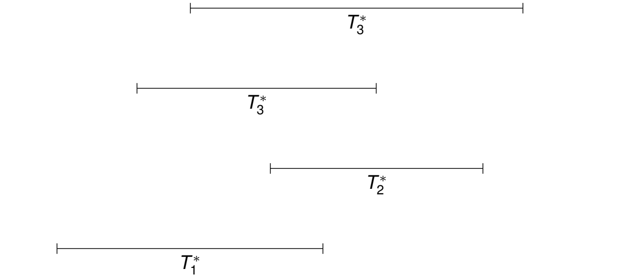
Estimating interaction effects in the age-cohort model

V. Goepp†, G. Nuel‡, O. Bouaziz†

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

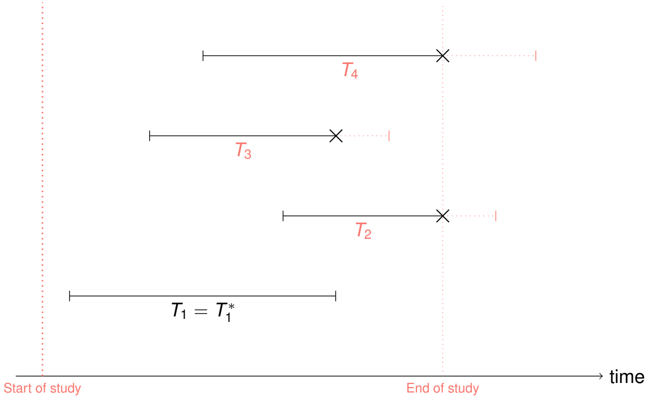
GDR “Statistiques et Santé”, 27-28 September 2018

Introduction: right-censored data



- We want to infer T^*, time before an event of interest.

Introduction: right-censored data



- We want to infer T^*, time before an event of interest.
- But we don't observe the T_i^*s, but

$$T_i = \min\left(T_i^*, C_i\right).$$

Introduction: survival analysis

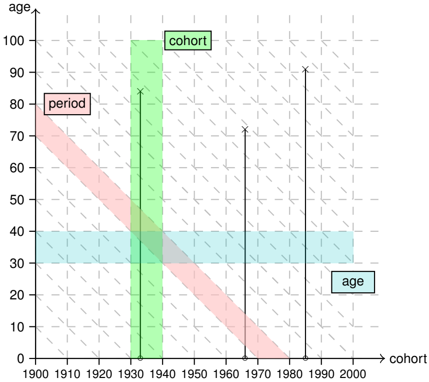
Framework of survival analysis:

- C is the *censoring* variable.
- We also observe $\Delta_i = \mathbb{1}_{T_i^* = T_i}$.
- We infer the conditional density called *hazard rate*:

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

- If $C \perp\!\!\!\perp T^*$, the likelihood writes:

$$L_n = \sum_{i=1}^{n} \left(\Delta_i \log\left(\lambda\left(T_i\right)\right) - \int_0^{T_i} \lambda\left(t\right) dt \right).$$



Age-Cohort Diagram

Additional variable u.

$\rightarrow$ Bi-dimensional hazard $\lambda(t|u)$

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

Parametric estimation

- The hazard λ is discretized into J age intervals and K cohort intervals:

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

Goal: infer $\lambda_{j,k}$

Parametric estimation

- The hazard λ is discretized into J age intervals and K cohort intervals:

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

Goal: infer $\lambda_{j,k}$

- Log-likelihood: $\ell_n = \sum O_{j,k} \log\left(\lambda_{j,k}\right) - R_{j,k}\lambda_{j,k}$, with exhaustive statistics

$$\begin{cases} O_{j,k} = \text{number of events} \\ R_{j,k} = \text{time at risk.} \end{cases}$$

Explicit MLE: $\lambda_{j,k} = \log \frac{O_{j,k}}{R_{j,k}} \quad \rightarrow$ **overfitting**.

Existing models

(i) 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.

(ii) In the AGE-COHORT model, we assume

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

- $J + K - 1$ parameters for JK variables $\rightarrow$ regularizing
- Strong *a priori* over λ

[1] 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:

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

  where $\delta_{j,k}$ is the interaction (with $\delta_{1,k} = \delta_{j,1} = 0$).
- We regularize over the differences of $\delta_{j,k}$.

