Study of the adaptive-ridge algorithm with applications to time to event data

Olivier Bouaziz[1]
with Rémy Abergel[1], Grégory Nuel[2]

[1]MAP5 (CNRS 8145), Université Paris Cité
[2]LPSM (CNRS 8001), Sorbonne Université, Paris

Séminaire Parisien de Statistique

Institut Henri Poincaré

Outline

1. Study of the adaptive ridge algorithm
2. Simulations
3. The adaptive ridge procedure for piecewise constant hazards
4. The adaptive ridge procedure for interval-censored data

Presentation of the problem

We consider the following penalised criterion :

$$\overline{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \mathcal{E}_\lambda(\beta) := C(\beta) + \lambda \, \mathcal{L}_0(\beta) \right\}$$

with

- $C : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, $\mathrm{dom}(C) := \{\beta \in \mathbb{R}^p \, , \; C(\beta) < +\infty\} \neq \emptyset$,
- $\mathcal{L}_0(\beta) := \#\{j \in \{1, 2, \ldots, p\}, \; \beta_j \neq 0\}$,
- $\lambda > 0$ is a regularisation parameter.

Examples of contrast functions :

- $C(\beta) = \| Y - X\beta \|^2$, $Y \in \mathbb{R}^n$ response variable, X design matrix (dim$= n \times p$).
- $C(\beta) = -\ell_n(Y_1, \ldots, Y_n; \beta)$ is minus a log-likelihood function.

The adaptive-ridge algorithm

Let $w^{(0)} \in (\mathbb{R}_+^*)^p$, $\delta > 0$, $q \in [0,2)$. The $\mathrm{AR}_{\lambda,q}^{\delta}$ scheme is an iterative algorithm : for $k = 1, 2, \ldots$

$$\begin{cases} \beta^{(k+1)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ C(\beta) + \frac{\lambda}{2} \sum_{j=1}^{p} w_j^{(k)} \beta_j^2 \right\} \\ w_j^{(k+1)} = \left(|\beta_j^{(k+1)}|^2 + \delta^2 \right)^{\frac{q-2}{2}}, \text{ j=1,...,p.} \end{cases}$$

We will study two scenarios :

- $q \in (0,2)$, $\delta \geq 0$
- $q = 0$, $\delta > 0$

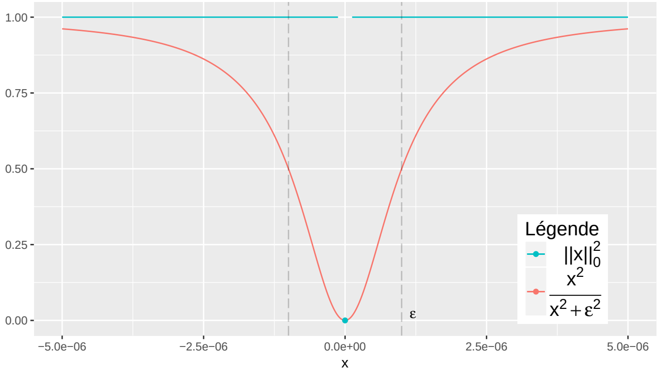
Rippe, R. C. A., Meulman, J. J. and Eilers, P. H. C. *Visualization of Genomic Changes by Segmented Smoothing Using an L_0 Penalty*. **PlosOne** (2012).

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

L_0 norm approximation - Heuristic

When $\delta \ll 1,\ q = 0$

$$\sum_{j=1}^{p} w_j^{(k)}\, \beta_j^2 = \sum_{j=1}^{p} \frac{\beta_j^2}{\beta_j^2 + \delta^2} \simeq \|\beta\|_0 = \begin{cases} 0 & \text{if } \beta_j = 0 \\ 1 & \text{if } \beta_j \neq 0 \end{cases}$$



Our main contribution

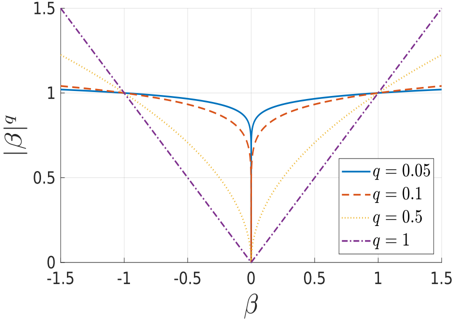
- In the case $q \in (0, 2)$, $\delta \geq 0$, we show that the AR algorithm is related to the following problem :

$$\tilde{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ E_{\lambda,q}(\beta) := C(\beta) + \lambda \|\beta\|_q^q \right\}$$

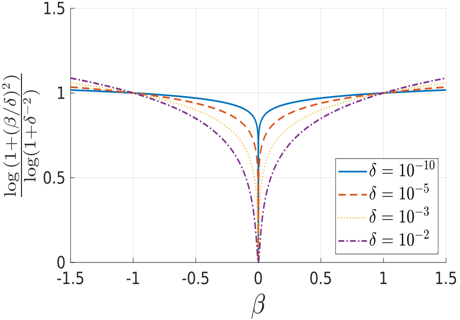
- In the case $q = 0$, $\delta > 0$, we show that the AR algorithm is related to the following problem :

$$\tilde{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ F_{\lambda,\delta}(\beta) := C(\beta) + \lambda \sum_{j=1}^{p} \underbrace{\frac{\log\left(1 + (\beta_j/\delta)^2\right)}{\log\left(1 + \delta^{-2}\right)}}_{\xrightarrow[\delta \to 0]{} \mathbb{1}_{\beta_j \neq 0}} \right\}$$

Two smooth approximations of the $\mathcal{L}_0$ penalty



ℓ^q penalty



square-log penalty

Variational formulation of the ℓ^q penalty

Proposition (R. Abergel, O. B., G. Nuel)

For all $\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$, for all $q > 0$ and for all $\nu > q$, we have

$$\|\beta\|_q^q = \inf_{\eta=(\eta_1,\eta_2,\dots,\eta_p)\in\left(\mathbb{R}_+^*\right)^p} \left(\mathcal{L}_q^\nu(\beta,\eta) := \sum_{j=1}^p \frac{q}{\nu} \cdot \frac{|\beta_j|^\nu}{\eta_j} + \frac{\nu - q}{\nu} \cdot \eta_j^{\frac{q}{\nu - q}} \right),$$

and when $\beta \in (\mathbb{R}^*)^p$, the infimum is attained at $\eta = |\beta|^{\nu - q}$.

