An Iterative Regularized Method for Segmentation with Applications to Statistics

Vivien Goepp

under the supervision of Pr. O. Bouaziz and Pr. G. Nuel

27 septembre 2019

MAP5, Université Paris Descartes

An introductory example

- Illustrative example : the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

with $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta}^* \in \mathbb{R}^p$, and $\mathbb{E}\left[\varepsilon_i\right] = 0$.

An introductory example

- Illustrative example : the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

with $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta}^* \in \mathbb{R}^p$, and $\mathbb{E}[\varepsilon_i] = 0$.

- Ordinary least square estimate (OLS) :

$$\hat{\boldsymbol{\beta}}^{\text{ols}} = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

An introductory example

- Illustrative example : the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

with $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta}^* \in \mathbb{R}^p$, and $\mathbb{E}[\varepsilon_i] = 0$.

- Ordinary least square estimate (OLS) :

$$\hat{\boldsymbol{\beta}}^{\text{ols}} = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$\rightarrow$ regularize $\hat{\boldsymbol{\beta}}^{\text{ols}}$

An introductory example

- Illustrative example : the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

with $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta}^* \in \mathbb{R}^p$, and $\mathbb{E}[\varepsilon_i] = 0$.

- Ordinary least square estimate (OLS) :

$$\hat{\boldsymbol{\beta}}^{\text{ols}} = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$\rightarrow$ regularize $\hat{\boldsymbol{\beta}}^{\text{ols}}$

Predictive approach :
Ridge regularization :

$$\text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \kappa\|\boldsymbol{\beta}\|^2$$

Explicative approach :
Lasso regularization :

$$\text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \kappa\|\boldsymbol{\beta}\|_1$$

Layout

- The adaptive ridge
- Segmentation in survival analysis
  - Bidimensional hazard rate estimation
  - Extension to age-period-cohort analysis
- Spline regression with knot selection

1. The adaptive ridge

Penalized likelihood-based variable selection methods

- The LASSO[1] estimate is sparse.
- Definition :

$$\boldsymbol{\beta}^{\text{lasso}} = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \kappa \|\boldsymbol{\beta}\|_1 \quad (\kappa > 0)$$

[1] :Tibshirani, R., Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, 1996.

Penalized likelihood-based variable selection methods

- The LASSO[1] estimate is sparse.
- Definition :

$$\boldsymbol{\beta}^{\text{lasso}} = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \kappa \|\boldsymbol{\beta}\|_1 \quad (\kappa > 0)$$

- Equivalent definition :

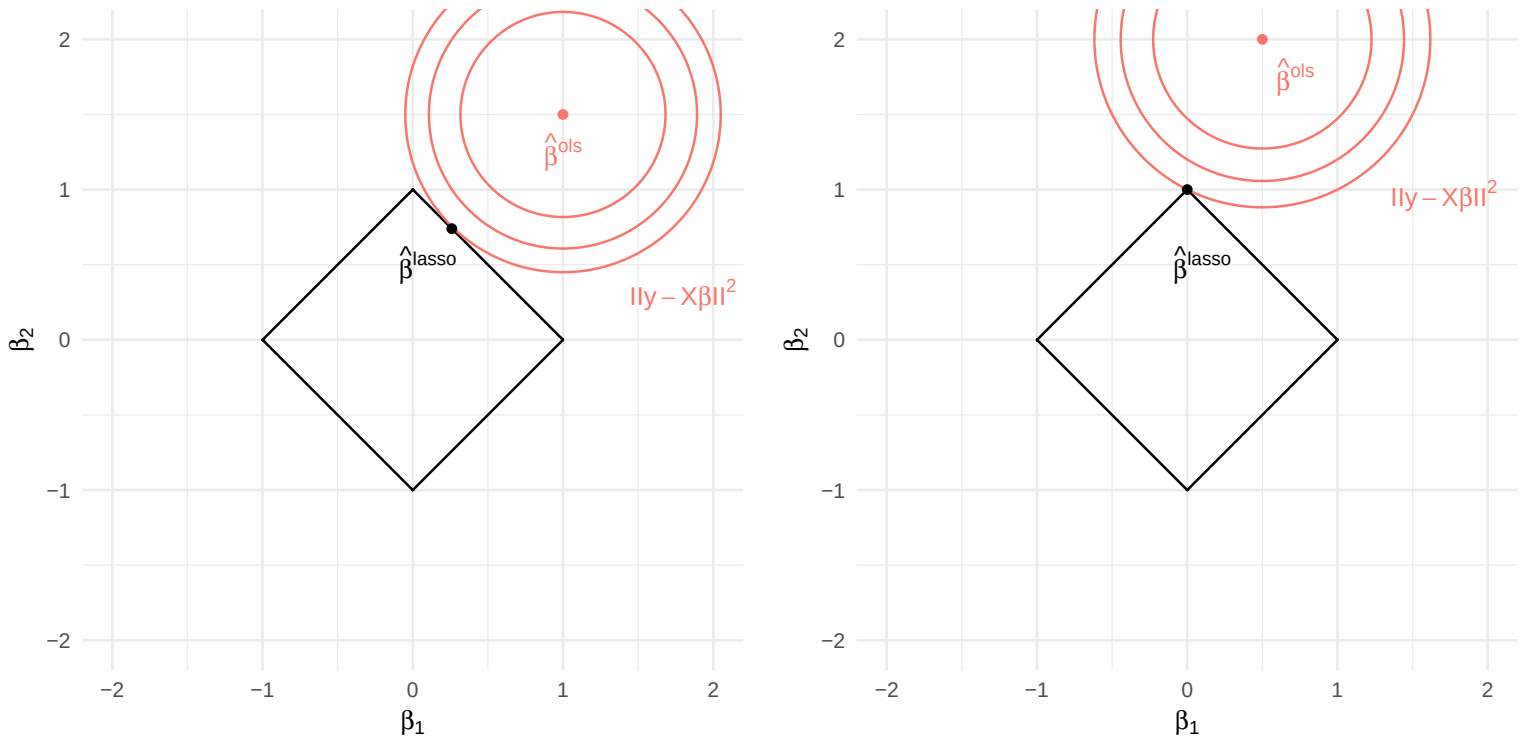
$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq t \quad (t > 0)$$

[1] :Tibshirani, R., Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, 1996.

Sparsity-inducing norms

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq t \quad (t > 0)$$

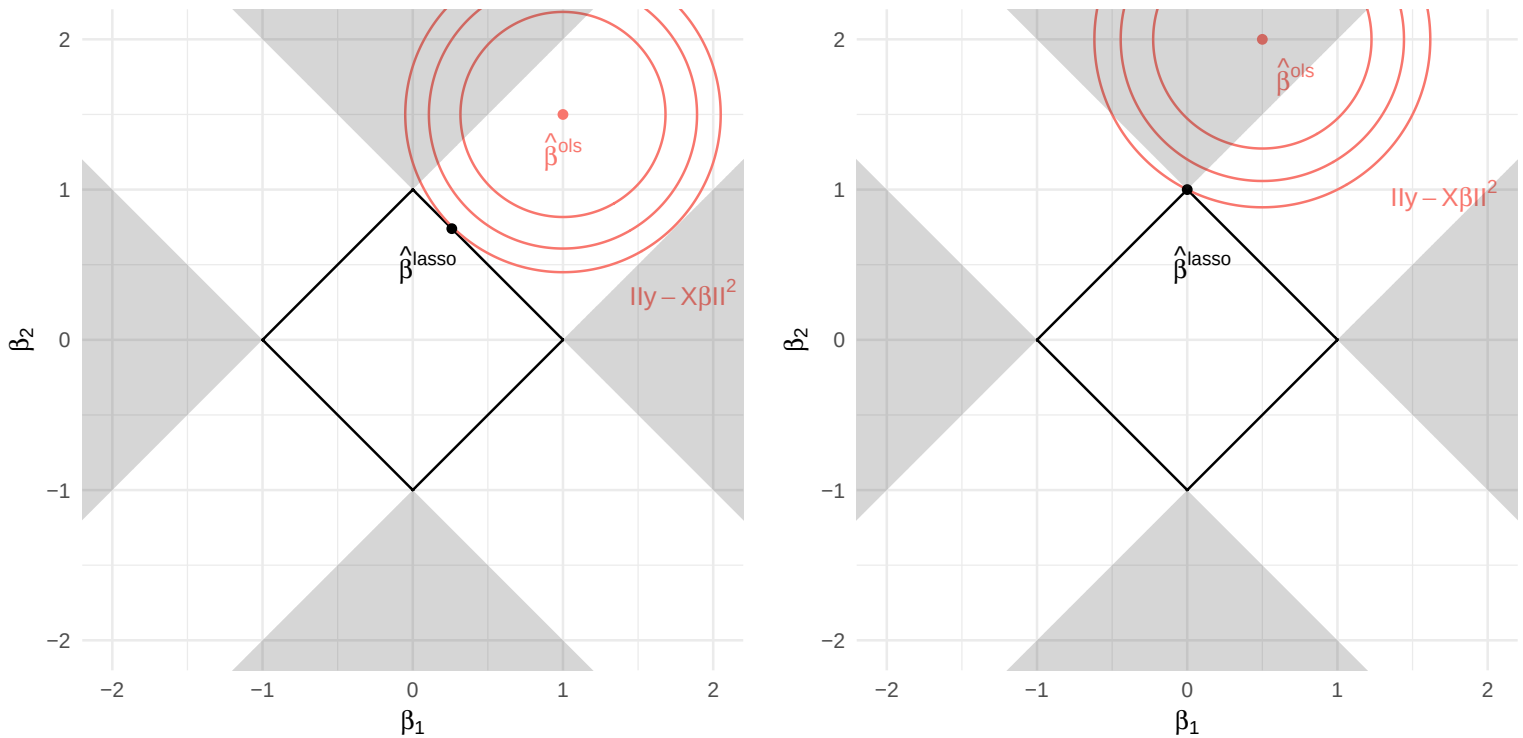
Illustration with $p = 2$:



Sparsity-inducing norms

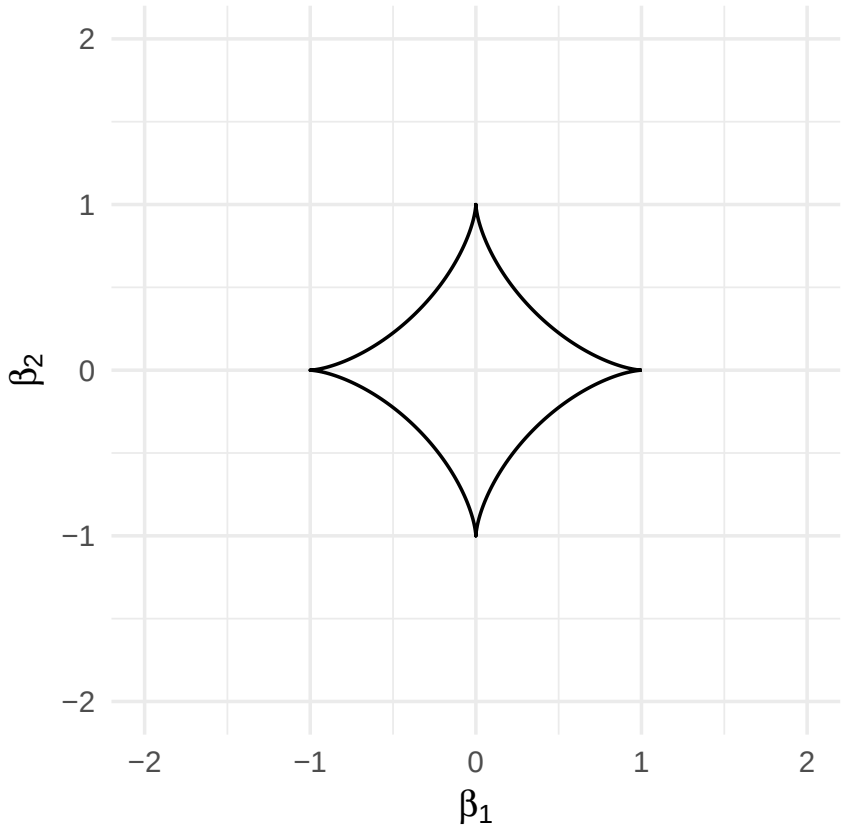
$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq t \quad (t > 0)$$

Illustration with $p = 2$:



Bridge penalty : L_q quasi-norm

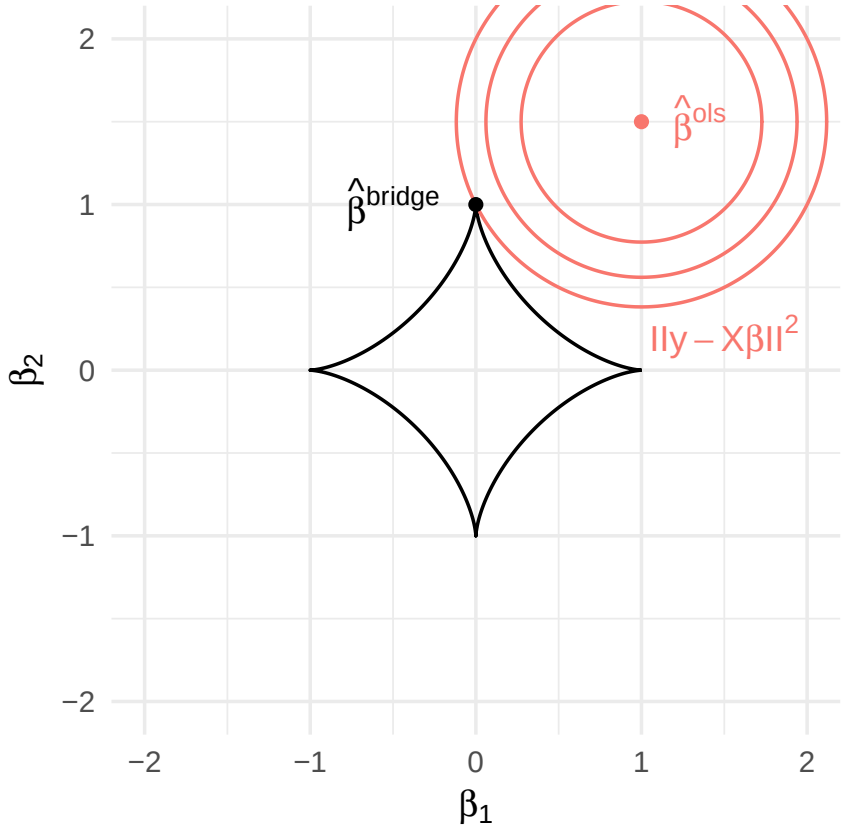
$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \kappa \|\boldsymbol{\beta}\|_q^q \quad \text{with} \quad 0 < q < 1.$$



L_q norm penalty with $q = 2/3$

Bridge penalty : L_q quasi-norm

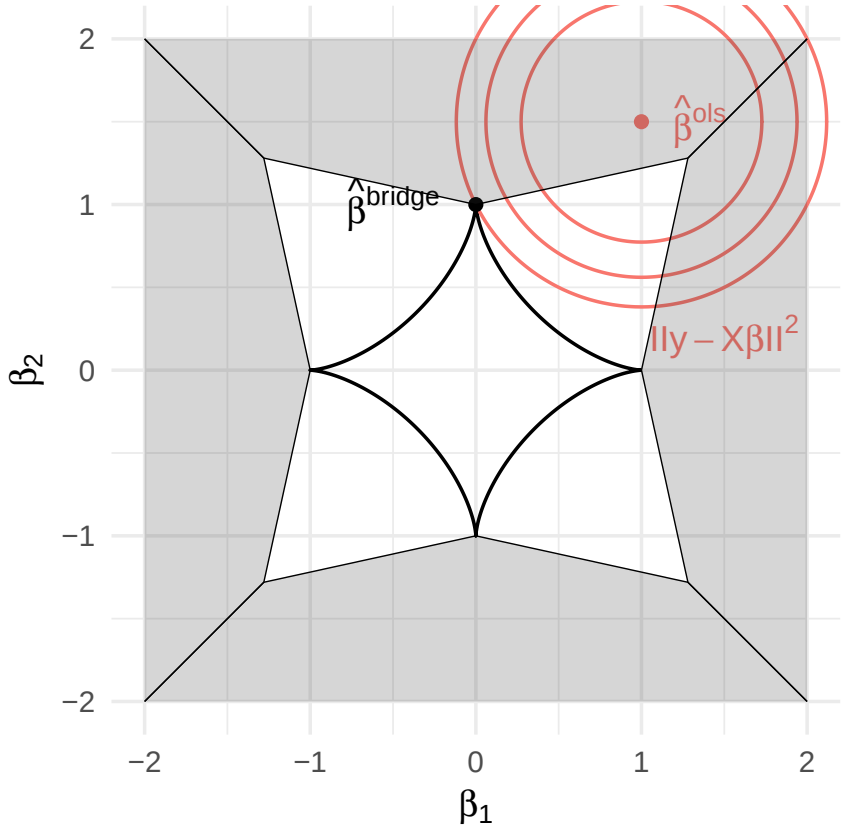
$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \kappa \|\boldsymbol{\beta}\|_q^q \quad \text{with} \quad 0 < q < 1.$$



L_q norm penalty with $q = 2/3$

Bridge penalty : L_q quasi-norm

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \kappa \|\boldsymbol{\beta}\|_q^q \quad \text{with} \quad 0 < q < 1.$$



L_q norm penalty with $q = 2/3$

Approximating L_q norms with the L_2 norm

For any $q \in (0, 2)$, we have[2] :

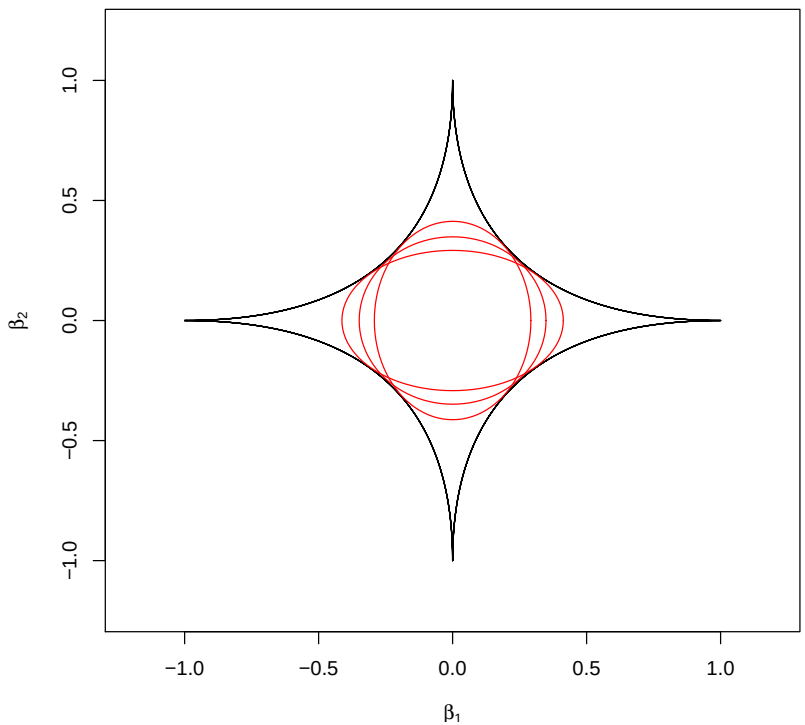
$$\frac{1}{q}\|\boldsymbol{\beta}\|_q^q = \inf_{\boldsymbol{w} \in \mathbb{R}_+^d} \left\{ \frac{1}{2} \sum_{j=1}^{d} w_j \beta_j^2 + \frac{2-q}{2q} \sum_{j=1}^{d} |w_j|^{\frac{q}{q-2}} \right\}$$

[2] : Mairal, J, Bach, F, and Ponce, J., Sparse Modeling for Image and Vision Processing, *Foundations and Trends in Computer Graphics and Vision*, 2014.

