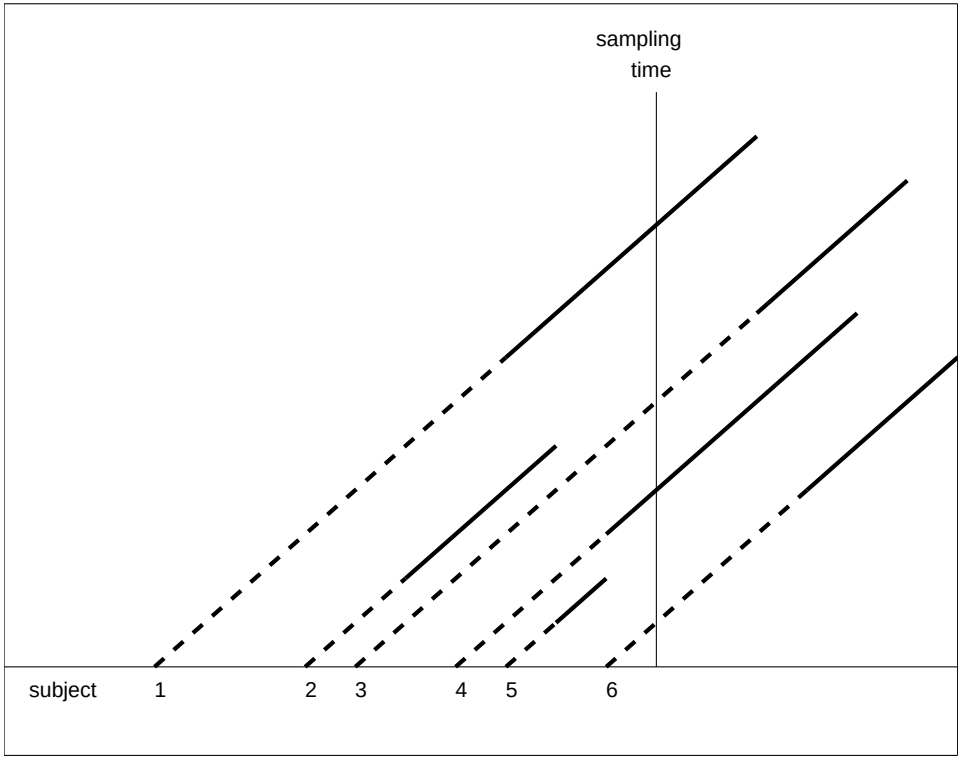
Nonparametric Estimation of a Distribution Function Under Biased Sampling and Multiplicative Censoring – A Multivariate Model –

Micha Mandel

The Hebrew University of Jerusalem

Joint with Yosi Rinott, Yehuda Vardi and Cun-Hui Zhang

The Story



Estimation of (i) total lifetime distribution (ii) joint distribution

Outline

- Biased sampling
- Multiplicative censoring
- Multivariate ordered event-time data
- The current story: questions, difficulties and a model
- Simulation
- Demonstration

Biased Sampling

A parameter of interest is a distribution function F. Data are realizations of X from the biased density

$$F^w(dx) = w(x)F(dx)/Ew(X)$$

If $w = 0$ on (a, b), then F is not identifiable on (a, b).
If $w(x) > 0$ whenever $F(dx) > 0$, then F is identifiable and

$$F(x) \propto \int_0^x [w(t)]^{-1} F^w(dt)$$

Biased Sampling - Example

Suppose that $X \sim F$ is right (randomly) censored by $C \sim G$. If we look only at the uncensored observations, they have the law

$$\begin{aligned} \text{“}P(X = x | uncensored)\text{”} &= P(X = x, C > x)/P(X < C) \\ &= \underbrace{F(dx)}_{\text{'}f(x)\text{'}} \underbrace{[1 - G(x)]}_{w(x)} \Big/ \underbrace{\int [1 - G(t)] F(dt)}_{Ew(X)} \end{aligned}$$

$$\widehat{F}(x) \propto \sum_{x_i \le x} [1 - G(x_i)]^{-1} F_n^w(dx_i)$$

This is an inverse probability of censoring weighted average

Biased Sampling - The Important Point

If there are no identifiability problems, one can estimate the biased distribution F^w using standard methods (empirical distribution) and then estimate the unbiased law using

$$\widehat{F}(dx) \propto \frac{\widehat{F}^w(dx)}{w(x)}$$

Multiplicative Censoring (Vardi, 1989)

$X \sim F$ independent of $U \sim \mathrm{U}(0,1)$. Multiplicative censored observations are realizations of $Z = XU$. This is a model of informative censoring.

$$f_Z(z) = \int_z^\infty \frac{1}{x} F(dx)$$

Multiplicative Censoring - NPMLE

$$L(F; z_1, \ldots, z_n) = \prod_{i=1}^{n} \int_{z_i}^{\infty} \frac{1}{x} F(dx)$$

- Estimation via the EM algorithm with complete data X_i.
- Values of z are support points.