- $\nu = 2$, $q \in (0, 2)$.

  Chan, R. H. and Liang, H.-X. *Half-Quadratic Algorithm for l_p-l_q Problems with Applications to TV-l1 Image Restoration and Compressive Sensing*. **Efficient Algorithms for Global Optimization Methods in Computer Vision**(2014).

- $\nu = 2$, $q = 1$.

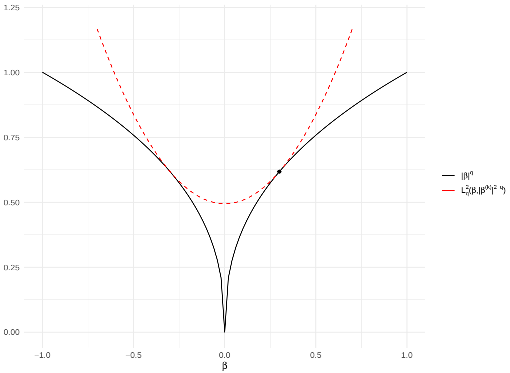
  Mairal, J., Bach, F. and Ponce, J. *Sparse Modeling for Image and Vision Processing*. **Foundations and TrendsR in Computer Graphics and Vision** (2014).

The adaptive ridge as a Majorize-Minimize (MM) algorithm

For $\beta_j^{(k)} \in \mathbb{R}^*$, set $\nu = 2$, $\eta_j^{(k)} = |\beta^{(k)}|^{2-q}$. For all $\beta_j \in \mathbb{R}$, we have :

$$\|\beta\|_q^q \leq \mathcal{L}_q^2(\beta, |\beta^{(k)}|^{2-q}) = \sum_{j=1}^{p} \frac{q}{2} \cdot \frac{|\beta_j|^2}{|\beta_j^{(k)}|^{2-q}} + \frac{2-q}{2} \cdot |\beta_j^{(k)}|^q,$$

with $\mathcal{L}_q^2(\beta^{(k)}, |\beta^{(k)}|^{2-q}) = \|\beta^{(k)}\|_q^q$. ($\beta^{(k)} = 0.3$ and $q = 0.4$ in the plot)



The adaptive ridge as a Majorize-Minimize (MM) algorithm

- For $\lambda > 0$, for all $\beta_j \in \mathbb{R}$ and for all $\beta_j^{(k)} \in \mathbb{R}^*$, we have

$$E_{\lambda,q}(\beta) := C(\beta) + \lambda \|\beta\|_q^q \leq \underbrace{C(\beta) + \lambda \mathcal{L}_q^2(\beta, |\beta^{(k)}|^{2-q})}_{g(\beta|\beta^{(k)})},$$

with $g(\beta^{(k)} \mid \beta^{(k)}) = E_{\lambda,q}(\beta^{(k)})$.

The adaptive ridge as a Majorize-Minimize (MM) algorithm

- For $\lambda > 0$, for all $\beta_j \in \mathbb{R}$ and for all $\beta_j^{(k)} \in \mathbb{R}^*$, we have

$$E_{\lambda,q}(\beta) := C(\beta) + \lambda \|\beta\|_q^q \leq \underbrace{C(\beta) + \lambda \mathcal{L}_q^2(\beta, |\beta^{(k)}|^{2-q})}_{g(\beta|\beta^{(k)})},$$

with $g(\beta^{(k)} \mid \beta^{(k)}) = E_{\lambda,q}(\beta^{(k)})$.

- Let $\beta^{(k+1)} = \arg\min_\beta g(\beta \mid \beta^{(k)})$. Then :

$$E_{\lambda,q}(\beta^{(k+1)}) \leq g(\beta^{(k+1)} \mid \beta^{(k)}) \leq g(\beta^{(k)} \mid \beta^{(k)}) = E_{\lambda,q}(\beta^{(k)}).$$

Properties of the adaptive ridge algorithm

$$\beta^{(k+1)} = \arg\min_{\beta} g(\beta \mid \beta^{(k)}) = \arg\min_{\beta} \left\{ C(\beta) + \mathcal{L}_q^2(\beta, |\beta^{(k)}|^{2-q}) \right\}$$
$$= \arg\min_{\beta} \left\{ C(\beta) + \frac{\lambda q}{2} \sum_{j=1}^{p} \frac{|\beta_j|^2}{|\beta_j^{(k)}|^{2-q}} \right\}.$$

- The $\mathrm{AR}^0_{\lambda q, q}$ algorithm minimises $E_{\lambda,q}$!
- But the procedure is only valid as long as the $(\beta^{(k)})$, $k = 0, 1 \ldots$ remain in $(\mathbb{R}_*)^p$.
- We introduce $r : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$ the function defined by

$$\forall (x,y) \in \mathbb{R}^2, \quad r(x,y) = \begin{cases} 0 & \text{if } x = y = 0 \\ +\infty & \text{if } x \neq 0 \text{ and } y = 0 \\ \frac{x}{y} & \text{otherwise.} \end{cases}$$

Properties of the adaptive ridge algorithm ($q > 0$)

Proposition (R. Abergel, O. B., G. Nuel) : $q > 0, \delta = 0$

The modified $\text{AR}^0_{\lambda q,q}$ algorithm defined by

$$\begin{cases} \beta^{(k+1)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ C(\beta) + \frac{\lambda q}{2} \sum_{j=1}^p r(|\beta_j|^2, \eta_j^{(k)}) \right\} \\ \eta_j^{(k+1)} = |\beta_j^{(k+1)}|^{2-q}, \text{ j=1,...,p.} \end{cases}$$

satisfies the property $E_{\lambda,q}(\beta^{(k+1)}) \leq E_{\lambda,q}(\beta^{(k)})$ $\forall k \in \mathbb{N}$, with

$$E_{\lambda,q}(\beta) = C(\beta) + \lambda \|\beta\|_q^q$$

Properties of the adaptive ridge algorithm ($q > 0$)