Approximating L_q norms with the L_2 norm

For any $q \in (0, 2)$, we have[2] :

$$\frac{1}{q}\|\boldsymbol{\beta}\|_q^q = \inf_{\boldsymbol{w}\in\mathbb{R}_+^d} \left\{ \frac{1}{2}\sum_{j=1}^{d} w_j \beta_j^2 + \frac{2-q}{2q}\sum_{j=1}^{d} |w_j|^{\frac{q}{q-2}} \right\}$$



[2] : Mairal, J, Bach, F, and Ponce, J., Sparse Modeling for Image and Vision Processing, *Foundations and Trends in Computer Graphics and Vision*, 2014.

Approximate any L_q norm

The problem becomes

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \kappa \|\boldsymbol{\beta}\|_q^q \right\} = \min_{\boldsymbol{\beta}} \inf_{\boldsymbol{w} \in \mathbb{R}_+^d} \{\ell(\boldsymbol{\beta}, \boldsymbol{w})\}$$

with

$$\ell(\boldsymbol{\beta}, \boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\kappa q}{2} \sum_j w_j \beta_j^2 + \frac{\kappa(2-q)}{2} \sum_j |w_j|^{\frac{q}{q-2}}$$

The L_q adaptive ridge : **minimize alternatively**

- $\arg\min_{\boldsymbol{w}} \{\ell(\boldsymbol{\beta}, \boldsymbol{w})\} = (|\beta_j|^2)^{\frac{q-2}{2}}$
- $\arg\min_{\boldsymbol{\beta}} \{\ell(\boldsymbol{\beta}, \boldsymbol{w})\} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\kappa q}{2} \sum_j w_j \beta_j^2$

Approximate any L_q norm

The problem becomes

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \kappa \|\boldsymbol{\beta}\|_q^q \right\} = \min_{\boldsymbol{\beta}} \inf_{\boldsymbol{w} \in \mathbb{R}_+^d} \{\ell(\boldsymbol{\beta}, \boldsymbol{w})\}$$

with

$$\ell(\boldsymbol{\beta}, \boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\kappa q}{2} \sum_j w_j \beta_j^2 + \frac{\kappa(2-q)}{2} \sum_j |w_j|^{\frac{q}{q-2}}$$

The L_q adaptive ridge : **minimize alternatively**

- $\arg\min_{\boldsymbol{w}} \{\ell(\boldsymbol{\beta}, \boldsymbol{w})\} = (|\beta_j|^2 + \delta^2)^{\frac{q-2}{2}}$
- $\arg\min_{\boldsymbol{\beta}} \{\ell(\boldsymbol{\beta}, \boldsymbol{w})\} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\kappa q}{2} \sum_j w_j \beta_j^2$

The L_0 norm Adaptive Ridge as logarithmic penalty

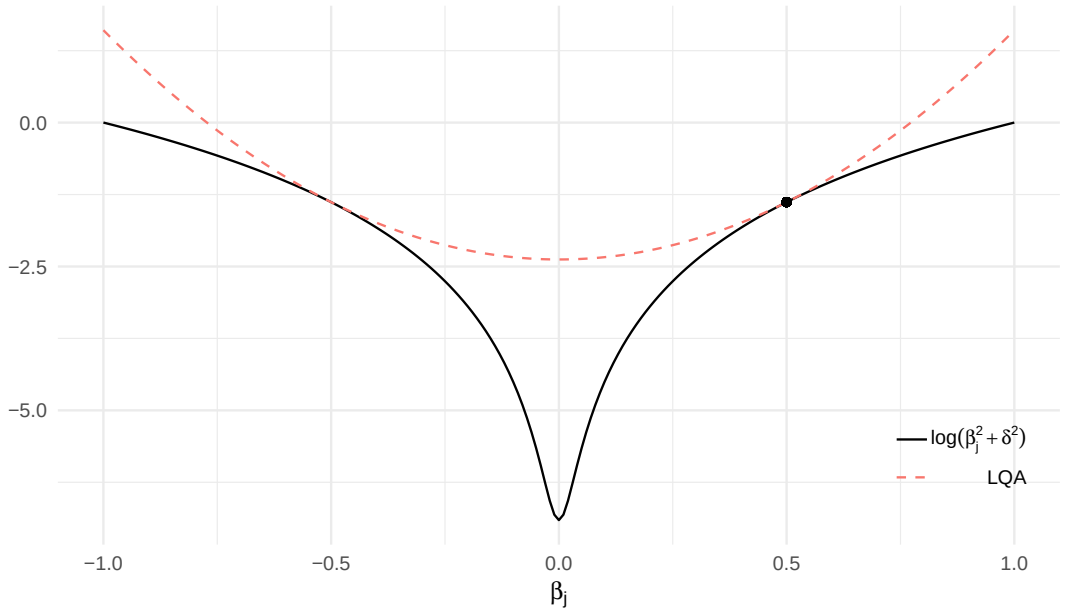
- The previous equation is defined when $q = 0$!
- The L_0 adaptive ridge estimate is the limit of the iterations :
  - $w_j \leftarrow (\beta_j^2 + \delta^2)^{-1}$
  - $\boldsymbol{\beta} \leftarrow \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \kappa \sum_j w_j \beta_j^2$
- But it does not correspond to the L_0 penalty.

The algorithm of the L_0 adaptive ridge

- It corresponds to the minimization of

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \kappa \sum_j \log(\beta_j^2 + \delta^2)$$

using the MM optimization :



- Define

$$q(\beta_j|\beta_j^{(k)}) = \log(\beta_j^{(k)2} + \delta^2) + (\beta_j^2 - \beta_j^{(k)2})$$

- Iterate over $\boldsymbol{\beta}^{(k+1)} \leftarrow \text{argmin}_{\boldsymbol{\beta}} q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$.

Practical implementation : L_0 adaptive ridge

Define $\ell(\boldsymbol{\beta}) = -\log(L(\boldsymbol{\beta}))$.

L_0 adaptive ridge :

$\boldsymbol{w}^{(0)} \leftarrow \mathbf{1}$
$k \leftarrow 1$
do
  $\boldsymbol{\beta}^{(k)} \leftarrow \text{argmin}_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) + \kappa \sum_j w_j^{(k-1)} \beta_j^{(k)2}$
  $w_j^{(k)} \leftarrow (\beta_j^{(k)2} + \delta^2)^{-1}$
  $k \leftarrow k + 1$
while $|w_j^{(k)} \beta_j^{(k)2} - 0.5| < 0.5 - \varepsilon$
For every j, we set $w_j^{(k)} \beta_j^{(k)2} \in \{0, 1\}$
If $w_j^{(k)} \beta_j^{(k)2} = 1$, we set $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}} \leftarrow \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{\text{mle}}$

Practical implementation : L_0 fused adaptive ridge

Fused L_0 adaptive ridge :

- Segmentation of $\boldsymbol{\beta}$.
- Penalize over $(\beta_j - \beta_{j-1})_j$.

Practical implementation : L_0 fused adaptive ridge

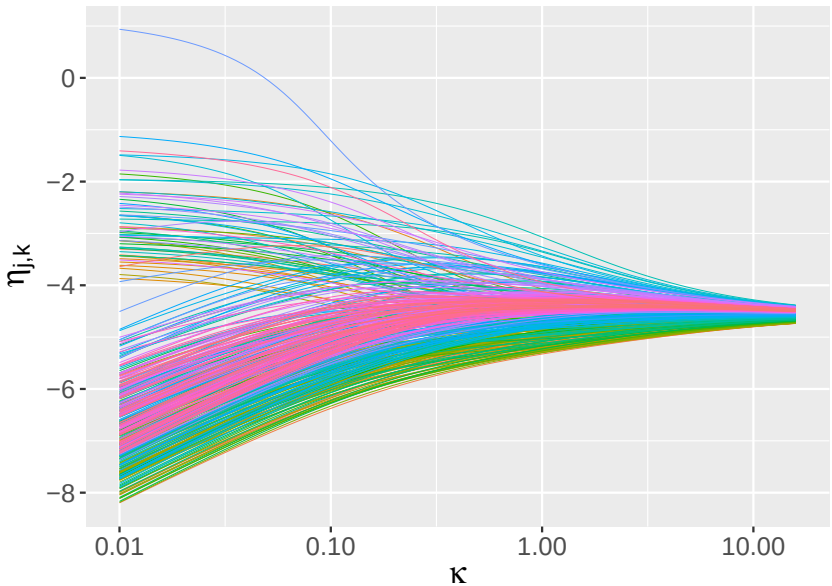
Fused L_0 adaptive ridge :

- Segmentation of $\boldsymbol{\beta}$.
- Penalize over $(\beta_j - \beta_{j-1})_j$.

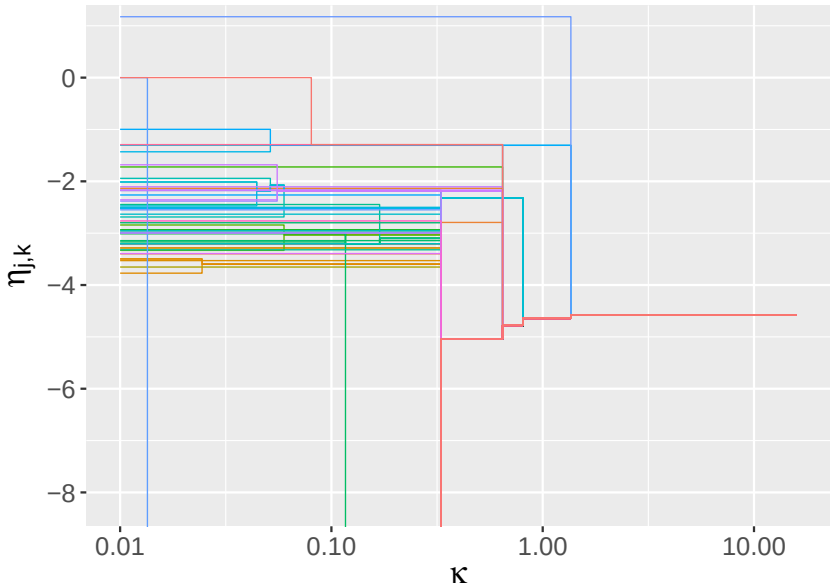
$\boldsymbol{w}^{(0)} \leftarrow \mathbf{1}$
$k \leftarrow 1$
do
  $\boldsymbol{\beta}^{(k)} \leftarrow \text{argmin}_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) + \kappa \sum_j w_j^{(k-1)} (\beta_j^{(k)} - \beta_{j-1}^{(k)})^2$
  $w_j^{(k)} \leftarrow ((\beta_j^{(k)} - \beta_{j-1}^{(k)})^2 + \delta^2)^{-1}$
  $k \leftarrow k + 1$
while $|w_j^{(k)} (\beta_j^{(k)} - \beta_{j-1}^{(k)})^2 - 0.5| < 0.5 - \varepsilon$

For every j, we set $w_j^{(k)} (\beta_j^{(k)} - \beta_{j-1}^{(k)})^2 \in \{0, 1\}$
On each segment : $\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{\beta}}^{\text{mle}}$

Comparison : smoothing vs fused L_0 adaptive ridge



Fused ridge :
Each κ yields an *estimate*



L_0 adaptive ridge segmentation :
Each κ yields a *model*

History and properties of the adaptive ridge

Selected papers :

- Grandvalet (1998) : Approximate L_1 norm
- Daubechies et al (2010) :
  - Study of the case $q > 0$ in sparse sensing
  - Speed of convergence is superlinear for $q > 0$.
- Frommlet and Nuel (2016) : Introduced L_q ($q \geq 0$) Adaptive Ridge
- Liu et al (2017) : The L_0 adaptive ridge : empirical study in high dimension ($p \gg n$).
- Dai et al (2018) : The L_0 adaptive ridge regression has the oracle properties.

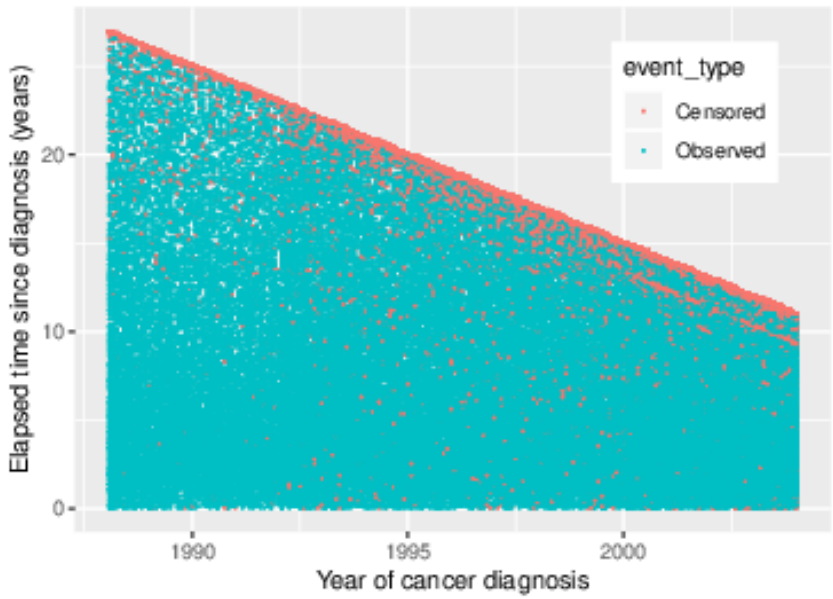
2. Application to survival analysis

2.1 Bidimensional estimation of the hazard rate

Motivating application : the SEER data

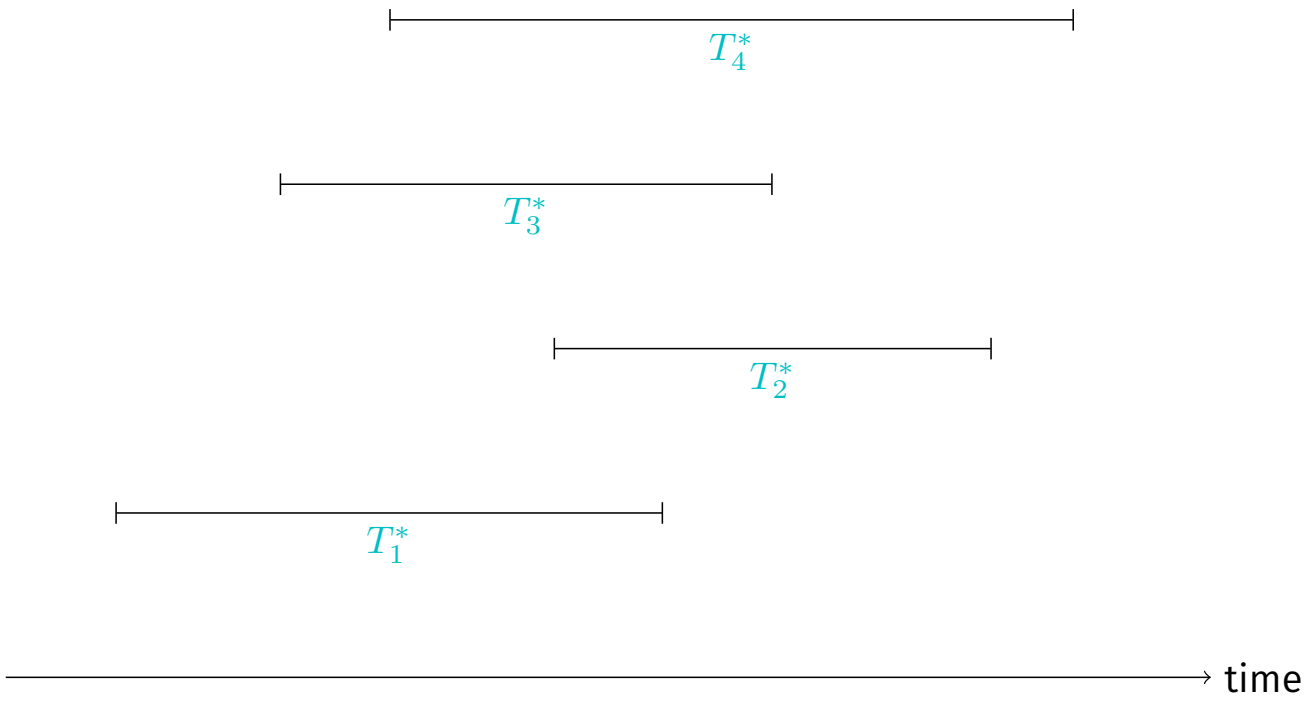
Study of mortality following breast cancer diagnosis

- US registry dataset of breast cancer
- Primary, unilateral, malignant and invasive cancers
- 1.2 million of patients, 60% of censoring
- The dates of diagnoses range from 1973 to 2014
- **Variable of interest** : the time from diagnosis until death (from cancer).



