Penalized estimation methods for time to event data based on the adaptive-ridge procedure

Olivier Bouaziz[1]
with Vivien Goepp[1], Eva Lauridsen[2], Grégory Nuel[3] and Jean-Christophe Thalabard[1]

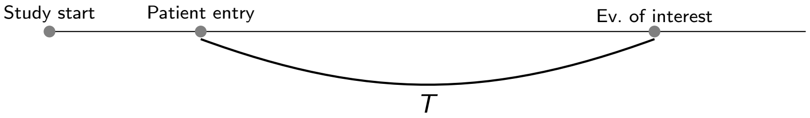
[1]MAP5 (CNRS 8145), Université de Paris
[2]Ressource Center for Rare Oral Diseases Copenhagen University Hospital, Rigshospitalet, Denmark
[3]LPSM (CNRS 8001), Sorbonne Université, Paris

Seminar at the Department of Mathematics and Computer Science
University of Southern Denmark, Odense

Outline

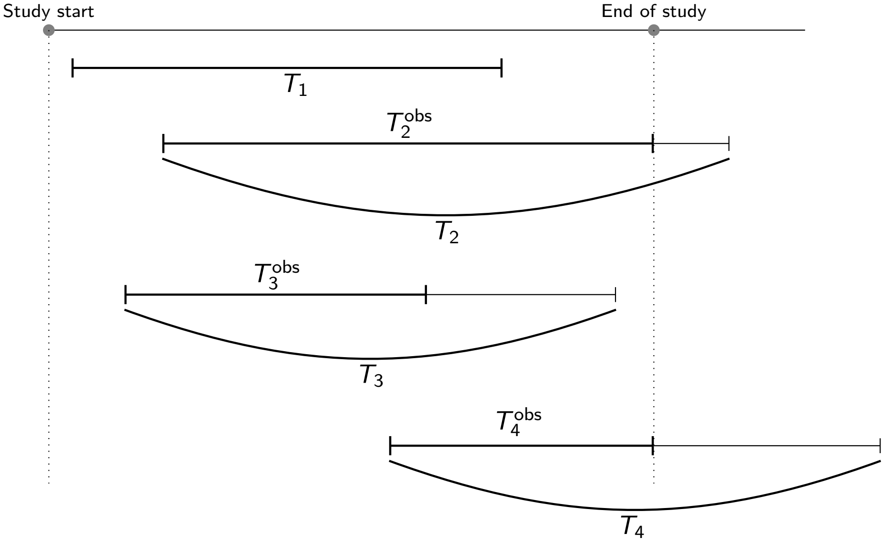
Background in time to event analysis

- We study a positive continuous time to event variable T.
- T represents the time difference between event of interest and patient entry.



- Examples : time to relapse of Leukemia patients, time to onset of cancer, time to death ...

Background in time to event analysis : right censoring



The hazard rate

- Observations :
$$\begin{cases} T_i^{\text{obs}} = T_i \wedge C_i \\ \Delta_i = \mathbb{1}_{T_i \leq C_i} \end{cases}$$

- Independent censoring : $T \perp\!\!\!\perp C$

- A key relation :

$$\begin{aligned} \lambda(t) &:= \lim_{\triangle t \to 0} \frac{\mathbb{P}[t \leq T < t + \triangle t \mid T \geq t]}{\triangle t} \\ &= \lim_{\triangle t \to 0} \frac{\mathbb{P}[t \leq T^{\text{obs}} < t + \triangle t, \Delta = 1 \mid T^{\text{obs}} \geq t]}{\triangle t}. \end{aligned}$$

Many estimators (Nelson Aalen, Kaplan-Meier, ...) are based on this relation.

Likelihood and the Cox model

- The likelihood of the observed data is equal to :

$$\prod_{i=1}^{n} f(T_i^{\text{obs}})^{\Delta_i} S(T_i^{\text{obs}})^{1-\Delta_i} = \prod_{i=1}^{n} \lambda(T_i^{\text{obs}})^{\Delta_i} \exp\left(-\int_0^{T_i^{\text{obs}}} \lambda(t)dt\right),$$

where f is the density of T and $S(t) = \mathbb{P}[T > t]$.

- Regression modelling : let $Z \in \mathbb{R}^d$ be a covariate.

$$\lambda(t \mid Z_i) = \lambda_0(t) \exp(\beta Z_i) \qquad \text{(Cox Model)}$$

For a binary covariate,

$$\frac{\lambda(t \mid Z_i = 1)}{\lambda(t \mid Z_i = 0)} = \exp(\beta).$$

Outline

The piecewise constant hazard model

- The model :

$$\lambda(t) = \sum_{k=1}^{K} \lambda_k \mathbb{1}_{c_{k-1} < t \leq c_k}$$

- Goal : estimate the λ_ks.

The log-likelihood is equal to :

$$\ell_n(\boldsymbol{\lambda}) = \sum_{k=1}^{K} \left\{ \bar{O}_k \log(\lambda_k) - \lambda_k \bar{R}_k \right\},$$

where

- $\bar{O}_k = \sum_i \Delta_i \mathbb{1}_{c_{k-1} < T_i^{\text{obs}} \leq c_k}$: number of observed events in interval $(c_{k-1}, c_k]$
- $\bar{R}_k = \sum_i (T_i^{\text{obs}} \wedge c_k - c_{k-1}) \mathbb{1}_{T_i^{\text{obs}} > c_{k-1}}$: total time at risk in interval $(c_{k-1}, c_k]$

The piecewise constant hazard model

- $\bar{O}_k$: number of observed events in interval $(c_{k-1}, c_k]$
- $\bar{R}_k$: total time at risk in interval $(c_{k-1}, c_k]$

The maximum likelihood estimator is explicit :

$$\hat{\lambda}_k^{\mathrm{mle}} = \frac{\bar{O}_k}{\bar{R}_k}$$

O. Aalen, Ø. Borgan, H. Gjessing, *Survival and Event History Analysis*. (2008)

The piecewise constant hazard model

- $\bar{O}_k$: number of observed events in interval $(c_{k-1}, c_k]$
- $\bar{R}_k$: total time at risk in interval $(c_{k-1}, c_k]$

The maximum likelihood estimator is explicit :

$$\hat{\lambda}_k^{\text{mle}} = \frac{\bar{O}_k}{\bar{R}_k}$$

O. Aalen, Ø. Borgan, H. Gjessing, *Survival and Event History Analysis*. (2008)

- We want to choose the number and location of the cuts from the data
- We start from a large grid of cuts ($K = 100, 1\,000, \ldots$)
- We use a penalization technique to constrain similar adjacent hazard values to be equal.