Proposition (R. Abergel, O. B., G. Nuel) : $q > 0, \delta = 0$

The modified $\mathrm{AR}^0_{\lambda q, q}$ algorithm defined by

$$\begin{cases} \beta^{(k+1)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ C(\beta) + \frac{\lambda q}{2} \sum_{j=1}^{p} r(|\beta_j|^2, \eta_j^{(k)}) \right\} \\ \eta_j^{(k+1)} = |\beta_j^{(k+1)}|^{2-q}, \text{ j=1,...,p.} \end{cases}$$

satisfies the property $E_{\lambda,q}(\beta^{(k+1)}) \leq E_{\lambda,q}(\beta^{(k)})$ $\forall k \in \mathbb{N}$, with

$$E_{\lambda,q}(\beta) = C(\beta) + \lambda \|\beta\|_q^q$$

Proposition (R. Abergel, O. B., G. Nuel) : $q > 0, \delta > 0$

The $\mathrm{AR}^\delta_{\lambda q, q}$ algorithm, $\delta > 0$, satisfies the property $E^\delta_{\lambda,q}(\beta^{(k+1)}) \leq E^\delta_{\lambda,q}(\beta^{(k)})$ $\forall k \in \mathbb{N}$, with

$$E^\delta_{\lambda,q}(\beta) = C(\beta) + \lambda \|\beta^2 + \delta^2\|_{q/2}^{q/2}$$

R. Abergel, O. Bouaziz, O., G. Nuel. *A Review on the Adaptive-Ridge Algorithm with several extensions*.
`https://helios2.mi.parisdescartes.fr/~obouaziz/adaptive-ridge_preprint2023.pdf`

Properties of the adaptive ridge algorithm ($q = 0$)

Proposition (R. Abergel, O. B., G. Nuel) : $q = 0, \delta > 0$

The $\mathrm{AR}^{\delta}_{\lambda',q}$ algorithm, $\delta > 0$, $\lambda' = 2\lambda / \log(1+\delta^{-2})$, satisfies the property $F_{\lambda,\delta}(\beta^{(k+1)}) \leq F_{\lambda,\delta}(\beta^{(k)})\ \forall k \in \mathbb{N}$, with

$$F_{\lambda,\delta}(\beta) := C(\beta) + \lambda \sum_{j=1}^{p} \underbrace{\frac{\log\left(1 + (\beta_j/\delta)^2\right)}{\log\left(1+\delta^{-2}\right)}}_{\xrightarrow[\delta \to 0]{} \mathbb{1}_{\beta_j \neq 0}}$$

R. Abergel, O. Bouaziz, O., G. Nuel. *A Review on the Adaptive-Ridge Algorithm with several extensions*.
https://helios2.mi.parisdescartes.fr/~obouaziz/adaptive-ridge_preprint2023.pdf

Outline

Simulations setting

- Linear regression model

$$Y = X\beta^\star + \varepsilon,$$

$X_{ij} \sim U(0,1)$, $\varepsilon_i \sim \mathcal{N}(0, 0.2^2)$, $i = 1, \ldots, n$, $j = 1, \ldots, p$.

-
$$\forall j = 1, \ldots, p, \quad \beta_j^\star = \begin{cases} 1 & \text{if } U_j > 0.95 \\ 0 & \text{otherwise} \end{cases}$$

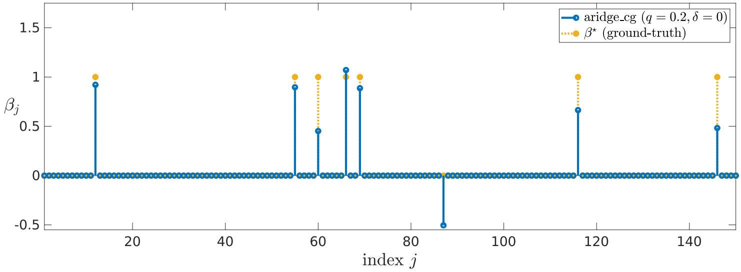
where $U_j \sim U(0,1)$, $j = 1, \ldots, p$.

- $n = 300$, $p = 150$.
- $C(\beta) = \|Y - X\beta\|_2^2/2$.

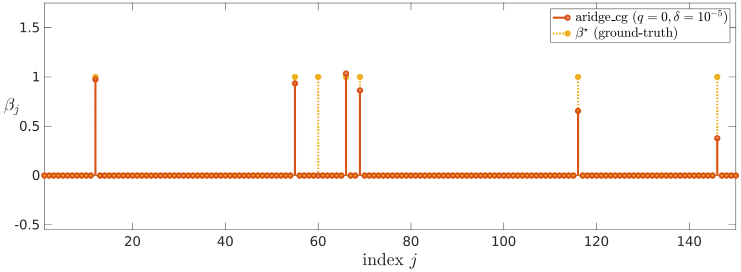
The AR algorithm is implemented using a conjugate-gradient based method.

- The algorithm is named `aridge_cg`
- Iterative algorithm : computation time is $\mathcal{O}(p^2)$ at each iteration.