Stage 2 cancer diagnoses

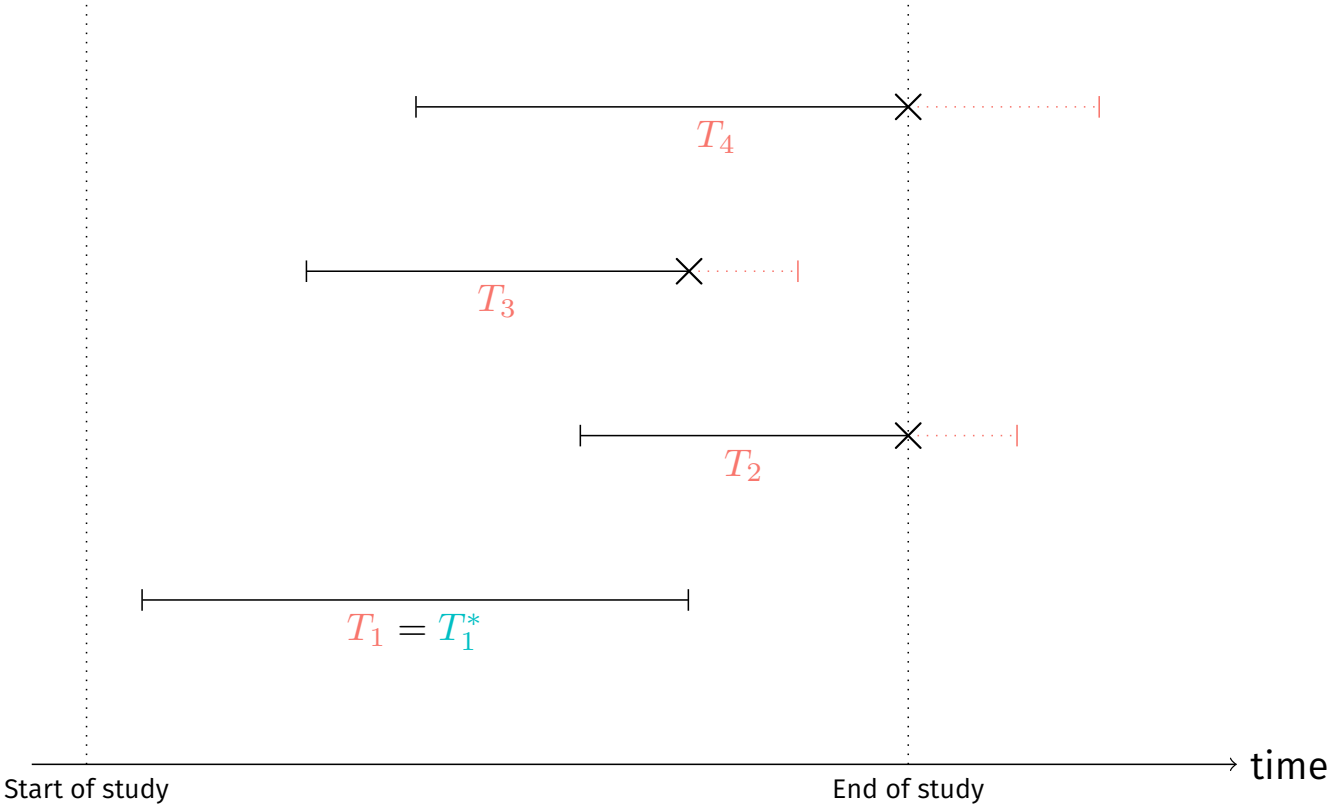
Right-censored data



- Variable of interest : T^*, time before an event of interest.
- But we don't observe the T_i^*s, but

$$\begin{cases} T_i = \min\left(T_i^*, C_i\right), \\ \Delta_i = \mathbb{1}_{T_i = T_i^*}. \end{cases}$$

Right-censored data



- Variable of interest : T^*, time before an event of interest.
- But we don't observe the T_i^*s, but

$$\begin{cases} T_i = \min\left(T_i^*, C_i\right), \\ \Delta_i = \mathbb{1}_{T_i = T_i^*}. \end{cases}$$

Survival analysis

- T_i : observed time
  C_i : censoring
  U_i : covariate
- Aim : infer the *hazard rate* :

$$\lambda(t|u) = \lim_{\delta t \to 0} \frac{\mathbb{P}\left(t \leq T^* \leq t + \delta t | T^* > t, U = u\right)}{\delta t}$$

- **Our approach** : discretize λ :

$$\lambda(t|u) = \sum_{j=1}^{J} \sum_{k=1}^{K} \lambda_{j,k} I_{[c_{j-1}, c_j) \times [d_{k-1}, d_k)}(t, u)$$

with segmentation over $\lambda_{j,k}$.

Fused adaptive ridge estimate

Exhaustive statistics :

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

The negative log-likelihood writes :

$$\ell_n(\boldsymbol{\eta}) = \sum_{j=1}^{J}\sum_{k=1}^{K}\{R_{j,k}\exp(\eta_{j,k}) - \eta_{j,k}O_{j,k}\} \quad \text{with} \quad \log\lambda_{j,k} = \eta_{j,k}.$$

MLE is explicit : $\lambda_{j,k} = O_{j,k}/R_{j,k}$.

.

Fused adaptive ridge estimate

Exhaustive statistics :

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

The negative log-likelihood writes :

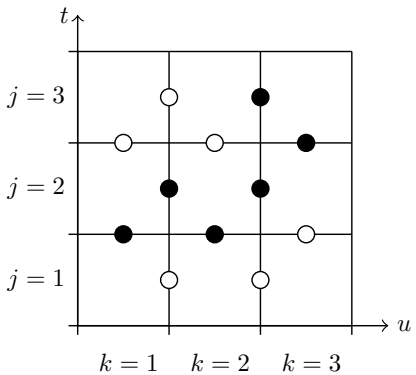
$$\ell_n(\boldsymbol{\eta}) = \sum_{j=1}^{J}\sum_{k=1}^{K}\left\{R_{j,k}\exp\left(\eta_{j,k}\right) - \eta_{j,k}O_{j,k}\right\} \quad \text{with} \quad \log\lambda_{j,k} = \eta_{j,k}.$$

MLE is explicit : $\lambda_{j,k} = O_{j,k}/R_{j,k}$.

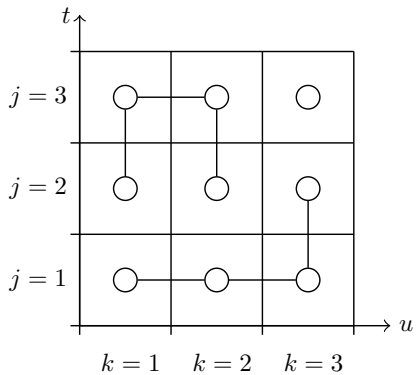
Bidimensional fused adaptive ridge :

$$\ell_n^{\text{pen}}(\boldsymbol{\eta}) = \ell_n(\boldsymbol{\eta}) + \kappa\sum_{j,k}\left\{v_{j,k}\left(\eta_{j+1,k} - \eta_{j,k}\right)^2 + w_{j,k}\left(\eta_{j,k+1} - \eta_{j,k}\right)^2\right\}.$$

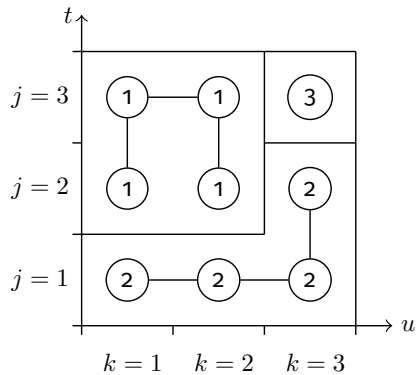
Segmentation into constant areas



(a) Representation of $v_{j,k}\left(\eta_{j+1,k}-\eta_{j,k}\right)^2$ et $w_{j,k}\left(\eta_{j,k+1}-\eta_{j,k}\right)^2$

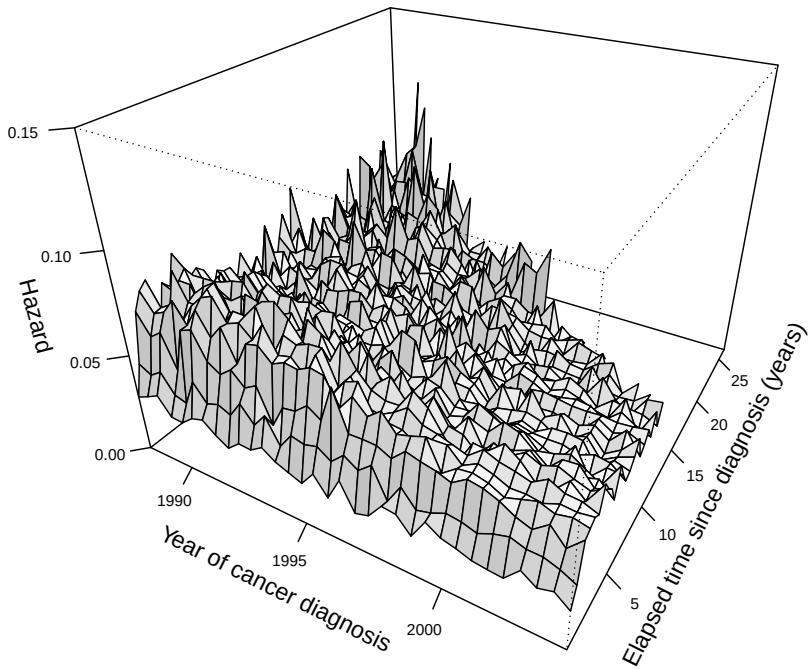


(b) Corresponding graph



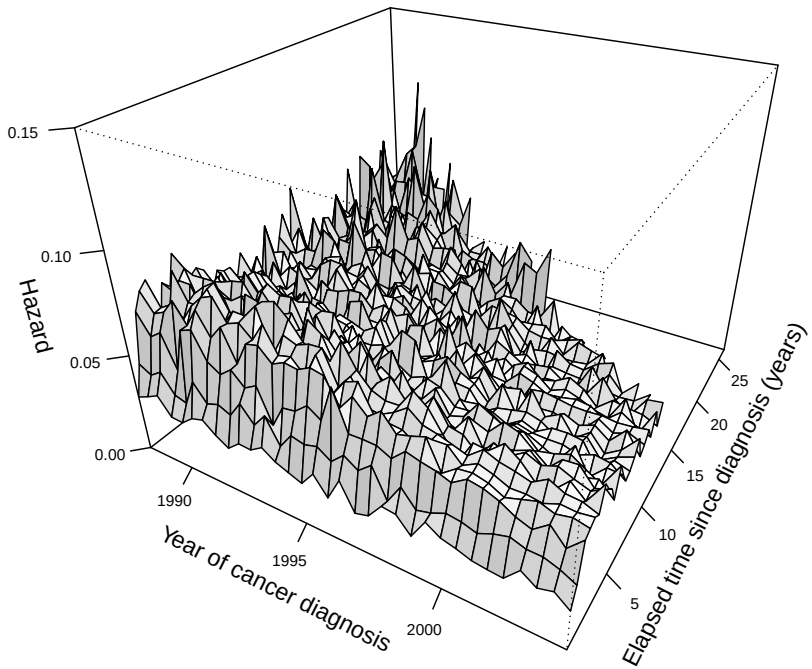
(c) Segmentation into connected components

Segmented estimate of the hazard rate

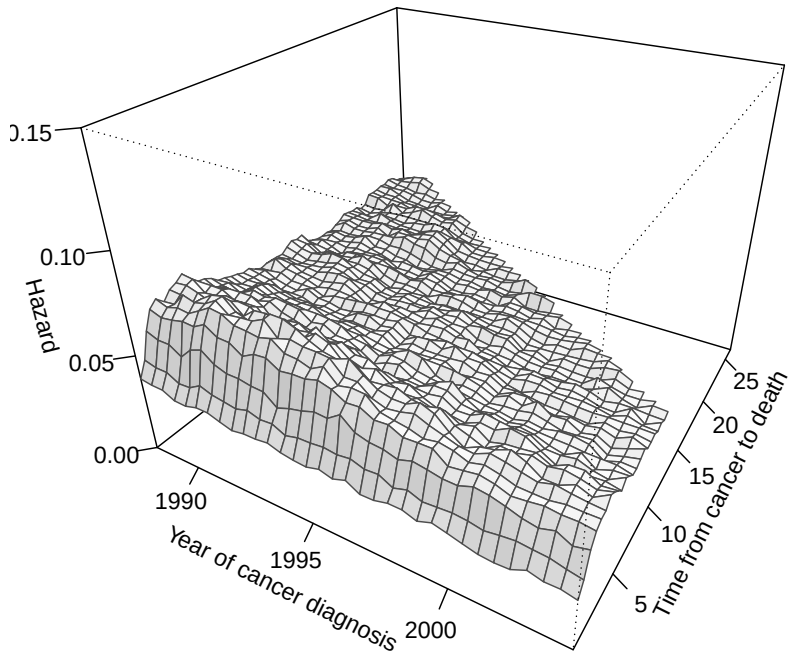


(Unpenalized) MLE

Segmented estimate of the hazard rate

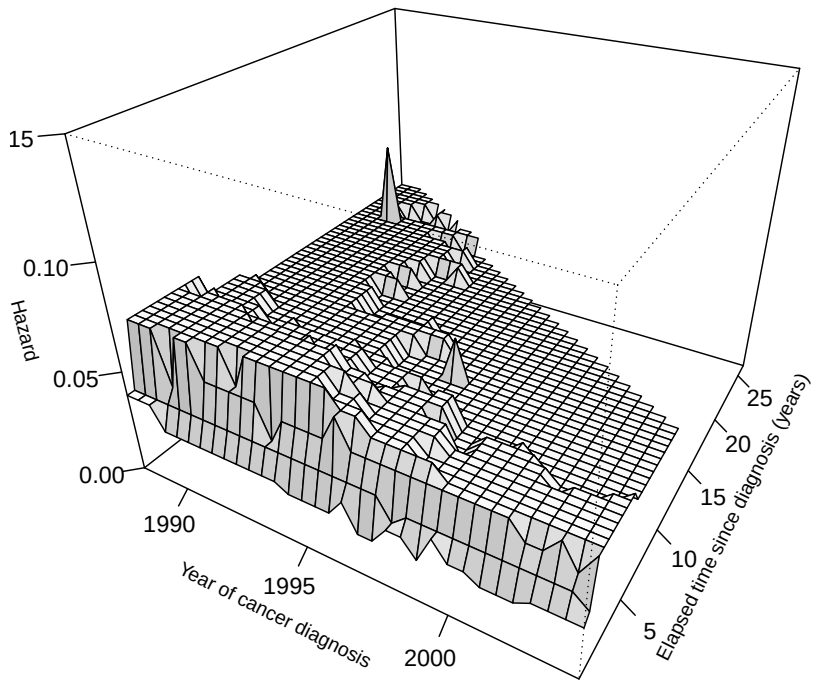


(Unpenalized) MLE

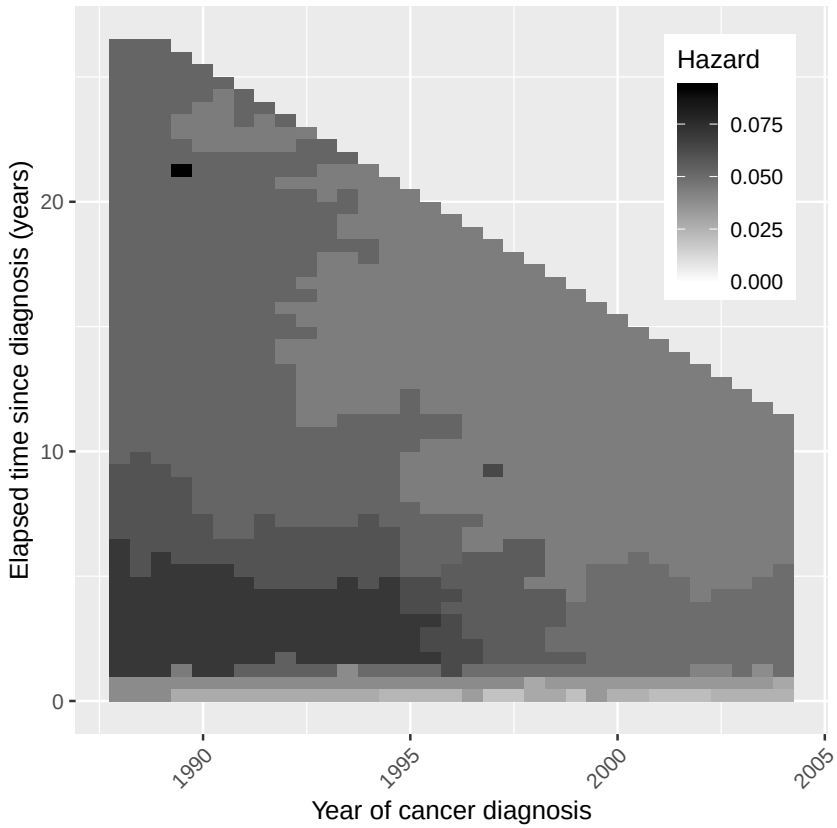


L_2 Regularization

Results with segmentation



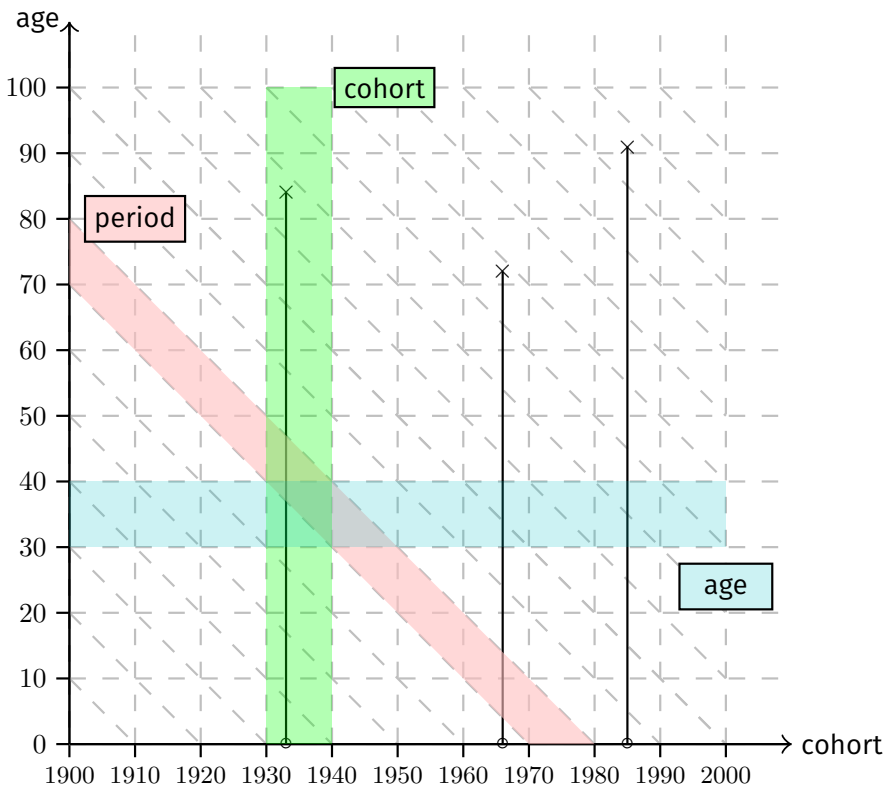
Fused Adaptive Ridge (perspective)



Fused Adaptive Ridge (greyscale)

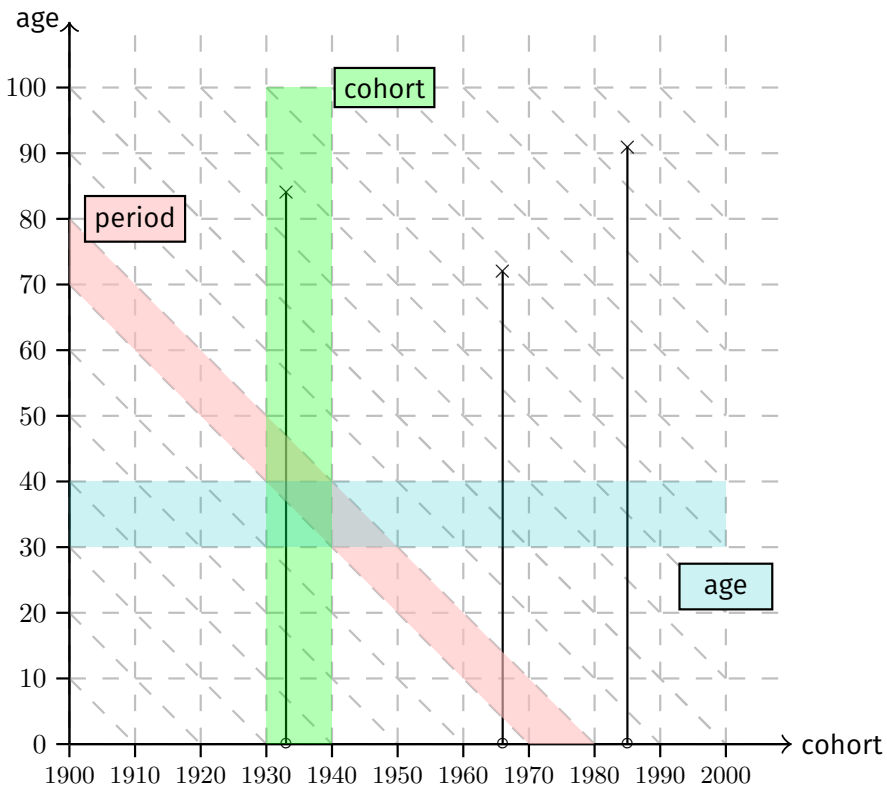
2.2 Extension to age-period-cohort analysis

