New methods based on the adaptive ridge procedure to take into account age, period and cohort effects

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

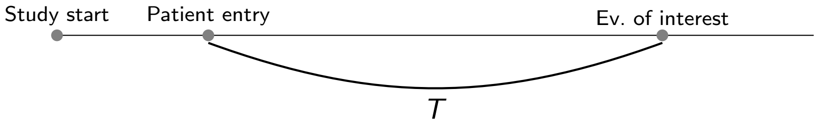
[1]MAP5 (CNRS 8145), Université Paris Descartes, Sorbonne Paris Cité
[3]LPSM (CNRS 8001), Sorbonne Université, Paris

Séminaire du CépiDc

Outline

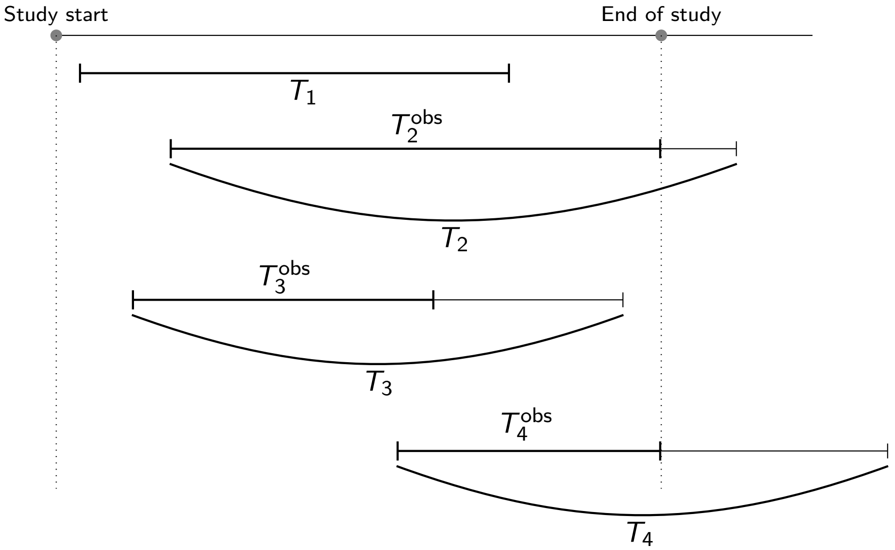
Background in time to event analysis

- 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 observations, the hazard rate and the likelihood

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

- The hazard rate is defined as :

$$\lambda(t) := \lim_{\triangle t \to 0} \frac{\mathbb{P}[t \leq T < t + \triangle t | T \geq t]}{\triangle t}$$

- 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 $S(t) = \mathbb{P}[T > t]$.

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 $(c_{k-1}, c_k]$
- $\bar{R}_k = \sum_i (T_i^{\text{obs}} \wedge c_k - c_{k-1})$: 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 :

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

- 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 adjacent cut 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 $\boldsymbol{w} = \mathbf{1}$
2. L_0 regularization with the adaptive ridge procedure.
   Iterative updates of the weights :

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

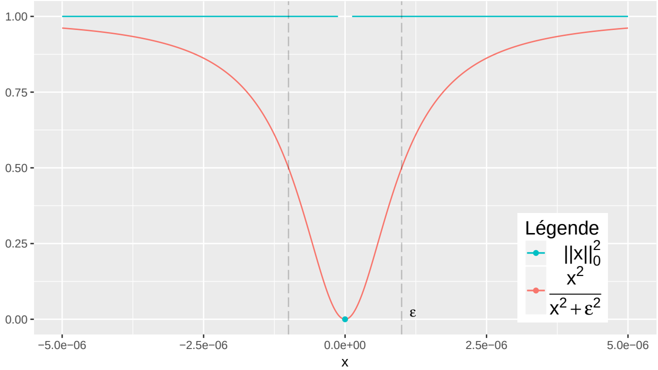
   with $\varepsilon \ll 1$.