Penalizing the maximum likelihood estimator

Set $\log \lambda_k = a_k$. Estimation of $\boldsymbol{a}$ is achieved through penalized log-likelihood :

$$\ell_n^{\text{pen}}(\boldsymbol{a}) = \underbrace{\ell_n(\boldsymbol{a})}_{\text{log-likelihood}}$$

Penalizing the maximum likelihood estimator

Set $\log \lambda_k = a_k$. Estimation of $\boldsymbol{a}$ is achieved through penalized log-likelihood :

$$\ell_n^{\text{pen}}(\boldsymbol{a}) = \underbrace{\ell_n(\boldsymbol{a})}_{\text{log-likelihood}} - \underbrace{\frac{\text{pen}}{2}\left\{\sum_{k=1}^{K-1} w_k \left(a_{k+1} - a_k\right)^2\right\}}_{\text{regularization term}},$$

- $\boldsymbol{w}$ represents a weight.
- pen is a penalty term.

Two types of regularization

1. L_2 regularization (Ridge) with $\boldsymbol{w} = 1$.
2. L_0 regularization with the iterative adaptive ridge procedure.
   At the m^{th} step, we update the weights

$$w_k^{(m-1)} = \left(\left(a_{k+1}^{(m-1)} - a_k^{(m-1)} \right)^2 + \varepsilon^2 \right)^{-1},$$

   with $\varepsilon \ll 1$, and we maximize with respect to $\boldsymbol{a}$

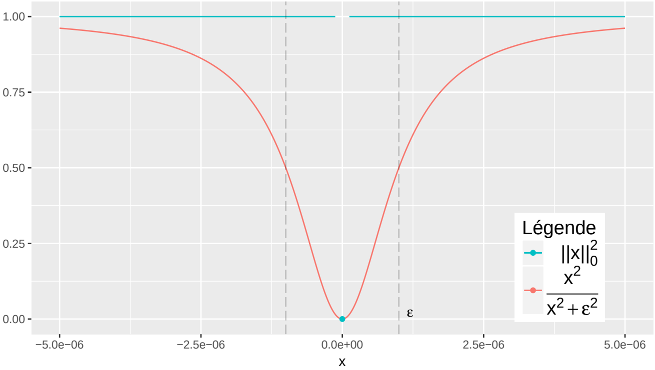
$$\ell_n^{\text{pen}}(\boldsymbol{a}) = \ell_n(\boldsymbol{a}) - \frac{\text{pen}}{2} \left\{ \sum_{k=1}^{K-1} w_k^{(m-1)} \left(a_{k+1} - a_k \right)^2 \right\}.$$

F. Frommlet and G. Nuel, *An Adaptive Ridge Procedure for L_0 Regularization*. **PlosOne** (2016).

L_0 norm approximation - Heuristic

When $\varepsilon \ll 1$,

$$\frac{(a_{k+1} - a_k)^2}{(a_{k+1} - a_k)^2 + \varepsilon^2} \simeq \|a_{k+1} - a_k\|_0^2 = \begin{cases} 0 & \text{if } a_{k+1} = a_k \\ 1 & \text{if } a_{k+1} \neq a_k \end{cases}$$



Maximization of the penalized log-likelihood

- The penalized estimator is no longer explicit.
- Maximization is performed from the Newton-Raphson algorithm. For a given sequence of weights $\boldsymbol{w}$, the ℓth Newton Raphson iteration step is obtained from the equation

$$\boldsymbol{a}^{(\ell)} = \boldsymbol{a}^{(\ell-1)} + I(\boldsymbol{a}^{(\ell-1)}, \boldsymbol{w})^{-1} U(\boldsymbol{a}^{(\ell-1)}, \boldsymbol{w}),$$

  where I is the opposite of the Hessian matrix, U is the score vector.

- The Hessian matrix is tri-diagonal.
- $\Longrightarrow$ computation time for the inversion of the Hessian is $\mathcal{O}(K)$

The *Adaptive Ridge* procedure for a given penalty

procedure $\textsc{Adaptive-Ridge}(\boldsymbol{O}, \boldsymbol{R}, \text{pen})$

  $(\boldsymbol{a}, \boldsymbol{w}, \textbf{sel}) \leftarrow (0,1,0)$

  while not converge **do**

    $\boldsymbol{a}^{\text{new}} \leftarrow \textsc{Newton-Raphson}(\boldsymbol{O}, \boldsymbol{R}, \text{pen}, \boldsymbol{a}, \boldsymbol{w})$

    $w_k^{\text{new}} \leftarrow \left((a_{k+1}^{\text{new}} - a_k^{\text{new}})^2 + \varepsilon^2\right)^{-1}$

    $\text{sel}_k^{\text{new}} \leftarrow w_k^{\text{new}} \left(a_{k+1}^{\text{new}} - a_k^{\text{new}}\right)^2$

    $(\boldsymbol{a}, \boldsymbol{w}, \textbf{sel}) \leftarrow (\boldsymbol{a}^{\text{new}}, \boldsymbol{w}^{\text{new}}, \textbf{sel}^{\text{new}})$

  end while

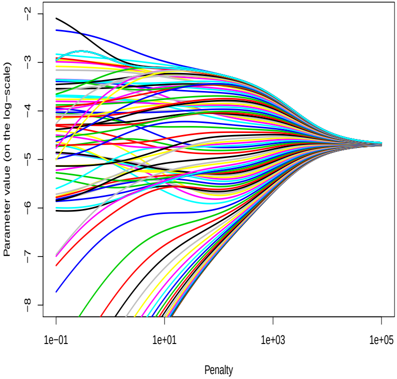
end procedure

The *Adaptive Ridge* procedure for a given penalty

procedure $\textsc{Adaptive-Ridge}(\boldsymbol{O}, \boldsymbol{R}, \text{pen})$
  $(\boldsymbol{a}, \boldsymbol{w}, \textbf{sel}) \leftarrow (0,1,0)$
  while not converge **do**
    $\boldsymbol{a}^{\text{new}} \leftarrow \textsc{Newton-Raphson}(\boldsymbol{O}, \boldsymbol{R}, \text{pen}, \boldsymbol{a}, \boldsymbol{w})$
    $w_k^{\text{new}} \leftarrow \left((a_{k+1}^{\text{new}} - a_k^{\text{new}})^2 + \varepsilon^2\right)^{-1}$
    $\text{sel}_k^{\text{new}} \leftarrow w_k^{\text{new}} (a_{k+1}^{\text{new}} - a_k^{\text{new}})^2$
    $(\boldsymbol{a}, \boldsymbol{w}, \textbf{sel}) \leftarrow (\boldsymbol{a}^{\text{new}}, \boldsymbol{w}^{\text{new}}, \textbf{sel}^{\text{new}})$
  end while
  $\widehat{\textbf{cuts}} \leftarrow \textbf{cuts}[\textbf{sel} > 0.99]$
  Compute $(\boldsymbol{O}^{\widehat{\textbf{cuts}}}, \boldsymbol{R}^{\widehat{\textbf{cuts}}})$
  $\exp(\hat{\boldsymbol{a}}^{\textbf{mle}}) \leftarrow \boldsymbol{O}^{\widehat{\textbf{cuts}}} / \boldsymbol{R}^{\widehat{\textbf{cuts}}}$
  return $\hat{\boldsymbol{a}}^{\textbf{mle}}$
end procedure