Age-period-cohort analysis

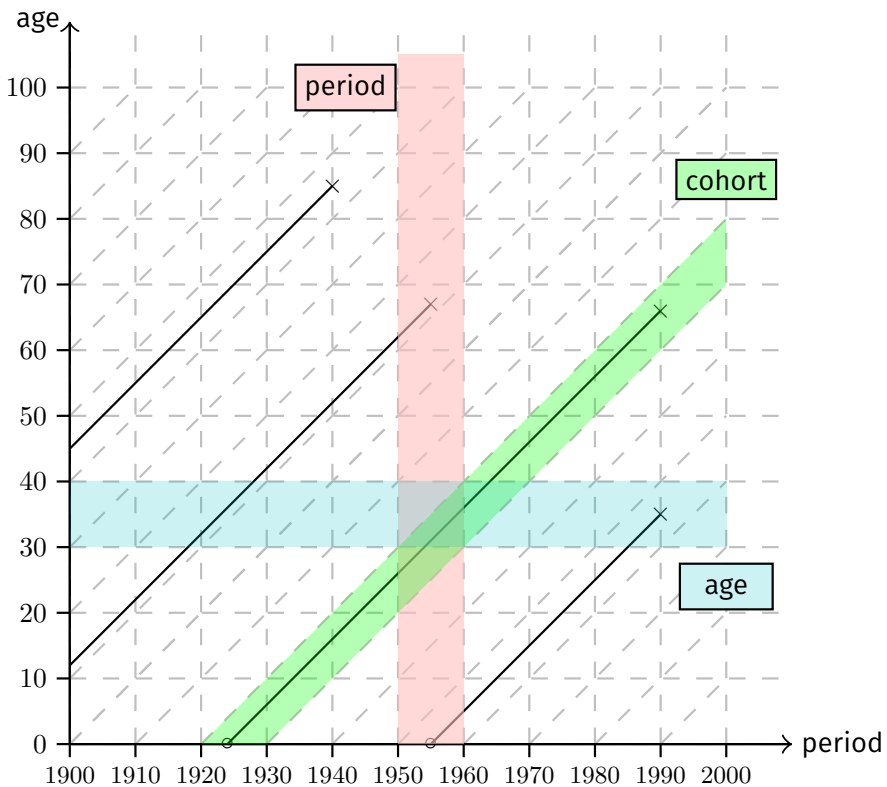


Lexis Diagram : Age-Cohort

Age-period-cohort analysis



Lexis Diagram : Age-Cohort



Lexis diagram : Age-Period

Key relation : period $=$ age $+$ cohort.

Age-period-cohort analysis

We want to infer the effects of age, period, and cohort.

Example : incidence of breast cancer :

- age effect : menopause
- cohort effect : carcinogenic baby food
- period effect : nuclear accident

We define one parameter vector for each effect : α, β et γ

Existing models

1. In the AGE-PERIOD-COHORT[3] model, we assume

$$\log \lambda_{j,k} = \alpha_j + \beta_k + \gamma_{j+k-1}.$$

   - Non-identifiable :
     - infer $\Delta^2\boldsymbol{\alpha}$, $\Delta^2\boldsymbol{\beta}$ et $\Delta^2\boldsymbol{\gamma}$.
     - or add a constraint to the model.

2. In the AGE-COHORT[3] model, we assume

$$\log \lambda_{j,k} = \alpha_j + \beta_k.$$

   - Outer product structure $\rightarrow$ regularizing
   - Additive effect of the variables : strong *a priori*

[3] : Carstensen, B., Age–period–cohort models for the Lexis diagram, *Statistics in medicine*, 2007.

Our approach : model the effects and their interactions

- We introduce an AGE-COHORT-INTERACTION model :

$$\log\left(\lambda_{j,k}\right) = \alpha_j + \beta_k + \delta_{j,k},$$

where $\delta_{j,k}$ is the interaction (with $\delta_{1,k} = \delta_{j,1} = 0$).

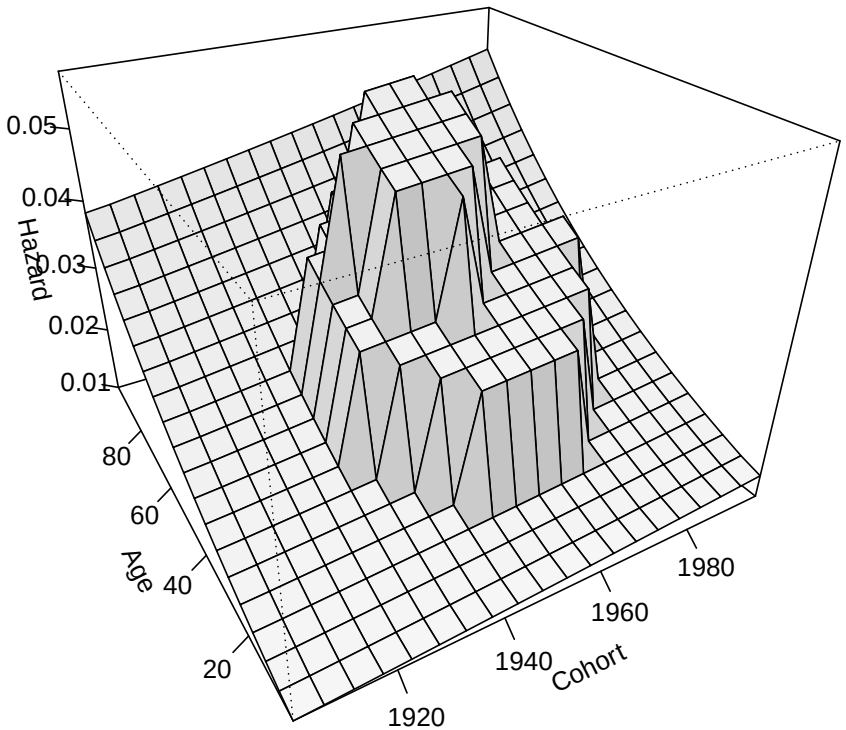
- We regularize over the differences of $\delta_{j,k}$:

$$\ell_n^{\text{pen}}(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}) + \kappa \sum_{j,k} \left\{ v_{j,k} \left(\delta_{j+1,k} - \delta_{j,k}\right)^2 + w_{j,k} \left(\delta_{j,k+1} - \delta_{j,k}\right)^2 \right\}.$$

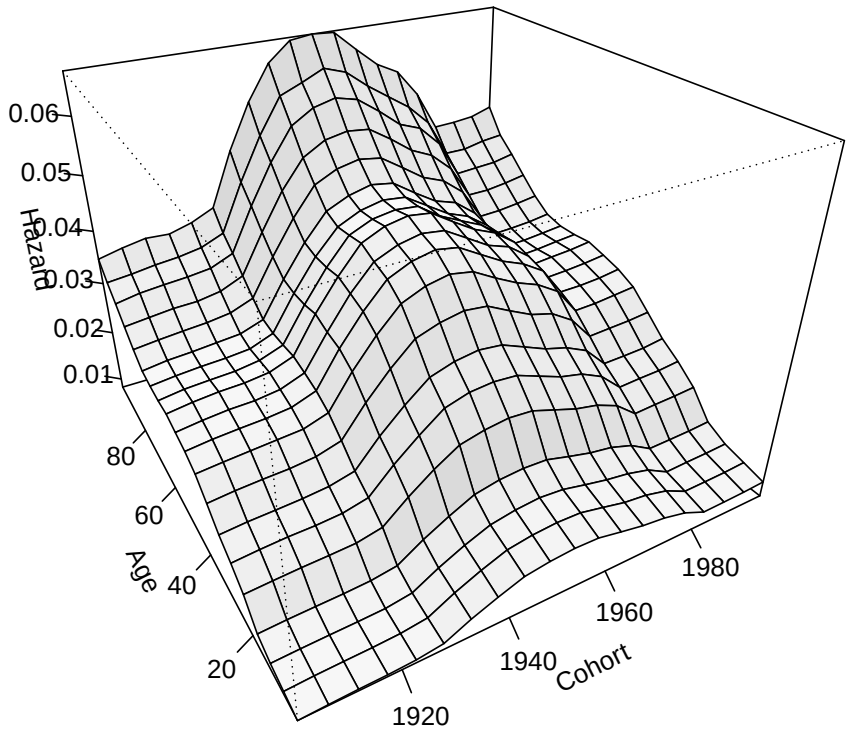
- Compromise between :
  - AC Model : $\kappa \to \infty$
  - MLE : $\kappa \to 0$

- $J = 20$ age intervals and $K = 20$ cohort intervals
- Sample the cohort uniformly
- Sample the age using the hazard rate $(\lambda_{j,k})$
- Uniform censoring over the age $[75, 100]$
- Infer $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$ in the ACI model.
- We represent medians over 100 repetitions

