Estimation régularisée du risque pour l'analyse age-period-cohort

V. Goepp[†], G. Nuel[‡], O. Bouaziz[†]

† : MAP5, Université Paris-Descartes
‡ : LPMA, Université Pierre et Marie Curie

Séminaire de Statistiques, MAP5, 15 décembre 2017

Plan de la présentation

Population: 91992 femmes adhérentes à la MGEN
Données récoltées par formulaire (2-3 ans)
Date calendaire $\in [1990, 2010]$
L'évènement observé est l'apparition du cancer du sein

Objectif : estimer le risque d'avoir un cancer du sein à chaque instant
Difficultés :

- Pourcentage de cancers observés: 7%
- Date de naissance $\in [1925, 1950]$: population hétérogène

[1] F. Clavel-Chapelon et al, Cohort profile: the French E3N cohort study, *International journal of epidemiology*, 2014.

- On veut estimer T, le temps passé avant l'apparition d'un évènement.
- On n'a pas accès à $(T_i)_i$ mais à

$$Y_i = \min(T_i, C_i)$$

où C est une variable de censure avec $C \perp\!\!\!\perp T$.
- On connaît aussi $\Delta_i = \mathbb{1}_{Y_i = T_i}$.
- On estime le risque instantané:

$$\lambda(t) = \lim_{\delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \delta t | T > t)}{\delta t}$$

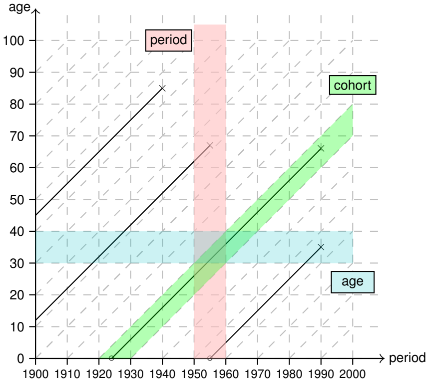


Diagramme de Lexis: Age-Period

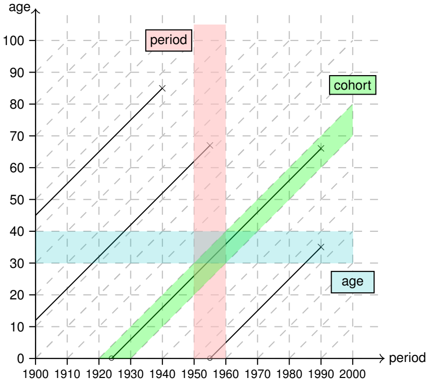


Diagramme de Lexis: Age-Period

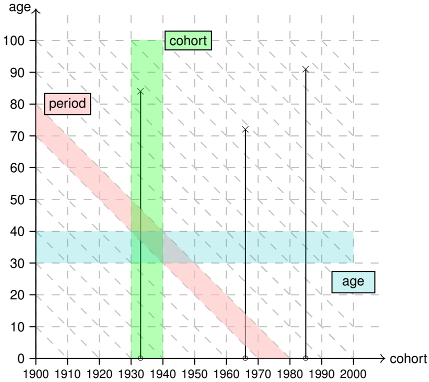


Diagramme Age-Cohort

Estimation paramétrique

Le risque instantané λ est discrétisé en J intervalles d'âge et K intervalles de cohorte :

$$\lambda(\text{age}, \text{cohorte}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \lambda_{j,k} \mathbb{1}_{[c_{j-1}, c_j) \times [d_{k-1}, d_k)}(\text{age}, \text{cohorte})$$

Objectif : estimer $\lambda_{j,k}$

Analyse age-period-cohort

On veut modéliser l'effet de l'âge, la cohorte et la période.

- effet de l'âge : ménopause
- effet de la cohorte : biberon cancérigène
- effet de la période : accident nucléaire

On définit un vecteur de paramètres par effet: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$

Modèles existants

(i) Dans le modèle AGE-COHORT, on suppose

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

- $J + K - 1$ paramètres pour JK variables: régularisation
- Fort *a priori* sur λ

(ii) Dans le modèle AGE-PERIOD-COHORT, on suppose

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

- Non identifiable : on peut soit
  - estimer $\Delta^2\boldsymbol{\alpha}$, $\Delta^2\boldsymbol{\beta}$ et $\Delta^2\boldsymbol{\gamma}$.
  - rajouter une contrainte.