Comparison of the two regularization methods

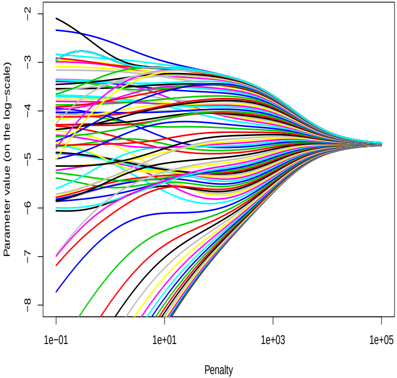
$$\text{pen} = 0 \quad \Longrightarrow \quad \widehat{\boldsymbol{a}} = \widehat{\boldsymbol{a}}^{\text{mle}}$$
$$\text{pen} = \infty \quad \Longrightarrow \quad \widehat{\boldsymbol{a}} = \text{ constant}$$



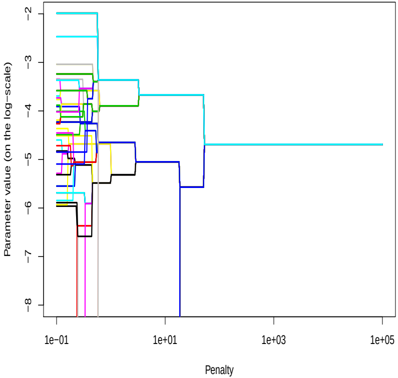
L_2 regularization

Comparison of the two regularization methods

$$\text{pen} = 0 \quad \Longrightarrow \quad \widehat{\boldsymbol{a}} = \widehat{\boldsymbol{a}}^{\text{mle}}$$
$$\text{pen} = \infty \quad \Longrightarrow \quad \widehat{\boldsymbol{a}} = \text{ constant}$$

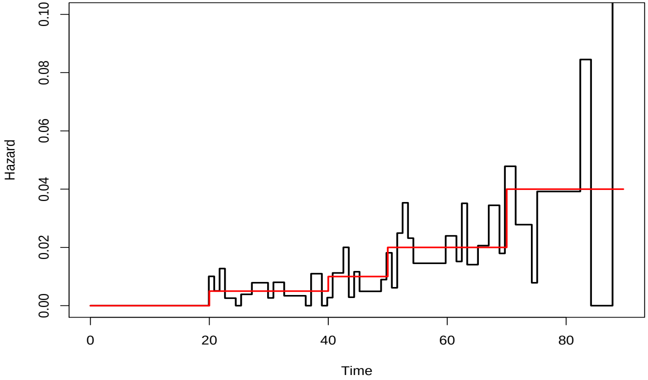


L_2 regularization



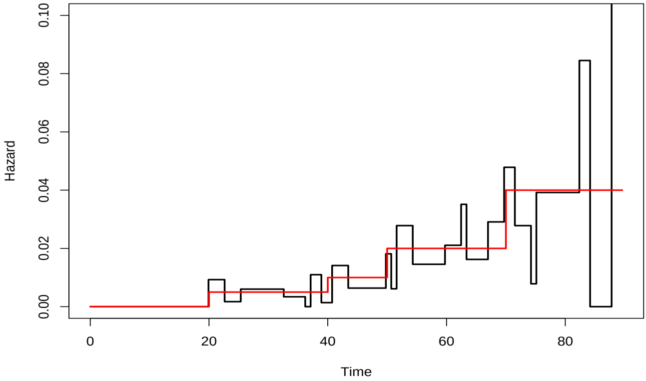
L_0 regularization

Model selection for the *Adaptive Ridge* estimator ($n = 400$)



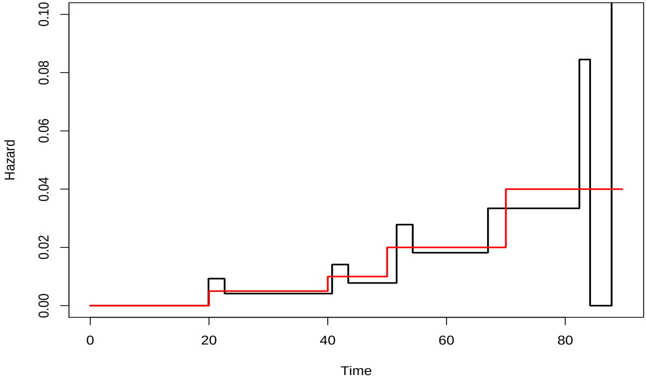
- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 0.1$

Model selection for the *Adaptive Ridge* estimator ($n = 400$)



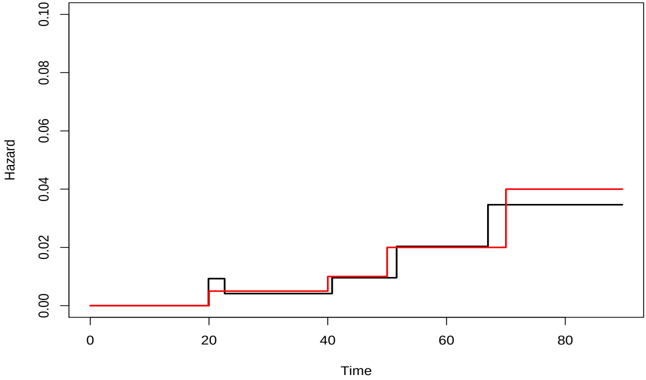
- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 0.27$

Model selection for the *Adaptive Ridge* estimator ($n = 400$)



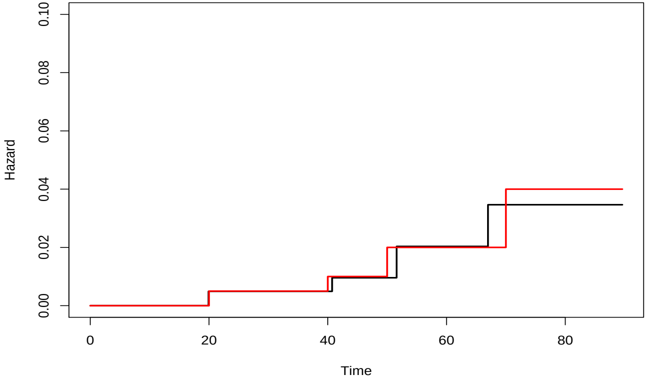
- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 0.55$

Model selection for the *Adaptive Ridge* estimator ($n = 400$)



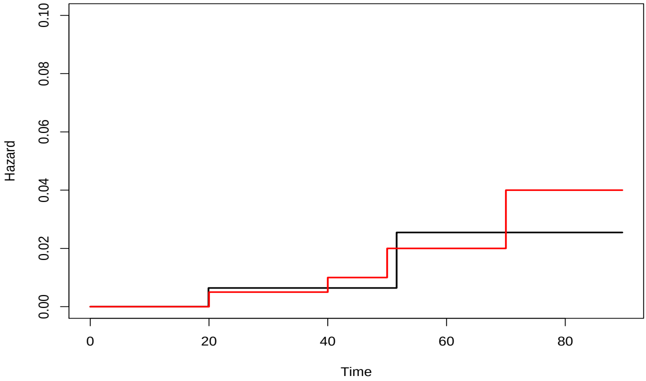
- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 0.77$