F. Frommlet and G. Nuel, *An Adaptive Ridge Procedure for L0 Regularization*. **PlosOne** (2016).

L_0 norm approximation

When $\varepsilon \ll 1$:

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



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 mth Newton Raphson iteration step is obtained from the equation

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

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

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

The *Adaptive Ridge* procedure for a given penalty

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

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

while not converge **do**

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

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

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

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

end while

end procedure

The *Adaptive Ridge* procedure for a given penalty

procedure ADAPTIVE-RIDGE($\boldsymbol{O}, \boldsymbol{R}$, pen)

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

while not converge **do**

$\boldsymbol{a}^{\text{new}} \leftarrow \text{NEWTON-RAPHSON}(\boldsymbol{O}, \boldsymbol{R}, \text{pen}, \boldsymbol{a}, \boldsymbol{w})$

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

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

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

end while

Compute $(\boldsymbol{O}^{\textbf{sel}}, \boldsymbol{R}^{\textbf{sel}})$

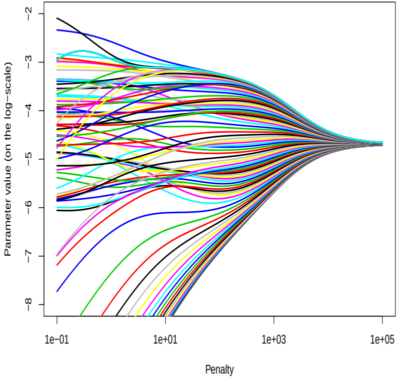
$\exp(\hat{\boldsymbol{a}}^{\textbf{mle}}) \leftarrow \boldsymbol{O}^{\textbf{sel}} / \boldsymbol{R}^{\textbf{sel}}$

return $\hat{\boldsymbol{a}}^{\textbf{mle}}$

end procedure

Comparison of the two regularization methods

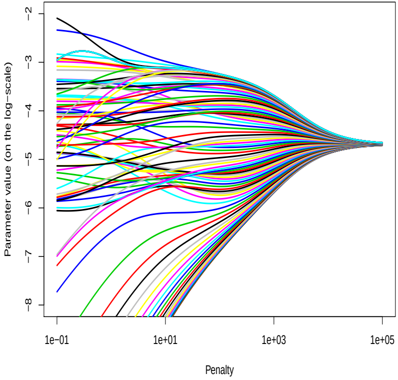
$$\text{pen} = 0 \quad \Longrightarrow \quad \widehat{\boldsymbol{a}} = \widehat{\boldsymbol{a}}^{\text{mle}}$$
$$\text{pen} = \infty \quad \Longrightarrow \quad \widehat{\boldsymbol{a}} = \text{ constant}$$