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

Estimateur du maximum de vraisemblance

On appelle :

- $O_{j,k}$: nombre d'évènements dans le (j,k)-ième rectangle
- $R_{j,k}$: temps à risque dans le (j,k)-ième rectangle

La log vraisemblance négative s'écrit

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

L'estimateur du maximum de vraisemblance est :

$$\lambda_{j,k}^{\text{mle}} = \frac{O_{j,k}}{R_{j,k}}$$

Estimateur du maximum de vraisemblance

On appelle :

- $O_{j,k}$: nombre d'évènements dans le (j, k)-ième rectangle
- $R_{j,k}$: temps à risque dans le (j, k)-ième rectangle

La log vraisemblance négative s'écrit

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

L'estimateur du maximum de vraisemblance est :

$$\lambda_{j,k}^{\text{mle}} = \frac{O_{j,k}}{R_{j,k}} \quad \rightarrow \quad \textbf{overfitting}$$

Notre approche : vraisemblance pénalisée

Aucun *a priori* :

$$\log \lambda_{j,k} = \eta_{j,k},$$

Notre approche : vraisemblance pénalisée

Aucun *a priori* :

$$\log \lambda_{j,k} = \eta_{j,k},$$

Mais l'estimation de $\boldsymbol{\eta}$ est faite par vraisemblance **pénalisée**:

$$\ell_n^{\text{pen}}(\boldsymbol{\eta}) = \underbrace{\ell_n(\boldsymbol{\eta})}_{\substack{\text{attache aux}\\ \text{données}}}$$

Notre approche : vraisemblance pénalisée

Aucun *a priori* :

$$\log \lambda_{j,k} = \eta_{j,k},$$

Mais l'estimation de $\boldsymbol{\eta}$ est faite par vraisemblance **pénalisée**:

$$\ell_n^{\text{pen}}(\boldsymbol{\eta}) = \underbrace{\ell_n(\boldsymbol{\eta})}_{\substack{\text{attache aux}\\ \text{données}}} + \underbrace{\frac{\text{pen}}{2} \sum_{j,k} v_{j,k} \left(\eta_{j+1,k} - \eta_{j,k}\right)^2 + w_{j,k} \left(\eta_{j,k+1} - \eta_{j,k}\right)^2}_{\text{régularisation}},$$

Notre approche : vraisemblance pénalisée

Aucun *a priori* :

$$\log \lambda_{j,k} = \eta_{j,k},$$

Mais l'estimation de $\boldsymbol{\eta}$ est faite par vraisemblance **pénalisée**:

$$\ell_n^{\text{pen}}(\boldsymbol{\eta}) = \underbrace{\ell_n(\boldsymbol{\eta})}_{\substack{\text{attache aux}\\ \text{données}}} + \underbrace{\frac{\text{pen}}{2} \sum_{j,k} v_{j,k} \left(\eta_{j+1,k} - \eta_{j,k}\right)^2 + w_{j,k} \left(\eta_{j,k+1} - \eta_{j,k}\right)^2}_{\text{régularisation}},$$

$\boldsymbol{v}$ et $\boldsymbol{w}$ sont des poids,

pen est une constante de régularisation.

Deux types de régularisation

(i) Régularisation L_2 (Ridge) avec $\boldsymbol{v} = \boldsymbol{w} = \mathbf{1}$

Deux types de régularisation

(i) Régularisation L_2 (Ridge) avec $\boldsymbol{v} = \boldsymbol{w} = \mathbf{1}$

(ii) Régularisation L_0 avec la procédure itérative **adaptive ridge**.
Les poids sont adaptés itérativement :

$$\begin{cases} v_{j,k} = \left(\left(\eta_{j+1,k} - \eta_{j,k}\right)^2 + \varepsilon^2\right)^{-1} \\ w_{j,k} = \left(\left(\eta_{j,k} - \eta_{j,k-1}\right)^2 + \varepsilon^2\right)^{-1} \end{cases},$$

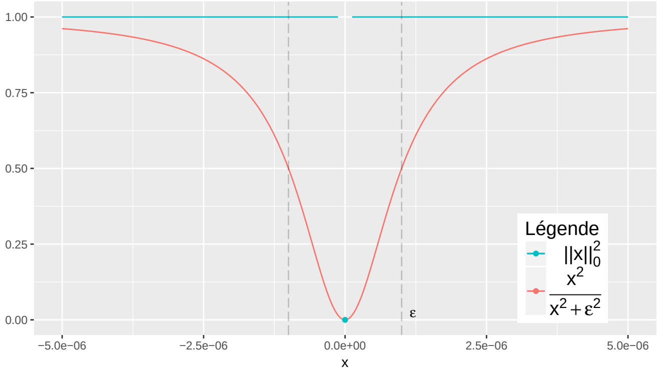
avec $\varepsilon \ll 1$.