Model selection for the *Adaptive Ridge* estimator ($n = 400$)



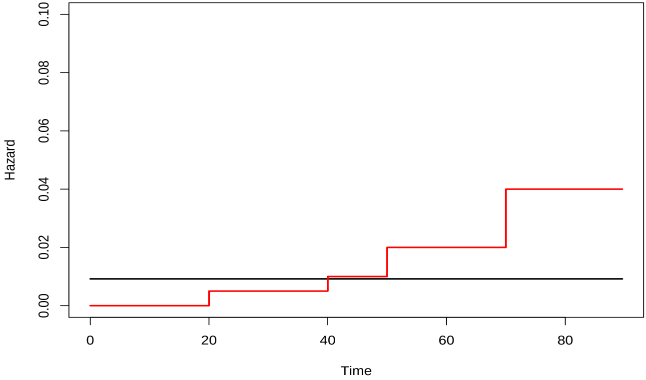
- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 1.54$

Model selection for the *Adaptive Ridge* estimator ($n = 400$)



- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 6.16$

Model selection for the *Adaptive Ridge* estimator ($n = 400$)



- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 52.70$

Model selection for the *Adaptive Ridge* estimator

Three different methods to perform model selection :

1. $\text{BIC}(D) = -2\ell_n(\widehat{\boldsymbol{a}}_D^{\text{mle}}) + D \log n$
2. $\text{AIC}(D) = -2\ell_n(\widehat{\boldsymbol{a}}_D^{\text{mle}}) + 2D$
3. K-fold Cross Validation (CV),

with D the dimension of the model :

$$D = \sum_{k=0}^{K-1} \mathbb{1}\{\hat{a}_{k+1,D}^{\text{mle}} - \hat{a}_{k,D}^{\text{mle}} \neq 0\}.$$

Model selection for the *Adaptive Ridge* estimator

Three different methods to perform model selection :

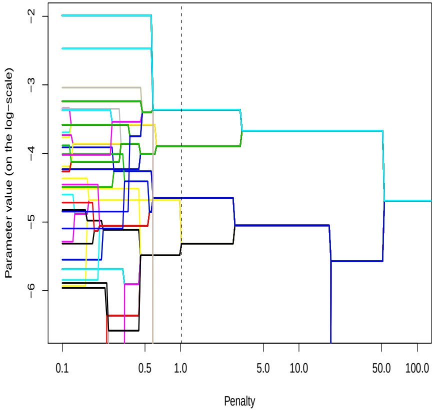
1. $\text{BIC}(D) = -2\ell_n(\widehat{\boldsymbol{a}}_D^{\text{mle}}) + D \log n$
2. $\text{AIC}(D) = -2\ell_n(\widehat{\boldsymbol{a}}_D^{\text{mle}}) + 2D$
3. K-fold Cross Validation (CV),

with D the dimension of the model :

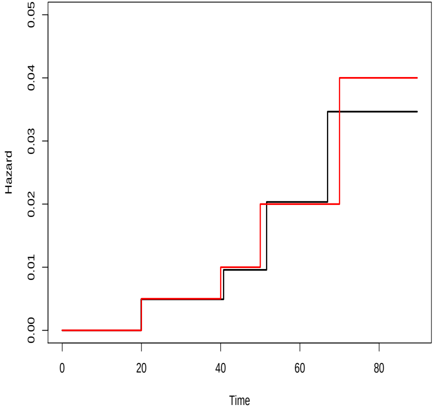
$$D = \sum_{k=0}^{K-1} \mathbb{1}\{\hat{a}_{k+1,D}^{\text{mle}} - \hat{a}_{k,D}^{\text{mle}} \neq 0\}.$$

- O. Bouaziz and G. Nuel, *L_0 regularization for the estimation of piecewise constant hazard rates in survival analysis*. **Applied Mathematics** (2017).
- Package **pchsurv** available on GitHub : `install_github("obouaziz/pchsurv")`

Model selection for the *Adaptive Ridge* estimator using the BIC ($n = 400$)



Regularization path



Hazard estimator (in black)

Outline

The adaptive ridge as an approximation for the L_0 "norm"

- Let $\Delta a_k := a_{k+1} - a_k$, $k = 1, \ldots, K-1$, and $\|\Delta \boldsymbol{a}\|_0 := \sum_{k=1}^{K-1} \mathbb{1}_{\Delta a_k \neq 0}$.
- For $\beta \in \mathbb{R}, \varepsilon > 0$, let

$$p(\beta) := \frac{\log(1 + \beta^2/\varepsilon^2)}{\log(1 + 1/\varepsilon^2)} \xrightarrow[\varepsilon \to 0]{} \mathbb{1}_{\beta \neq 0}.$$

We have : $\sum_{k=1}^{K-1} p(\Delta a_k) \xrightarrow[\varepsilon \to 0]{} \|\Delta \boldsymbol{a}\|_0$.

Theorem (V. Goepp, J-C. Thalabard, G. Nuel and O. Bouaziz)

The adaptive-ridge algorithm solves the maximization problem :

$$\hat{\boldsymbol{a}} = \arg\max_{\boldsymbol{a}} \tilde{\ell}_n^{\text{pen}}(\boldsymbol{a}),$$

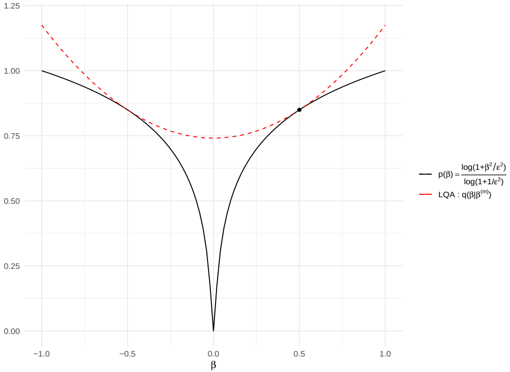
with $\tilde{\ell}_n^{\text{pen}}(\boldsymbol{a}) := \ell_n(\boldsymbol{a}) - \kappa \sum_{k=1}^{K-1} p(\Delta a_k)$, $\kappa > 0$.