Illustration : simulated data

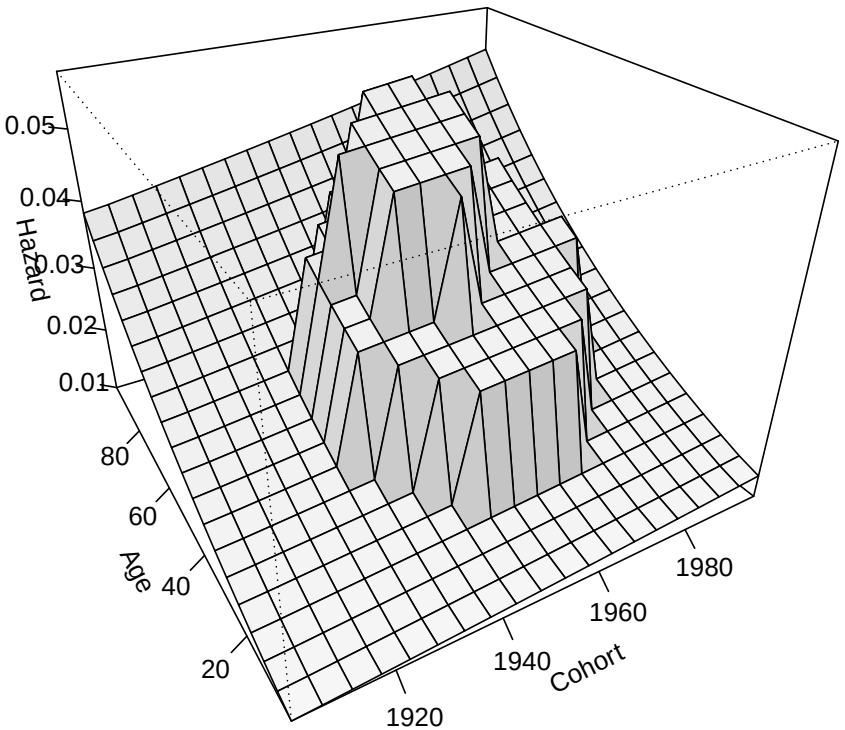


True hazard $\lambda^*_{j,k}$

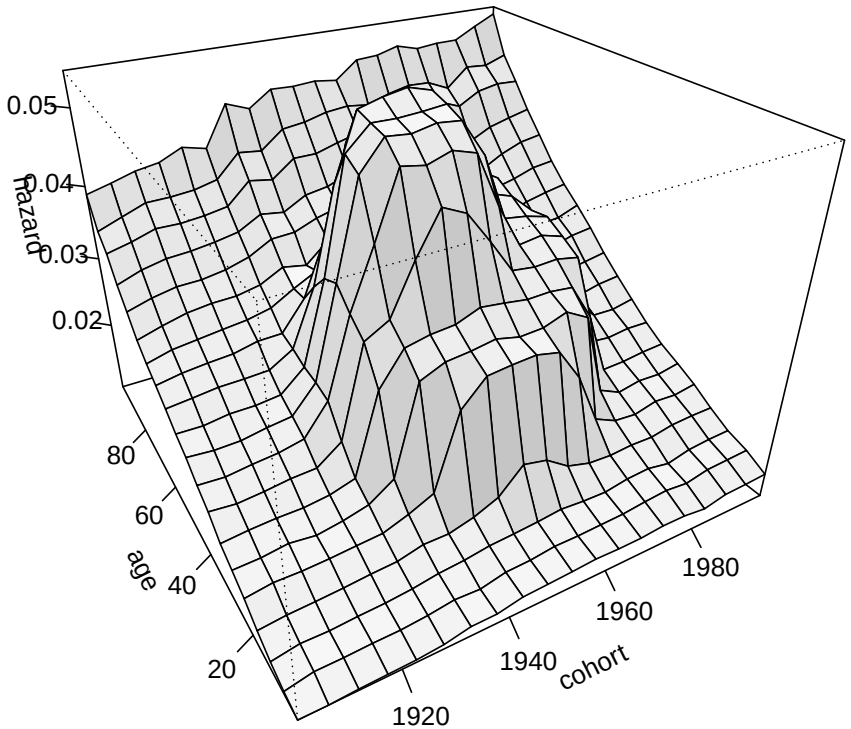


Age-cohort model
$(\log \lambda_{j,k} = \alpha_j + \beta_k)$

Results with the ACI model

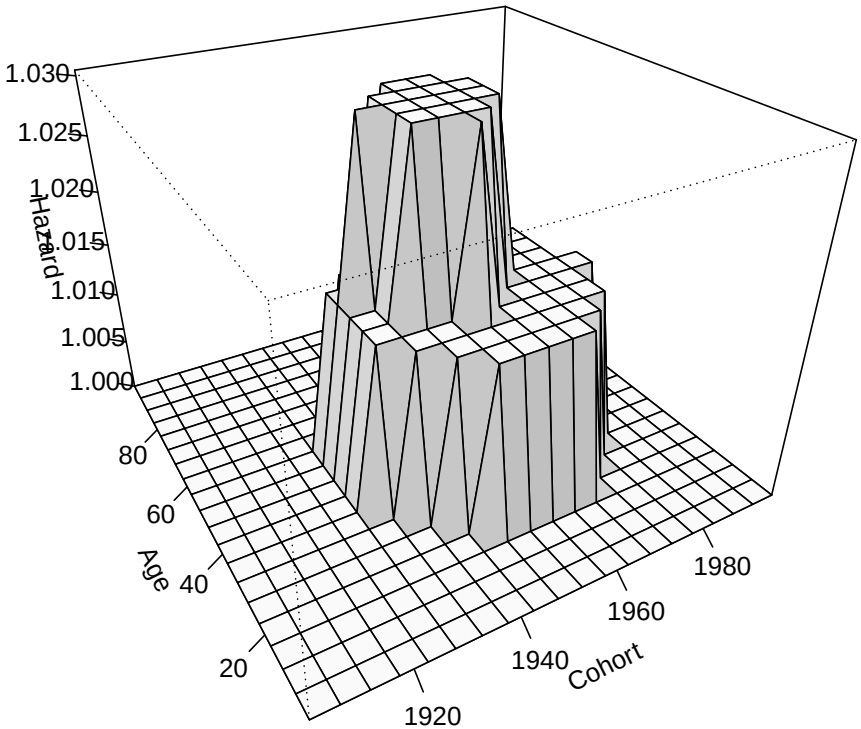


True hazard $\lambda^*_{j,k}$

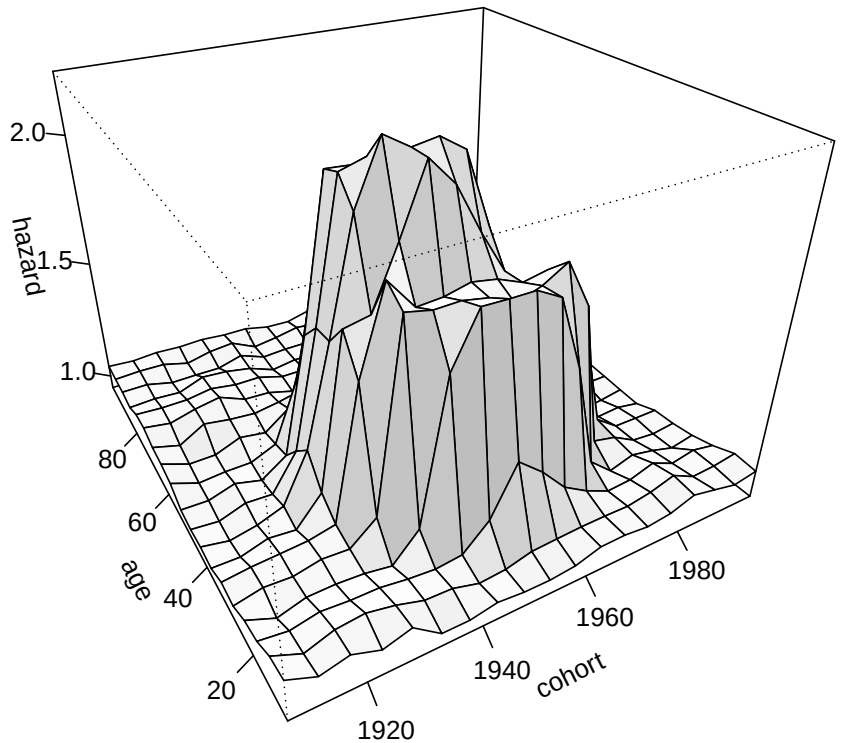


ACI model : estimated $\hat{\lambda}_{j,k}$

Results with the ACI model

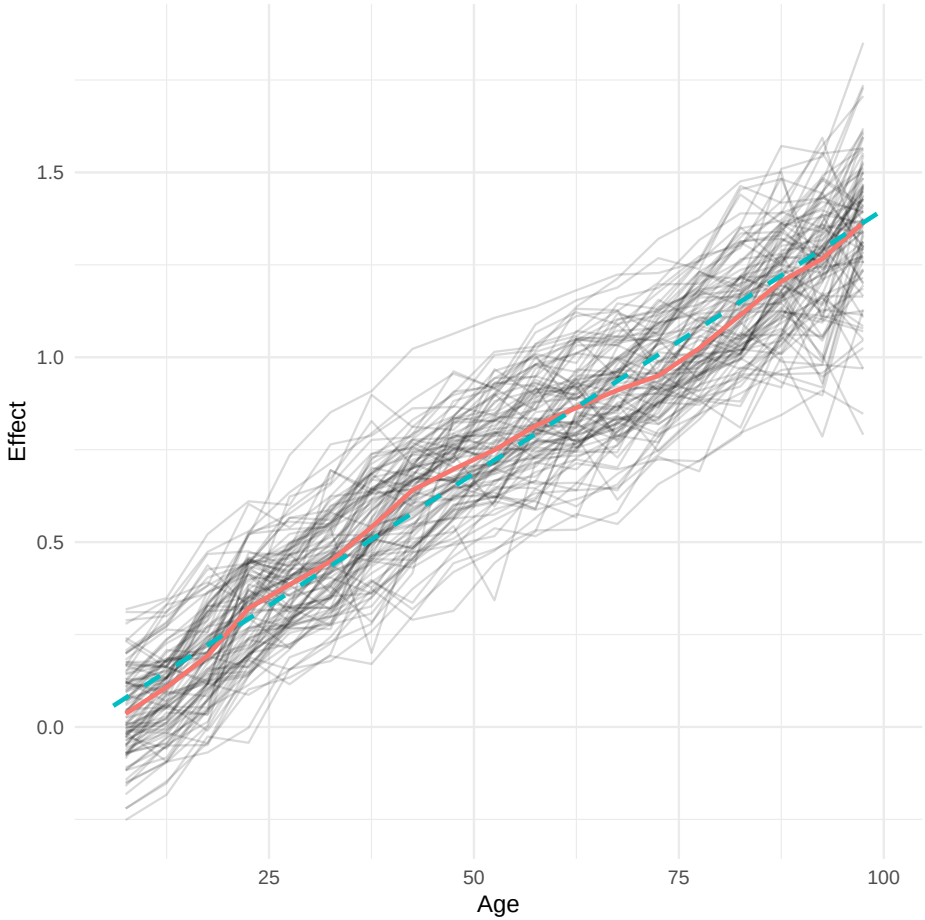


True interaction : $\delta^*_{j,k}$

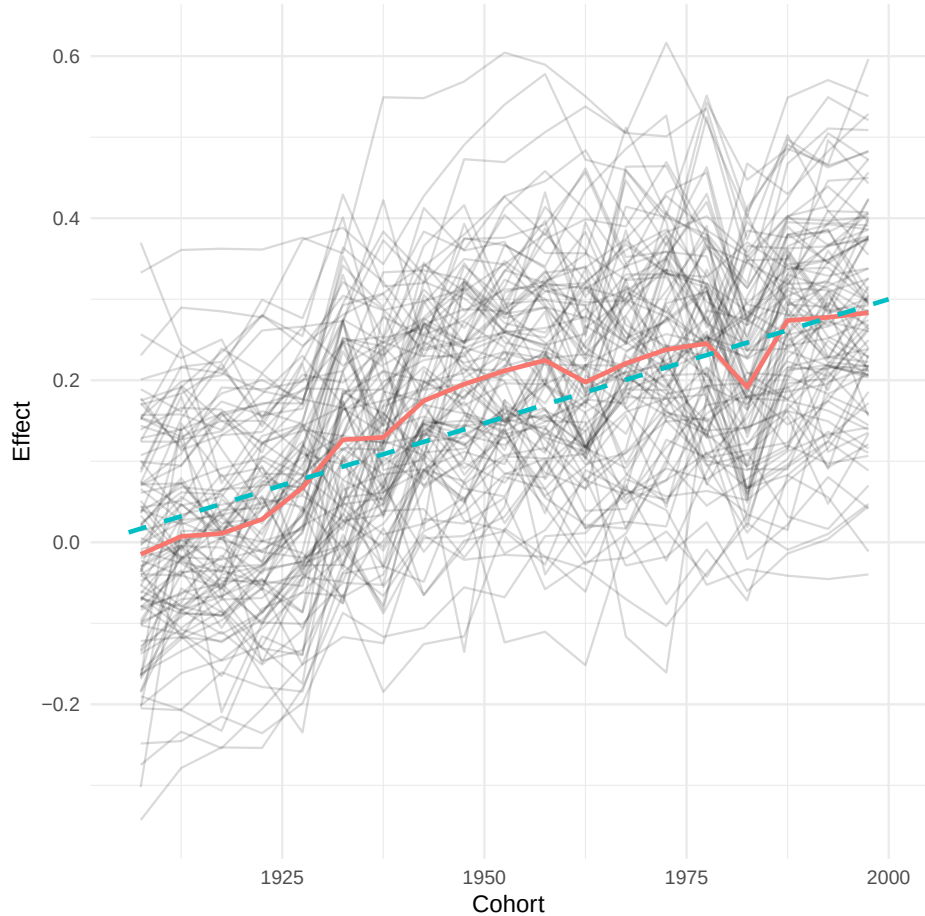


ACI model : estimated $\hat{\delta}_{j,k}$

Results with the ACI model



ACI model : age effect $\hat{\alpha}_j$



ACI model : cohort effect $\hat{\beta}_j$

3. Spline Regression with Automatic Knot Selection

Functional regression

Let $(x_i, y_i) \in \mathbb{R}$:

$$y_i = f(x_i) + \varepsilon_i, \quad 1 \leq i \leq n,$$

- f “smooth”
- ε_i i.i.d and centered
- $x_i \in [a, b]$

Aim : infer f.

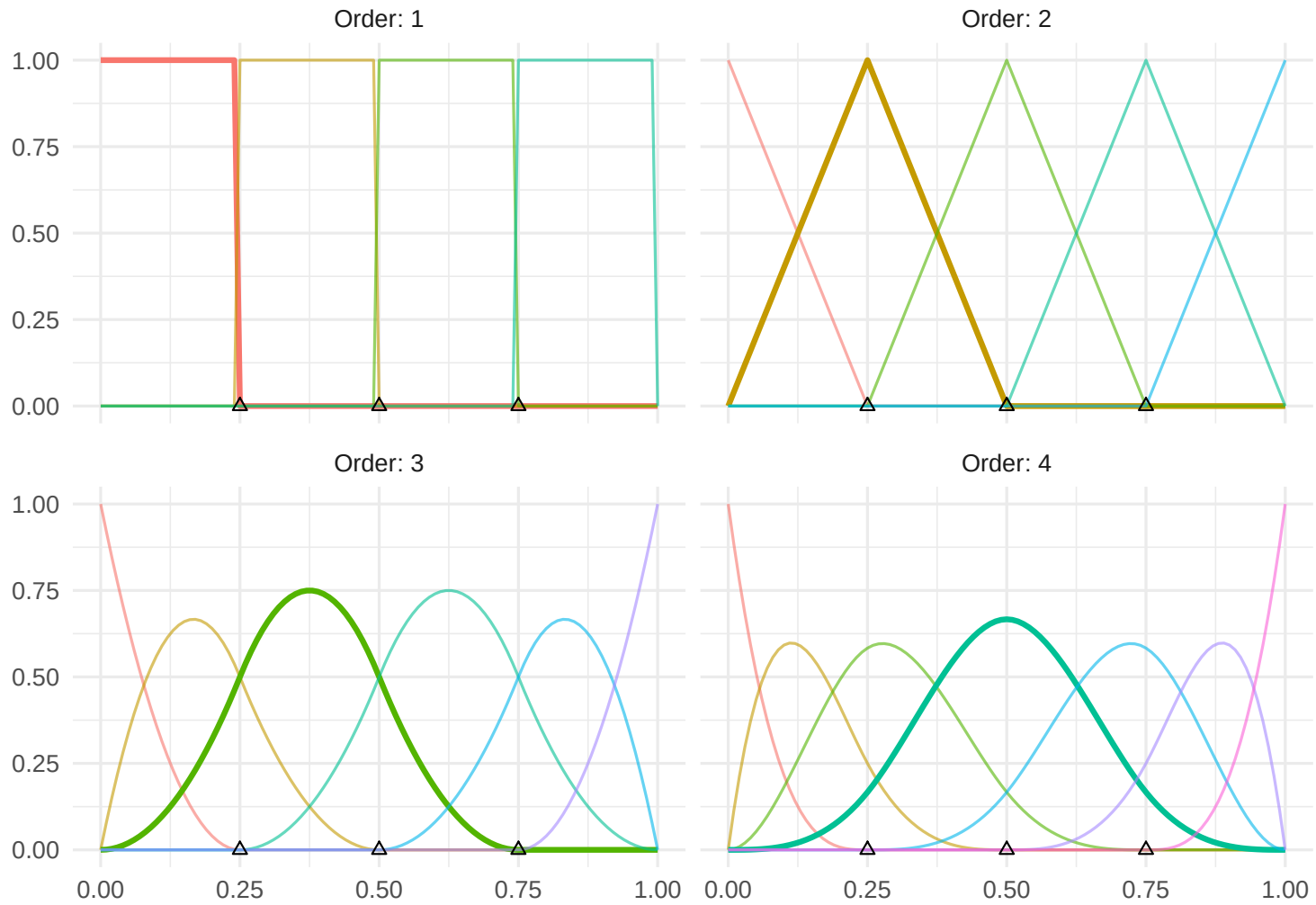
Spline regression

- Define the spline order $q \geq 1$ and knots $(t_1, \cdots, t_k) \in [a, b]$.
- The estimate is the spline $\hat{f}(x) = \sum_{j=1}^{q+k} a_j B_{j,q}(x)$, where $B_{j,q}$ is the B-spline of order q.
- We minimize in $\boldsymbol{a} \in \mathbb{R}^{q+k}$:

$$\mathrm{SS}(\boldsymbol{a}, \boldsymbol{t}) = \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{q+k} a_j B_{j,q}(x_i) \right\}^2 = \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}\|^2,$$

B-spline regression

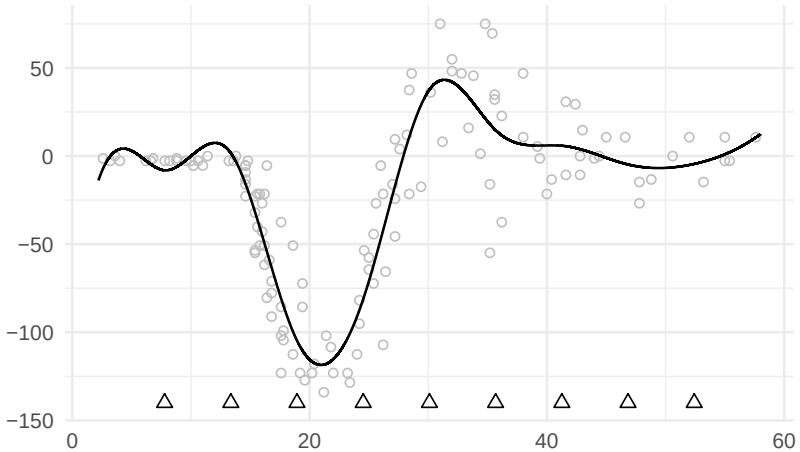
$(B_{j,q})_{j=1}^{k+q}$ is a basis of splines of order q.



B-splines bases with three knots : $(0.25, 0.5, 0.75)$.

Knot placement

Where to place knots ?



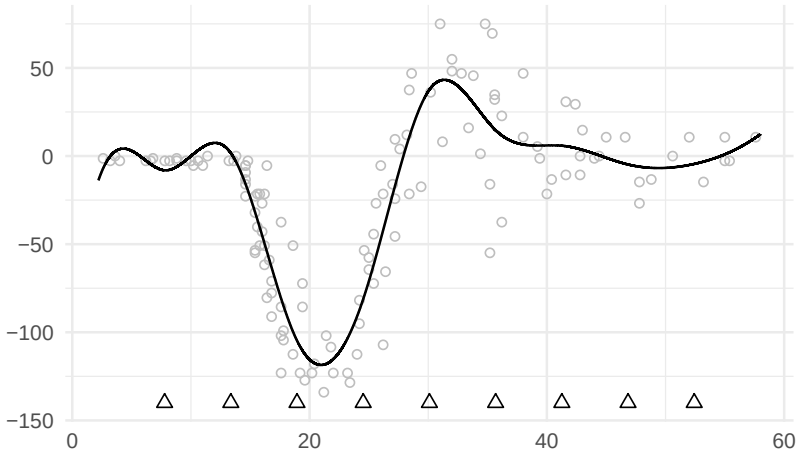
Uniform knots

Helmet crash test data[5]

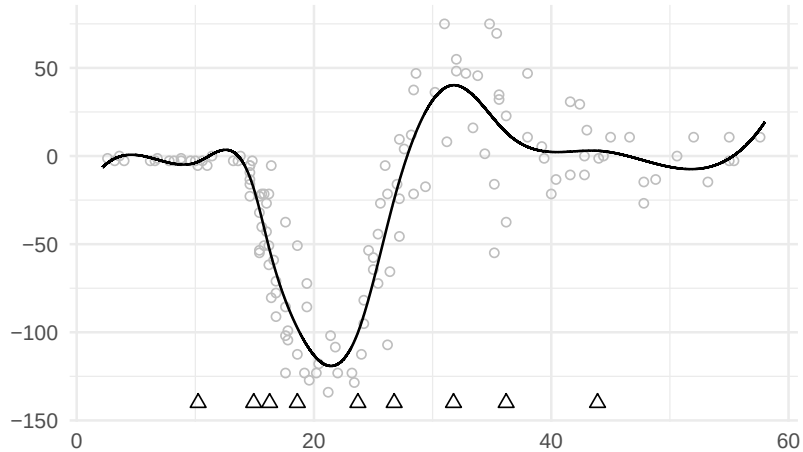
[5] : Silverman, B.W., Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting, *Journal of the Royal Statistical Society*, 1985.

Knot placement

Where to place knots ?



Uniform knots



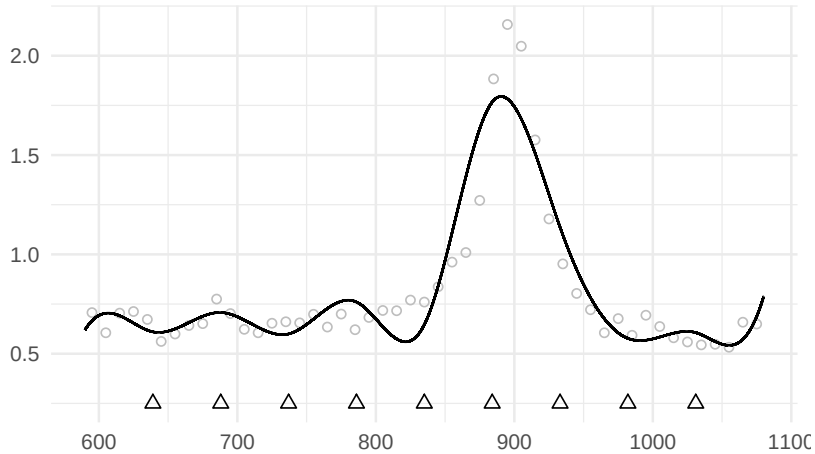
Knots at quantiles

Helmet crash test data[5]

[5] : Silverman, B.W., Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting, *Journal of the Royal Statistical Society*, 1985.

Knot placement

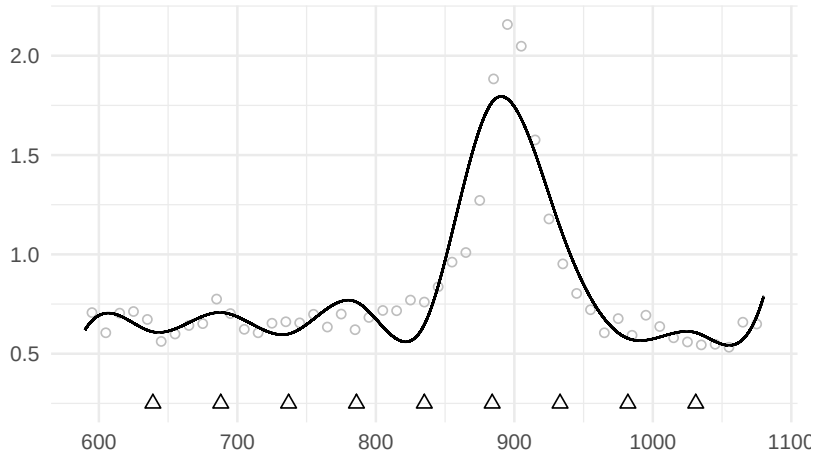
It is only based on the x-distribution of the points :



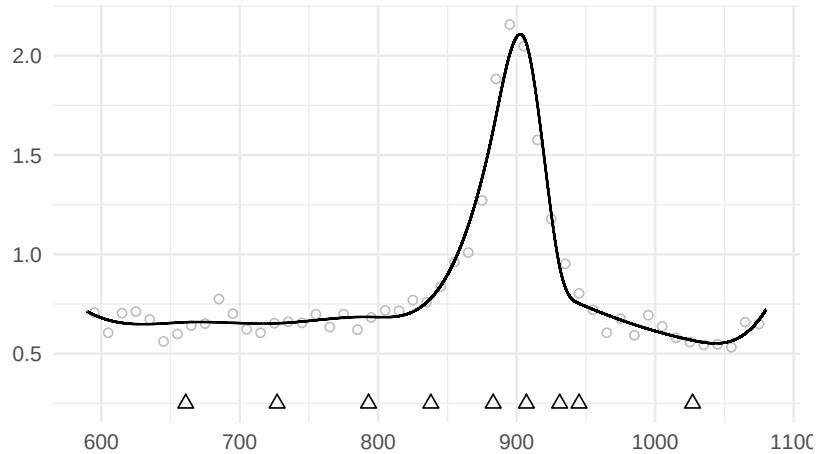
Uniform = quantile knots

Knot placement

It is only based on the x-distribution of the points :



Uniform = quantile knots



Manually set knots

- Need to find optimal knot position
- Their optimal position is informative about breakpoints in f.

Two different approaches

- **Predictive approach** : place too many knots and regularize $\boldsymbol{a}$: P-splines[4].
- **Explicative approach** :
  - Jointly optimize w.r.t. $\boldsymbol{t}$ and $\boldsymbol{a}$: Jupp (1978), Lindstrom (1999)
  - Monte-Carlo-based Approach : Denison et al. (1998), DiMatteo et al. (2001)

[4] : Eilers, P. and Marx, D., Flexible Smoothing with B-splines and Penalties, *Statistical Science*, 1996.

