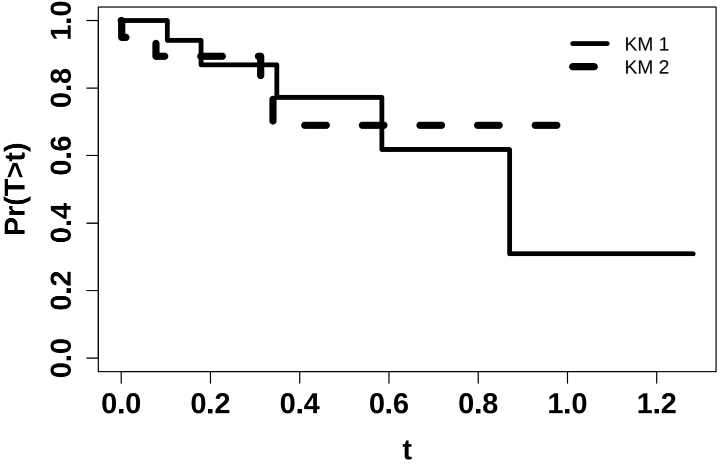
The C-NPMLE of $S_1(\cdot)$ and the MC-NPMLE of $S_2(\cdot)$ are given by $S_1(t) = \exp(\sum_{a_i \le t} h_{1i})$ and $S_2(t) = \exp(\sum_{a_i \le t} h_{2i})$, where

$$\hat{h}_{1i} = \log(1 - \frac{d_{1i}}{n_{1i} + \hat{k}^i})$$

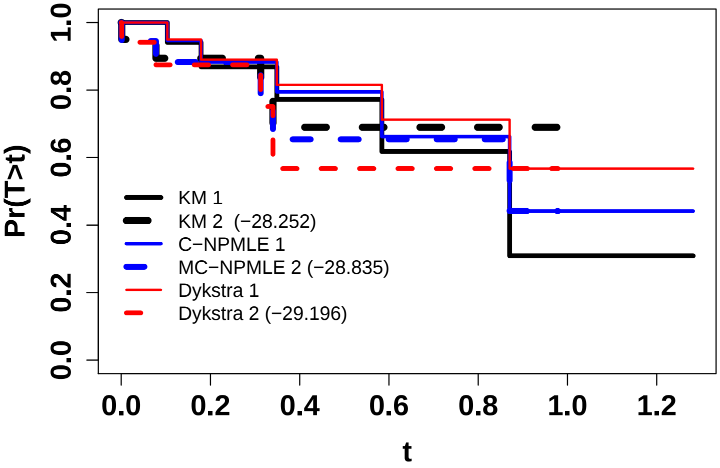
$$\hat{h}_{2i} = \begin{cases} \log(1 - \dfrac{d_{2i}}{n_{2i} - \hat{k}^i}) & d_{2i} > 0 \\ \min\left[0,\ \sum_{j=1}^{i} h_{1j} - \sum_{j=1}^{i-1} \hat{h}_{2j}\right] & d_{2i} = 0 \end{cases}$$

and $\hat{k}^i = \min_{a \le i} \max_{b \ge i} \min(K_2^+(a,b), n_{2b})$,
(Dykstra 1982: $\hat{k}^i = \min_{a \le i} \max_{b \ge i} K_2^+(a,b)$) where
$K_2^+(a,b) = \max(K_2(a,b), 0)$ and $K_2(a,b)$ is the solution of
$\sum_{j=a}^{b}(\log(1 - \frac{d_{1j}}{n_{1j}+k}) = \sum_{j=a}^{b}(\log(1 - \frac{d_{2j}}{n_{2j}-k})$.

Example - Two sample



Example - C-NPMLE, Dykstra



Simulation in Two-sample Case

Simulation - Two-sample case

Finite sample property

- MSE $= (\hat{S}(t) - S(t))^2$; Pointwise criteria

Event distributions and sample sizes

- $S_1(t) = \exp(-t)$, $n_1 = 100$;
- $S_2(t) = \exp(-1.2t)$, $n_2 = 40$.

Scenarios

1. Same censoring: $S_1^c(t) = S_2^c(t) = \exp(-1.5t)$;
2. Excessive censoring 1: $S_1^c(t) = \exp(-3t), S_2^c(t) = 1$;
3. Excessive censoring 2: $S_1^c(t) = 1, S_2^c(t) = \exp(-3t)$.