Local Quadratic Approximation (LQA)

Local Quadratic Approximation (see **J. Fan and R. Li 2001, D. R. Hunter and R. Li 2005**) of $\mathrm{p}(\beta)$. We prove that for all $\beta^{(m)} \in \mathbb{R}$, for all $\beta \in \mathbb{R}$,

$$\mathrm{p}(\beta) \leq \mathrm{q}(\beta \mid \beta^{(\mathrm{m})}) := \frac{\log(1 + (\beta^{(\mathrm{m})})^2/\varepsilon^2)}{\log(1 + 1/\varepsilon^2)} + \frac{\beta^2 - (\beta^{(\mathrm{m})})^2}{\varepsilon^2 + (\beta^{(\mathrm{m})})^2} \cdot \frac{1}{\log(1 + 1/\varepsilon^2)},$$

with $q(\beta^{(m)} \mid \beta^{(m)}) = \mathrm{p}(\beta^{(\mathrm{m})})$. $\qquad (\beta^{(m)} = 0.5$ and $\varepsilon = 10^{-2}$ in the plot)



The adaptive ridge as an MM algorithm

Minorize-Maximization (MM) algorithm

- For $\kappa > 0$, we have

$$\tilde{\ell}_n^{\text{pen}}(\boldsymbol{a}) = \ell_n(\boldsymbol{a}) - \kappa \sum_{k=1}^{K-1} p(\Delta a_k) \geq \underbrace{\ell_n(\boldsymbol{a}) - \kappa \sum_{k=1}^{K-1} q(\Delta a_k \mid \Delta a_k^{(m)})}_{g(\boldsymbol{a}|\boldsymbol{a}^{(m)})},$$

  with $g(\boldsymbol{a}^{(m)} \mid \boldsymbol{a}^{(m)}) = \tilde{\ell}_n^{\text{pen}}(\boldsymbol{a}^{(m)})$.
- Let $\boldsymbol{a}^{(m+1)} = \arg\max_{\boldsymbol{a}} g(\boldsymbol{a} \mid \boldsymbol{a}^{(m)})$. Then :

$$\tilde{\ell}_n^{\text{pen}}(\boldsymbol{a}^{(m+1)}) \geq g(\boldsymbol{a}^{(m+1)} \mid \boldsymbol{a}^{(m)}) \geq g(\boldsymbol{a}^{(m)} \mid \boldsymbol{a}^{(m)}) = \tilde{\ell}_n^{\text{pen}}(\boldsymbol{a}^{(m)}).$$

- $\boldsymbol{a}^{(m+1)} = \arg\max_{\boldsymbol{a}} g(\boldsymbol{a} \mid \boldsymbol{a}^{(m)})$ is the update obtained from our adaptive-ridge algorithm ! The adaptive-ridge algorithm solves the maximization problem :

$$\hat{\boldsymbol{a}} = \arg\max_{\boldsymbol{a}} \tilde{\ell}_n^{\text{pen}}(\boldsymbol{a}).$$

Outline

The SEER data

- Huge american registry dataset of breast cancer **https ://seer.cancer.gov**
- Primary, unilateral, malignant and invasive cancers
- 1.2 million of patients, 60% of censoring
- The cancer diagnosis range from 1973 to 2014
- The time from cancer diagnosis to death or censoring ranges from 0 to 41 years.
- The variable of interest is the time from cancer diagnosis until death.

Aim : estimate the hazard of death as a function of both date of cancer diagnosis and time since diagnosis.

- We use the adaptive ridge procedure
- Penalization over the two directions.

The SEER data

- Huge american registry dataset of breast cancer **https ://seer.cancer.gov**
- Primary, unilateral, malignant and invasive cancers
- 1.2 million of patients, 60% of censoring
- The cancer diagnosis range from 1973 to 2014
- The time from cancer diagnosis to death or censoring ranges from 0 to 41 years.
- The variable of interest is the time from cancer diagnosis until death.

Aim : estimate the hazard of death as a function of both date of cancer diagnosis and time since diagnosis.

- We use the adaptive ridge procedure
- Penalization over the two directions.

- V. Goepp, J-C. Thalabard, G. Nuel and O. Bouaziz. *Regularized Bidimensional Estimation of the Hazard Rate*. **To appear in International Journal of Biostatistics**.
- Package **hazreg** available on GitHub : `install_github("goepp/hazreg")`

The SEER data

- $\lambda_{j,k}$: true hazard in rectangle (j,k)
- $O_{j,k}$: number of observed events in rectangle (j,k)
- $R_{j,k}$: total time at risk in rectangle (j,k)

The log-likelihood is equal to :

$$\ell_n(\boldsymbol{\lambda}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \left\{ O_{j,k} \log\left(\lambda_{j,k}\right) - \lambda_{j,k} R_{j,k} \right\}$$

Set $\log \lambda_{j,k} = a_{j,k}$. Estimation of $\boldsymbol{a}$ through penalized log-likelihood :

$$\ell_n^{\text{pen}}(\boldsymbol{a}) = \underbrace{\ell_n(\boldsymbol{a})}_{\text{log-likelihood}}$$

The SEER data

- $\lambda_{j,k}$: true hazard in rectangle (j,k)
- $O_{j,k}$: number of observed events in rectangle (j,k)
- $R_{j,k}$: total time at risk in rectangle (j,k)

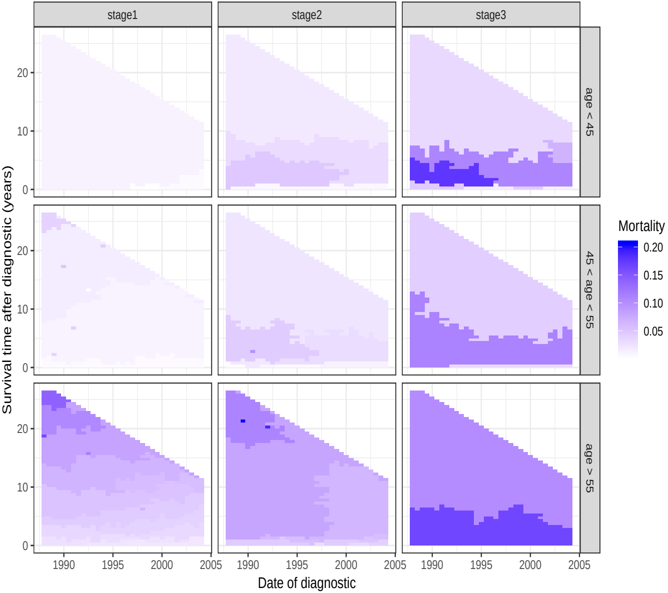
The log-likelihood is equal to :

$$\ell_n(\boldsymbol{\lambda}) = \sum_{j=1}^{J}\sum_{k=1}^{K}\left\{O_{j,k}\log\left(\lambda_{j,k}\right) - \lambda_{j,k}R_{j,k}\right\}$$

Set $\log\lambda_{j,k} = a_{j,k}$. Estimation of $\boldsymbol{a}$ through penalized log-likelihood :

$$\ell_n^{\text{pen}}(\boldsymbol{a}) = \underbrace{\ell_n(\boldsymbol{a})}_{\text{log-likelihood}} \underbrace{- \frac{\text{pen}}{2}\sum_{j,k}\left\{v_{j,k}\left(a_{j+1,k} - a_{j,k}\right)^2 + w_{j,k}\left(a_{j,k+1} - a_{j,k}\right)^2\right\}}_{\text{regularization term}}.$$

The SEER data



Outline

The dental dataset

Data collected from Eva Lauridsen at the hospital Rigshospitalet (Denmark).