Simulations : illustration of AR estimates

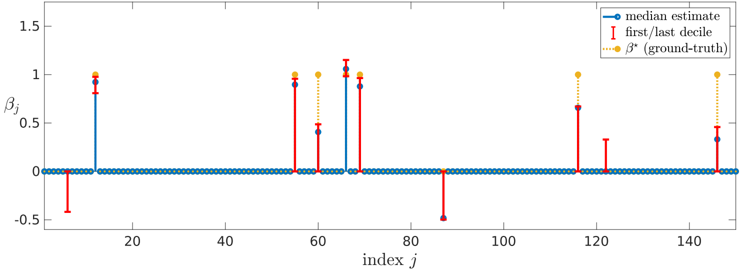


(a) estimated (local) minimizer of the ℓ^q penalized energy

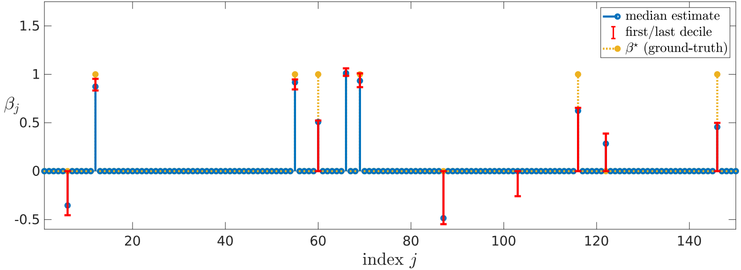


(b) estimated (local) minimizer of the log-square penalized energy

Simulations : sensitivity to initialisation

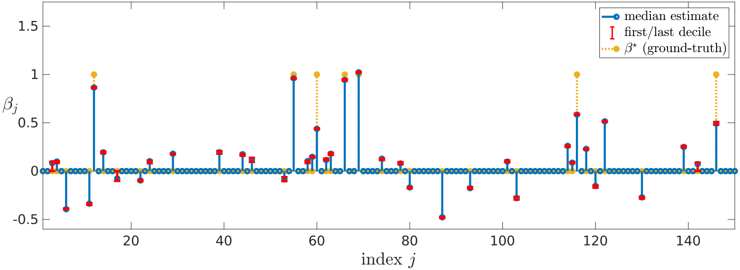


(a) module aridge_cg with $q = 0.1$ and $\delta = 0$



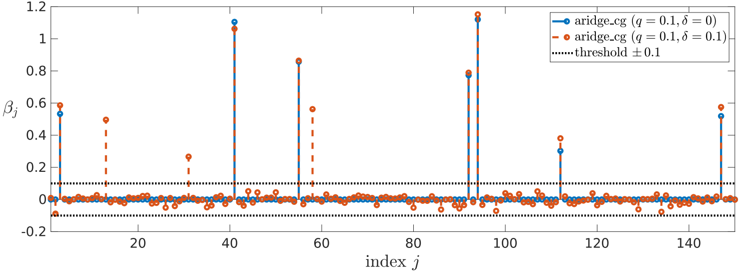
(b) module aridge_cg with $q = 0.3$ and $\delta = 0$

Simulations : sensitivity to initialisation

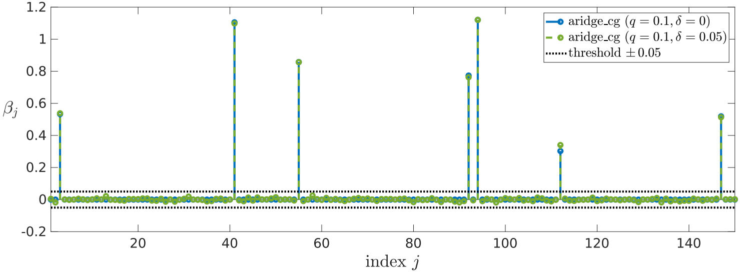


(c) module aridge_cg with $q = 0.8$ and $\delta = 0$

Simulations : influence of the δ parameter

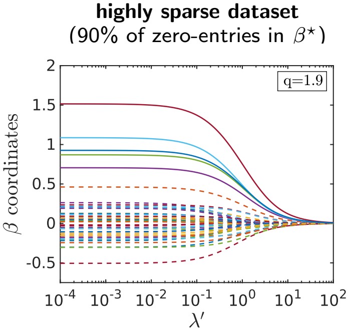


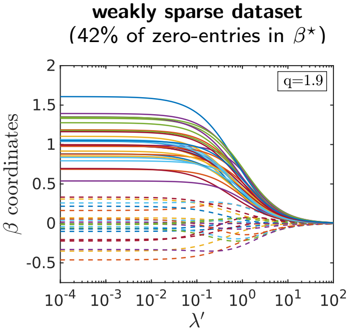
(a) module aridge_cg with $\delta = 0$ or $\delta = 0.1$



(b) module aridge_cg with $\delta = 0$ or $\delta = 0.05$

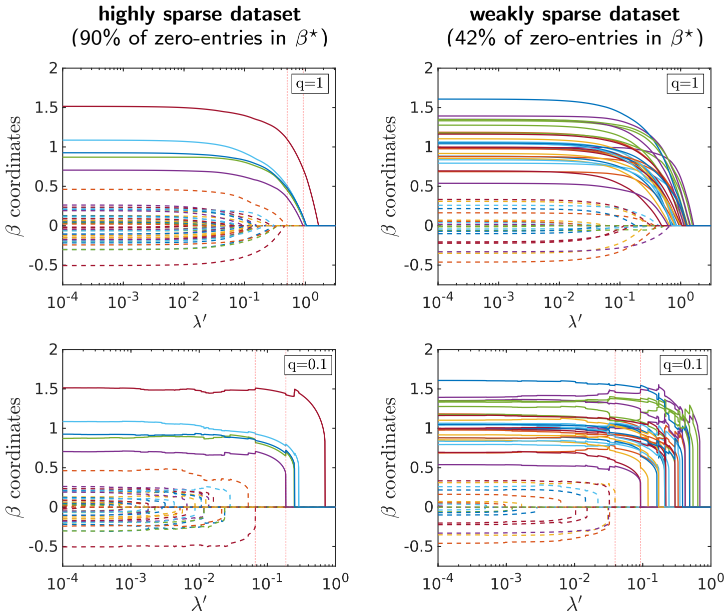
Simulations : regularisation paths





- Plain curves : active coordinates.
- Dashed curves : coordinates equal to 0.

Simulations : regularisation paths



Outline

Background in time to event data : right-censoring

- Positive time variable of interest : T.
- Observations :
  $$\begin{cases} T_i^{\text{obs}} = T_i \wedge C_i \\ \Delta_i = \mathbb{1}_{T_i \le C_i} \end{cases}$$
- Independent censoring : $T \perp\!\!\!\perp C$
- The hazard rate and a key relation :

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

  Many estimators (Nelson Aalen, Kaplan-Meier, ...) are based on this relation.
- The likelihood of the observed data is equal to :

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

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

The piecewise constant hazard model

- The model :

$$h(t) = \sum_{l=1}^{L} \alpha_l \mathbb{1}_{c_{l-1} < t \leq c_l}$$

- Goal : estimate the α_ls.

The log-likelihood is equal to :

$$\ell_n(\boldsymbol{h}) = \sum_{l=1}^{L} \left\{ \bar{O}_l \log(\alpha_l) - \alpha_l \bar{R}_l \right\},$$

where

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

The piecewise constant hazard model

- The model :

$$h(t) = \sum_{l=1}^{L} \alpha_l \mathbb{1}_{c_{l-1} < t \leq c_l}$$

- Goal : estimate the α_ls.

The log-likelihood is equal to :

$$\ell_n(\boldsymbol{h}) = \sum_{l=1}^{L} \left\{ \bar{O}_l \log\left(\alpha_l\right) - \alpha_l \bar{R}_l \right\},$$

where

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

The maximum likelihood estimator is explicit :

$$\hat{\alpha}_l^{\text{mle}} = \frac{\bar{O}_l}{\bar{R}_l}$$

The piecewise constant hazard model

- The model :

$$h(t) = \sum_{l=1}^{L} \alpha_l \mathbb{1}_{c_{l-1} < t \leq c_l}$$

- Goal : estimate the α_ls.