[3] F. Frommlet and G. Nuel, An Adaptive Ridge Procedure for L0 Regularization, *Public Library of Science*, 2016.

Approximation de la norme L_0

Lorsque $\varepsilon \ll 1$:

$$v_{j,k}\left(\eta_{j+1,k}-\eta_{j,k}\right)^2 \simeq \|\eta_{j+1,k}-\eta_{j,k}\|_0^2 = \begin{cases} 0 & \text{si } \eta_{j+1,k}=\eta_{j,k} \\ 1 & \text{si } \eta_{j+1,k}\neq\eta_{j,k} \end{cases}$$



La procédure *Adaptive Ridge*

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

end procedure

La procédure *Adaptive Ridge*

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

$\eta \leftarrow \mathbf{0}$

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

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

end procedure

La procédure *Adaptive Ridge*

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

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

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

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

while not converge **do**

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

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

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

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

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

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

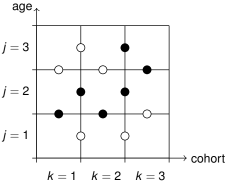
end while

end procedure

La procédure *Adaptive Ridge*

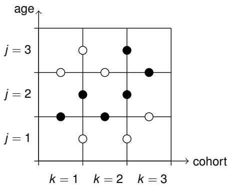
procedure $\textsc{Adaptive-Ridge}(\boldsymbol{O}, \boldsymbol{R}, \text{pen})$
$\boldsymbol{\eta} \leftarrow \mathbf{0}$
$\boldsymbol{v} \leftarrow \mathbf{1}$
$\boldsymbol{w} \leftarrow \mathbf{1}$
while not converge **do**
$\boldsymbol{\eta}^{\text{new}} \leftarrow \textsc{Newton-Raphson}(\boldsymbol{O}, \boldsymbol{R}, \text{pen}, \boldsymbol{v}, \boldsymbol{w})$
$v_{j,k}^{\text{new}} \leftarrow \left(\left(\eta_{j+1,k}^{\text{new}} - \eta_{j,k}^{\text{new}}\right)^2 + \varepsilon^2\right)^{-1}$
$w_{j,k}^{\text{new}} \leftarrow \left(\left(\eta_{j,k}^{\text{new}} - \eta_{j,k-1}^{\text{new}}\right)^2 + \varepsilon^2\right)^{-1}$
$\boldsymbol{\eta} \leftarrow \boldsymbol{\eta}^{\text{new}}$
$\boldsymbol{v} \leftarrow \boldsymbol{v}^{\text{new}}$
$\boldsymbol{w} \leftarrow \boldsymbol{w}^{\text{new}}$
end while
Compute $(\boldsymbol{O}^{\text{sel}}, \boldsymbol{R}^{\text{sel}})$ **from** $(\boldsymbol{\eta}^{\text{new}}, \boldsymbol{v}^{\text{new}}, \boldsymbol{w}^{\text{new}})$
$\boldsymbol{\eta}^{\text{mle}} \leftarrow \boldsymbol{O}^{\text{sel}} / \boldsymbol{R}^{\text{sel}}$
return $\boldsymbol{\eta}^{\text{mle}}$
end procedure

Adaptive Ridge permet de sélectioner un modèle

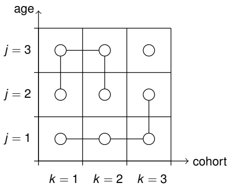


(a) Représentation de $v_{j,k}\left(\eta_{j+1,k} - \eta_{j,k}\right)^2$ et $w_{j,k}\left(\eta_{j,k+1} - \eta_{j,k}\right)^2$

Adaptive Ridge permet de sélectioner un modèle

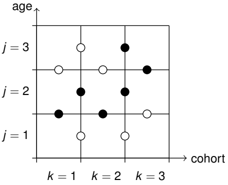


(a) Représentation de $v_{j,k}\left(\eta_{j+1,k} - \eta_{j,k}\right)^2$ et $w_{j,k}\left(\eta_{j,k+1} - \eta_{j,k}\right)^2$