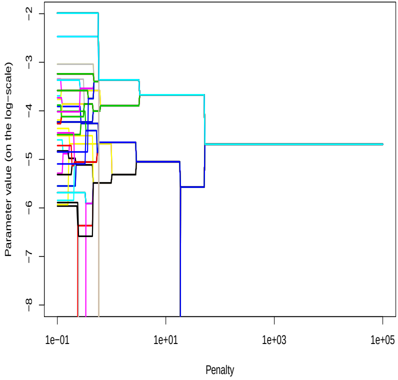
L_2 regularization

Comparison of the two regularization methods

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

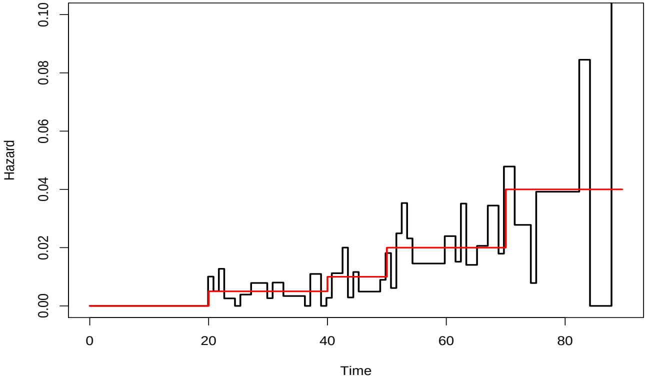


L_2 regularization



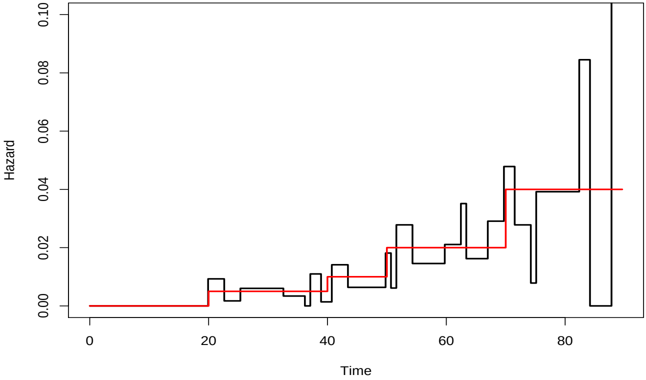
L_0 regularization

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



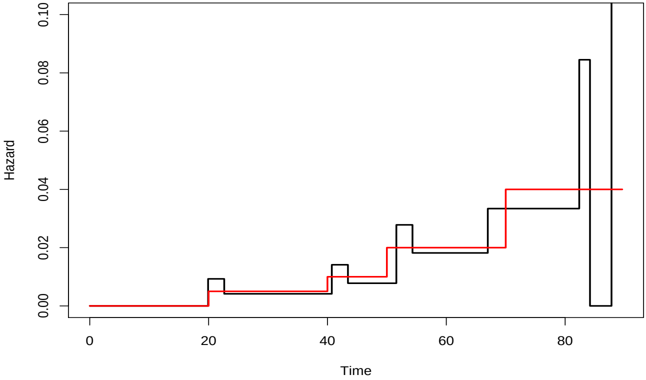
- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 0.1$

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



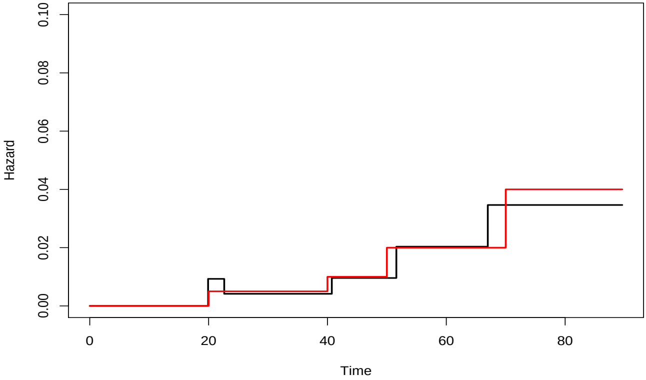
- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 0.27$

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



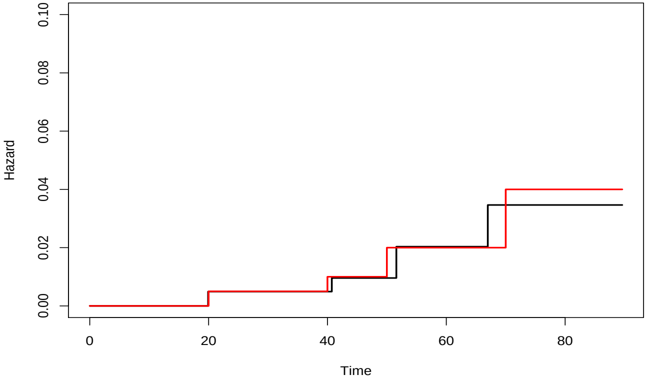
- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 0.55$

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



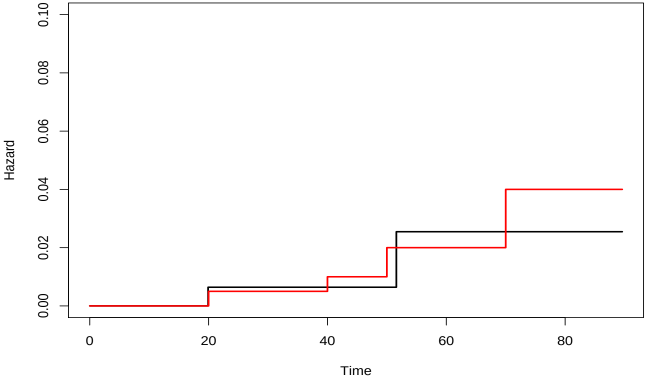
- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 0.77$