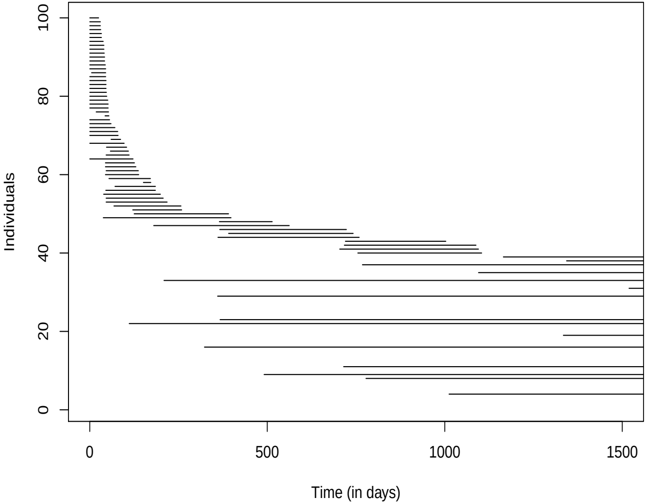
- Study of 322 patients with 400 avulsed and replanted permanent teeth from 1965 to 1988.
- The variable of interest is time from replantation until the ankylosis complication.
- Patients are examined at intermittent visits to the dentist.
  - Left-censoring (28%) if ankylosis occurred before the first visit.
  - Interval-censoring (35.75%) if ankylosis occurred between two visits.
  - Right-censoring (36.25%) if ankylosis did not occur yet after the last visit.

The dental dataset

Data collected from Eva Lauridsen at the hospital Rigshospitalet (Denmark).

- Study of 322 patients with 400 avulsed and replanted permanent teeth from 1965 to 1988.
- The variable of interest is time from replantation until the ankylosis complication.
- Patients are examined at intermittent visits to the dentist.
  - Left-censoring (28%) if ankylosis occurred before the first visit.
  - Interval-censoring (35.75%) if ankylosis occurred between two visits.
  - Right-censoring (36.25%) if ankylosis did not occur yet after the last visit.
- Covariates :
  - stage of root formation : 72.5% mature teeth, 27.5% immature teeth
  - length of extra-alveolar storage : mean time is 30.9 minutes
  - type of storage media : 85.25% physiologic, 14.75% non physiologic
  - age of the patient : mean age for mature teeth is 16.81 years

The raw data on a subsample of size 100



The observed likelihood

The observations are L_i, R_i, $i = 1, \ldots, n$.

- $0 = L_i < R_i < +\infty$ for left-censored observation ($\delta_i = 1$)
- $0 < L_i < R_i < +\infty$ for interval-censored observation ($\delta_i = 1$)
- $0 < L_i < R_i = +\infty$ for right-censored observation ($\delta_i = 0$)

With these types of data, the observed likelihood is equal to :

$$\mathrm{L}^{\mathrm{obs}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})\}^{\delta_i} \times \Big\{S(L_i \mid Z_i, \boldsymbol{\theta})\Big)\Big\}^{1-\delta_i}.$$

The observed likelihood

The observations are L_i, R_i, $i = 1, \ldots, n$.

- $0 = L_i < R_i < +\infty$ for left-censored observation ($\delta_i = 1$)
- $0 < L_i < R_i < +\infty$ for interval-censored observation ($\delta_i = 1$)
- $0 < L_i < R_i = +\infty$ for right-censored observation ($\delta_i = 0$)

With these types of data, the observed likelihood is equal to :

$$\mathrm{L}^{\text{obs}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \exp\left(- \int_0^{L_i} \lambda_0(t) dt e^{\beta Z_i}\right) \left(1 - \exp\left(- \int_{L_i}^{R_i} \lambda_0(t) dt e^{\beta Z_i}\right)\right) \right\}^{\delta_i} \times \left\{ \exp\left(- \int_0^{L_i} \lambda_0(t) dt e^{\beta Z_i}\right) \right\}^{1-\delta_i},$$

for the Cox model $\lambda(t \mid Z_i) = \lambda_0(t) \exp(\beta Z_i)$.

The observed likelihood

- The piecewise constant model for the baseline :

$$\lambda_0(t) = \sum_{k=1}^{K} \exp(a_k) \mathbb{1}_{c_{k-1} < t \leq c_k}$$

- The model parameter is : $\boldsymbol{\theta} = (a_1, \ldots, a_K, \beta) \in \mathbb{R}^{K+d}$

Maximization of :

$$\mathrm{L}^{\mathrm{obs}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \exp\left(-\int_0^{L_i} \lambda_0(t) dt e^{\beta Z_i}\right) \left(1 - \exp\left(-\int_{L_i}^{R_i} \lambda_0(t) dt e^{\beta Z_i}\right)\right) \right\}^{\delta_i} \times \left\{ \exp\left(-\int_0^{L_i} \lambda_0(t) dt e^{\beta Z_i}\right) \right\}^{1-\delta_i},$$

requires to use the Newton-Raphson algorithm.

- The Hessian is of full rank !
- Intractable solution if K is large !

The EM algorithm

The complete likelihood is defined as

$$\mathrm{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(T_i \mid Z_i, \boldsymbol{\theta}).$$

Introduce data $= (L_i, R_i, Z_i)$.

- E-step :

$$\mathbb{E}[\log(f(T_i \mid Z_i, \boldsymbol{\theta}))|\text{data}, \boldsymbol{\theta}_{\text{old}}] = \int f(t \mid \text{data}, \boldsymbol{\theta}_{\text{old}}) \log f(t \mid Z_i, \boldsymbol{\theta}) dt$$

- Under the assumptions
  - $\mathbb{P}(T \in [L, R]) = 1$,
  - $\mathbb{P}(T \le t \mid L = l, R = r, Z) = \mathbb{P}(T \le t \mid l \le T \le r, Z)$ (see Zhang, Sun, Zhao, and Sun, **Canadian J. of Stat.**, 2005),

  we have

$$f(t \mid \text{data}, \boldsymbol{\theta}_{\text{old}}) = \frac{f(t \mid Z_i, \boldsymbol{\theta}_{\text{old}}) \mathbb{1}(L_i < t < R_i)}{S(L_i \mid Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i \mid Z_i, \boldsymbol{\theta}_{\text{old}})}.$$

Using the EM algorithm

- The M-step corresponds of maximizing, with respect to $\boldsymbol{\theta}$,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\mathrm{old}}) &:= \mathbb{E}_{T_{1:n}|\mathrm{data},\boldsymbol{\theta}_{\mathrm{old}}}[\log(\mathrm{L}(\boldsymbol{\theta}))] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\left\{\left(a_{i,k} - \sum_{j=1}^{k-1}(c_j - c_{j-1})e^{a_{i,j}}\right)A_{k,i}^{\mathrm{old}} - e^{a_{i,k}}B_{k,i}^{\mathrm{old}}\right\}, \end{aligned}$$

  with $a_{i,k} := a_k + \beta Z_i$ and with explicit expressions of $A_{k,i}^{\mathrm{old}}$ and $B_{k,i}^{\mathrm{old}}$.