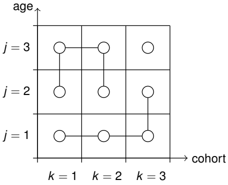


(b) Graphe correspondant

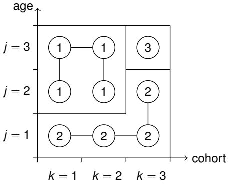
Adaptive Ridge permet de sélectioner un modèle



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



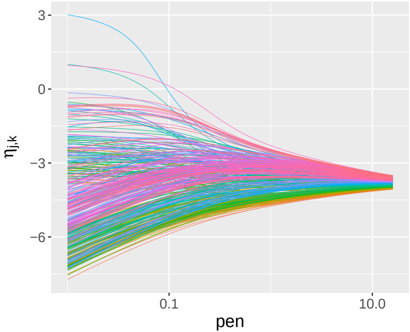
(b) Graphe correspondant



(c) Segmentation selon les composantes connexes

Comparaison des deux régularisations

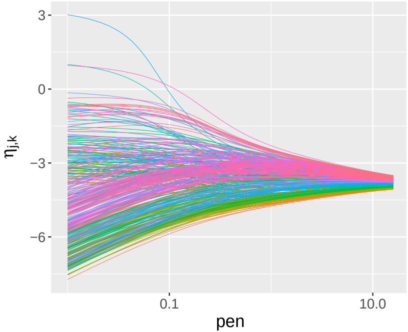
pen $\to 0$: $\widehat{\lambda} \to \widehat{\lambda}^{\text{mle}}$
pen $\to \infty$: $\widehat{\lambda}$ constant



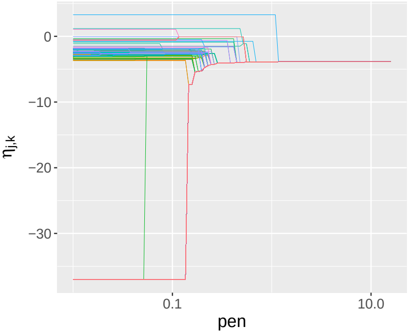
Régularisation L_2 : à chaque pen correspond un estimateur

Comparaison des deux régularisations

$$\text{pen} \to 0 \quad : \quad \widehat{\lambda} \to \widehat{\lambda}^{\text{mle}}$$
$$\text{pen} \to \infty \quad : \quad \widehat{\lambda} \text{ constant}$$



Régularisation L_2 : à chaque pen correspond un estimateur



Régularisation L_0 : à chaque pen correspond un *modèle*

Choix du modèle pour *Adaptive Ridge*

Critères bayesiens

- Problème: choisir entre M modèles $\mathcal{M}_1, \ldots, \mathcal{M}_M$ de dimensions $q_1, \ldots, q_M$.
- Solution: maximiser $\mathbb{P}(\mathcal{M}_m | \boldsymbol{R}, \boldsymbol{O}) \propto \mathbb{P}(\boldsymbol{R}, \boldsymbol{O} | \mathcal{M}_m) \pi(\mathcal{M}_m)$.
- Par approximation :

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

- Comment choisir la distribution *a priori* $\pi\left(\mathcal{M}_m\right)$?

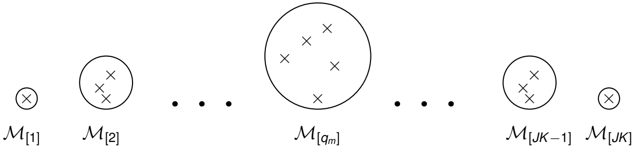
Choix du modèle pour *Adaptive Ridge*

BIC: $\pi\left(\mathcal{M}_m\right) = 1$
Tous les $\mathcal{M}_m$ sont équiprobables

Choix du modèle pour *Adaptive Ridge*

BIC: $\pi\left(\mathcal{M}_m\right) = 1$
Tous les $\mathcal{M}_m$ sont équiprobables

Ensembles de modèles



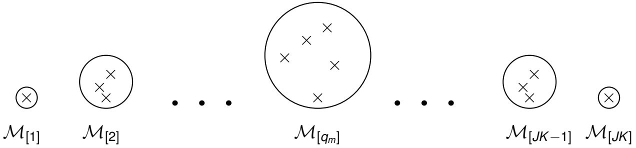
$\mathcal{M}_{[q_m]}$ est l'ensemble des modèles de dimension q_m