The log-likelihood is equal to :

$$\ell_n(\boldsymbol{h}) = \sum_{l=1}^{L} \left\{ \bar{O}_l \log\left(\alpha_l\right) - \alpha_l \bar{R}_l \right\},$$

where

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

The maximum likelihood estimator is explicit :

$$\hat{\alpha}_l^{\text{mle}} = \frac{\bar{O}_l}{\bar{R}_l}$$

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

Penalising the maximum likelihood estimator with the *fused* AR

Set $\log \alpha_l = a_l$. Implement the AR with $q = 0$ and $\delta > 0$.

$$\begin{cases} \boldsymbol{a}^{(k+1)} \in \underset{\boldsymbol{a}\in\mathbb{R}^L}{\arg\min} \left\{ \ell_n(\boldsymbol{a}) - \frac{\lambda}{2} \sum_{l=1}^{L-1} w_l^{(k)} \left(a_{l+1} - a_l \right)^2 \right\} \\ w_l^{(k+1)} = \left(\left(a_{l+1}^{(k+1)} - a_l^{(k+1)} \right)^2 + \delta^2 \right)^{-1}, l = 1, \ldots, L. \end{cases}$$

- The penalized estimator is no longer explicit.

Penalising the maximum likelihood estimator with the *fused* AR

Set $\log \alpha_l = a_l$. Implement the AR with $q = 0$ and $\delta > 0$.

$$\begin{cases} \boldsymbol{a}^{(k+1)} \in \underset{\boldsymbol{a}\in\mathbb{R}^L}{\arg\min} \left\{ \ell_n(\boldsymbol{a}) - \frac{\lambda}{2} \sum_{l=1}^{L-1} w_l^{(k)} \left(a_{l+1} - a_l\right)^2 \right\} \\ w_l^{(k+1)} = \left(\left(a_{l+1}^{(k+1)} - a_l^{(k+1)} \right)^2 + \delta^2 \right)^{-1}, l = 1, \dots, L. \end{cases}$$

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

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

where $\mathcal{I}$ is the opposite of the Hessian matrix, U is the score vector.

Penalising the maximum likelihood estimator with the *fused* AR

Set $\log \alpha_l = a_l$. Implement the AR with $q = 0$ and $\delta > 0$.

$$\begin{cases} \boldsymbol{a}^{(k+1)} \in \underset{\boldsymbol{a} \in \mathbb{R}^L}{\arg\min} \left\{ \ell_n(\boldsymbol{a}) - \frac{\lambda}{2} \sum_{l=1}^{L-1} w_l^{(k)} \left(a_{l+1} - a_l \right)^2 \right\} \\ w_l^{(k+1)} = \left(\left(a_{l+1}^{(k+1)} - a_l^{(k+1)} \right)^2 + \delta^2 \right)^{-1}, l = 1, \ldots, L. \end{cases}$$

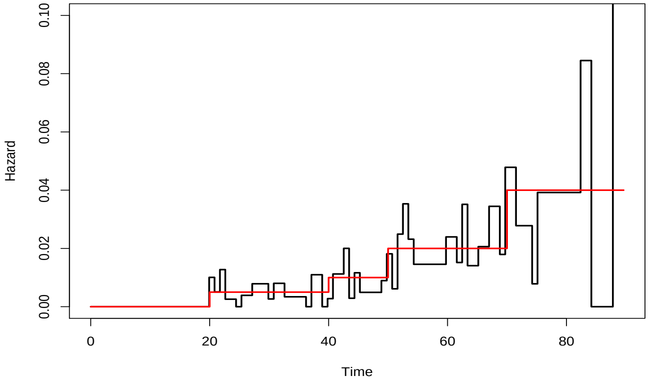
- The penalized estimator is no longer explicit.
- Maximization is performed from the Newton-Raphson algorithm. For a given sequence of weights $\boldsymbol{w}$, the mth Newton Raphson iteration step is obtained from the equation

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

  where $\mathcal{I}$ is the opposite of the Hessian matrix, U is the score vector.

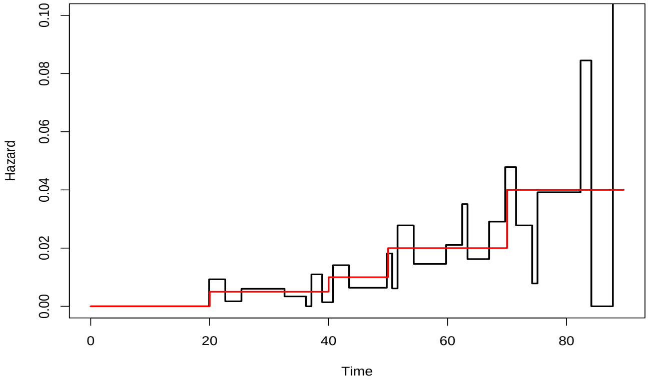
- The Hessian matrix is tri-diagonal.
- $\Longrightarrow$ computation time for the inversion of the Hessian is $\mathcal{O}(L)$

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



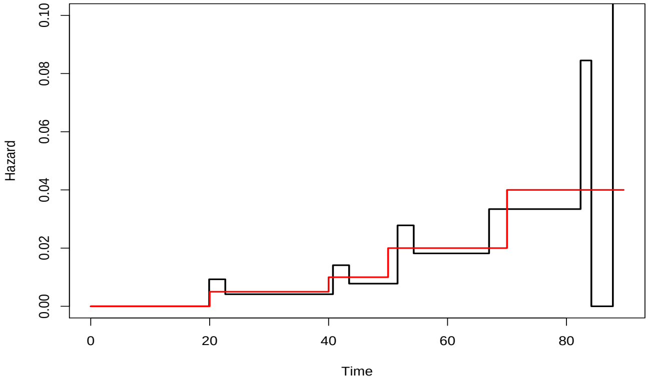
- In red the true hazard function
- In black the hazard estimator for $\lambda = 0.1$

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



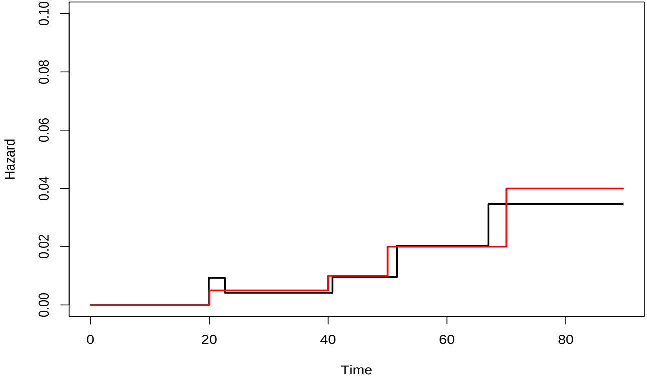
- In red the true hazard function
- In black the hazard estimator for $\lambda = 0.27$