- Suppose that $z_1 < z_2 < \cdots < z_n$ and let p_j be the estimated mass at z_j. An EM step:

$$p_j^{new} = \frac{1}{n} \sum_{k \le j} \frac{z_j^{-1} p_j^{old}}{\sum_{l \ge k} z_l^{-1} p_l^{old}}$$

Multiplicative Censoring - Remarks

I) F is identifiable and the NPMLE is consistent even if all observations are censored.

II) Vardi presented the model in a more general situation where there is an additional sample of uncensored observations.

III) Cross-sectional sampling designs result in size biased and multiplicative censored lifetime data.

Multivariate Ordered Event-Time Data

$0 \equiv Y_0 < Y_1 < Y_2 < \cdots < Y_m$ are ordered events such as phases of a disease, ranks in the university, recurrent events, etc.
We also look at the duration at each phase $X_j = Y_j - Y_{j-1}$.

Censoring (a variable C) usually acts on the total lifetime Y_m. A typical observation has the form $\{Y_1, \ldots, Y_k, C, I(Y_k < C < Y_{k+1})\}$

Note that X_{k+1} is censored by $C - X_1 - \cdots - X_k$, so the censoring model is not 'random' in the usual sense.

Non-parametric Estimation of Ordered Event-Time Models

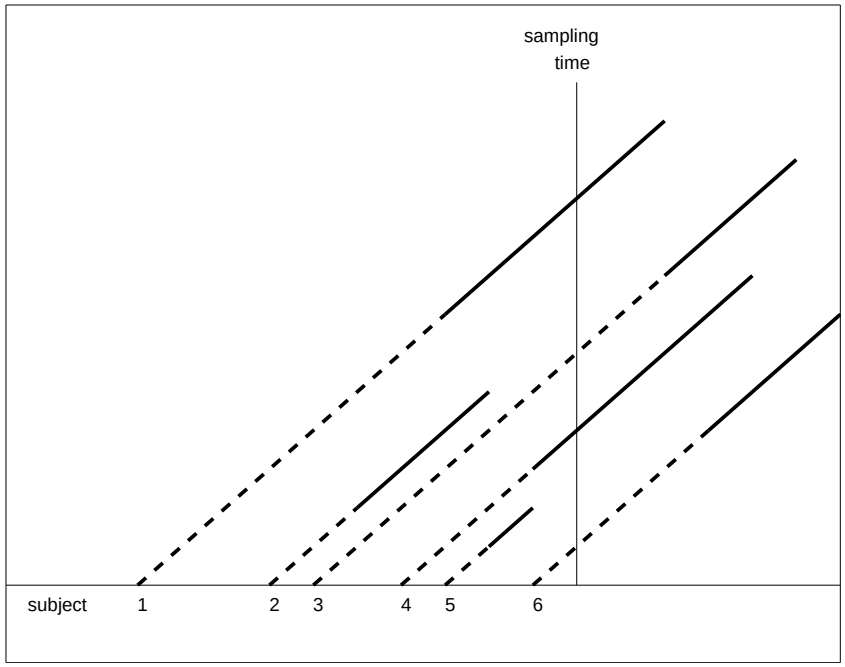
- NPMLE problematic (not unique).

- Estimation of the laws of Y_1 and $Y_2|Y_1$ (Visser, 1996*). Limited to Y_1 discrete.

- Inverse probability of censoring weighting (Lin, Sun and Ying, 1999; van der Laan, Hubbard and Robins, 2002; Chang and Tzeng, 2006*). Estimate may assign negative mass.

