Regression modeling of interval censored data with a cure fraction

Olivier Bouaziz (1) and Grégory Nuel (2)

(1) MAP5, Université Paris Descartes, Sorbonne Paris Cité, Paris, France
(2) LPSM, CNRS 7599, UPMC, Sorbonne Université, Paris, France

Interval-censored data with a cure fraction

- T is the time to event of interest.
- We observe a random interval $[L, R]$ where $\mathbb{P}(T \in [L, R]) = 1$.
- Mixed case interval censored observations (with $\delta \in \{0, 1\}$):
  - Left censoring if $0 = L < R < \infty$ $(\delta = 1)$
  - Interval censoring if $0 < L < R < \infty$ $(\delta = 1)$
  - Exact observation if $L = R = T$ $(\delta = 1)$
  - Right censoring if $0 < L < R = \infty$ $(\delta = 0)$.
- Y is a latent variable: $Y = 1$ for susceptibles, $Y = 0$ for non-susceptibles.

Modeling the hazard rate

- The hazard model is:
$$\lambda(t|Y_i = 1, Z_i) = \lambda_0(t)\exp(\beta_0 Z_i),$$
with $\lambda_0(t) = \sum_{\ell=1}^{L} \mathbb{1}_{c_{\ell-1}<t\leq c_\ell}\exp(a_\ell)$ and $c_0 = 0 < c_1 < \cdots < c_L = \infty$.
- The model for susceptibility is:
$$p_i = \mathbb{P}[Y_i = 1|X_i] = \frac{\exp(\gamma_0 X_i)}{1 + \exp(\gamma_0 X_i)}.$$
- The observations are data$=(L_i, R_i, \delta_i, Z_i, X_i)_{i=1,\dots,n}$.
- The unobserved data are $(T_i, Y_i)_{i=1,\dots,n}$.
- The goal is to estimate $\theta = (a_1, \dots, a_L, \beta, \gamma)$.

The EM algorithm

- E-step: The complete likelihood is
$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i}\prod_{i=1}^{n}\{f(T_i|Y_i = 1, Z_i; \boldsymbol{\theta})\}^{Y_i}.$$
Let $\pi_i^{\text{old}} = \mathbb{E}[Y_i|\text{data}, \boldsymbol{\theta}_{\text{old}}]$. We have:
$$\pi_i^{\text{old}} = \delta_i + \frac{(1-\delta_i)p_i^{\text{old}}S(L_i|Y_i = 1, Z_i, \boldsymbol{\theta}_{\text{old}})}{1 - p_i^{\text{old}} + p_i^{\text{old}}S(L_i|Y_i = 1, Z_i, \boldsymbol{\theta}_{\text{old}})},$$
and
$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) &= \mathbb{E}_{T_{1:n},Y_{1:n}|\text{data},\boldsymbol{\theta}_{\text{old}}}[\log(L(\boldsymbol{\theta}))]\\ &= \sum_{i=1}^{n}\left\{\pi_i^{\text{old}}\log(p_i) + (1-\pi_i^{\text{old}})\log(1-p_i)\right\}\\ &+ \sum_{i\text{ not exact}}\pi_i^{\text{old}}\sum_{\ell=1}^{L}\left\{\left(a_{i,\ell} - \sum_{j=1}^{\ell-1}(c_j - c_{j-1})e^{a_{i,j}}\right)A_{\ell,i}^{\text{old}} - e^{a_{i,\ell}}B_{\ell,i}^{\text{old}}\right\}\\ &+ \sum_{i\text{ exact}}\sum_{\ell=1}^{L}\left\{O_{i,l}a_{i,\ell} - \exp(a_{i,\ell})R_{i,\ell}\right\}.\end{aligned}$$
$O_{i,\ell}$ is number of observed events and $R_{i,\ell}$ is total time at risk in cut $(c_{\ell-1}, c_\ell]$ for individual i. $A_{\ell,i}^{\text{old}}, B_{\ell,i}^{\text{old}} = 0$ if $[L_i, R_i] \cap (c_{\ell-1}, c_\ell] = \emptyset$.
- M-step: Newton-Raphson algorithm. The block Hessian for λ_0 is diagonal.