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



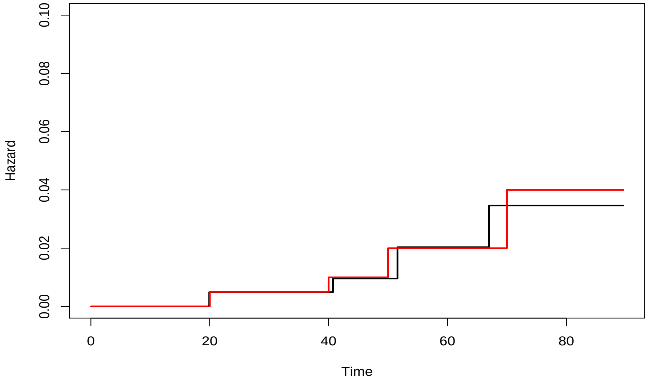
- In red the true hazard function
- In black the hazard estimator for $\lambda = 0.55$

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



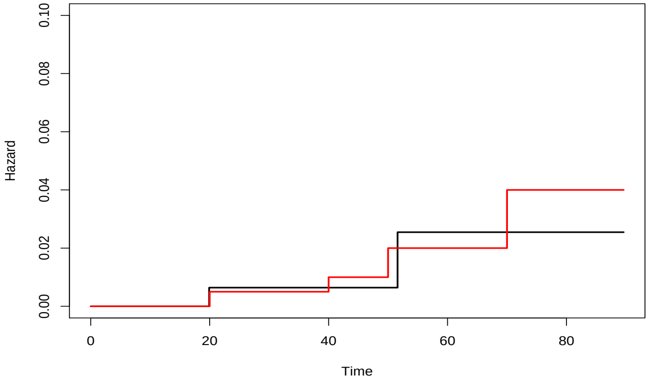
- In red the true hazard function
- In black the hazard estimator for $\lambda = 0.77$

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



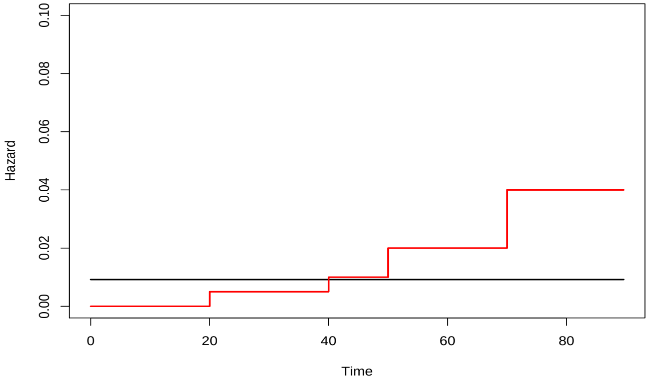
- In red the true hazard function
- In black the hazard estimator for $\lambda = 1.54$

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



- In red the true hazard function
- In black the hazard estimator for $\lambda = 6.16$

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



- In red the true hazard function
- In black the hazard estimator for $\lambda = 52.70$

Model selection for the *fused 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),

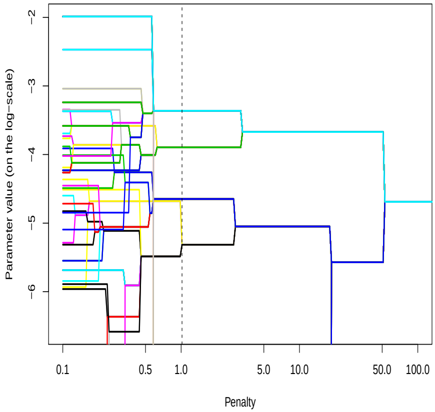
with D the dimension of the model :

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

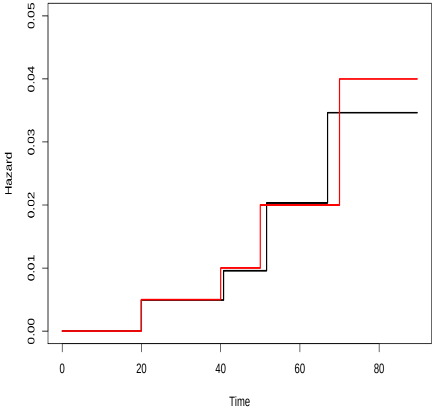
Bouaziz, O. and Nuel, G. L_0 *regularization for the estimation of piecewise constant hazard rates in survival analysis*. **Applied Mathematics** (2017).

Package **pchsurv** available on GitHub : `install_github("obouaziz/pchsurv")`

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



Regularization path



Hazard estimator (in black)

Outline

The dental dataset

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

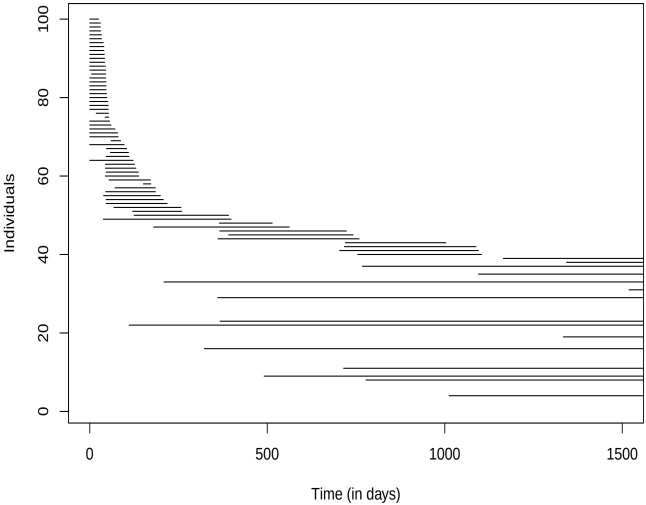
- Study of 322 patients with 400 avulsed and replanted permanent teeth from 1965 to 1988.
- The variable of interest is time from replantation until the ankylosis complication.
- Patients are examined at intermittent visits to the dentist.
  - Left-censoring (28%) if ankylosis occurred before the first 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, $i = 1, \ldots, n$.