Inference using penalized likelihood

Let $\boldsymbol{\theta} = (\alpha, \beta, \delta)$ be the parameter. We infer using the penalized negative log-likelihood:

$$\ell_n^{\text{pen}}(\boldsymbol{\theta}) = \underbrace{\ell_n(\boldsymbol{\theta})}_{\substack{\text{goodness}\\ \text{of fit}}} + \underbrace{\frac{\text{pen}}{2} \sum_{j,k} v_{j,k} \left(\delta_{j+1,k} - \delta_{j,k}\right)^2 + w_{j,k} \left(\delta_{j,k+1} - \delta_{j,k}\right)^2}_{\text{regularization}},$$

Inference using penalized likelihood

Let $\boldsymbol{\theta} = (\alpha, \beta, \delta)$ be the parameter. We infer using the penalized negative log-likelihood:

$$\ell_n^{\text{pen}}(\boldsymbol{\theta}) = \underbrace{\ell_n(\boldsymbol{\theta})}_{\substack{\text{goodness} \\ \text{of fit}}} + \underbrace{\frac{\text{pen}}{2} \sum_{j,k} v_{j,k} \left(\delta_{j+1,k} - \delta_{j,k}\right)^2 + w_{j,k} \left(\delta_{j,k+1} - \delta_{j,k}\right)^2}_{\text{regularization}},$$

where the weights are iteratively adapted:

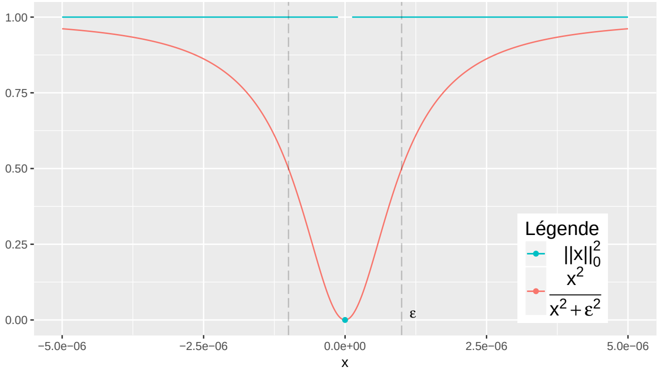
$$\begin{cases} v_{j,k} = \left(\left(\delta_{j+1,k} - \delta_{j,k}\right)^2 + \varepsilon^2\right)^{-1} \\ w_{j,k} = \left(\left(\delta_{j,k} - \delta_{j,k-1}\right)^2 + \varepsilon^2\right)^{-1} \end{cases}, \qquad \text{with } \varepsilon \ll 1.$$

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

Approximating the L_0 norm

We use a L_0 norm regularization using the iterative procedure *adaptive ridge*:

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



The *Adaptive Ridge* algorithm

procedure ADAPTIVE-RIDGE($\boldsymbol{O}, \boldsymbol{R}$, pen)
  while not converge **do**
    $\theta^{\text{new}} \leftarrow$ NEWTON-RAPHSON($\boldsymbol{O}, \boldsymbol{R}$, 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

} Model selection

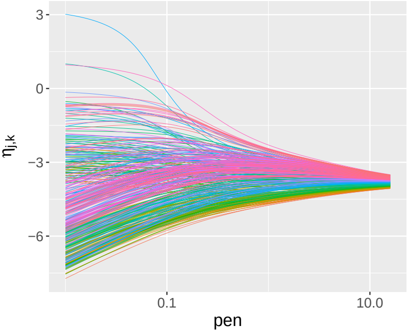
The *Adaptive Ridge* algorithm

procedure ADAPTIVE-RIDGE($\boldsymbol{O}, \boldsymbol{R}$, pen)
    while not converge **do**
        $\theta^{\text{new}} \leftarrow$ NEWTON-RAPHSON($\boldsymbol{O}, \boldsymbol{R}$, 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** ($\delta^{\text{new}}, \boldsymbol{v}^{\text{new}}, \boldsymbol{w}^{\text{new}}$)
    $\theta^{\text{mle}} \leftarrow$ **NEWTON-RAPHSON**($\boldsymbol{O}^{\text{sel}}, \boldsymbol{R}^{\text{sel}}$, **pen** $= 0$)
    return θ^{mle}
end procedure