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



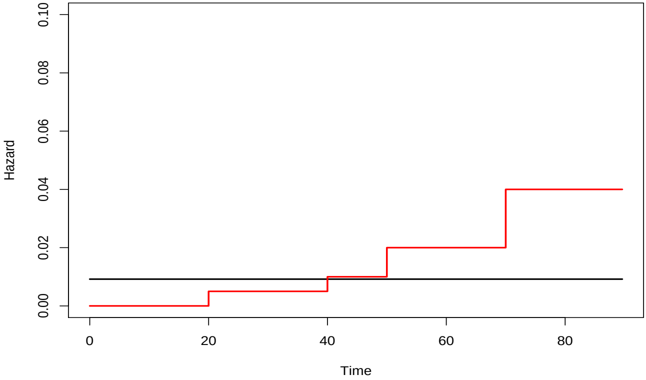
- In red the true hazard function
- In black the hazard estimator for $\mathrm{pen} = 1.54$

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

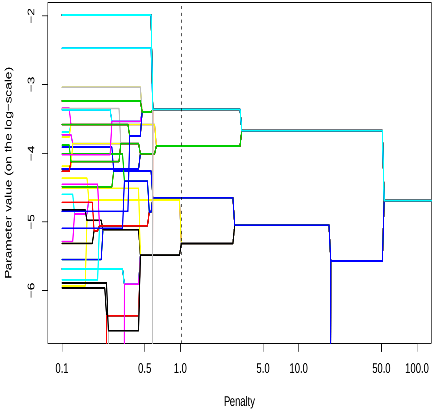
Three different methods to perform model selection :

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

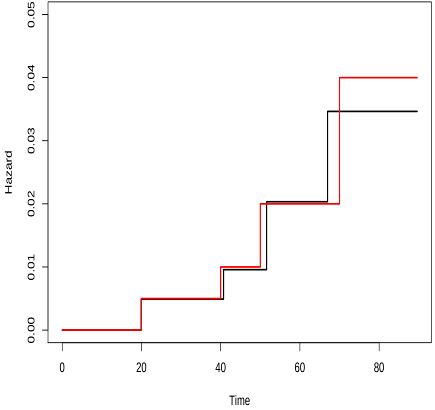
with m the dimension of the model :

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

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



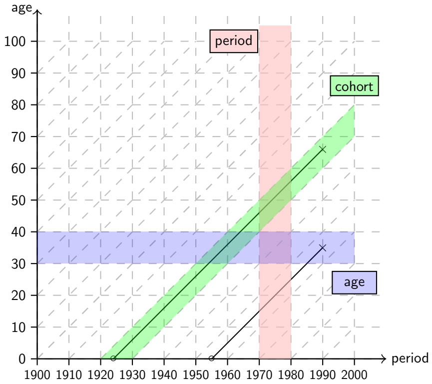
Regularization path



Hazard estimator (in black)

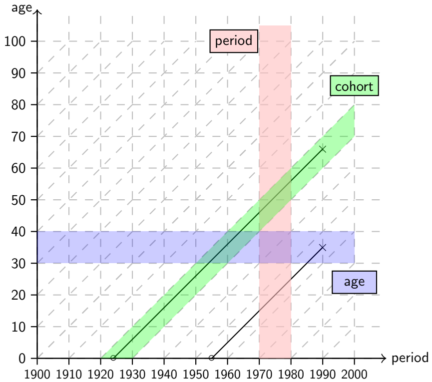
Outline

The Lexis diagram

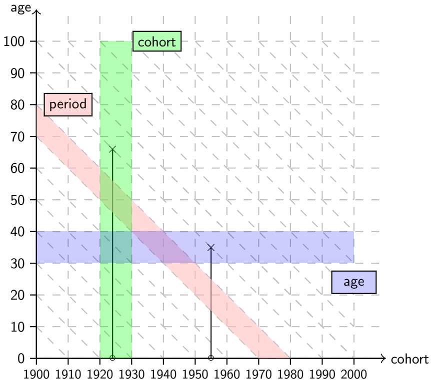


Age-Period Lexis diagram

The Lexis diagram



Age-Period Lexis diagram



Age-Cohort Lexis diagram

Key relation : cohort+age=period

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*. **Submitted**.