- $A_{k,i}^{\mathrm{old}}$ and $B_{k,i}^{\mathrm{old}}$ depend only on $\boldsymbol{\theta}_{\mathrm{old}}, L_i, R_i, Z_i$.

Using the EM algorithm

- The M-step corresponds of maximizing, with respect to $\boldsymbol{\theta}$,

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) &:= \mathbb{E}_{T_{1:n}|\text{data},\boldsymbol{\theta}_{\text{old}}}[\log(\mathrm{L}(\boldsymbol{\theta}))] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\left\{\left(a_{i,k} - \sum_{j=1}^{k-1}(c_j - c_{j-1})e^{a_{i,j}}\right)A_{k,i}^{\text{old}} - e^{a_{i,k}}B_{k,i}^{\text{old}}\right\},
\end{aligned}
$$

  with $a_{i,k} := a_k + \beta Z_i$ and with explicit expressions of $A_{k,i}^{\text{old}}$ and $B_{k,i}^{\text{old}}$.
- $A_{k,i}^{\text{old}}$ and $B_{k,i}^{\text{old}}$ depend only on $\boldsymbol{\theta}_{\text{old}}, L_i, R_i, Z_i$.
- In the absence of covariates ($Z_i = 0$, $a_{i,k} = a_k$, $\boldsymbol{\theta} = (a_1, \ldots, a_K)$) : the M-step is explicit.

Using the EM algorithm

- The M-step corresponds of maximizing, with respect to $\boldsymbol{\theta}$,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) := \mathbb{E}_{T_{1:n}|\text{data},\boldsymbol{\theta}_{\text{old}}}[\log(\mathrm{L}(\boldsymbol{\theta}))]$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K}\left\{\left(a_{i,k} - \sum_{j=1}^{k-1}(c_j - c_{j-1})e^{a_{i,j}}\right)A_{k,i}^{\text{old}} - e^{a_{i,k}}B_{k,i}^{\text{old}}\right\},$$

  with $a_{i,k} := a_k + \beta Z_i$ and with explicit expressions of $A_{k,i}^{\text{old}}$ and $B_{k,i}^{\text{old}}$.
- $A_{k,i}^{\text{old}}$ and $B_{k,i}^{\text{old}}$ depend only on $\boldsymbol{\theta}_{\text{old}}, L_i, R_i, Z_i$.
- In the absence of covariates ($Z_i = 0$, $a_{i,k} = a_k$, $\boldsymbol{\theta} = (a_1, \ldots, a_K)$) : the M-step is explicit.
- In the general regression framework : the M-step is solved using the Newton-Raphson procedure.
  - The block matrix of the Hessian for the a_ks is diagonal !
  - Using the Schurr complement, inversion of the Hessian is of order $\mathcal{O}(K)$ in the case $K >> d$.

A penalized EM algorithm

- We want to choose the number and location of the cuts from the data
- We start from a large grid of cuts ($K = 100, 1\,000, \ldots$)
- We use a penalization technique : the adaptive ridge (see Frommlet and Nuel, **PloS one**, 2016).

A penalized EM algorithm

- We want to choose the number and location of the cuts from the data
- We start from a large grid of cuts ($K = 100, 1\,000, \ldots$)
- We use a penalization technique : the adaptive ridge (see Frommlet and Nuel, **PloS one**, 2016).
- The adaptive ridge procedure consists in maximizing at the m^{th} step

$$\ell(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) - \frac{\text{pen}}{2} \sum_{k=1}^{K-1} w_k^{(m-1)} (a_{k+1} - a_k)^2,$$

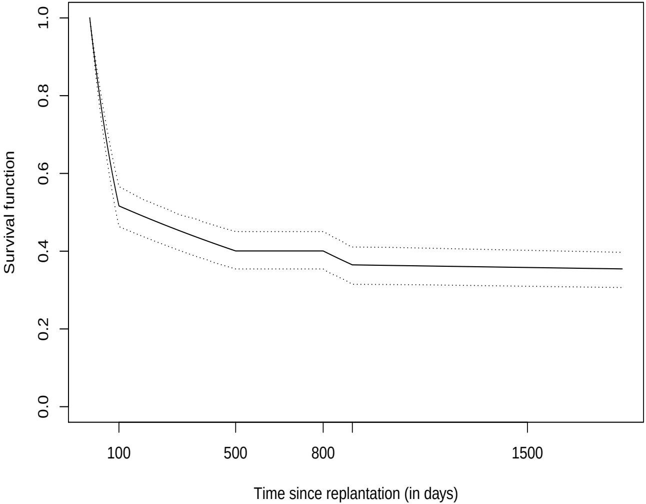
  with

$$w_k^{(m-1)} = \left(\left(a_{k+1}^{(m-1)} - a_k^{(m-1)} \right)^2 + \varepsilon^2 \right)^{-1},$$

  and $\varepsilon \ll 1$.
- The block matrix of the Hessian for the a_ks is now tri-diagonal !
- Using the Schurr complement, inversion of the Hessian is still of order $\mathcal{O}(K)$ in the case $K >> d$.

Dental dataset - without covariates

- The adaptive ridge method finds four cuts : $100, 500, 800, 900$.
- 95% confidence intervals computed using the bootstrap.



Dental dataset - Cox model

Covariates	HR= $e^{\hat{\beta}}$	95% CI	p-value
Mature	2.00	[1.74; 2.29]	1.89×10^{-5}
Storage time (hours)	1.23	[1.11; 1.34]	0.0017
Physiologic storage	0.93	[0.81; 1.06]	0.6980
Age>20 (mature teeth)	1.27	[0.99; 1.61]	0.1272

Dental dataset - Cox model

Covariates	HR= $e^{\hat{\beta}}$	95% CI	p-value
Mature	2.00	$[1.74; 2.29]$	1.89×10^{-5}
Storage time (hours)	1.23	$[1.11; 1.34]$	0.0017
Physiologic storage	0.93	$[0.81; 1.06]$	0.6980
Age>20 (mature teeth)	1.27	$[0.99; 1.61]$	0.1272

Risk of ankylosis of 400 avulsed and replanted human teeth in relation to length of dry storage. A re-evaluation of a previous long-term clinical study.
E. Lauridsen, J. Andreasen, O. Bouaziz, L. Andersson.
Dental Traumatology (2019).

Asymptotic results

We consider the following estimator.