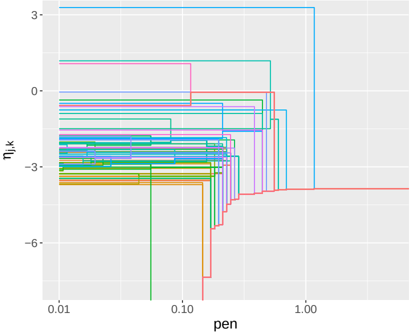
Model selection

Unpenalized estimation

Comparision: smoothed *vs* segmented estimate

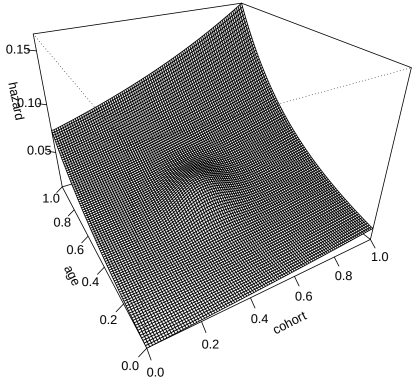


L_2 regularization:
Each penalty yields an estimate

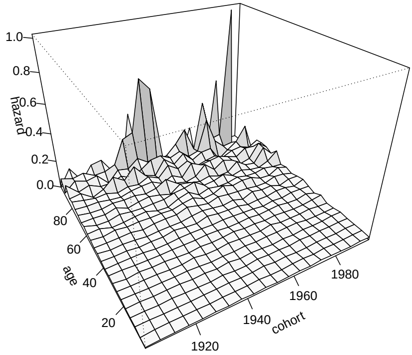


L_0 regularization:
Each penalty yields a *model*

Illustration: simulated data

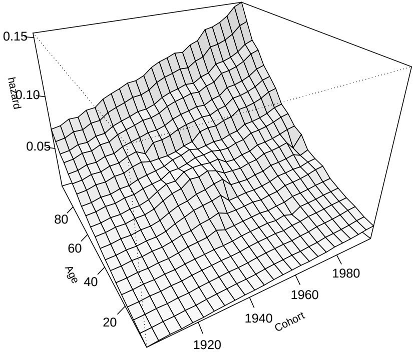


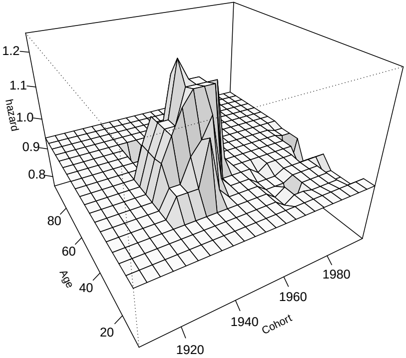
True hazard



MLE

Results





Estimated hazard $\lambda_{j,k}$

Estimated interaction $\delta_{j,k}$

Thank you for your attention

Conclusion:

- Extends the age-cohort model
- More general than the age-period-cohort model
- Interaction term is important in epidemiology

Perspectives:

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

More info:

- Code: `github.com/goepp/hazreg`
- Website: `www.math-info.univ-paris5.fr/~vgoepp/`

Bonus: Model selection

Model selection for *Adaptive Ridge*

Bayesian criteria

- Problem: choose between M models $\mathcal{M}_1, \dots, \mathcal{M}_M$ of dimensions $q_1, \dots, q_M$.
- 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:

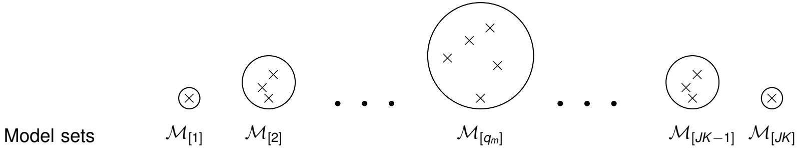
$$-2\log\left(\mathbb{P}(\mathcal{M}_m|\boldsymbol{R}, \boldsymbol{O})\right) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m \log n - 2\log \pi(\mathcal{M}_m) + \mathcal{O}_{\mathbb{P}}(1)$$

- 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

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

Model selection for *Adaptive Ridge*

We compare different model selection criteria:

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

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

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

(iv) K-fold Cross validation (CV)