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 :

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

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

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

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)

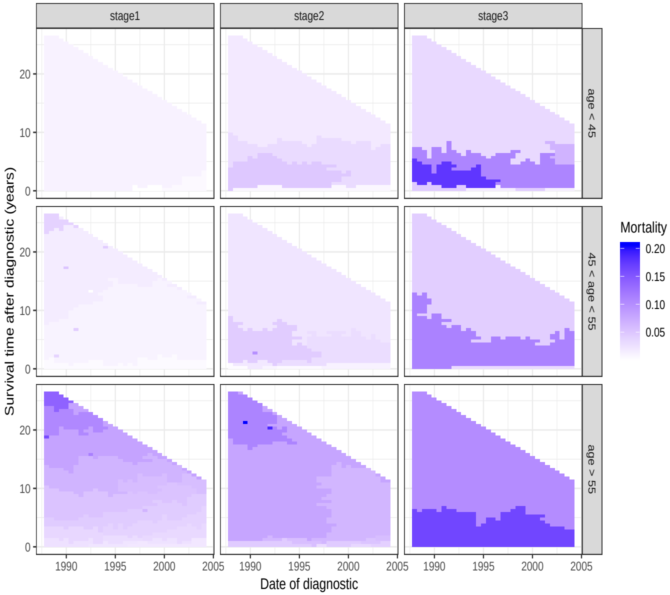
The log-likelihood is equal to :

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

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

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

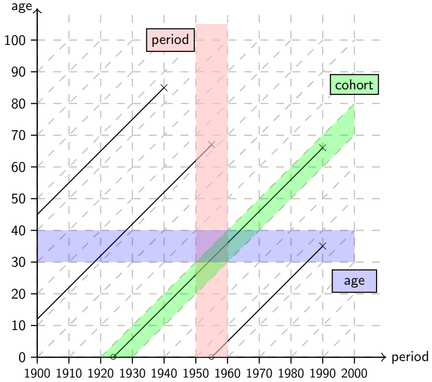
The SEER data



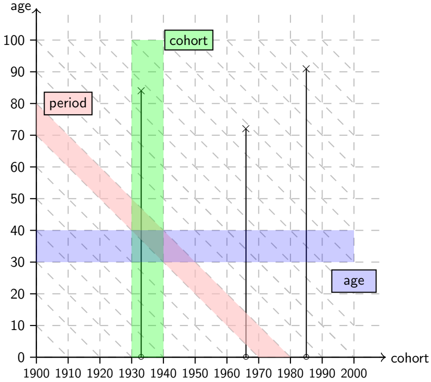
Outline

Recap

Age-period-cohort analysis



Lexis diagram : Age-Period



Lexis Diagram : Age-Cohort

Age-period-cohort analysis

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

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

We define one parameter vector for each effect : $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$

Existing models

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

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

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

2. In the AGE-COHORT model, we assume

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

- $J + K - 1$ parameters for JK variables $\rightarrow$ regularizing
- Additive effect of the variables : strong *a priori*

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

Our approach : model the effects **and** their interactions

- The AGE-COHORT and AGE-PERIOD-COHORT models do not infer interactions between effects.
- We introduce an *age-cohort-interaction* model :

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

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

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

Estimation in the ACI model

- The model parameter is $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$.
- Estimation using the adaptive ridge :

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

- With adaptive ridge procedure over the interaction term $\boldsymbol{\delta}$.

Adaptive ridge procedure

procedure ADAPTIVE-RIDGE($\boldsymbol{O}, \boldsymbol{R}, \text{pen}$)

end procedure

Adaptive ridge procedure

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

$\boldsymbol{\theta} \leftarrow \mathbf{0}$

$\boldsymbol{v} \leftarrow \mathbf{1}$

$\boldsymbol{w} \leftarrow \mathbf{1}$

end procedure

Adaptive ridge procedure

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

$\boldsymbol{\theta} \leftarrow \mathbf{0}$

$\boldsymbol{v} \leftarrow \mathbf{1}$

$\boldsymbol{w} \leftarrow \mathbf{1}$

while not converge **do**

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

$$v_{j,k}^{\text{new}} \leftarrow \left(\left(\delta_{j+1,k}^{\text{new}} - \delta_{j,k}^{\text{new}} \right)^2 + \varepsilon^2 \right)^{-1}$$

$$w_{j,k}^{\text{new}} \leftarrow \left(\left(\delta_{j,k}^{\text{new}} - \delta_{j,k-1}^{\text{new}} \right)^2 + \varepsilon^2 \right)^{-1}$$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{\text{new}}$

