Inference for the Logconcave NPMLE for Interval Censored Data

Clifford Anderson-Bergman

advised by

Professor Yaming Yu

Department of Statistics

University of California, Irvine

Interval Censored Data

- Interval censored data occurs when an event time is known only up to an interval for each subject
- Define L_i to be the last time known for the event not to have occurred yet occur and R_i to be the first time event is known to have happened
    - Example: subject i has a doctor's visit at $t = 20$ and tests negative for a disease, next visit is at $t = 30$ and tests positive. $L_i = 20$ and $R_i = 30$.

Interval Censored Data

- Case I interval censoring: current status data
  - Subject i is monitored at time C_i
    - If event has already occurred, $L_i = 0$ and $R_i = C_i$
    - If event has not occurred, $L_i = C_i$ and $R_i = \infty$
    - All data is either left censored or right censored
  - Each observation is very cheap (no need to follow subjects)
  - Each observation is fairly uninformative
- Case II interval censored data
  - More than one possible observation time (i.e. doctor visits example)
  - Data may be left censored, right censored or within a window

Motivating Example

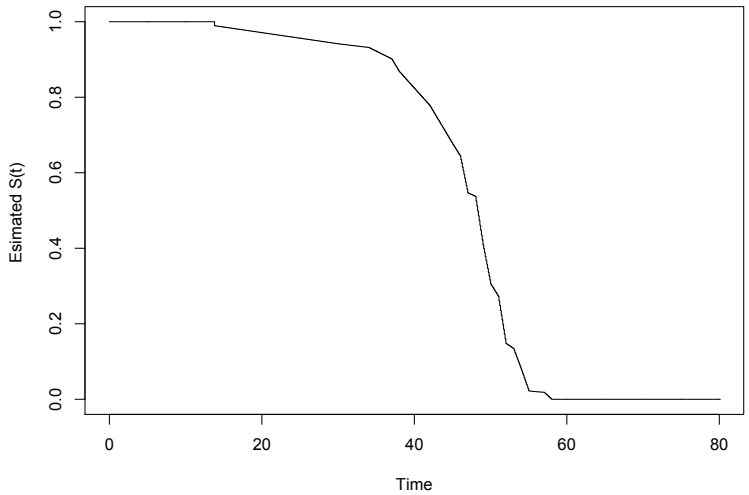
- Study preformed by MacMahon and Worcestor (1966)
- Outcome of interest: age at menopause
- Interviewed 2,423 subjects, asked whether they had experienced menopause and if so, when
    - Leads to right censored data
- Researchers noted that there was excessive clustering at digits 0 and 5, likely from recall bias
- Krailo and Pike (1983) recommended using only menopause status of women at time of the questionnaire
- Leads to current status data

Demand for Non Parametric Estimators for Interval Censored Data

- Non parametric estimators often make fewer assumptions about data structure, reducing potential bias
- Especially important for interval censored data
  - Typically very difficult to asses model fit: no histograms!
- Downside: reduction in bias often comes with increase in variance

Unconstrained NPMLE

- Classic solution: (unconstrained) NPMLE
  - Turbull (1976)
  - Find $\hat{F}(t) = \arg\max_F \left(\sum_{i=1}^{n} \log(F(R_i) - F(L_i)) \right)$
  - Solution can be written as a discrete probability function



Issues with Unconstrained NPMLE

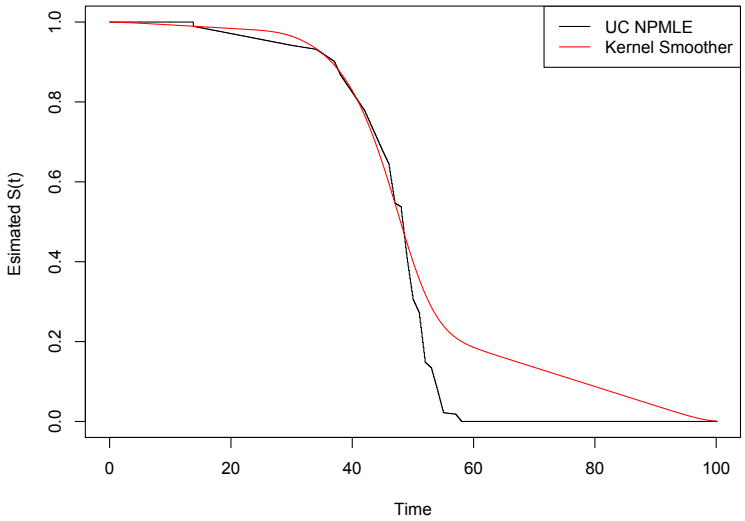
- Solution assigns probability mass to isolated intervals
  - Does not dictate how probability mass is assigned within interval
  - No density estimates!
- Solution often jumps erratically, causing excessively high variance of quantile estimates
- Both problems could be solved by assuming $f_T(t)$ is "smooth"