A-Spline

Our approach :

- Set many initial knots
- Successively remove the least relevant knots
- Using the fused adaptive ridge

Enforce the successive values of $\boldsymbol{a}$ to be equal.

$$\mathrm{PSS}\left(\boldsymbol{a}, \boldsymbol{w}\right) = \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}\|_2^2 + \kappa \sum_{j=q+1}^{q+k} w_j \left(\Delta^q a_j\right)^2,$$

$$\text{with} \quad \Delta a_j = a_j - a_{j-1} \quad \text{and} \quad \Delta^q = \Delta^{q-1} \circ \Delta$$

Our approach :

- Set many initial knots
- Successively remove the least relevant knots
- Using the fused adaptive ridge

Enforce the successive values of $\boldsymbol{a}$ to be equal.

$$\mathrm{PSS}\left(\boldsymbol{a}, \boldsymbol{w}\right) = \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}\|_2^2 + \kappa \sum_{j=q+1}^{q+k} w_j \left(\Delta^q a_j\right)^2,$$

$$\text{with} \quad \Delta a_j = a_j - a_{j-1} \quad \text{and} \quad \Delta^q = \Delta^{q-1} \circ \Delta$$

Explicit iteration step :

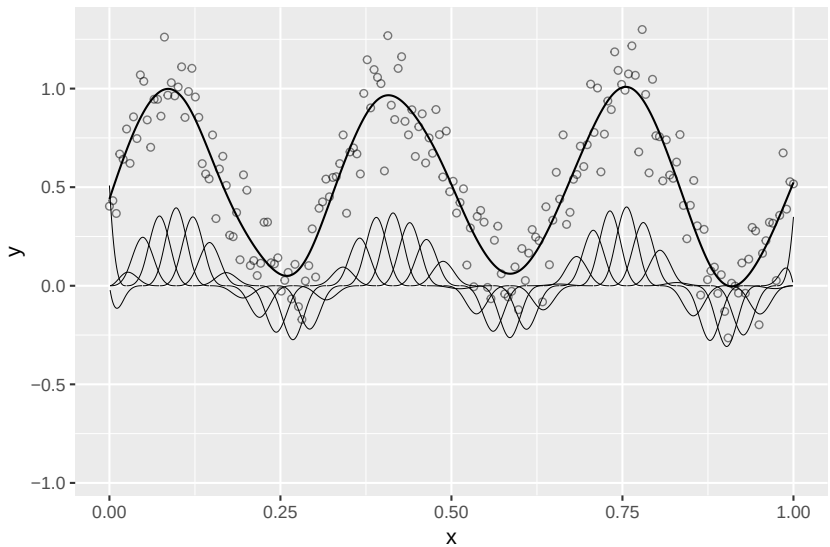
$$\text{argmin}_{\boldsymbol{a}} \mathrm{PSS}(\boldsymbol{a}, \boldsymbol{w}) = (\boldsymbol{B}^T\boldsymbol{B} + \kappa \boldsymbol{D}_{\boldsymbol{q}}^T \boldsymbol{W} \boldsymbol{D}_{\boldsymbol{q}})^{-1} \boldsymbol{B}^T \boldsymbol{y}$$

$$\text{with} \quad \boldsymbol{D_1} = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{bmatrix}$$

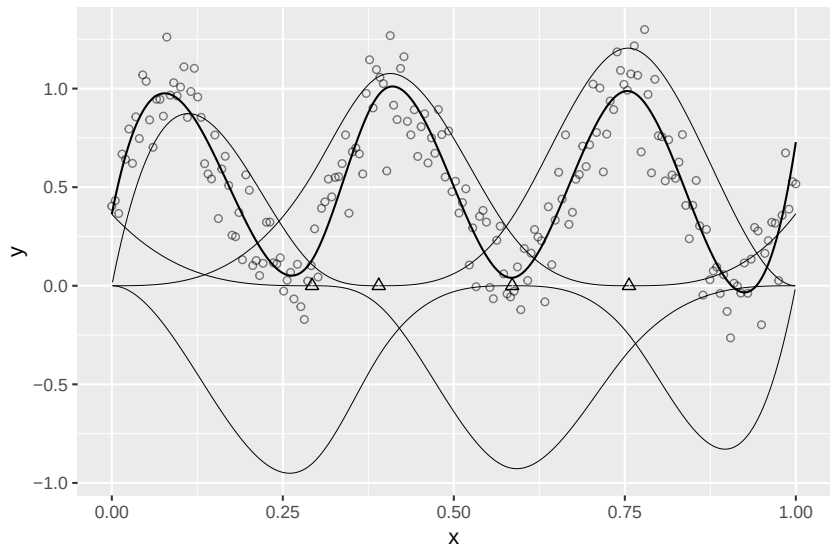
Results on synthetic data

Illustration on simulated data :

- $f(x) = \sin(6\pi x) * 0.5 + 0.5$
- $\varepsilon \sim \mathcal{N}(0, 0.15^2)$ and $n = 200$.
- $q = 4$
- 40 uniform initial knots



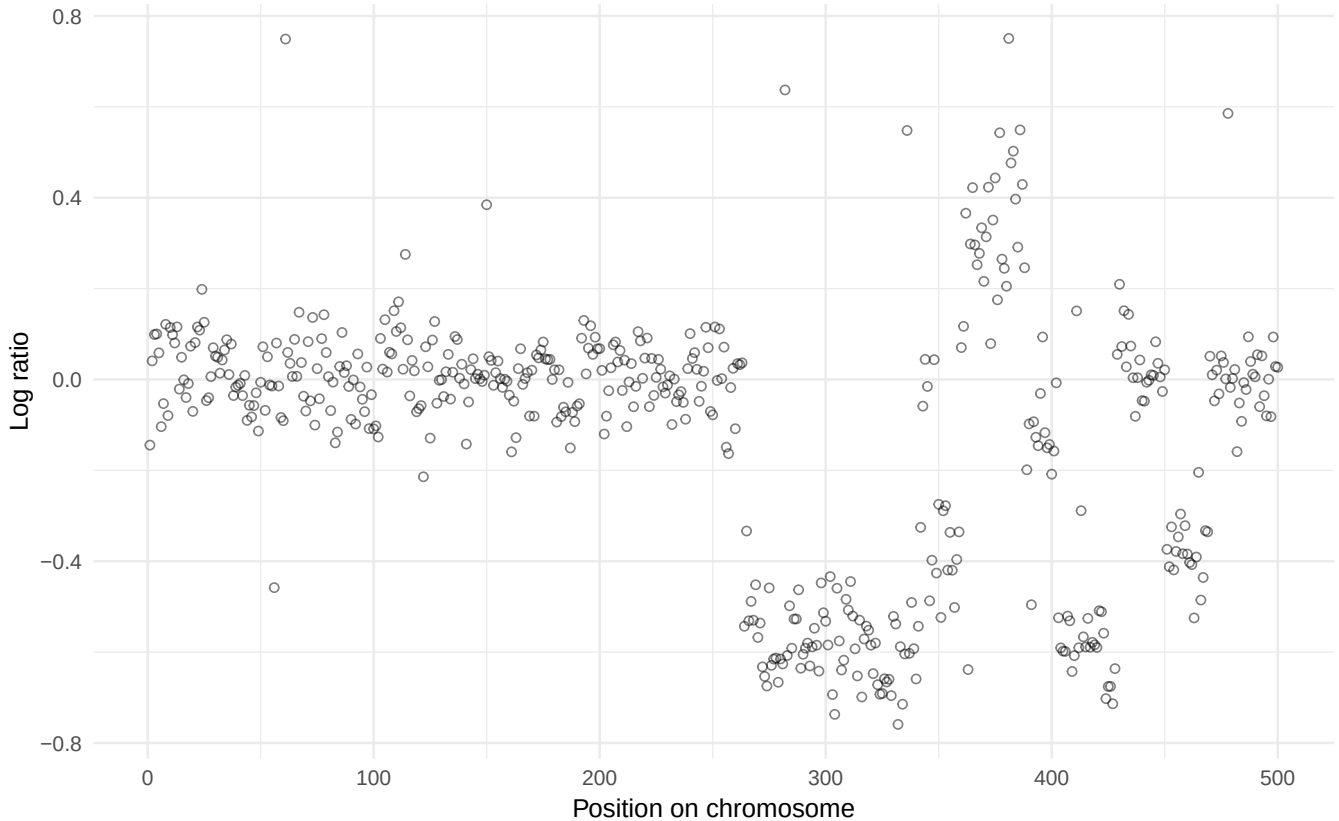
P-splines



A-splines

Results on real data : detecting changes in mean

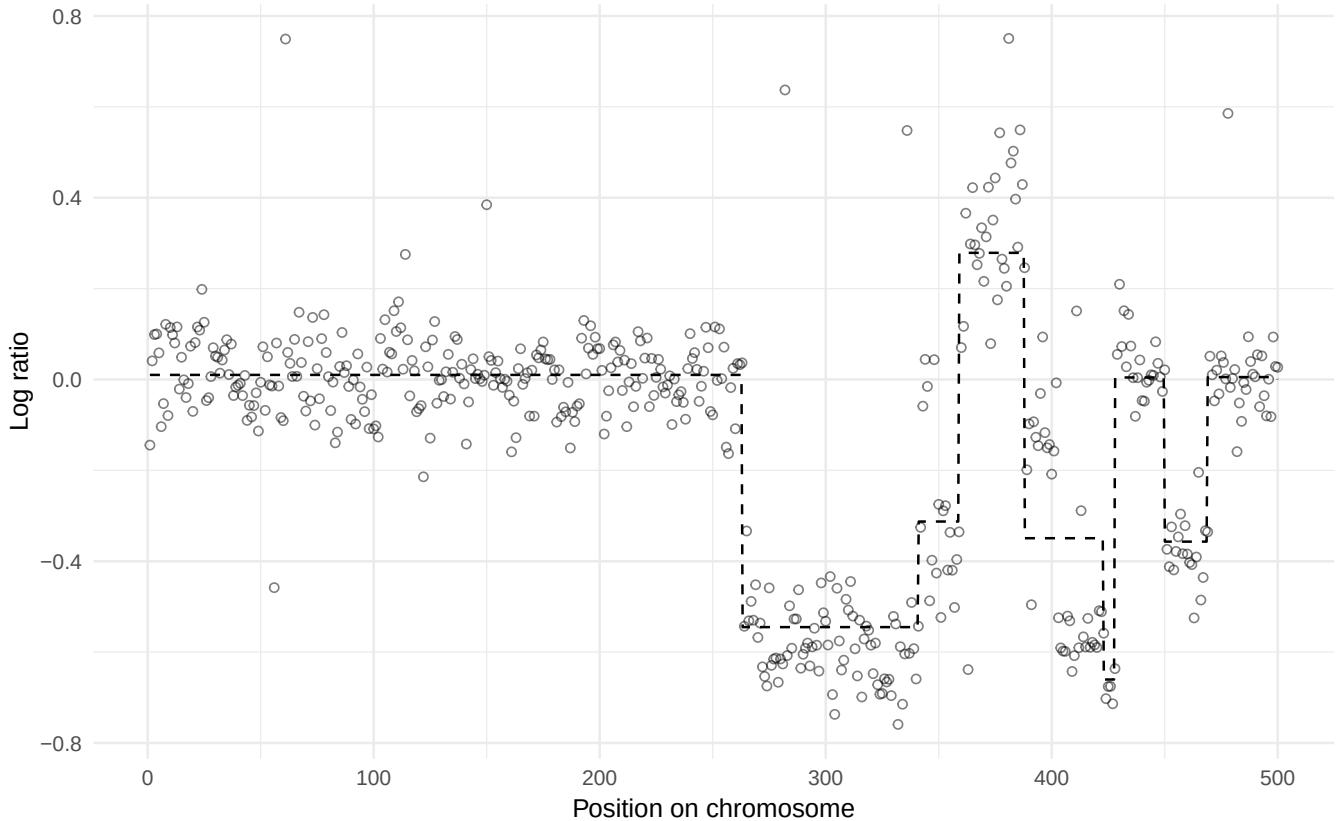
aCGH profile of bladder tumor samples[6] :



[6] : Stransky, N. et al, Regional Copy Number–Independent Deregulation of Transcription in Cancer, *Nature Genetics*, 2006.
[7] : Killick, R. et al, Optimal detection of changepoints with a linear computational cost, *Journal of the American Statistical Association*, 2006.

Results on real data : detecting changes in mean

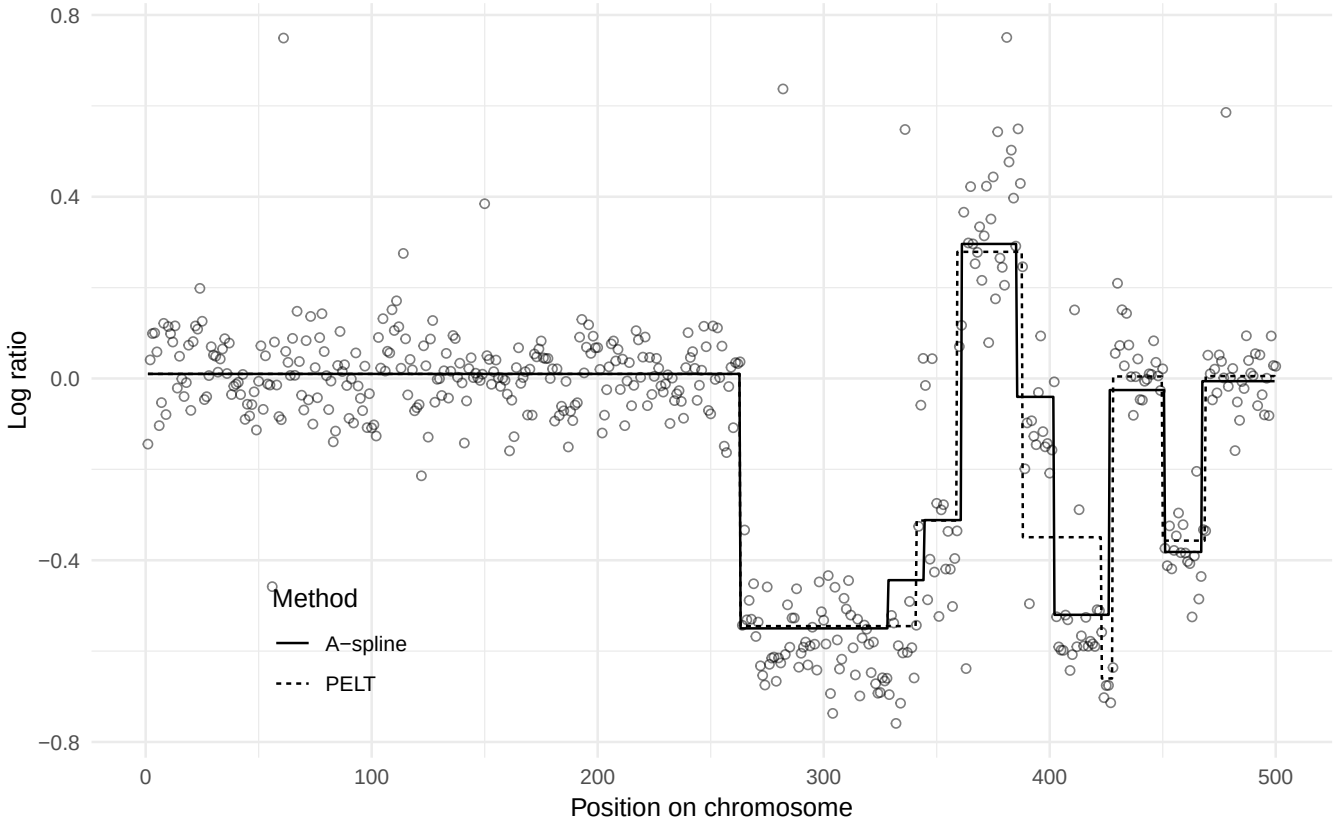
aCGH profile of bladder tumor samples[6] :



[6] : Stransky, N. et al, Regional Copy Number–Independent Deregulation of Transcription in Cancer, *Nature Genetics*, 2006.
[7] : Killick, R. et al, Optimal detection of changepoints with a linear computational cost, *Journal of the American Statistical Association*, 2006.

Results on real data : detecting changes in mean

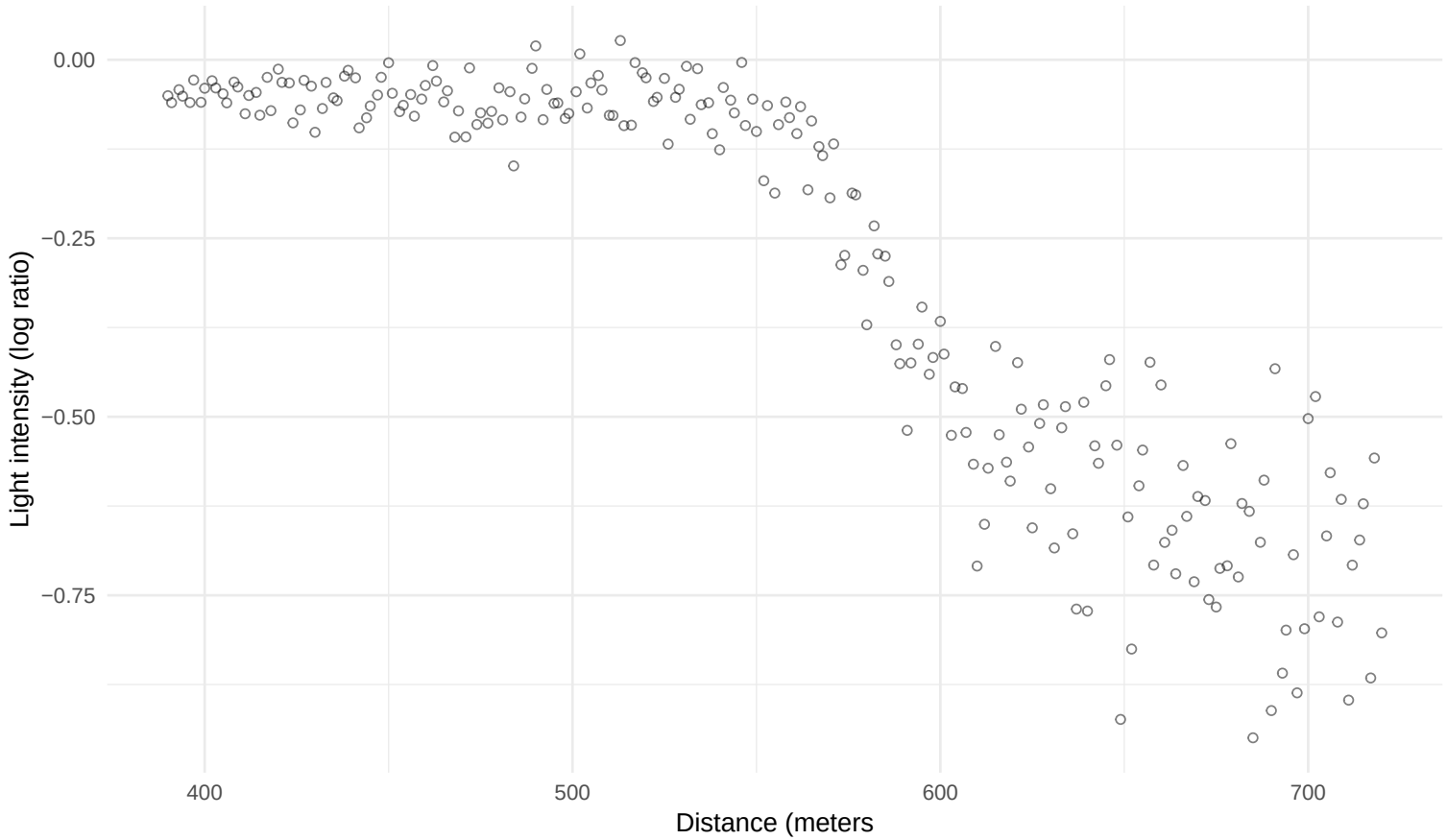
aCGH profile of bladder tumor samples[6] :



[6] : Stransky, N. et al, Regional Copy Number–Independent Deregulation of Transcription in Cancer, *Nature Genetics*, 2006.
[7] : Killick, R. et al, Optimal detection of changepoints with a linear computational cost, *Journal of the American Statistical Association*, 2006.