$\boldsymbol{v} \leftarrow \boldsymbol{v}^{\text{new}}$

$\boldsymbol{w} \leftarrow \boldsymbol{w}^{\text{new}}$

end while

end procedure

Adaptive ridge procedure

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

$\boldsymbol{\theta} \leftarrow \mathbf{0}$

$\boldsymbol{v} \leftarrow \mathbf{1}$

$\boldsymbol{w} \leftarrow \mathbf{1}$

while not converge **do**

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

$$v_{j,k}^{\text{new}} \leftarrow \left(\left(\delta_{j+1,k}^{\text{new}} - \delta_{j,k}^{\text{new}} \right)^2 + \varepsilon^2 \right)^{-1}$$

$$w_{j,k}^{\text{new}} \leftarrow \left(\left(\delta_{j,k}^{\text{new}} - \delta_{j,k-1}^{\text{new}} \right)^2 + \varepsilon^2 \right)^{-1}$$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{\text{new}}$

$\boldsymbol{v} \leftarrow \boldsymbol{v}^{\text{new}}$

$\boldsymbol{w} \leftarrow \boldsymbol{w}^{\text{new}}$

end while

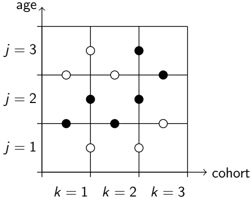
Compute $(\boldsymbol{O}^{\text{sel}}, \boldsymbol{R}^{\text{sel}})$ **from** $(\theta^{\text{new}}, \boldsymbol{v}^{\text{new}}, \boldsymbol{w}^{\text{new}})$

$\theta^{\text{mle}} \leftarrow \boldsymbol{O}^{\text{sel}} / \boldsymbol{R}^{\text{sel}}$

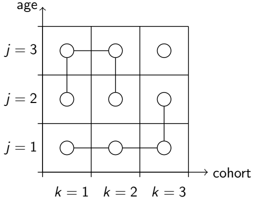
return θ^{mle}

end procedure

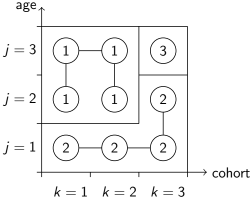
Adaptive ridge procedure



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



(b) Corresponding graph



(c) Segmentation into connected components

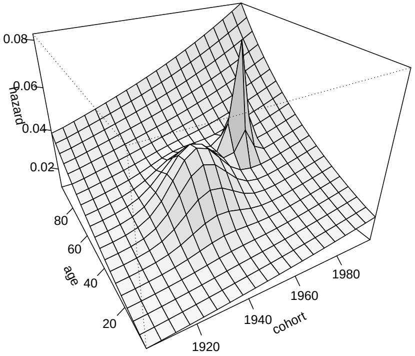
Illustration : simulated data

Simulation setting

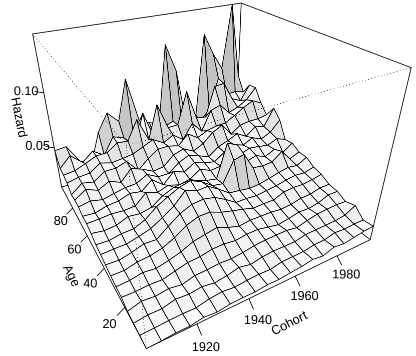
- $J = 20$ age intervals and $K = 20$ cohort intervals
- Sample the cohort uniformly
- Sample the age using the hazard rate $(\lambda_{j,k})$
- Uniform censoring over the age $[75, 100]$
- Infer $(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$ in the ACI model.
- We represent medians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