Modern Non-Parametric Alternatives

- Logspline density estimator (Kooperberg and Stone, 1992)
- Kernel Smoother (Betensky et al 1999)
- It has been shown that both these estimators can reduce variance of estimates when compared to unconstrained NPMLE (Pan, W. 2000)

Issues with Modern Non-Parametric Alternatives

- While both estimators work well for light censoring (i.e. narrow intervals), both do poorly under heavy censoring, such as current status data
  - We found kernel smoother often heavily biased
  - Logspline algorithm often gives degenerate estimates or fails!
  - We found these problems to get *worse* as n increases



Log Concave NPMLE

- We enforce "smoothness" by assuming logconcavity
  - $f_T(t) = e^{\phi(t)}$ where $\phi(t)$ is a concave function
    - Insures $f_T(t)$ is unimodal
    - Insures $f_T(t)$ does not have heavier tails than an exponential distribution (exponential distribution is log linear)
    - Insures non-decreasing hazard
  - Fairly flexible assumption
    - Normal, gamma with shape ≥ 1, Weibull with exponent ≥ 1, beta with both parameters ≥ 1 and logistic distribution all log concave
    - t, lognormal and multi-modal (common in mixture distributions) are not log concave

Logconcave NPMLE

- Dümbgen et al (2011) present a very efficient active set algorithm for finding the LC NPMLE with exact observations
- Chang and Walther (2007) use log concave components for clustering with mixture models
- Dümbgen et al (2011) purpose EM algorithm for censored data which discretizes support space, giving approximation of LC NPMLE
    - Current implementation (CRAN package "logconcens") too slow for moderate sized data sets

Computing LC NPMLE for Interval Censored Data

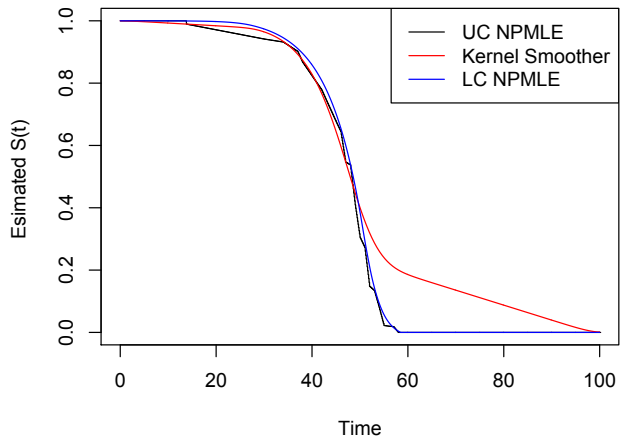
- Theorem: an MLE can be achieved via log linear spline with at most $2u - 1$ knots, where u = number of unique times in dataset
- Algorithm: Constructed active set algorithm, similar to Dümbgen et al 2011, with additional tools to cope with $\ell(\phi)$ not being concave

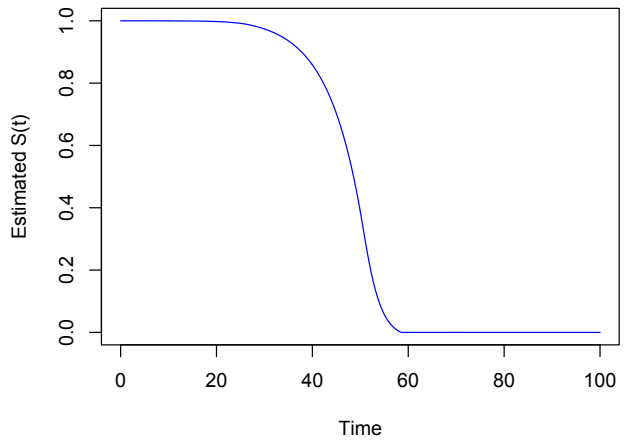
Table: Average Computation Times (in seconds)

	Unique Times						
n	10	50	100	500	1000	2000	5000
100	0.07	0.15	0.17	NA	NA	NA	NA
500	0.19	0.27	0.46	0.86	2.44	NA	NA
5000	0.42	0.86	1.11	5.07	7.59	31.6	156

Motivating Example

- LC NPMLE appears to agree with UC NPMLE
- LC NPMLE solution is once differentialable, UC NPMLE is not





Simulations

Table: Quantile Estimation Gamma(100, 2)