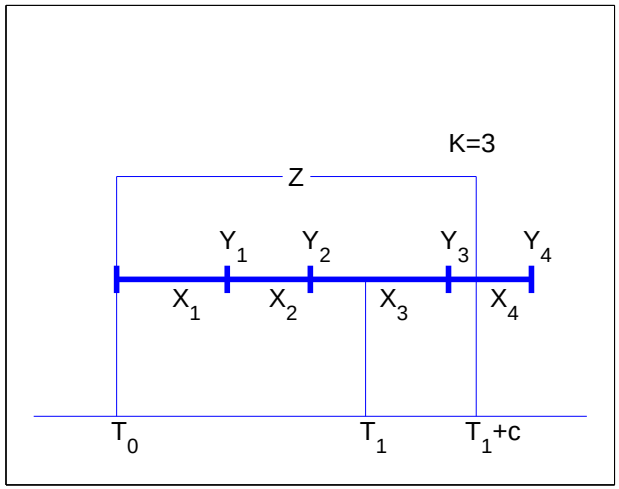
* For left truncated data

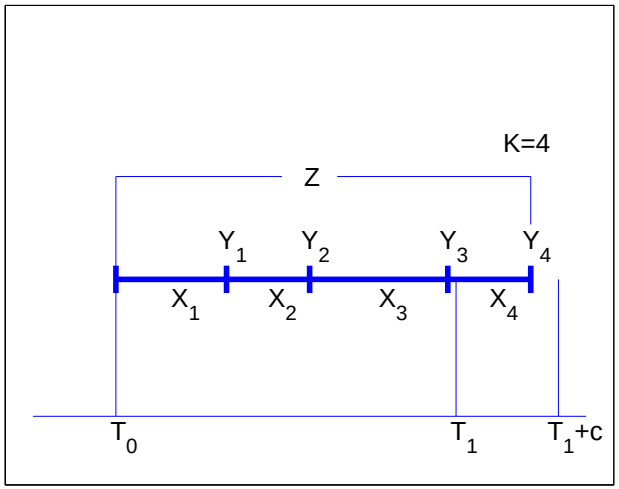
Multivariate Ordered Event-Time Data Under Biased Sampling and Multiplicative Censoring



The process is size biased by Y_m and multiplicative censored. We hope to exploit all the previous techniques to do something here.

Notations





$Y_1, \ldots, Y_m$ - event times, $X_1, \ldots, X_m$ - duration times, T_0 - beginning of the process, T_1 - sampling time, c - Follow-up, Z - observed total lifetime, K - last phase observed.

Specific Questions

1. Do phase data improve estimation of the total lifetime CDF?

2. Estimation of the joint CDF of phases' lengths.

Estimation of the Joint CDF F

In the univariate multiplicative censoring model the NPMLE is unique and it is consistent, even when there are no uncensored observations.

In multivariate survival models the NPMLE is not unique and it is not clear how to construct a consistent sequence of NPMLEs.

In multivariate multiplicative censoring problems....

NPMLE of F

The NPMLE is unique, but inconsistent.

The idea: (1) Change the measure to the biased law

$$\begin{aligned} lik &= \int_{z<y_{k+1}<\cdots<y_m} \frac{F(dy_1,\ldots,dy_k,dy_{k+1},\ldots,dy_m)}{E(Y_m)} \\ &= \int_{z<y_{k+1}<\cdots<y_m} \frac{1}{y_m} F^*(dy_1,\ldots,dy_k,dy_{k+1},\ldots,dy_m) \end{aligned}$$

(2) Show that the NPMLE assigns mass only to observed values (i.e., $y_{k+1} = \cdots = y_m = z$).

(3) Claim that this give positive mass to $P(Y_m = Y_{m-1})$.

A Model for F

Let $\mathbf{y}^{(k)} = (y_1, \ldots, y_k)$. We assume that

$$F(dy_1, \cdots, dy_m) = g_\theta(\mathbf{y}^{(m-1)}|y_m)dy_1 \cdots dy_{m-1}F_{Y_m}(dy_m)$$

First term is the conditional density of event-times given total lifetime, and it is known up to a finite dimensional parameter θ.

Second term is the unspecified CDF of total lifetime.

The Schur Constant Model

Caramellino and Spizzichino 1996; Spizzichino 2001; Nelsen 2006

$$(X_1, \ldots, X_m)|Y_m = y_m \sim \text{Unif}\left\{(x_1, \ldots, x_m) : \textstyle\sum_i x_i = y_m, x_i \geq 0\right\}$$

$$F(dx_1, \cdots, dx_m) = \frac{(m-1)!}{(x_1 + \cdots + x_m)^{m-1}} dF_{Y_m}(x_1 + \cdots + x_m) dx_1 \cdots dx_{m-1}$$