Choix du modèle pour *Adaptive Ridge*

BIC: $\pi\left(\mathcal{M}_m\right) = 1$
Tous les $\mathcal{M}_m$ sont équiprobables

EBIC$_0$: $\mathbb{P}\left(\mathcal{M}_m \in \mathcal{M}_{[q_m]}\right) = 1$
Tous les $\mathcal{M}_{[q_m]}$ sont équiprobables

Ensembles de modèles



$\mathcal{M}_{[q_m]}$ est l'ensemble des modèles de dimension q_m

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

Choix du modèle pour *Adaptive Ridge*

Critères utilisés

On compare différents critères de sélection :

(i) $\mathrm{BIC}(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m \log n$

(ii) $\mathrm{EBIC}_0(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m \log n - 2\log\binom{JK}{q_m}$

(iii) $\mathrm{AIC}(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + 2q_m$

(iv) K-fold Cross validation (CV)

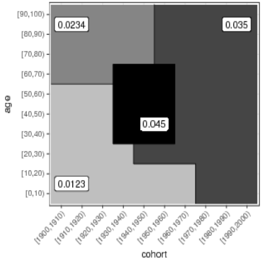
Illustration sur données simulées

On simule des données selon :

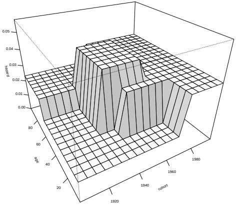
- λ constant par morceaux
- λ lisse

On compare :

- Modèle AGE-COHORT : $\log \lambda_{j,k} = \alpha_j + \beta_k$
- Régularisation L_2 avec CV
- Régularisation L_0 avec AIC, BIC, $EBIC_0$ et CV.