- $0 = \mathrm{L}_i < \mathrm{R}_i < +\infty$ for left-censored observation ($\Delta_i = 1$)
- $0 < \mathrm{L}_i < \mathrm{R}_i < +\infty$ for interval-censored observation ($\Delta_i = 1$)
- $0 < \mathrm{L}_i < \mathrm{R}_i = +\infty$ for right-censored observation ($\Delta_i = 0$)

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

$$\mathcal{L}^{\text{obs}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \{S(\mathrm{L}_i \mid Z_i, \boldsymbol{\theta}) - S(\mathrm{R}_i \mid Z_i, \boldsymbol{\theta})\}^{\Delta_i} \times \Big\{S(\mathrm{L}_i \mid Z_i, \boldsymbol{\theta})\Big)\Big\}^{1-\Delta_i}.$$

The observed likelihood

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

- $0 = \mathrm{L}_i < \mathrm{R}_i < +\infty$ for left-censored observation ($\Delta_i = 1$)
- $0 < \mathrm{L}_i < \mathrm{R}_i < +\infty$ for interval-censored observation ($\Delta_i = 1$)
- $0 < \mathrm{L}_i < \mathrm{R}_i = +\infty$ for right-censored observation ($\Delta_i = 0$)

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

$$\mathcal{L}^{\text{obs}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \exp\left(- \int_0^{\mathrm{L}_i} h_0(t) dt e^{\beta Z_i}\right) \left(1 - \exp\left(- \int_{\mathrm{L}_i}^{\mathrm{R}_i} h_0(t) dt e^{\beta Z_i}\right)\right) \right\}^{\Delta_i}$$
$$\times \left\{ \exp\left(- \int_0^{\mathrm{L}_i} h_0(t) dt e^{\beta Z_i}\right) \right\}^{1-\Delta_i},$$

for the Cox model $h(t \mid Z_i) = h_0(t) \exp(\beta Z_i)$.

The observed likelihood

- The piecewise constant model for the baseline :

$$h_0(t) = \sum_{l=1}^{L} \exp(a_l) \mathbb{1}_{c_{l-1} < t \leq c_l}$$

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

Maximization of :

$$\mathcal{L}^{\text{obs}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \exp\left(-\int_0^{\mathrm{L}_i} h_0(t) dt e^{\beta Z_i} \right) \left(1 - \exp\left(-\int_{\mathrm{L}_i}^{\mathrm{R}_i} h_0(t) dt e^{\beta Z_i} \right) \right) \right\}^{\Delta_i} \times \left\{ \exp\left(-\int_0^{\mathrm{L}_i} h_0(t) dt e^{\beta Z_i} \right) \right\}^{1-\Delta_i},$$

requires to use the Newton-Raphson algorithm.

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

The EM algorithm

The complete likelihood is defined as

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

Introduce data $= (\mathrm{L}_i, \mathrm{R}_i, Z_i)$.

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

  we have

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

Using the EM algorithm

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

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

  with $a_{i,l} := a_l + \beta Z_i$ and with explicit expressions of $A_{l,i}^{\text{old}}$ and $B_{l,i}^{\text{old}}$.

- $A_{l,i}^{\text{old}}$ and $B_{l,i}^{\text{old}}$ depend only on $\boldsymbol{\theta}_{\text{old}}, \mathrm{L}_i, \mathrm{R}_i, Z_i$.

Using the EM algorithm

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

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

  with $a_{i,l} := a_l + \beta Z_i$ and with explicit expressions of $A_{l,i}^{\text{old}}$ and $B_{l,i}^{\text{old}}$.

- $A_{l,i}^{\text{old}}$ and $B_{l,i}^{\text{old}}$ depend only on $\boldsymbol{\theta}_{\text{old}}, \mathrm{L}_i, \mathrm{R}_i, Z_i$.
- In the absence of covariates ($Z_i = 0$, $a_{i,l} = a_l$, $\boldsymbol{\theta} = (a_1, \ldots, a_L)$) : the M-step is explicit.

Using the EM algorithm

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

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

  with $a_{i,l} := a_l + \beta Z_i$ and with explicit expressions of $A_{l,i}^{\text{old}}$ and $B_{l,i}^{\text{old}}$.
- $A_{l,i}^{\text{old}}$ and $B_{l,i}^{\text{old}}$ depend only on $\boldsymbol{\theta}_{\text{old}}, \mathrm{L}_i, \mathrm{R}_i, Z_i$.
- In the absence of covariates ($Z_i = 0$, $a_{i,l} = a_l$, $\boldsymbol{\theta} = (a_1, \ldots, a_L)$) : 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_ls is diagonal !
  - Using the Schurr complement, inversion of the Hessian is of order $\mathcal{O}(L)$ in the case $L \gg d$.

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 ($L = 100, 1\,000, \ldots$)

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 ($L = 100, 1\,000, \dots$)
- We use the adaptive ridge. At the k^{th} step we maximise

$$\ell(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) - \frac{\lambda}{2}\sum_{l=1}^{L-1} w_l^{(k-1)}(a_{l+1} - a_l)^2,$$

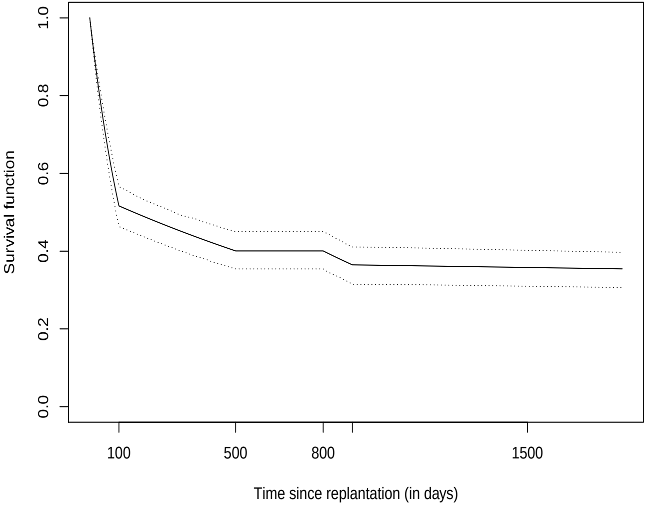
  with

$$w_l^{(k-1)} = \left(\left(a_{l+1}^{(k-1)} - a_l^{(k-1)}\right)^2 + \delta^2\right)^{-1},$$