The Schur Constant Model - Properties

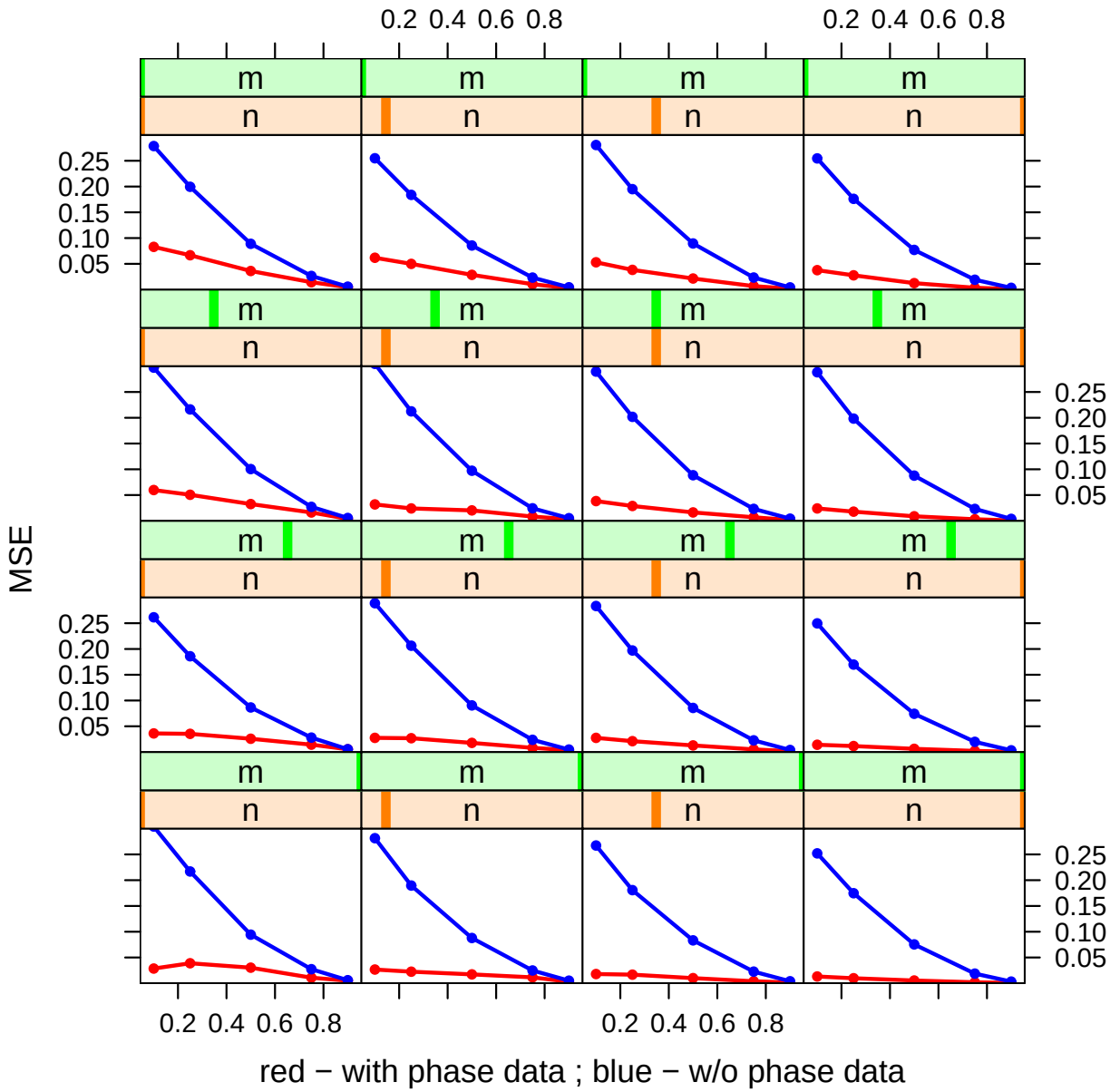
- $\{(Z_i, K_i)\}_{i=1}^n$ are the sufficient statistics (data are $\{(X_{1i}, \ldots, X_{K_i i}, Z_i, K_i)\}_{i=1}^n$)

- $K \sim \mathsf{U}\{0, 1, \ldots, m-1\}$

- The density of $Z|K = k$ is equivalent to the density of $U_k Y_m^*$, where $U_k \sim \mathsf{Beta}(k+1, m-k) \rightarrow$ estimation by an EM algorithm

Simulation

Comparison of MSE at the quantiles 0.1,0.25,0.5,0.75,0.9 between the NPMLE with (red line) and without (blue line) information on phases. Rows top to bottom - different number of phases, 2,3,4,5. Column left to right - different sample sizes, 50,100,200,500. $Y_m \sim$ Gamma$(6, 1)$. Shown are averages of 100 replications.