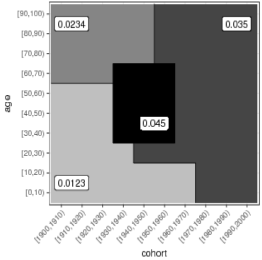


Vrai λ

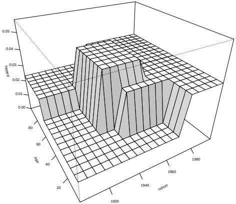


Vrai λ

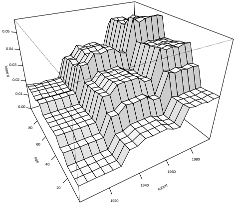
Simulations : cas n°1



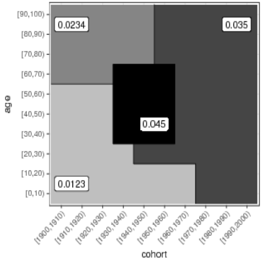
Vrai λ



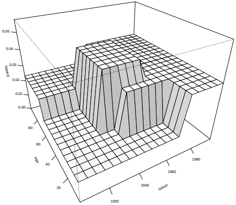
Vrai λ



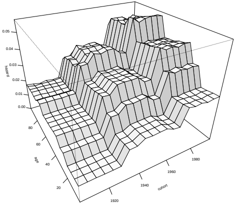
Modèle AGE-COHORT



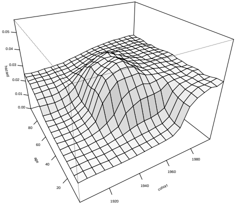
Vrai λ



Vrai λ

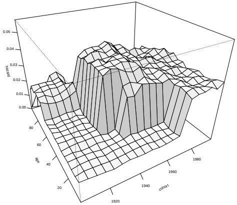


Modèle AGE-COHORT



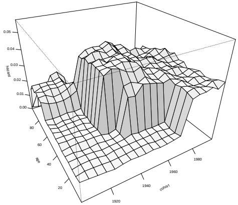
Régularisation L_2 avec CV

Simulations : cas n°1

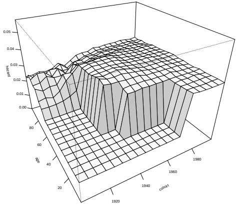


Régularisation L_0 avec AIC

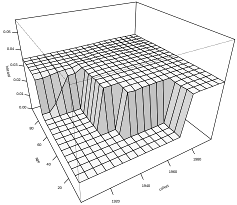
Simulations : cas n°1



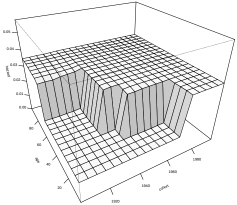
Régularisation L_0 avec AIC



Régularisation L_0 avec BIC

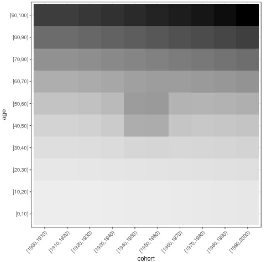


Régularisation L_0 avec $EBIC_0$

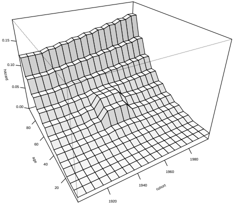


Régularisation L_0 avec CV

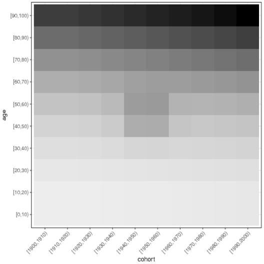
Simulations : cas n°2



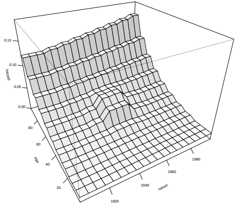
Vrai λ



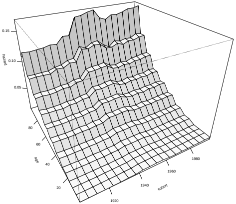
Vrai λ



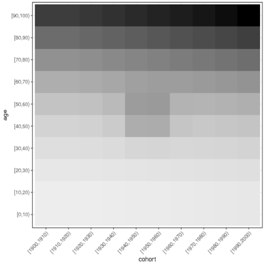
Vrai λ



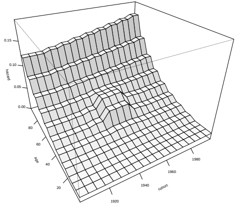
Vrai λ



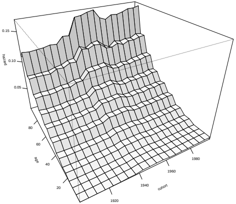
Modèle AGE-COHORT



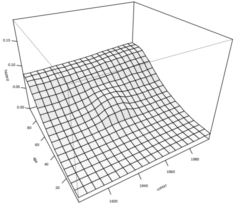
Vrai λ



Vrai λ

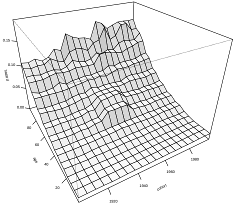


Modèle AGE-COHORT



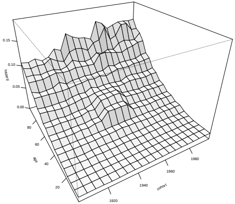
Régularisation L_2 avec CV

Simulations : cas n°2

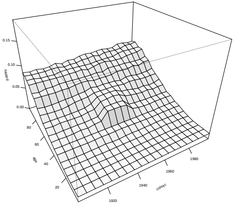


Régularisation L_0 avec AIC

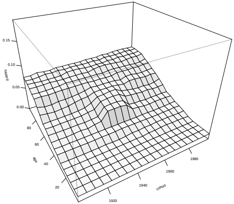
Simulations : cas n°2