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

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

- O. Bouaziz, E. Lauridsen, G. Nuel. Regression modelling of interval-censored data based on the adaptive-ridge procedure. **Journal of Applied Statistics** (2022)
- E. Lauridsen, J. Andreasen, O. Bouaziz, L. Andersson. *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.* **Dental Traumatology** (2019)

Discussion and extensions

- Connections with similar works :
  - The Iteratively Reweighted Least Squares (IRLS) algorithm : AR algorithm with $\nu = 2$ and update of δ

    I. Daubechies, R. DeVore, M. Fornasier, C. S. Gunturk. *Iteratively reweighted least squares minimization for sparse recovery.* **Communications on Pure and Applied Mathematics** (2010).
  - The IRL1 algorithm corresponds : AR algorithm with $\nu = 1$ and update of δ

    E. J. Candes, M. B. Wakin, S. P. Boyd. *Enhancing sparsity by reweighted l1 minimization.* **Journal of Fourier analysis and applications** (2008)

    D. Needell. Noisy signal recovery via iterative reweighted L1- minimization. **Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers** (2009)
  - The adaptive Lasso algorithm : AR algorithm with two steps, $\nu = 1$ ($\delta = 0$).

    H. Zhou. *The Adaptive Lasso and Its Oracle Properties.* **Journal of the American Statistical Association** (2006)

Discussion and extensions

- Connections with similar works :
  - The Iteratively Reweighted Least Squares (IRLS) algorithm : AR algorithm with $\nu = 2$ and update of δ

    I. Daubechies, R. DeVore, M. Fornasier, C. S. Gunturk. *Iteratively reweighted least squares minimization for sparse recovery.* **Communications on Pure and Applied Mathematics** (2010).
  - The IRL1 algorithm corresponds : AR algorithm with $\nu = 1$ and update of δ

    E. J. Candes, M. B. Wakin, S. P. Boyd. *Enhancing sparsity by reweighted l1 minimization.* **Journal of Fourier analysis and applications** (2008)

    D. Needell. Noisy signal recovery via iterative reweighted L1- minimization. **Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers** (2009)
  - The adaptive Lasso algorithm : AR algorithm with two steps, $\nu = 1$ $(\delta = 0)$.

    H. Zhou. *The Adaptive Lasso and Its Oracle Properties.* **Journal of the American Statistical Association** (2006)
- An AR type algorithm can also be derived as a ℓ^q constrained problem.

  R. Abergel, O. Bouaziz, O., G. Nuel. *A Review on the Adaptive-Ridge Algorithm with several extensions.*

  https://helios2.mi.parisdescartes.fr/~obouaziz/adaptive-ridge_preprint2023.pdf

Discussion and extensions

- Connections with similar works :
  - The Iteratively Reweighted Least Squares (IRLS) algorithm : AR algorithm with $\nu = 2$ and update of δ

    I. Daubechies, R. DeVore, M. Fornasier, C. S. Gunturk. *Iteratively reweighted least squares minimization for sparse recovery.* **Communications on Pure and Applied Mathematics** (2010).
  - The IRL1 algorithm corresponds : AR algorithm with $\nu = 1$ and update of δ

    E. J. Candes, M. B. Wakin, S. P. Boyd. *Enhancing sparsity by reweighted l1 minimization.* **Journal of Fourier analysis and applications** (2008)

    D. Needell. Noisy signal recovery via iterative reweighted L1- minimization. **Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers** (2009)
  - The adaptive Lasso algorithm : AR algorithm with two steps, $\nu = 1$ ($\delta = 0$).

    H. Zhou. *The Adaptive Lasso and Its Oracle Properties.* **Journal of the American Statistical Association** (2006)
- An AR type algorithm can also be derived as a ℓ^q constrained problem.

  R. Abergel, O. Bouaziz, O., G. Nuel. *A Review on the Adaptive-Ridge Algorithm with several extensions.*

  `https://helios2.mi.parisdescartes.fr/~obouaziz/adaptive-ridge_preprint2023.pdf`
- In time to event data, use of the fused Adaptive Ridge for a piecewise constant baseline hazard provides a flexible model and interpretable results.

Discussion and extensions

- Connections with similar works :
  - The Iteratively Reweighted Least Squares (IRLS) algorithm : AR algorithm with $\nu = 2$ and update of δ

    I. Daubechies, R. DeVore, M. Fornasier, C. S. Gunturk. *Iteratively reweighted least squares minimization for sparse recovery.* **Communications on Pure and Applied Mathematics** (2010).
  - The IRL1 algorithm corresponds : AR algorithm with $\nu = 1$ and update of δ

    E. J. Candes, M. B. Wakin, S. P. Boyd. *Enhancing sparsity by reweighted l1 minimization.* **Journal of Fourier analysis and applications** (2008)

    D. Needell. Noisy signal recovery via iterative reweighted L1- minimization. **Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers** (2009)
  - The adaptive Lasso algorithm : AR algorithm with two steps, $\nu = 1$ $(\delta = 0)$.

    H. Zhou. *The Adaptive Lasso and Its Oracle Properties.* **Journal of the American Statistical Association** (2006)
- An AR type algorithm can also be derived as a ℓ^q constrained problem.

  R. Abergel, O. Bouaziz, O., G. Nuel. *A Review on the Adaptive-Ridge Algorithm with several extensions.*

  `https://helios2.mi.parisdescartes.fr/~obouaziz/adaptive-ridge_preprint2023.pdf`
- In time to event data, use of the fused Adaptive Ridge for a piecewise constant baseline hazard provides a flexible model and interpretable results.
- For interval-censored data, the EM algorithm + piecewise constant baseline hazard leads to tractable solutions !

Bibliography

[1] Rémy Abergel, Olivier Bouaziz, and Grégory Nuel. A review on the adaptive-ridge algorithm with several extensions. *Submitted*.

[2] Olivier Bouaziz, Eva Lauridsen, and Grégory Nuel. Regression modelling of interval censored data based on the adaptive ridge procedure. *Journal of Applied Statistics*, 49(13) :3319–3343, 2022.

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

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

[5] Vivien Goepp, Olivier Bouaziz, and Grégory Nuel. Spline regression with automatic knot selection. *Submitted*.

[6] Vivien Goepp, Jean-Christophe Thalabard, Grégory Nuel, and Olivier Bouaziz. Regularized bidimensional estimation of the hazard rate. *The international journal of biostatistics*, 18(1) :263–277, 2021.

[7] 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