Illustration : simulated data

Simulation design 1



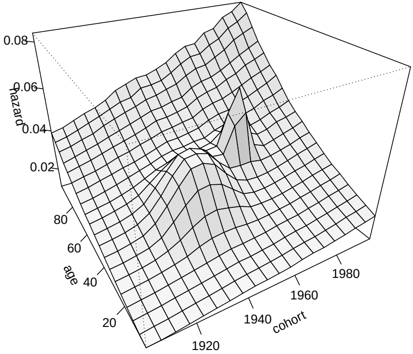
True hazard



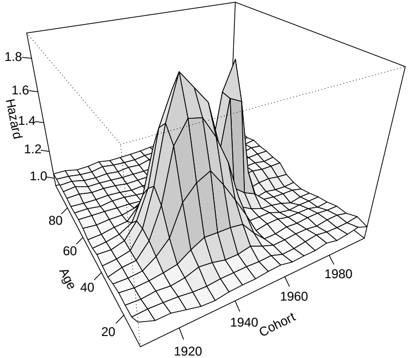
Hazard MLE

Results with the ACI model

Simulation design 1



Estimated hazard $\lambda_{j,k}$



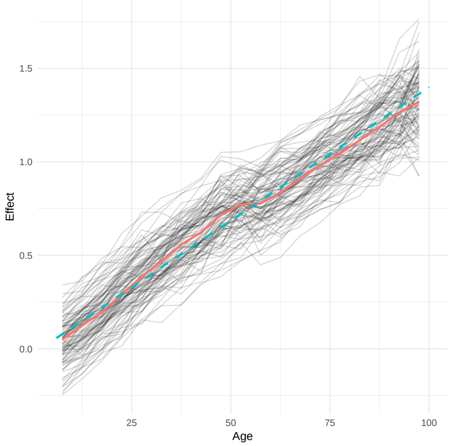
Estimated interaction $\delta_{j,k}$

Results with the ACI model

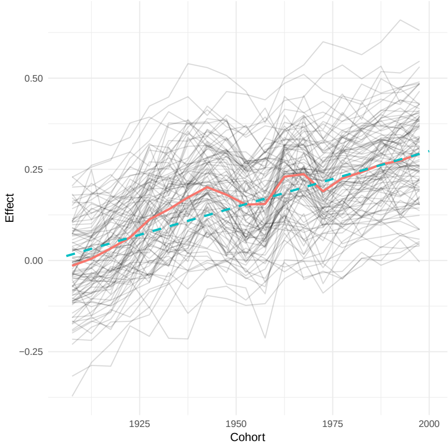
Simulation design 1

Blue : true values

Red : median estimate over 100 repetitions



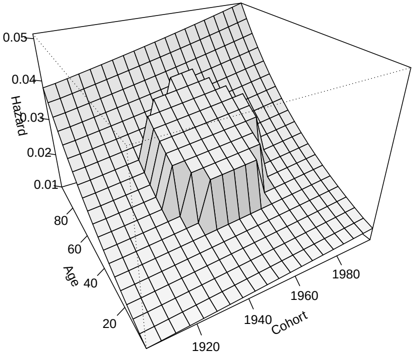
Estimated age effect α_j



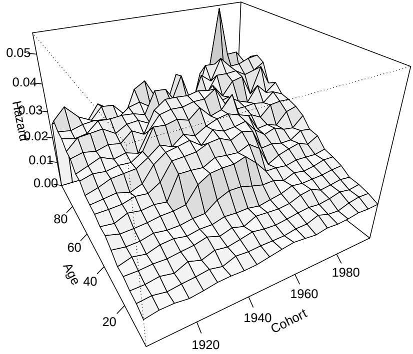
Estimated cohort effect β_k

Illustration : simulated data

Simulation design 2



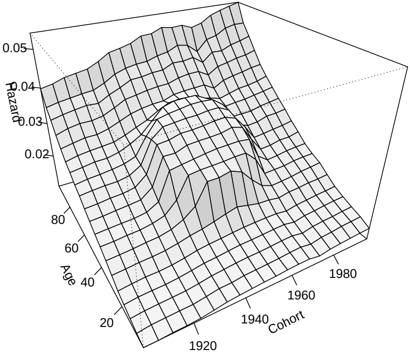
True hazard



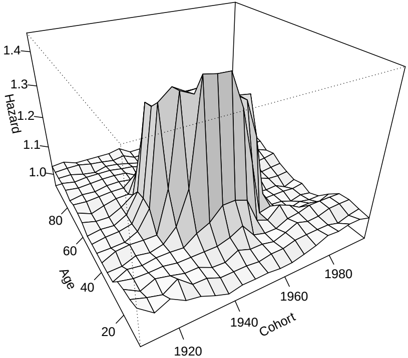
Hazard MLE estimate

Results with the ACI model

Simulation design 2



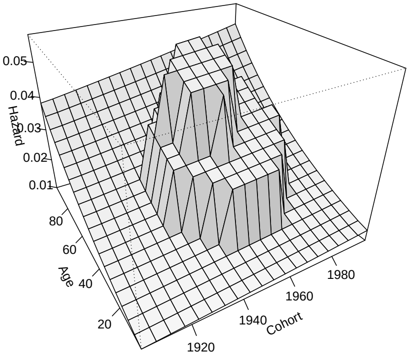
Estimated hazard $\lambda_{j,k}$



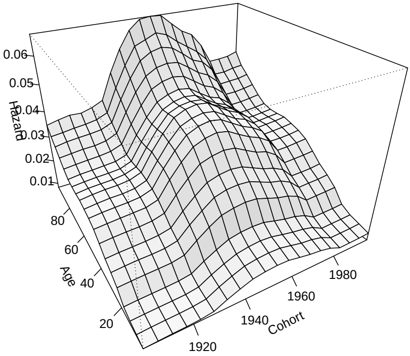
Estimated interaction $\delta_{j,k}$

Illustration : simulated data

Simulation design 3



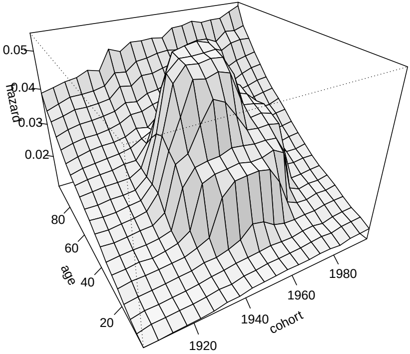
True hazard



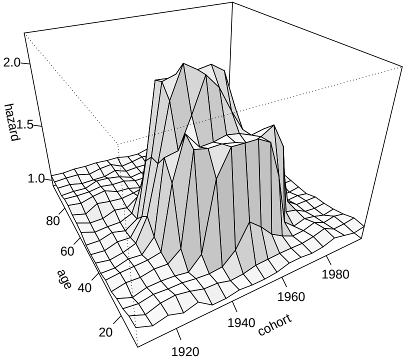