	Q(p)	43.7 (0.1)	46.5 (0.25)	49.8 (0.5)	53.3 (0.75)	56.5 (0.9)
n		Bias / Standard Deviation				
50	LC	-0.4/2.55	-0.08/1.54	-0.05/1.14	0.06/1.44	0.72/2.14
	UC	-0.67/3.73	-0.23/2.08	-0.21/1.75	0.03/2.03	0.93/2.41
	KS	11.48/7.22	3.36/3.15	-0.1/1.51	-2.24/1.56	-3.89/1.92
200	LC	-0.1/1.13	0.08/0.73	-0.02/0.67	-0.1/0.81	0.18/1.19
	UC	-0.17/1.51	0.04/1.1	-0.11/1.06	-0.01/1.23	0.14/1.82
	KS	12.64/5.73	1.88/1.33	-0.22/0.67	-1.69/0.93	-3.83/1.6
800	LC	0.14/0.68	0.04/0.44	-0.07/0.33	-0.12/0.44	0/0.61
	UC	0.1/0.97	-0.03/0.64	-0.1/0.51	-0.05/0.75	-0.04/0.98
	KS	15.94/3.23	1.18/0.66	-0.25/0.3	-1.22/0.48	-4.57/1.35

- SD of LC NPMLE typically about 2/3 SD of UC NPMLE
- Kernel smoother does not appear consistent!

Inference for LC NPMLE

- Further simulations suggest LC NPMLE is estimator of choice when log concavity assumption is correct
- Often still best estimator even if mild violations of log concavity
- Need some way of evaluating log concavity assumption

Unconstrained Likelihood Ratio Test

- Natural test to consider: Likelihood ratio, comparing LC NPMLE to UC NPMLE
- Properly nested
- Problem: very low power!
  - Intuition: UC NPMLE allows for non-log concave and "unsmooth" estimates
  - Would prefer model which allows for non-log concave but not unsmooth estimates

Mixture LC NPMLE

- After considering several other possible tests, most promising is LR ratio test based on a mixture of 2 LC NPMLE components

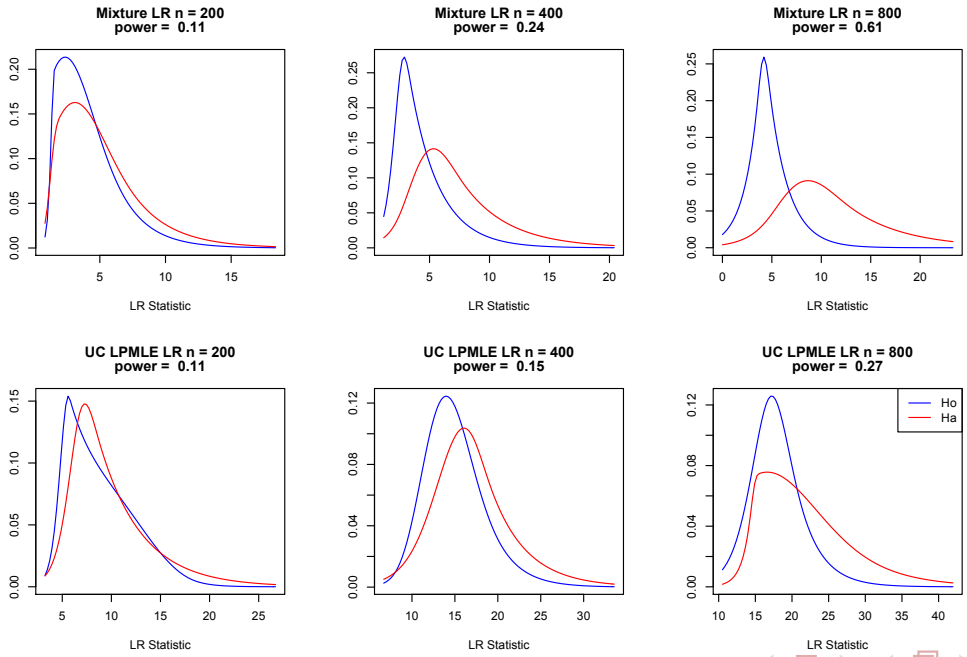
$$\arg\max_{f_1,f_2,p} \sum_{i=1}^{n} \log(p \int_{L_i}^{R_i} f_1(t)dt + (1-p) \int_{L_i}^{R_i} f_2(t)dt)$$

$$f_1, f_2 \text{ log concave}, 0 \leq p \leq 1$$

- Solution can be found via combination ECM algorithm, using Newton's method to estimate p and earlier algorithm to estimate f_i

Simulations

- Simulations show greater power for mixture LC NPMLE LR test compared to UC NPMLE LR test
  - Strongest gains when distribution really bimodal
  - Still more powerful if heavily skewed
  - Much more powerful for case II interval censored data than current status data



Future Work: Difficulties in Implementing Test

- The family of single component LC distributions are on the boundary of 2 component LC mixture distributions
- Typically it is suggested to sample under the null hypothesis to find p-value
- Problem: Censoring distribution affects likelihood!
  - For current status data, very easy to model censoring distribution
  - For general case II interval censoring, much more difficult...

Future Work: Computational Issues

- Mixture models are known to have several local maximum points, making finding the true MLE difficult
- Random starting points in our algorithm demonstrates this can be a problem for the mixture LC NPMLE
  - Differences in likelihoods occasionally observed to be higher than 1.0!
  - Fixing this may increase power of test
- Try applying methods used for general mixture models to efficiently deal with this