The adaptive ridge procedure

- Penalized log-likelihood:
$$l(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) - \frac{\text{pen}}{2}\sum_{\ell=1}^{L-1}w_l(a_{\ell+1} - a_\ell)^2,$$
where $(w_1, \dots, w_{L-1})$ are non-negative weights, pen is a tuning parameter.
  - $\text{pen} = 0$ corresponds to unpenalized log-likelihood.
  - $\text{pen} = \infty$ corresponds to exponential baseline (no cuts).
- Update of weights (mth step):
$$w_\ell^{(m)} = \left((\hat{a}_{\ell+1}^{(m)} - \hat{a}_\ell^{(m)})^2 + \varepsilon^2\right)^{-1},$$
for $\ell = 1, \dots, L-1$ with $\varepsilon = 10^{-5}$. $\hat{a}_\ell^{(m)}$ is the estimate of a_ℓ obtained from the Newton-Raphson algorithm.
  - $|\hat{a}_{\ell+1}^{(m)} - \hat{a}_\ell^{(m)}| < \varepsilon \implies w_\ell^{(m)}(\hat{a}_{\ell+1}^{(m)} - \hat{a}_\ell^{(m)})^2 \approx 0$.
  - $|\hat{a}_{\ell+1}^{(m)} - \hat{a}_\ell^{(m)}| > \varepsilon \implies w_\ell^{(m)}(\hat{a}_{\ell+1}^{(m)} - \hat{a}_\ell^{(m)})^2 \approx 1$.

  Approximation of the L0 norm!
- The block Hessian for λ_0 is tri-diagonal. Using the R **bandsolve** package, the total complexity for the inversion of the Hessian matrix is $O(L)$.
- pen is chosen from the Bayesian Information Criteria.
- For the adaptive ridge, see: F. Frommlet and G. Nuel. *An adaptive ridge procedure for L0 regularization*. PLoS ONE, 11(2):1-23, 2016.

Data example 1: HIV infection in Danish homosexual men

- 297 people were followed up at six different dates: December 1981, April 1982, February 1983, September 1984, April 1987 and May 1989.
- T is time to HIV infection in calendar days.
- Observations in percentage

exact	left-censored	interval-censored	right-censored
0.00	08.75	13.13	78.12

- Results from standard Cox model (6 fixed cuts for the baseline, no cure)

Covariates	Hazard ratio ($e^{\hat\beta}$)	p-value
Nb. of partner/year	1.01	0.0498
Contact with USA	1.66	0.0207

  - Non-parametric survival probability in 1990 is estimated to: 71%.
  - See B. Carstensen. *Regression models for interval censored survival data: application to HIV infection in Danish homosexual men*. Statistics in Medicine, 15:2177-2189, 1996.
- Results from the adaptive ridge Cox model with cure fraction

Covariates	Hazard ratio ($e^{\hat\beta}$)	p-value	Odd ratio ($e^{\hat\gamma}$)	p-value
Nb. of partner/year	1.00	0.5658	1.02	0.0096
Contact with USA	1.62	0.2296	1.57	0.2310

  - The adaptive ridge selects the exponential baseline (no cuts)!
  - Nb. of partner/year is highly significant for the probability to be susceptible!
  - No significant effect on nb. of partner/year and visiting the USA on the hazard risk of HIV for the susceptibles!
  - Non-parametric probability of being susceptible: $\hat{p} = 0.29$.

Data example 2: replantation of 400 avulsed permanent teeth

- T is time from replantation to ankylosis.
- The goal is to study the effect on T of
  - stage of root formation
    72.5% mature teeth, 27.5% immature teeth
  - length of extra-alveolar storage
    Mean time: 30.86 seconds
  - type of storage media
    85.25% physiologic storage, 14.75% non-physiologic storage
  - age of the patient (in interaction with mature teeth only)
    Mean age for mature teeth: 16.81.
- Observations in percentage

