Regularized Hazard Estimation for Age-Period-Cohort Analysis

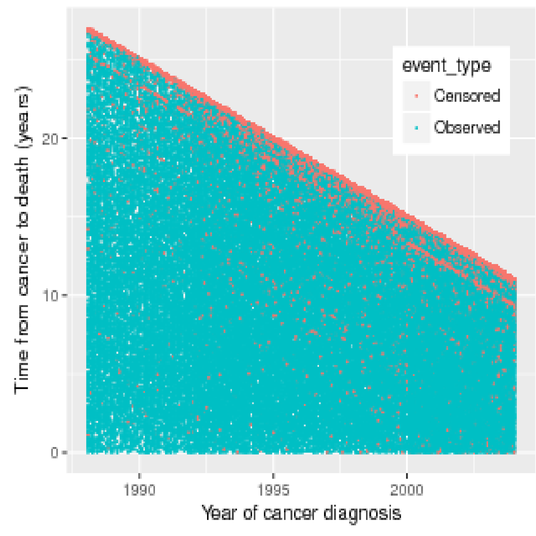
Vivien Goepp[1,*], Grégory Nuel[2], and Olivier Bouaziz[2]

XXIXth International Biometric Conference, 8-13 July 2018, Barcelona

Presentation of the Problem

SEER breast cancer mortality data:

- US cancer survey
- Period: 1986 –.
- Sample size ≃ 400,000
- **Cohort** = Time of diagnosis
- **Age** = Time after diagnosis
- Cancer stage is registered



Death after diagnosis of stage 1 breast cancer

Question: has the mortality of breast cancer evolved with time?

Right-Censoring

The death from cancer is observed for only a fraction of individuals.

- T_i is the age of cancer onset.
- We do not observe $(T_i)_i$ but

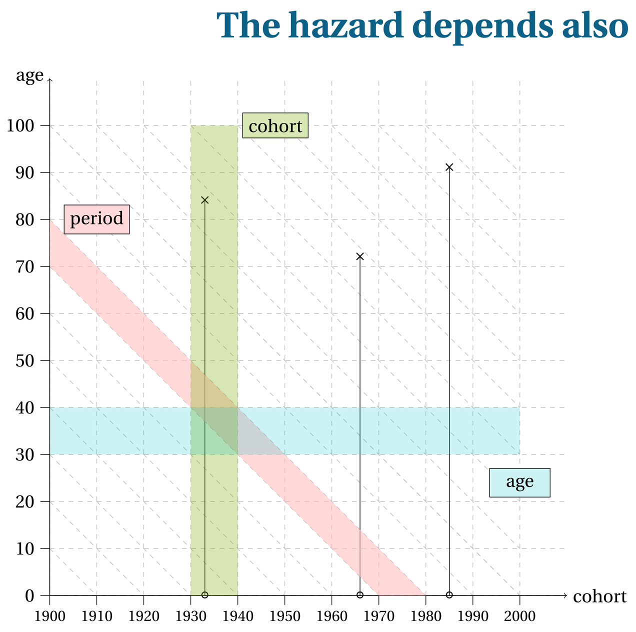
$$Y_i = \min(T_i, C_i)$$

where C is a censoring r.v. independent from T.
- We observe $\Delta_i = 1_{T_i = Y_i}$.
- We infer the instantaneous hazard rate

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \le T \le t + dt | T \ge t)}{dt}$$

Age-Period-Cohort analysis

The hazard depends also on the date of birth



period = calendar time
cohort = date of birth

Age-Cohort Diagram

New Approach: Penalized Likelihood

The unpenalized negative log-likelihood ℓ_n takes the form

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

where
- $O_{j,k}$ = number of observed events in the (j,k)-th rectangle
- $R_{j,k}$ = time at risk in the (j,k)-th rectangle

The MLE is explicit:

$$\hat{\eta}^{\text{mle}}_{j,k} = \log\left(\frac{O_{j,k}}{R_{j,k}}\right) \quad \rightarrow \quad \text{overfitting.}$$

Our model has no *a priori*. But the inference is made by minimizing the **penalized likelihood** [2]

$$\ell^{\text{pen}}_n(\eta) = \underbrace{\ell_n(\eta)}_{\text{goodness of fit}} + \underbrace{\frac{\text{pen}}{2}\sum_{j,k} v_{j,k}\left(\eta_{j+1,k} - \eta_{j,k}\right)^2 + w_{j,k}\left(\eta_{j,k+1} - \eta_{j,k}\right)^2}_{\text{regularization}}$$

- $\boldsymbol{v}$ et $\boldsymbol{w}$ are weights
- pen is a **tradeoff parameter**.