- Only one step for the AR procedure :

$$\hat{\boldsymbol{\theta}} = (\hat{a}_1, \dots, \hat{a}_K, \hat{\beta}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^{K+d}}{\arg\max} \left\{ \log(\mathrm{L}_n^{\mathrm{obs}}(\boldsymbol{\theta})) - \frac{\mathrm{pen}}{2} \sum_{k=1}^{K-1} \hat{w}_k^{(1)} (a_{k+1} - a_k)^2 \right\},$$

  with $\hat{w}_k^{(1)} = \left((\hat{a}_{k+1}^{(1)} - \hat{a}_k^{(1)})^2 + \varepsilon^2 \right)^{-1}$ and $\hat{\boldsymbol{a}}^{(1)}$ is a consistent estimator. Using a hard-thresholding, we obtain an estimated set of cuts $\mathcal{A}_n = \{\hat{c}_1, \dots, \hat{c}_{\hat{K}}\}$.

- The final estimator is the unpenalised MLE with set of cuts $\mathcal{A}_n$, $\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}_n} = (\hat{\hat{a}}_{1,\mathcal{A}_n}, \dots, \hat{\hat{a}}_{\hat{K},\mathcal{A}_n}, \hat{\hat{\beta}}_{\mathcal{A}_n})$.

Asymptotic results

O. Bouaziz, E. Lauridsen, G. Nuel. *Regression modelling of interval-censored data based on the adaptive-ridge procedure.* To appear in **Journal of Applied Statistics**.

We define the true parameter $\boldsymbol{\theta}^* = (a_1^*, \ldots, a_{K^*}^*, \beta^*)$ with true cuts $\mathcal{A}^* = \{c_1^*, \ldots, c_{K^*}^*\}$.

Theorem

Assume that $\mathcal{A}^* \subset \{c_1, \ldots, c_K\}$, and some standard conditions. Then, if pen$/n \to 0$ as $n \to \infty$ we have :

1. $\lim_{n\to\infty} \mathbb{P}[\mathcal{A}_n = \mathcal{A}^*] = 1$.
2. $\sqrt{n}(\hat{\hat{\beta}}_{\mathcal{A}_n} - \beta^*)$ converges in distribution toward a centered Gaussian variable with variance equal to $(\Sigma_{\beta^*})^{-1}$,

where Σ_{β^*} is the optimal variance obtained from the maximum likelihood estimator with true cuts.

Proof is inspired from H. Zou, *The adaptive Lasso and its oracle properties*. **JASA** (2006).

Extensions : inclusion of exact observations

For an exact observation i,

$$\begin{aligned}\mathbb{E}[\log(f(T_i \mid Z_i;\boldsymbol{\theta}))|\text{data},\boldsymbol{\theta}_{\text{old}}] &= \log(f(T_i \mid Z_i;\boldsymbol{\theta})) \\ &= \sum_{k=1}^{K}\left\{O_{i,k}a_{i,k} - \exp(a_{i,k})R_{i,k}\right\}.\end{aligned}$$

Q can be decomposed as

$$\begin{aligned}Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = &\sum_{i \text{ not exact}}\sum_{k=1}^{K}\left\{\left(a_{i,k} - \sum_{j=1}^{k-1}(c_j - c_{j-1})e^{a_{i,j}}\right)A_{k,i}^{\text{old}} - e^{a_{i,k}}B_{k,i}^{\text{old}}\right\} \\ &+ \sum_{i \text{ exact}}\sum_{k=1}^{K}\left\{O_{i,k}a_{i,k} - \exp(a_{i,k})R_{i,k}\right\}.\end{aligned}$$

Extensions : the cure model

- Latent variable $Y \in \{0, 1\}$.
- Cox model for the susceptible individuals :

$$\begin{aligned} \lambda(t \mid Y, Z) &= Y\lambda(t \mid Y = 1, Z) \\ &= Y\lambda_0(t)\exp(\beta Z). \end{aligned}$$

- Logistic link for the probability of being cured :

$$\mathbb{P}[Y = 1 \mid X] = \frac{\exp(\gamma X)}{1 + \exp(\gamma X)}.$$

Extensions : the cure model

- Latent variable $Y \in \{0, 1\}$.
- Cox model for the susceptible individuals :

$$\begin{aligned} \lambda(t \mid Y, Z) &= Y\lambda(t \mid Y = 1, Z) \\ &= Y\lambda_0(t)\exp(\beta Z). \end{aligned}$$

- Logistic link for the probability of being susceptible (cured) :

$$p_i := \mathbb{P}[Y_i = 1 \mid X_i] = \frac{\exp(\gamma X_i)}{1 + \exp(\gamma X_i)}.$$

- The complete likelihood is defined as

$$\mathrm{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1 - Y_i} \prod_{i=1}^{n} \{f(T_i \mid Y_i = 1, Z_i; \boldsymbol{\theta})\}^{Y_i}.$$

Summary and take-away messages

- The Adaptive Ridge algorithm is fast and easy to implement : derivatives of the penalized criterion can be computed.

Summary and take-away messages

- The Adaptive Ridge algorithm is fast and easy to implement : derivatives of the penalized criterion can be computed.
- For interval-censored data, the EM algorithm + piecewise constant baseline hazard leads to tractable solutions !

Summary and take-away messages

- The Adaptive Ridge algorithm is fast and easy to implement : derivatives of the penalized criterion can be computed.
- For interval-censored data, the EM algorithm + piecewise constant baseline hazard leads to tractable solutions !
- Use of the Adaptive Ridge for a piecewise constant baseline hazard provides a flexible model and interpretable results.

Summary and take-away messages

- The Adaptive Ridge algorithm is fast and easy to implement : derivatives of the penalized criterion can be computed.
- For interval-censored data, the EM algorithm + piecewise constant baseline hazard leads to tractable solutions !
- Use of the Adaptive Ridge for a piecewise constant baseline hazard provides a flexible model and interpretable results.
- In several time to event situations it is no longer possible to consider a non-parametric baseline. For example :
  - Mixture model (Y is a latent variable) :

    $$\lambda(t \mid Z, Y = k) = \lambda_k(t \mid Z) \exp(\beta_k Z).$$

    This model is not identifiable when using the non-parametric baseline.
  - Frailty models.
  - Joint models.
  - ...

Bibliography

[1] Olivier Bouaziz, Eva Lauridsen, and Grégory Nuel. Regression modelling of interval censored data based on the adaptive ridge procedure. *To appear in Journal of Applied Statistics arXiv :1812.09158*, 2021.

[2] Olivier Bouaziz and Grégory Nuel. L_0 regularization for the estimation of piecewise constant hazard rates in survival analysis. *Applied Mathematics*, 8(3), 2017.

[3] Florian Frommlet and Grégory Nuel. An adaptive ridge procedure for l_0 regularization. *PloS one*, 11(2), 2016.

[4] Vivien Goepp, Grégory Nuel, and Olivier Bouaziz. Regularized bidimensional estimation of the hazard rate. *To appear in International Journal of Biostatistics arXiv :1803.04853*, 2021.

[5] Eva Lauridsen, Jens O Andreasen, Oliver Bouaziz, and Lars Andersson. Risk of ankylosis of 400 avulsed and replanted human teeth in relation to length of dry storage : A re-evaluation of a long-term clinical study. *Dental Traumatology*, 2019.

Thank you for your attention