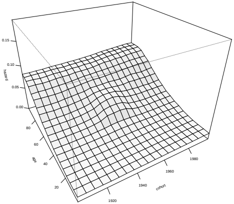
Régularisation L_0 avec AIC



Régularisation L_0 avec BIC



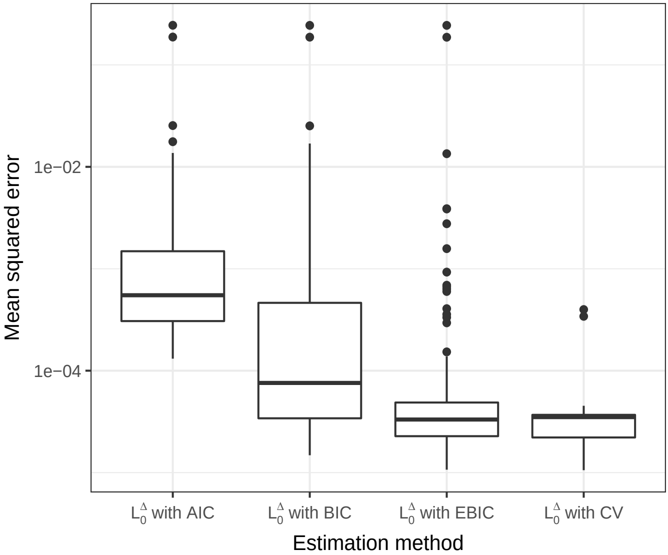
Régularisation L_0 avec $EBIC_0$



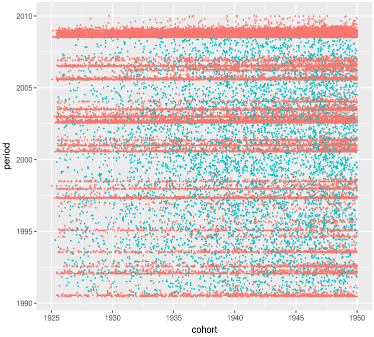
Régularisation L_0 avec CV

Simulations : comparaison quantitative

Erreur quadratique moyenne, pour λ constant par morceaux.

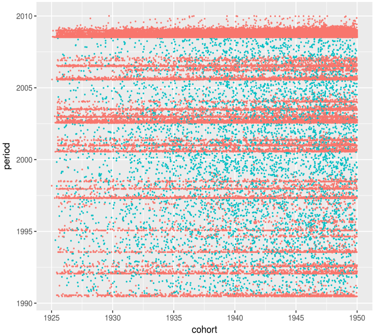


Cohorte $\in [1925, 1950]$
Période $\in [1990, 2010]$

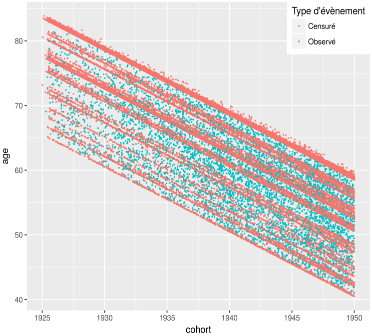


Plan période-cohorte

Cohorte $\in [1925, 1950]$
Période $\in [1990, 2010]$



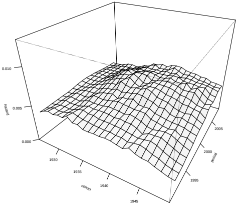
Plan période-cohorte



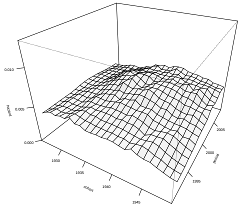
Plan âge-cohorte

Application : données de l'étude E3N

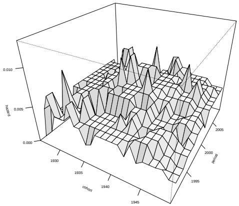
Résultats



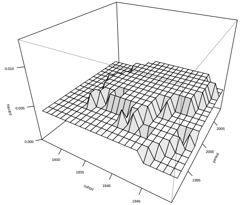
Régularisation L_2 avec CV



Régularisation L_2 avec CV

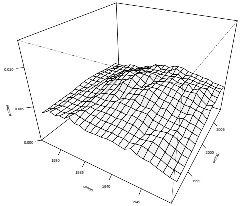


Régularisation L_0 avec AIC

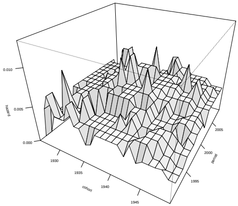


Régularisation L_0 avec $EBIC_0$

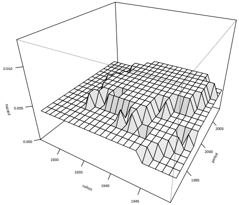
Application : données de l'étude E3N
Résultats



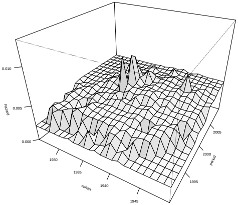
Régularisation L_2 avec CV



Régularisation L_0 avec AIC



Régularisation L_0 avec $EBIC_0$



Régularisation L_0 avec $EBIC_0$: bootstrapé

Conclusion et perspectives

- Segmentation du risque instantané
- EBIC_0 plus performant que les autres critères
- Amélioration possible : différences d'ordre supérieur
- Le modèle peut s'étendre :

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

avec régularisation de $\delta_{j,k}$

Merci