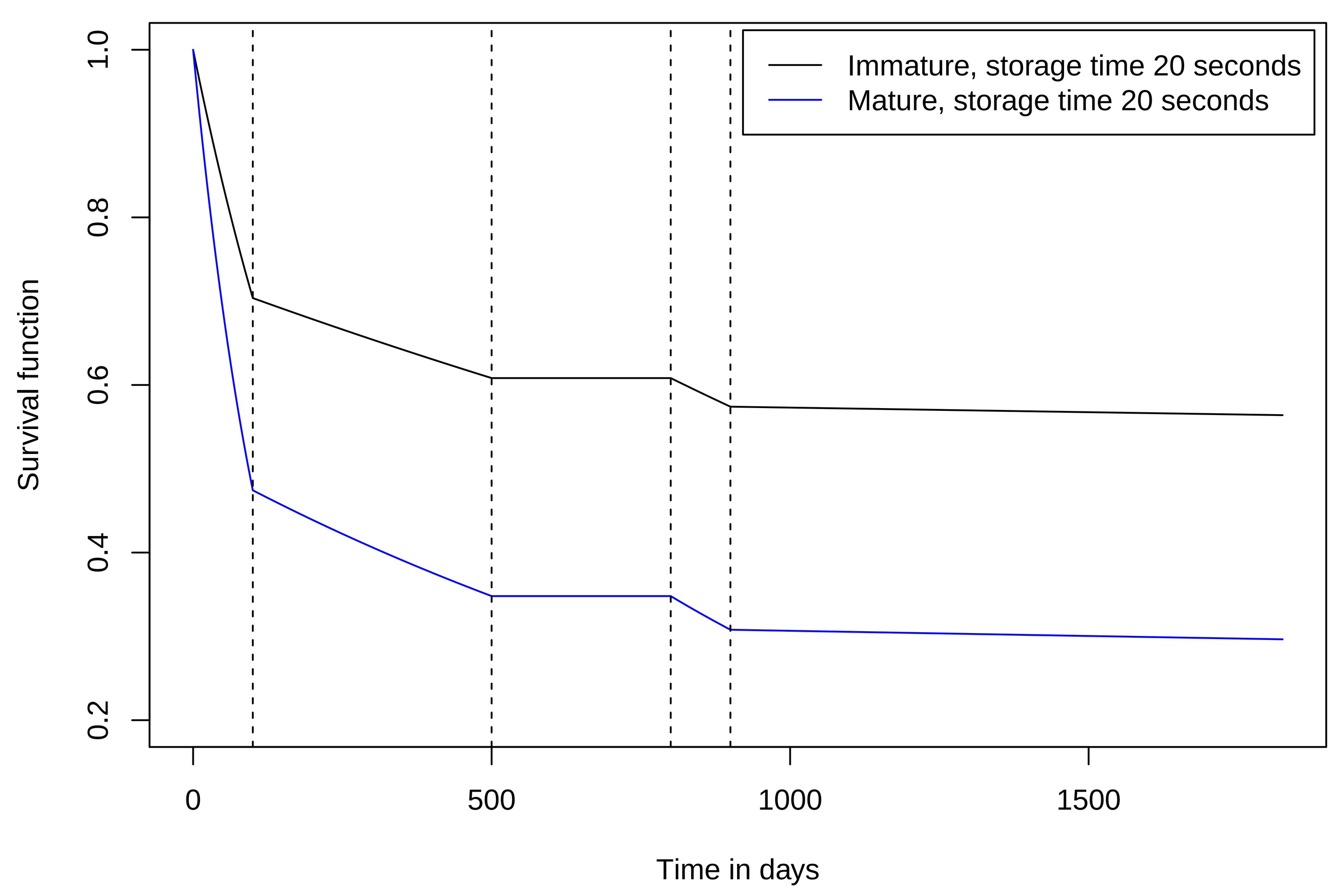
exact	left-censored	interval-censored	right-censored
0.00	28.00	35.75	36.25

- Results from the Cox model

Covariates	Hazard ratio	p-value
Mature	2.00	1.89×10^{-5}
Storage time (in min)	1.23	0.0017
Physiologic	0.93	0.6980
Age>20 (for mature teeth)	1.27	0.1272

- $\hat{p} = 1$: all patients are susceptible to ankylosis!
- The cuts found from the adaptive ridge method are: 100, 500, 800, 900.

Survival estimate of time to ankylosis for mature and immature teeth



Contact: olivier.bouaziz@parisdescartes.fr