Results on real data : detecting changes of slope

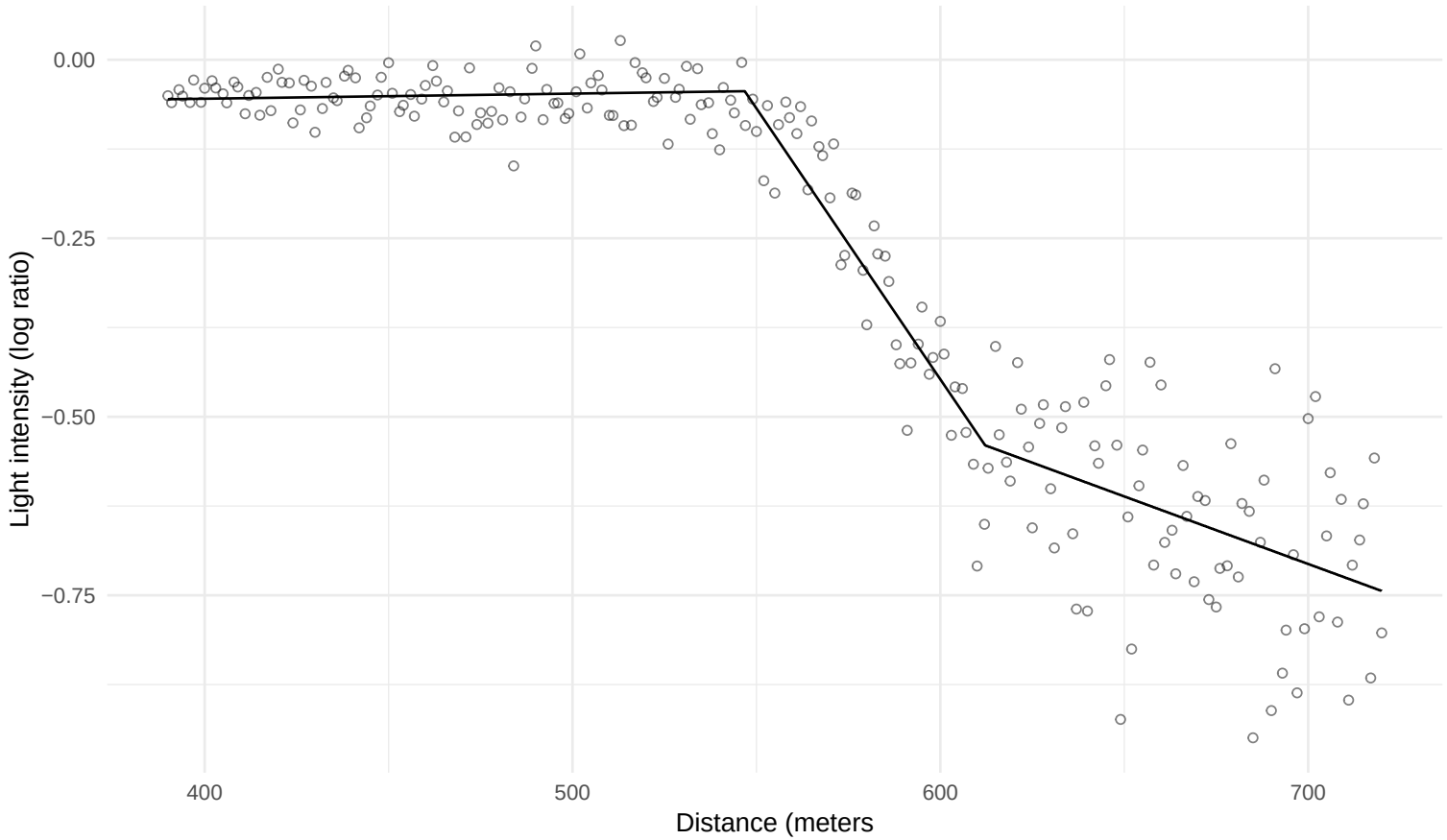
Light measurement data set[8] :



[8] : Sigrist, M. et al, Air monitoring by spectroscopic techniques, *John Wiley & Sons*, 1994.
[9] : Friedman ; J., Multivariate Adaptive Regression Splines, *Journal of the American Statistical Association*, 1991.

Results on real data : detecting changes of slope

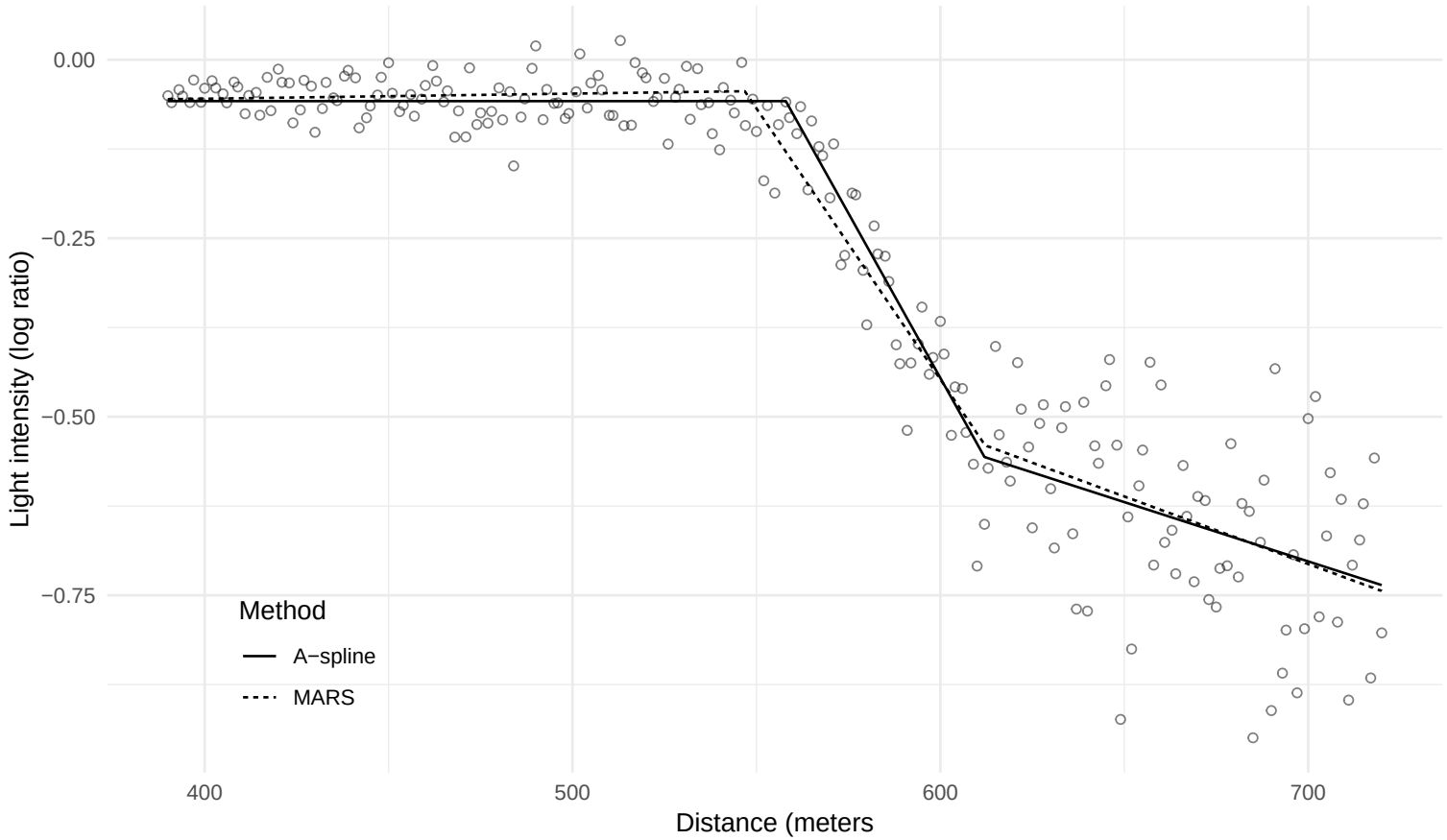
Light measurement data set[8] :



[8] : Sigrist, M. et al, Air monitoring by spectroscopic techniques, *John Wiley & Sons*, 1994.
[9] : Friedman ; J., Multivariate Adaptive Regression Splines, *Journal of the American Statistical Association*, 1991.

Results on real data : detecting changes of slope

Light measurement data set[8] :



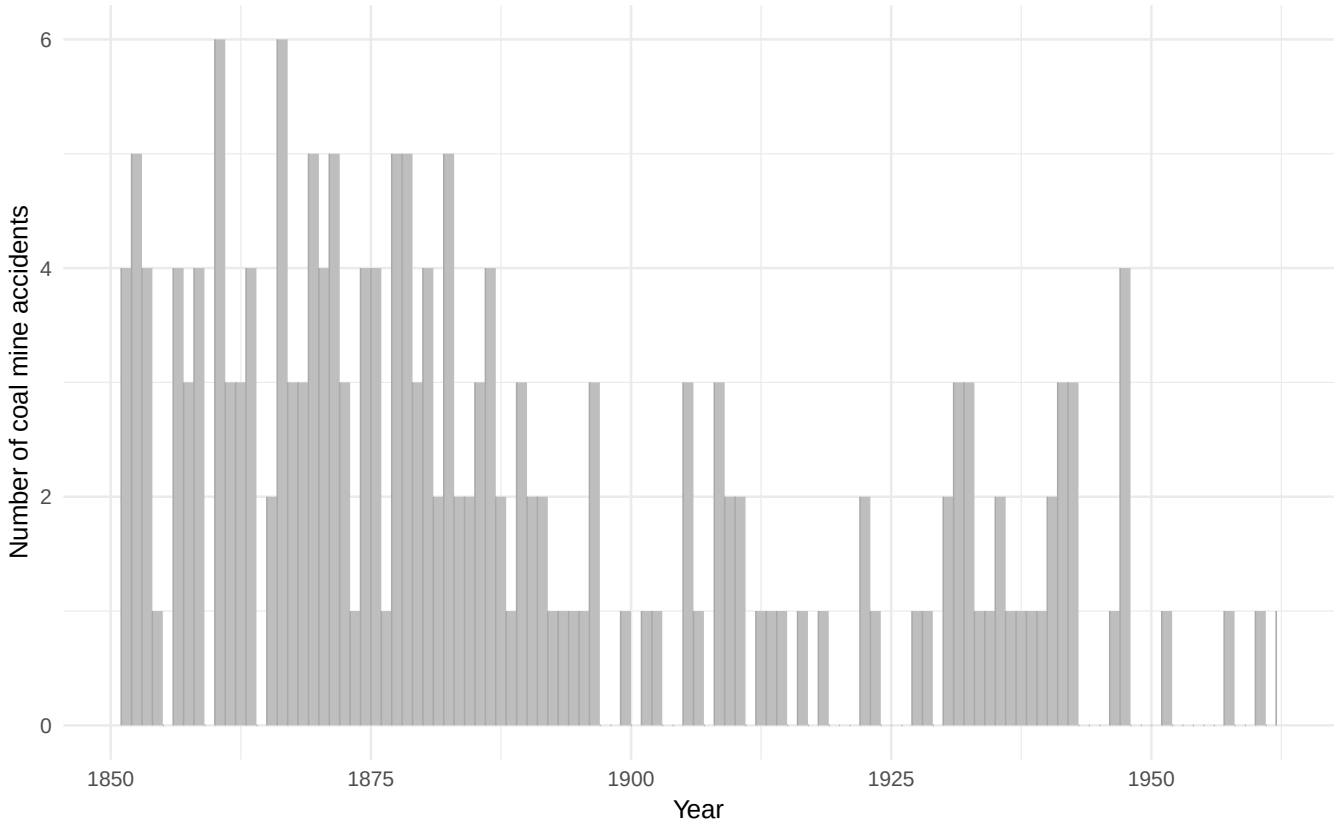
[8] : Sigrist, M. et al, Air monitoring by spectroscopic techniques, *John Wiley & Sons*, 1994.

[9] : Friedman ; J., Multivariate Adaptive Regression Splines, *Journal of the American Statistical Association*, 1991.

Results on real data : splines of higher order

Number of disasters in coal mines in the UK[10] :

Poisson model : $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{\mu}$ with $\hat{\boldsymbol{\mu}} = \exp(\boldsymbol{B}\hat{\boldsymbol{a}})$

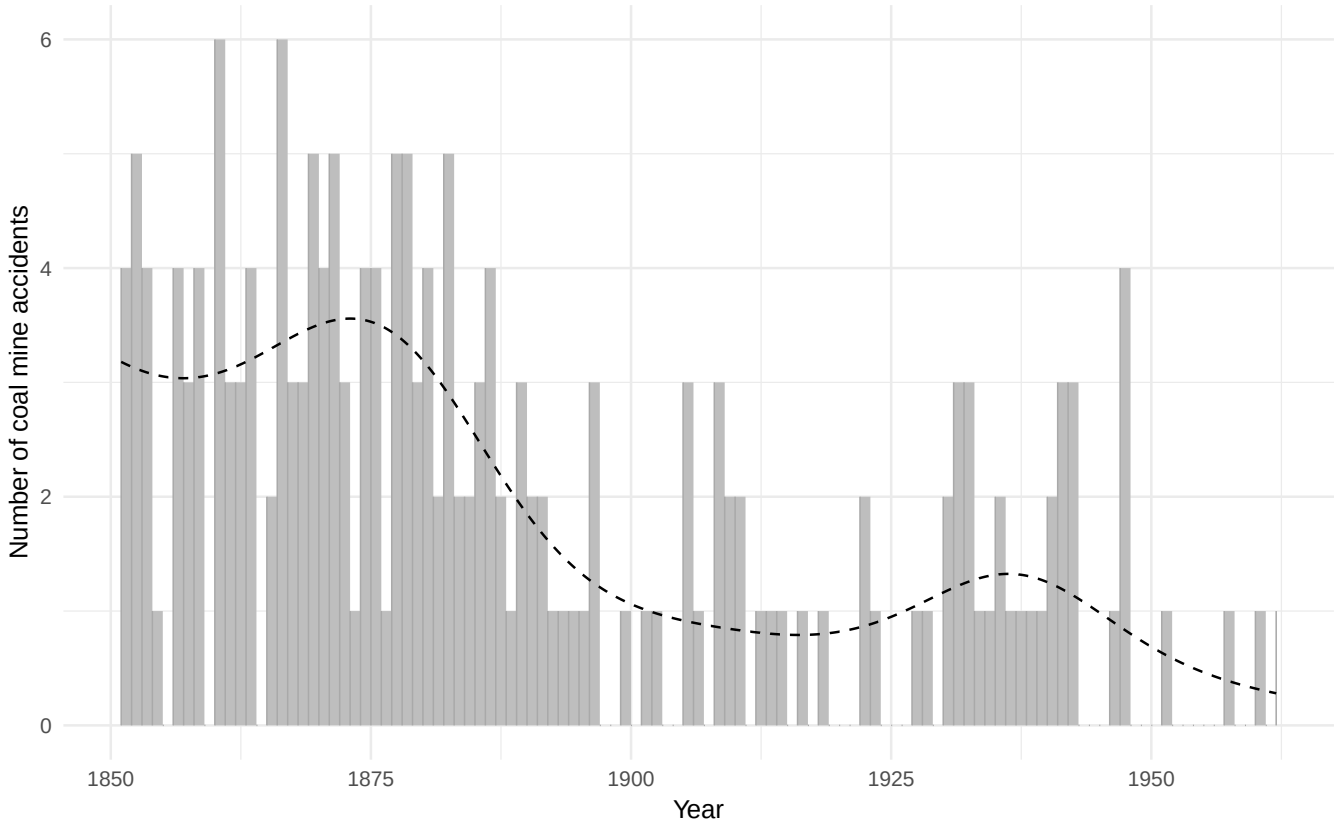


[10] : Diggle, P and Marron, J., Equivalence of Smoothing Parameter Selectors in Density and Intensity Estimation, *JASA*, 1988.

Results on real data : splines of higher order

Number of disasters in coal mines in the UK[10] :

Poisson model : $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{\mu}$ with $\hat{\boldsymbol{\mu}} = \exp(\boldsymbol{B}\hat{\boldsymbol{a}})$

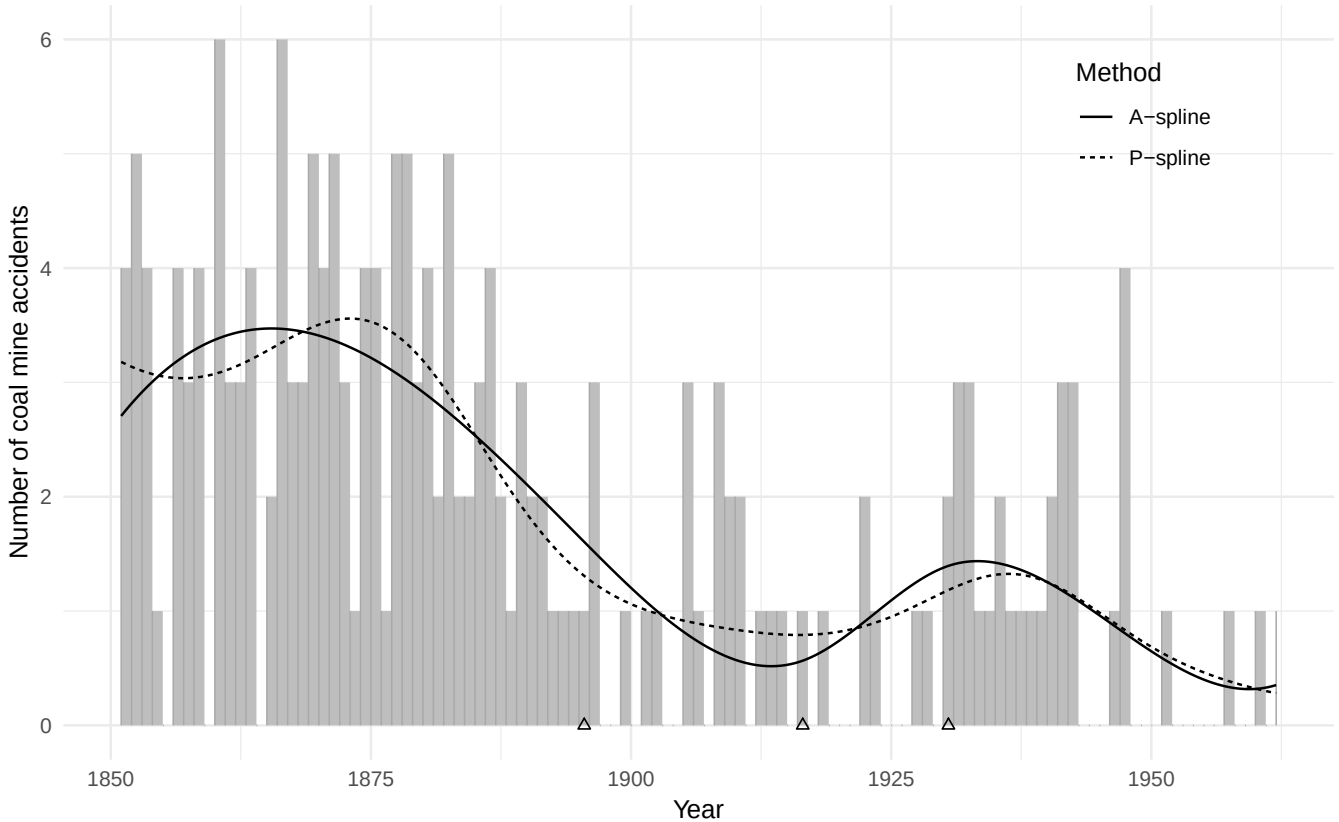


[10] : Diggle, P and Marron, J., Equivalence of Smoothing Parameter Selectors in Density and Intensity Estimation, *JASA*, 1988.