Age-cohort model

Results with the ACI model

Simulation design 3



Estimated hazard $\lambda_{j,k}$



Estimated interaction $\delta_{j,k}$

Conclusion and perspectives

Conclusion :

- Extends the age-cohort model
- More general than the age-period-cohort model

Perspectives :

- Bootstrapping to reduce sensitivity to outliers
- Application : Incidence of breast cancer in Norway (NOWAC cohort)

More info :

- R package : `github.com/goepp/hazreg`
- Website : `www.math-info.univ-paris5.fr/~obouaziz` and `goepp.github.io`

Merci de votre attention

Bonus : 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 :

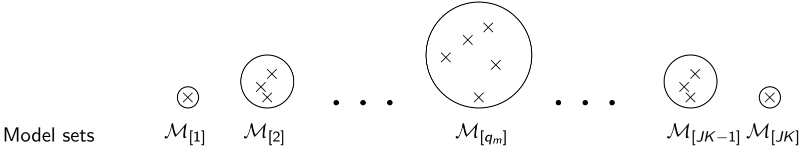
$$-2 \log \left(\mathbb{P}(\mathcal{M}_m | \boldsymbol{R}, \boldsymbol{O}) \right) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m \log n - 2 \log \pi(\mathcal{M}_m) + \mathcal{O}_{\mathbb{P}}(1)$$

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

Model selection for *Adaptive Ridge*

BIC : $\pi(\mathcal{M}_m) = 1$
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

J. Chen and Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, *Biometrika*, 2008.