0.2 0.4 0.6 0.8
m
n
0.25
0.20
0.15
0.10
0.05
MSE
red − with phase data ; blue − w/o phase data

Likelihood of General Model

$$p_c(k, z, y_1, \ldots, y_k; F, \theta) = \begin{cases} \frac{\int_z^\infty g_\theta(k, \mathbf{y}^{(k)} | y_m, z) F(dy_m)}{(c + EY_m)}, & z > 0, k < m \\ \frac{c g_\theta(\mathbf{y}^{(m-1)} | z) F(dz)}{(c + EY_m)}, & z > 0, k = m \end{cases}$$

where $g_\theta(k, \mathbf{y}^{(k)} | y_m, z)$ is the density of $(K, \mathbf{Y}^{(K)})$ conditionally on Y_m and Z

$$g(k, \mathbf{y}^{(k)} | y_m, z) \equiv \begin{cases} I\{y_k \leq z\} \int \cdots \int_{z < y_{k+1} < \cdots < y_m} g(\mathbf{y}^{(m-1)} | y_m) dy_{k+1} \cdots dy_{m-1} & k < m-1 \\ I\{y_k \leq z\} g(\mathbf{y}^{(m-1)} | y_m) & k = m-1 \end{cases}$$

Estimation when g is known

Let δ_i be the indicator of censoring ($k < m$), $w(y) = y + c$, and $F^w(dy) \propto (y + c)F(dy)$, then the likelihood has the form

$$L(F) \propto \prod_{i=1}^{n} \left\{ F^w(dz_i) \right\}^{\delta_i} \prod_{i=1}^{n} \left\{ \int_{z_i}^{\infty} [\tilde{g}_i(y)/w(y)] F^w(dy) \right\}^{1-\delta_i}$$

for a given support, an EM step is

$$F^{w(\text{new})}(dt) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i I\{z_i = t\} + (1 - \delta_i) \frac{[\tilde{g}_i(t)/w(t)] I\{z_i \le t\}}{\int_{z_i}^{\infty} [\tilde{g}_i(u)/w(u)] dF^{w(\text{old})}(u)} F^{w(\text{old})}(dt) \right\}$$

Estimation when θ is unknown

The likelihood has the form

$$L(\theta) \propto \prod_{i=1}^{n} \left\{ [\tilde{g}_{\theta i}(z_i)/w(z_i)] F^w(dz_i) \right\}^{\delta_i} \left\{ \int_{z_i}^{\infty} [\tilde{g}_{\theta i}(y)/w(y)] F^w(dy) \right\}^{1-\delta_i}$$

Assuming F is known, $\widehat{\theta}$ can be found by a numerical search.

Iterating between 'known g' and 'known F' yields the joint estimate.

(After estimating F^w we estimate F using the inverse transformation.)

An Example: The Dirichlet Distribution

$$g_\alpha(\mathbf{y}^{(m-1)}|y_m) = C(\alpha) y_m^{m-1} \prod_{i=1}^{n} \left(\frac{y_i - y_{i-1}}{y_m}\right)^{\alpha_i - 1}$$

where $C(\alpha) = \Gamma(\sum_i \alpha_i) / \prod_i \Gamma(\alpha_i)$ and $y_0 = 0$.

The conditional density is

$$\begin{aligned} g(k, \mathbf{y}^{(k)}|y_m, z) &= \frac{1}{y_m^k} \frac{\Gamma(\sum_{i=1}^m \alpha_i)}{\Gamma(\sum_{i=k+1}^m \alpha_i) \prod_{i=1}^k \Gamma(\alpha_i)} \left(\frac{y_m - y_k}{y_m}\right)^{\alpha_{k+1}+\cdots+\alpha_m-1} \\ &\quad \times \prod_{i=1}^k \left(\frac{y_i - y_{i-1}}{y_m}\right)^{\alpha_i - 1} \\ &\quad \times \left[1 - F_{\text{Beta}(\alpha_{k+1}, \alpha_{k+2}+\cdots+\alpha_m)}\left(\frac{z - y_k}{y_m - y_k}\right)\right] \end{aligned}$$

Summary and Topics for Further Research

We model the joint distribution by specifying a parametric model to the conditional distribution of phases given total lifetime so we can (i) easily deal with the bias, and (ii) improve estimation of the total lifetime distribution.

We still need to work on (i) Unknown bias, (ii) Support of the estimator, (iii) Properties of the estimator, and (iv) Random number of phases m.