СтатМод

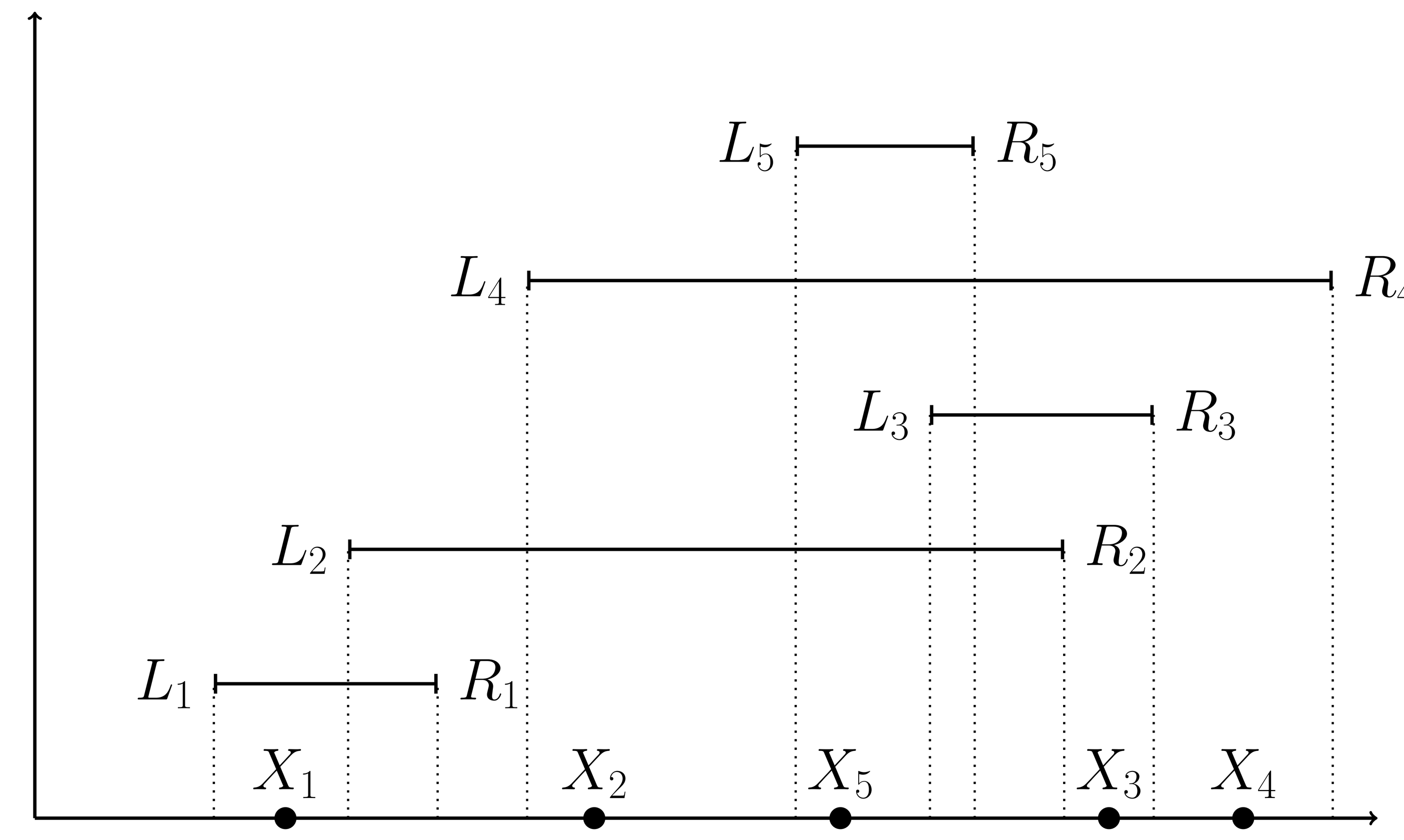
Comparison of Parameter Estimates for One Special Model of Survival Curves in Case of Interval Censoring

Anton Korobeynikov

Department of Statistical Modelling, Saint Petersburg State University

Survival Curves

- X — failure time, time to some event
- Survival curve $S(x) = P(X > x)$
- Survival (time to event) data is usually incomplete — *censored*.
- For example we might only know some interval, X belongs to:



Mixed Case Interval Censoring Model

- K — number of observations of subject
- Vector of observation times $T_K = (T_{K,1}, \ldots, T_{K,K})$ with
$$0 = T_{0,K} < T_{1,K} < \cdots < T_{K,K} < T_{K+1,K} = +\infty$$
- The status of subject is known only at observation times:
$$\Delta_K = (\Delta_{1,K}, \ldots, \Delta_{K+1,K}), \quad \Delta_{j,K} = \mathbb{I}_{[T_{j-1}, T_j)}(X)$$
- Observed variable is
$$(K, T_K, \Delta_K)$$

Maximum Likelihood Estimates

- Denote $X_1, \ldots, X_n$ the sample of i.i.d random variables with distribution function F. We assume that each X_i is censored. Observed variables are $\left(K^{(i)}, T_K^{(i)}, \Delta_K^{(i)}\right)$, $i = 1, \ldots, n$.
- Then we can introduce the *log-likelihood function* to estimate F:
$$l_n(F) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K^{(i)}+1} \Delta_{j,K}^{(i)} \log\left[F(T_{j,K}^{(i)}) - F(T_{j-1,K}^{(i)})\right].$$
- And define the ML-estimate of F via
$$\hat{F}_n = \arg\max_{F \in \mathcal{F}} l_n(F),$$
here $\mathcal{F}$ denotes some family of distribution functions:
  - Parametric models: $\mathcal{F} = \{F_\theta, \theta \in \Theta \subset \mathbb{R}^m\}$.
  - Non-parametric models: $\mathcal{F} =$ all distribution functions with $\operatorname{supp} = [0, +\infty)$.

