Non-parametric estimation of the survivor function for misclassified failure time data

Andrew Titman

Lancaster University

27 August 2013
Overview

- Interval censored survival data
- Misclassified failure statuses
- Characterization of the NPMLE
- Suitable algorithm for maximization
- Illness-death model extension
Interval censored survival data

- Study in which the status of patients is monitored intermittently
- Observations $Y_{i1}, \ldots, Y_{in_i}$ at examination times $X_{i1}, \ldots, X_{in_i}$
  where $Y_{ij} = I(T_i \leq X_{ij})$
- Event time $T_i$ is only known to have occurred within $(L_i, R_i)$
  - $L_i$ - last time at which patient known to not to have had failure
  - $R_i$ - first time at which patient known to have had failure
Interval censored survival data

- Parametric inference is straightforward
- Non-parametric inference
  - NPMLE unique only up to increments in $F$ on intervals

- Self-consistent estimation or iterative convex minorant algorithm (Jongbloed, 1998) for computation
Misclassified failure time data

- Data based on observations of status at a series of time points
- But now the observed status (e.g. presence of disease) is subject to classification error:

\[
P(Y_{ij} = 1|T_i > X_{ij}) = \alpha, \]

\[
P(Y_{ij} = 0|T_i \leq X_{ij}) = \beta, \]

i.e. at time \(X_{ij}\) the subject undergoes a diagnostic test which has sensitivity \(1 - \beta\) and specificity \(1 - \alpha\)

- Either
  - Continue to monitor after the first positive test
  - Or, stop monitoring after the first positive test
Examples of misclassified failure time data

- Time to first infection of gonorrhoea in Kenyan women (Richardson and Hughes, 2000)
  - Patients treated after first positive diagnosis
- Onset of puberty in adolescents (Espeland et al, 1989)
- Development of tooth cavities in children (Garcia-Zattera et al, 2010)
  - Observations continued independent of assessments
- Existing approaches to this problem have either been parametric or used a discrete-time approximation
Link with mixture models

- The likelihood can be expressed as
  \[ L_i = \int_{0}^{\infty} P(Y_i|X_i, T_i = t) dF(t) \]

- Any failure time in \([0, \infty)\) is consistent with the data
- \(\hat{F}\) unique up to probability \(\pi_k\) of being in particular intervals

\[ L_i = \sum_{k=1}^{m} \gamma_{ik} \pi_k \]

where \(\gamma_{ik} = P(Y_i|X_i, T_i \in I_k, \alpha, \beta)\)
Computation of NPMLE

Can show that the NPMLE can only increase in time intervals \((X_k, X_{k+1})\) in which \(Y_k = 0\) and \(Y_{k+1} = 1\) where \(X_k\) are the ordered observation times across all subjects.

For fixed \((\alpha, \beta)\) the log-likelihood is concave and the directional derivatives

\[
D(\pi)_j = \sum_{i=1}^{N} \frac{\gamma_{ij}}{\sum_{k=1}^{m} \gamma_{ik}\pi_k} - N \leq 0
\]

for \(j = 1, \ldots, m\), give necessary and sufficient conditions for optimality.
Analogy with non-parametric mixing models means can adapt algorithms for that purpose

- Wang (2007): active-set algorithm involving a sequence of non-negative least squares problems
- Convergence when \( \max_j D(\pi)_j = 0 \)

Profile likelihood approach can be used for joint estimation of \( F \) & \((\alpha, \beta)\)

- Provided observations continue after first positive result
- Algorithm for fixed \((\alpha, \beta)\) sufficiently stable to apply standard quasi-Newton type maximization
Confidence interval construction

- Confidence intervals for \((\hat{\alpha}, \hat{\beta})\) can be found via the profile-likelihood ratio

\[ 2\{\log L(\hat{\alpha}, \hat{\beta}) - \log L(\alpha_0, \beta_0)\} \sim \chi^2_2 \]

- For \(\hat{F}\) depends on the process generating the observation times
  - If continuous:
    - Non-standard asymptotics
    - Sub-sampling rather than bootstrapping required
  - If discrete, i.e. finite set of possible examination times:
    - Standard \(n^{1/2}\) asymptotics apply (but unlikely to be useful)
    - Bootstrapping consistent
Example: Onset of CAV

- Cardiac allograft vasculopathy (CAV): progressive deterioration of heart in post-heart-transplantation patients
- Follow-up continues after first observation of disease
- $\alpha = 0.022(0.005, 0.05)$, $\beta = 0.073(0.01, 0.16)$

![Graph of Onset of CAV in 6 years post-transplantation]

Andrew Titman
Lancaster University

Non-parametric estimation of the survivor function for misclassified failure time data
Practical problems

- Onset of CAV increases hazard of death
- Censoring via death is therefore informative for determining onset rate
- Requires joint modelling of onset rate and respective hazard of death with or without onset of disease through an illness-death model
  - More appropriate to estimate either disease-free survival or the cumulative incidence of disease
Possible approaches

- Parametric approach e.g. 3 state homogeneous hidden Markov model
- Teeple (2013): Bayesian B-spline model allowing time dependent sensitivity
- Model the hazards of death non-parametrically
  - Extension of the model by Frydman and Szarek (2009)
  - Maximization problem less straightforward
  - Different criteria for determining support intervals
  - Self-consistency algorithm easy to implement but has slow convergence rate
Illness-death model estimates

- Use complete follow-up of patients
- Naive estimator underestimates onset rate because of informative censoring.

![Onset of CAV](attachment:Onset_of_CAV.png)

Andrew Titman
Lancaster University

Non-parametric estimation of the survivor function for misclassified failure time data
Conclusion

- Non-parametric estimation for misclassified failure time data
  - Concave maximization problem
  - Of comparable complexity to standard interval censored data.
- If disease onset affects hazard of death
  - Need to consider as an illness-death model
  - More work needed on finding better approaches to maximization


Non-parametric estimation of the survivor function for misclassified failure time data
Covariates on the failure time distribution

- Cox proportional hazards model: extension of Huang (Annals of Stats, 1996)
  - Regions of support of baseline $F$ independent of regression parameters
- Accelerated failure time model: extension of Rabinowitz et al. (Biometrika, 1995)
  - Support regions depend on the regression parameters
  - But straightforward to compute $\hat{F}_b$ and hence the score $S(b)$