Results on real data : splines of higher order

Number of disasters in coal mines in the UK[10] :

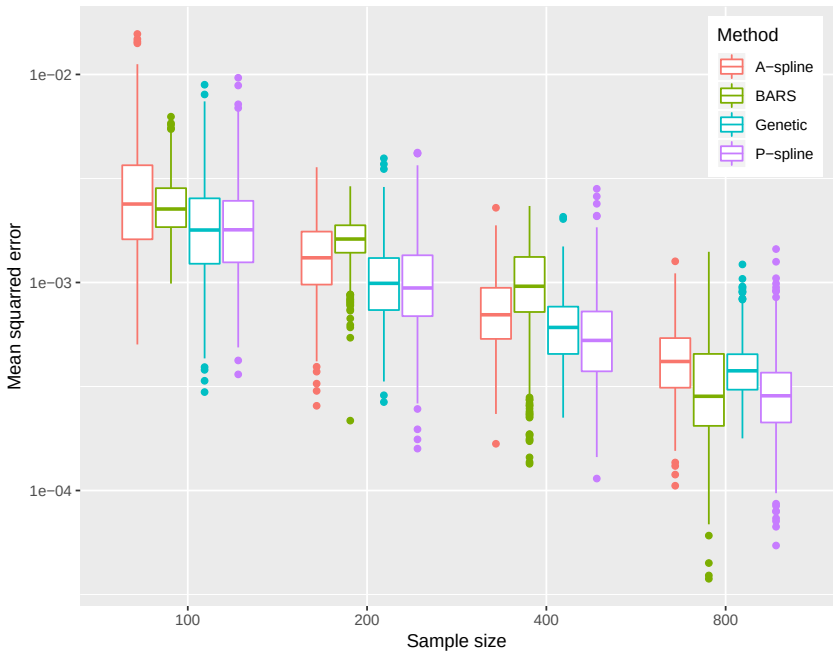
Poisson model : $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{\mu}$ with $\hat{\boldsymbol{\mu}} = \exp(\boldsymbol{B}\hat{\boldsymbol{a}})$



[10] : Diggle, P and Marron, J., Equivalence of Smoothing Parameter Selectors in Density and Intensity Estimation, *JASA*, 1988.

Comparison of predictive performance

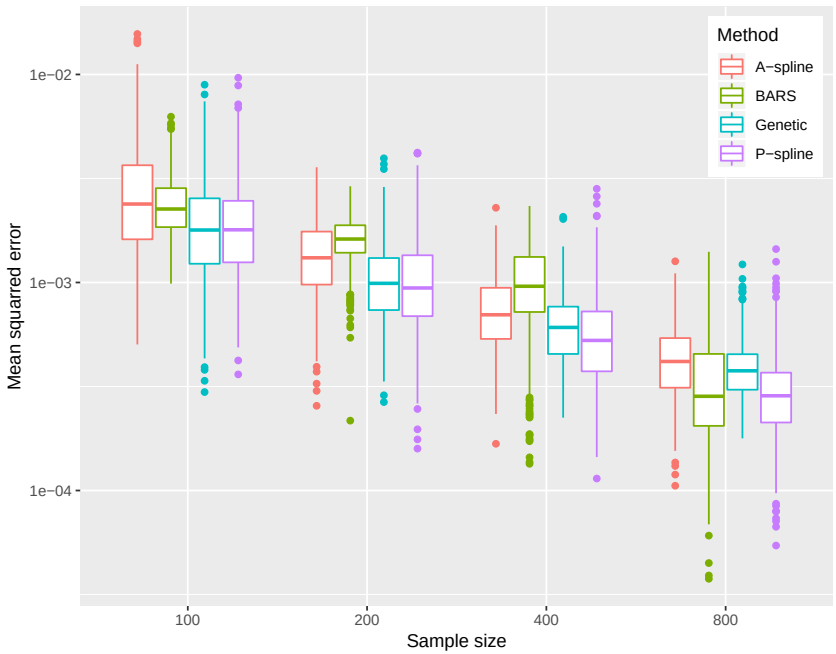
Comparison of mean squarred error (MSE) on simulated data :



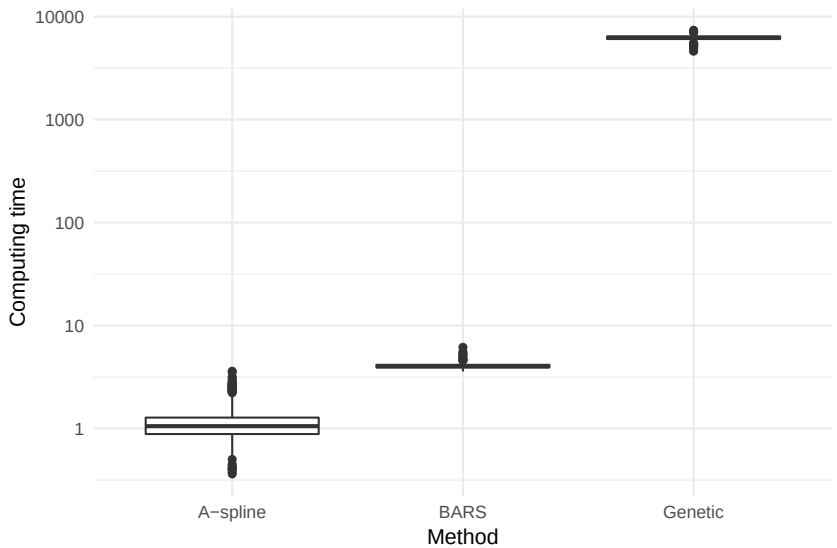
MSE

Comparison of predictive performance

Comparison of mean squarred error (MSE) on simulated data :



MSE



Computing time

Conclusion & perspectives

Conclusion :

- Papers :
  - Bidimensional estimation of the hazard rate [submitted]
  - Spline regression with automatic knot selection [submitted]
  - Age-cohort-interaction model [under preparation]
  - Segmentation of geographic-based data [under preparation]
- Communications :
  - Conference talks : SAM 2017, IWAP 2018
  - Conference posters : SMPGD 2018, IBC 2018, SMPGD 2019
  - Three invited seminars
- Three `R` packages : `hazreg`, `aspline`, and `graphseg`.

Perspectives :

- Study of consistency of the segmentation (ACI model).
- Use of splines in age-period-cohort models.
- Application to study of evolution of mortality causes in France.

Annex 1 : Adaptive ridge and MM optimization

Local Quadratic Approximation : Solve iteratively

$$\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) + \kappa \sum_j p(|\beta_j|),$$

where p is even, concave on $\mathbb{R}_+$, and nondecreasing.

$$p(|\beta_j^{(k)}|) + (\beta_j^2 - \beta_j^{(k)2}) \frac{p'(|\beta_j^{(k)}|)}{|\beta_j^{(k)}|} \succeq p(|\beta_j|)$$

MM Optimization :

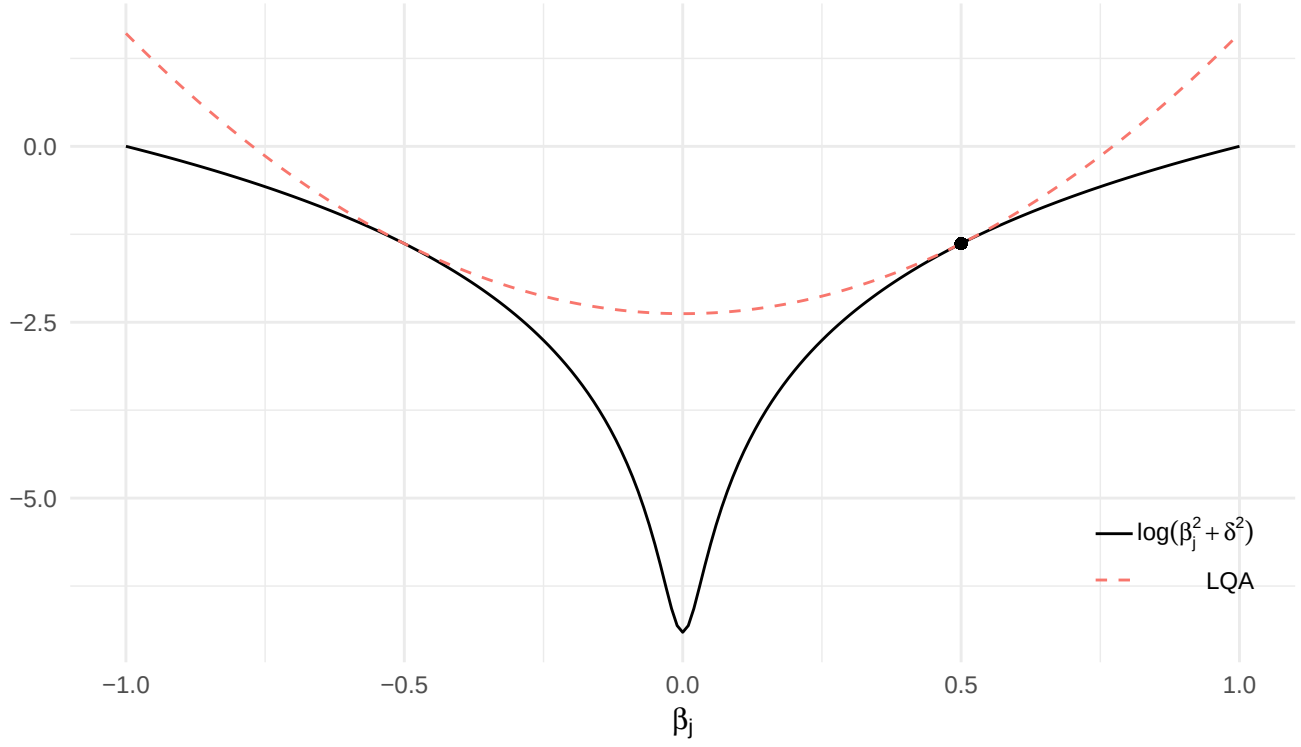
- $\min_{\boldsymbol{\beta}} \ell^{\text{pen}}(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$
- Update $\boldsymbol{\beta}^{(k)}$ to $\boldsymbol{\beta}^{(k+1)}$

The sequence $\boldsymbol{\beta}^{(k)}$ tends to a local minimum of $\ell^{\text{pen}}(\boldsymbol{\beta})$.

Fan, J. and Li, R., Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association*, 1996.

LQA Approximation

When $p(|\beta_j|) = \log(\beta_j^2 + \delta^2)$, this is the L_0 adaptive ridge.



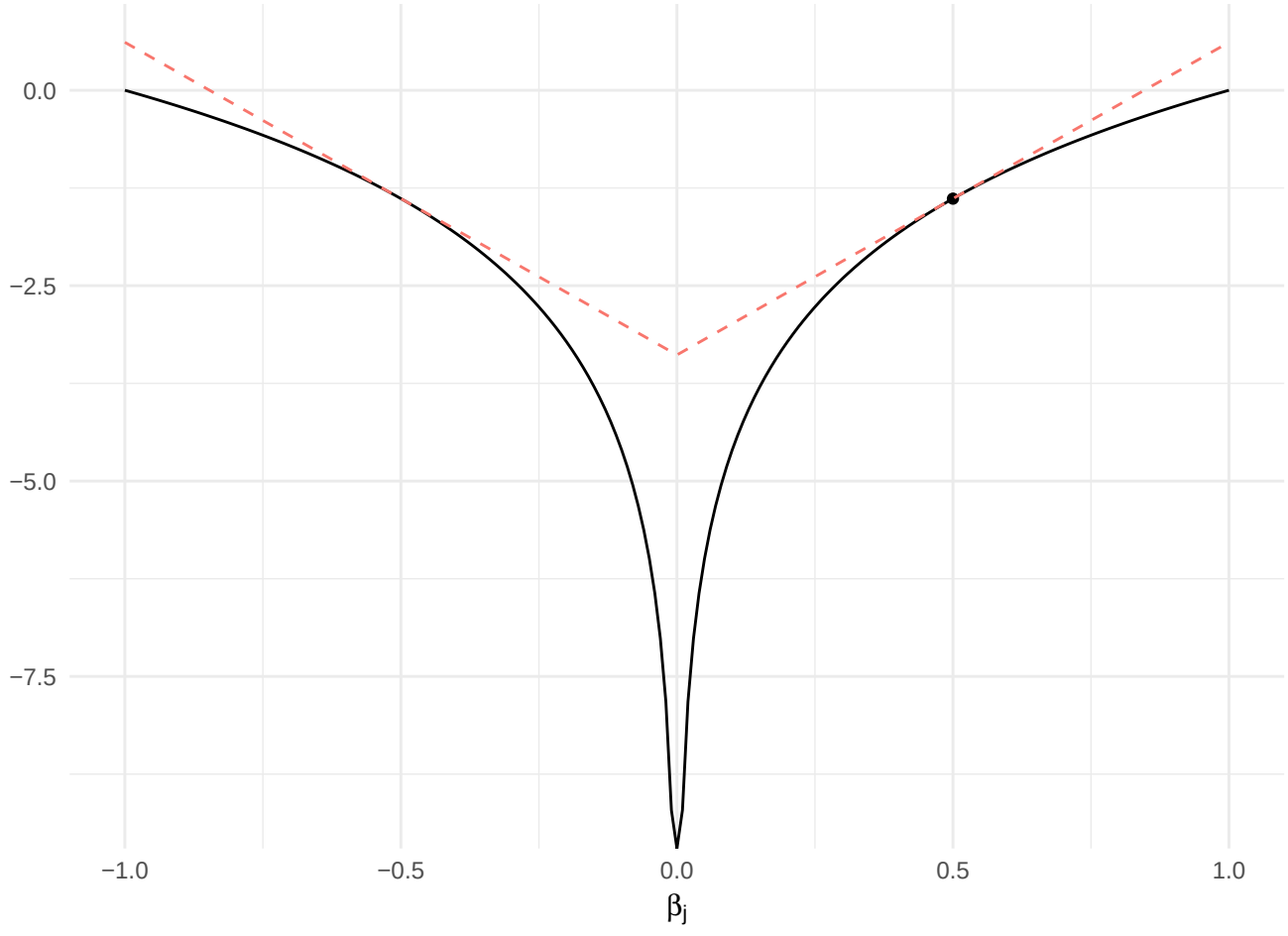
Annex 2 : Connexion with adaptive lasso (1)

The Adaptive Lasso comes from the LLA of a penalty.

$$p(|\beta_j|) + (|\beta_j| - |\beta_j^{(k)}|)p'(|\beta_j^{(k)}|) \succeq p(|\beta_j|)$$

Annex 2 : Connexion with adaptive lasso (2)

Example : with $p(|\beta_j|) = \log(|\beta_j|)$.



- Each iteration is sparse
- Zou and Li (2008) offer to stop at one iteration

Annex 2 : Connexion with adaptive lasso (3)

Compared with the adaptive ridge :

- It takes less iterations to converge.
- Each iteration is slower : gradient methods versus newton methods.

Model selection for *Adaptive Ridge*
Bayesian criteria

- Problem : choose between M models $\mathcal{M}_1, \ldots, \mathcal{M}_M$ of dimensions $q_1, \ldots, q_M$.
- Solution : maximize $\mathbb{P}(\mathcal{M}_m|\boldsymbol{R}, \boldsymbol{O}) \propto \mathbb{P}(\boldsymbol{R}, \boldsymbol{O}|\mathcal{M}_m)\pi(\mathcal{M}_m)$.
- By approximation :

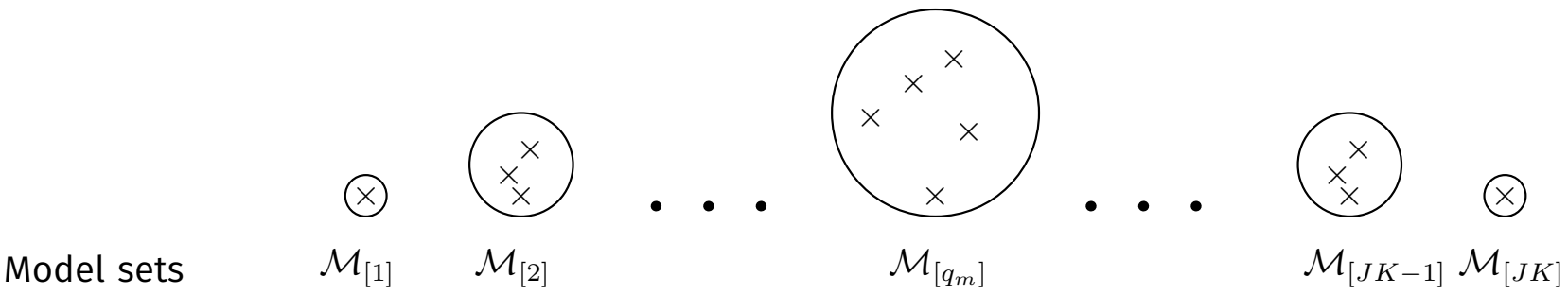
$$-2\log\left(\mathbb{P}(\mathcal{M}_m|\boldsymbol{R}, \boldsymbol{O})\right) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m \log n - 2\log \pi(\mathcal{M}_m) + \mathcal{O}_{\mathbb{P}}(1)$$

- We must choose the prior *a priori* $\pi\left(\mathcal{M}_m\right)$

Model selection for *Adaptive Ridge*

BIC : $\pi\left(\mathcal{M}_m\right) = 1$
All the $\mathcal{M}_m$ are equiprobable

EBIC$_0$: $\mathbb{P}\left(\mathcal{M}_m \in \mathcal{M}_{[q_m]}\right) = 1$
All the $\mathcal{M}_{[q_m]}$ are equiprobable



$\mathcal{M}_{[q_m]}$ is the set of models with q_m parameters

Model selection for *Adaptive Ridge*

We compare different model selection criteria :

1. $\mathrm{BIC}(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m \log n$
2. $\mathrm{EBIC}_0(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m \log n + 2 \log \binom{JK}{q_m}$
3. $\mathrm{AIC}(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + 2q_m$
4. K-fold Cross validation (CV)