Simulation - estimators in comparison

1. C-NPMLE from this paper:
2. Dykstra (1982):
3. Lo (1987): $\hat{S}_1^L(t) = \max(S_1^*(t), S_2^*(t))$
   $\hat{S}_2^L(t) = \min(S_1^*(t), S_2^*(t))$;
4. Rojo (2004): $\hat{S}_1^R(t) = \max(S_1^*(t), \frac{n_1 S_1^*(t)+n_2 S_2^*(t)}{n_1+n_2})$
   $\hat{S}_2^R(t) = \min(\frac{n_1 S_1^*(t)+n_2 S_2^*(t)}{n_1+n_2}, S_2^*(t))$;
5. Park et al (2010): PC-NPMLE (pointwise C-NPMLE)
   $\hat{S}_1^P(t) = \tilde{S}_1(t;t)$, $\hat{S}_2^P(t) = \tilde{S}_2(t;t)$
   where $\tilde{S}_1(t;x)$ and $\tilde{S}_2(t;x)$ are the MLE
   subject to $S_1(x) \geq S_2(x)$ at fixed time x.

Pointwise C-NPMLE

- Fix a time x of interest
- Find NPMLE $\hat{S}_1(t)$ and $\hat{S}_2(t)$ such that $\hat{S}_1(x) \geq \hat{S}_2(x)$
- This gives $\hat{S}_1(t)$ and $\hat{S}_2(t)$ at $t = x$
- Repeat for all x

Simulation - same censoring distributions

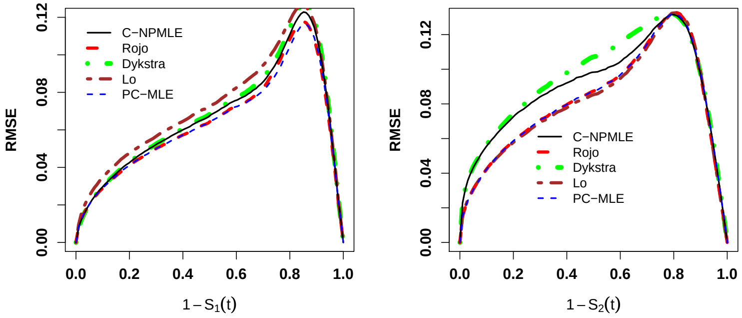


Figure: Same censoring distributions

Simulation - different censoring dist'n

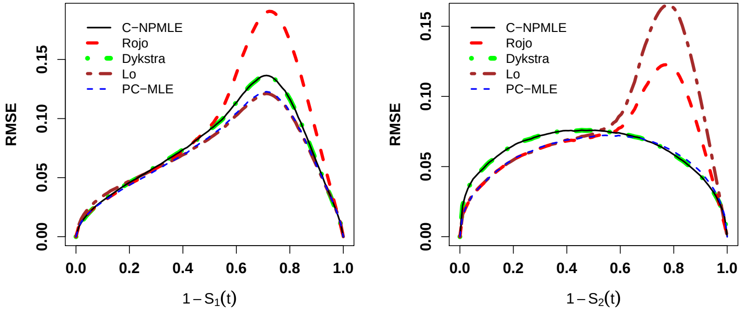


Figure: Excessive censoring group 1

Simulation - different censoring dist'n

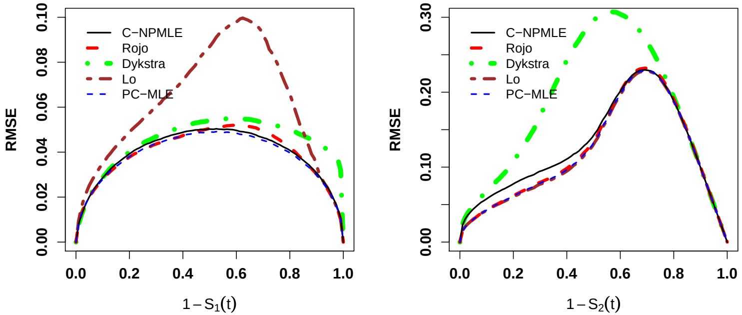


Figure: Excessive censoring group 2

Conclusion

1. Developed methods to obtain the C-NPMLE in the one- and two-sample cases;
   Including a correction of Dykstra's(1982) estimator and computationally efficient algorithms;
2. Developed a simple method to obtain the MC-NPMLE in the one-sample situation with a bounded above constraint when the constraint survivor function is continuous;
3. C-NPMLE is better than Dykstra's estimator;
   C-NPMLE and Rojo's estimator outperform each other at different situations;
   Pointwise C-NPMLE performs better in all cases considered.

Related Problems

1. 3 groups. $S_1(t) \geq S_2(t) \geq S_3(t)$
2. 4 groups. $S_1(t) \geq S_2(t) \geq S_4(t)$ and $S_1(t) \geq S_3(t) \geq S_4(t)$
3. Inference: Confidence Intervals