Parameter Estimates

We assume parametric model for F and consider two types of estimators:

- Ordinary maximum likelihood estimates:
$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} l_n(F_\theta).$$
- Estimates based on non-parametric estimate of distribution function. The underlying idea is same as in (Oakes, 1986):
  1. Obtain non-parametric estimate of distribution function:
  $$\tilde{F}_n = \arg\max_{F \in \mathcal{F}} l_n(F).$$
  2. Use it instead of ordinary empirical distribution function in log-likelihood:
  $$\tilde{\theta}_n = \arg\max_{\theta \in \Theta} \int \log f_\theta \, d\tilde{F}_n,$$
  here f_θ is density of F_θ under some dominating measure.

Robustness and Kullback-Leibler Optimality

Maximum likelihood estimates for the sample without censoring $\hat{\theta}_n^{nc}$ possess very important Kullback-Leibler optimality propery:
$$\hat{\theta}_n^{nc} \to_P \theta^* = \arg\min_{\theta \in \Theta} \int \log \frac{g}{f_\theta} \, dG,$$
here G denotes true underlying distribution function, g is the corresponding density and f_θ is the density of assumed parametric model.

If $f_{\theta_0} = g$ for some θ_0 then $\theta^* = \theta_0$ and MLE $\hat{\theta}_n^{nc}$ are consistent. Otherwise (case of *misspecified model*) MLE are still optimal in the sense of *minimum Kullbak-Leibler distance* between probability measures.

This is not true in general for the sample with censoring!

Even for "easy cases" like right censoring. See (Suzukawa et all, 2001).

However, we expect KL-optimality property to be held for modified estimators $\tilde{\theta}_n$, based on nonparametric estimate of distribution function. This fact is known to be true for special right censoring case (Suzukawa et all, 2001).

Special Parametric Model of Survival Curves

- We consider the following model of survival curves (Bart, 1980):
$$S(x) = \exp(-\eta x)\cos\left(\frac{\pi}{2\tau}x\right), \quad \eta > 0, \; 0 < x < \tau,$$
which was successfully used to describe the survival dynamics of the chronic glomerulonephritis patients (Bart, 1980), wound processes (Bart, 2003), hypertension (Bart, 2005), generalized severe periodontitis (Madai, 2006).
- The typical example of mixed case interval censoring model in clinical studies is the situation when an examination is performed at the start of the study and follow-ups are scheduled one at a time till the end of the study. If Z_i denote the times between consecutive follow-ups and L the total duration of the study, then
$$T_{j,k} = \sum_{i=1}^{j-1} Z_i, \quad K = \sup_{j \geqslant 1}\left\{\sum_{i=1}^{j-1} Z_i < L\right\}.$$
- We modelled the sample with the following parameters: $\eta = 0.125$, $\tau = 16$ (they corresponds to the estimates obtained for real cardiology data in (Korobeynikov, 2008)).
- Censoring scheme: Z_i were i.i.d $Exp(1)$ and $L = 8$.

Results

In order to test robustness of estimates we modelled the location mixture of the distributions: 30% of the sample was shifted by 1.

Sample Size		SD	MSE		SD	MSE
1000	$\hat{\eta}_n$	$0.12 \cdot 10^{-1}$	$1.33 \cdot 10^{-4}$	$\hat{\tau}_n$	2.84	8.02
2000		$0.68 \cdot 10^{-2}$	$4.53 \cdot 10^{-5}$		1.80	3.54
5000		$0.47 \cdot 10^{-2}$	$2.30 \cdot 10^{-5}$		1.22	1.54
10000		$0.36 \cdot 10^{-2}$	$1.35 \cdot 10^{-5}$		0.94	0.97
1000	$\tilde{\eta}_n$	$0.89 \cdot 10^{-2}$	$7.82 \cdot 10^{-5}$	$\tilde{\tau}_n$	2.20	4.94
2000		$0.63 \cdot 10^{-2}$	$3.90 \cdot 10^{-5}$		1.38	1.94
5000		$0.36 \cdot 10^{-2}$	$1.35 \cdot 10^{-5}$		0.81	0.66
10000		$0.28 \cdot 10^{-2}$	$8.17 \cdot 10^{-6}$		0.58	0.34

One can easily see that estimates $(\tilde{\eta}_n.\tilde{\tau}_n)$ outperforms MLE $(\hat{\eta}_n.\hat{\tau}_n)$ in terms of both SD and MSE in case of misspecified model.