Types of regularization

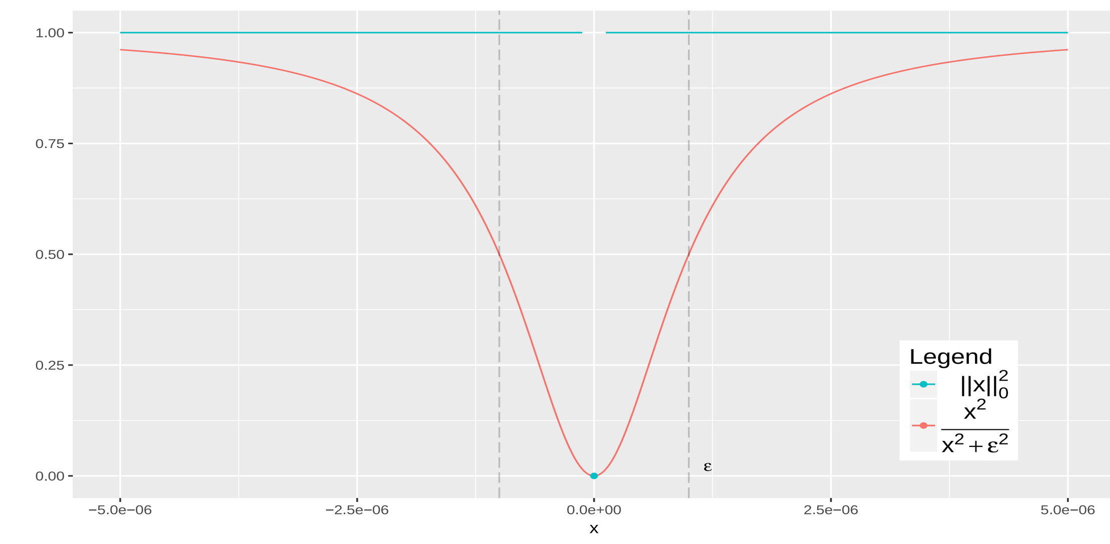
- L_2 norm (Ridge) Regularization with $\boldsymbol{v} = \boldsymbol{w} = 1$ → Smoothing
- L_0 norm Regularization with the iterative **Adaptive Ridge** [3] procedure → Segmented estimation

The weights are iteratively adapted:

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

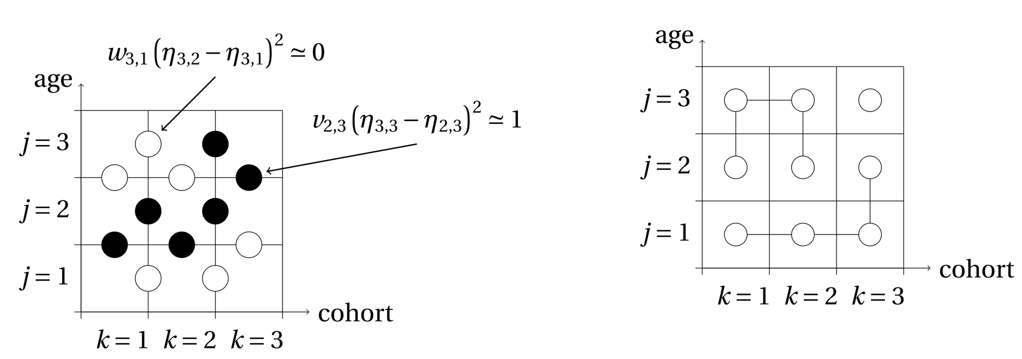
Approximation of the L_0 norm:

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



Principle of Model Selection using the L_0 norm

1. We alternate until convergence between
   - Minimizing $\ell^{\text{pen}}_n(\eta)$ for fixed **v** and **w**.
   - Adapting **v** and **w** using $\boldsymbol{\eta}$.
2. The weighted differences of $\boldsymbol{\eta}$ are used to **select areas over which the hazard is constant**:



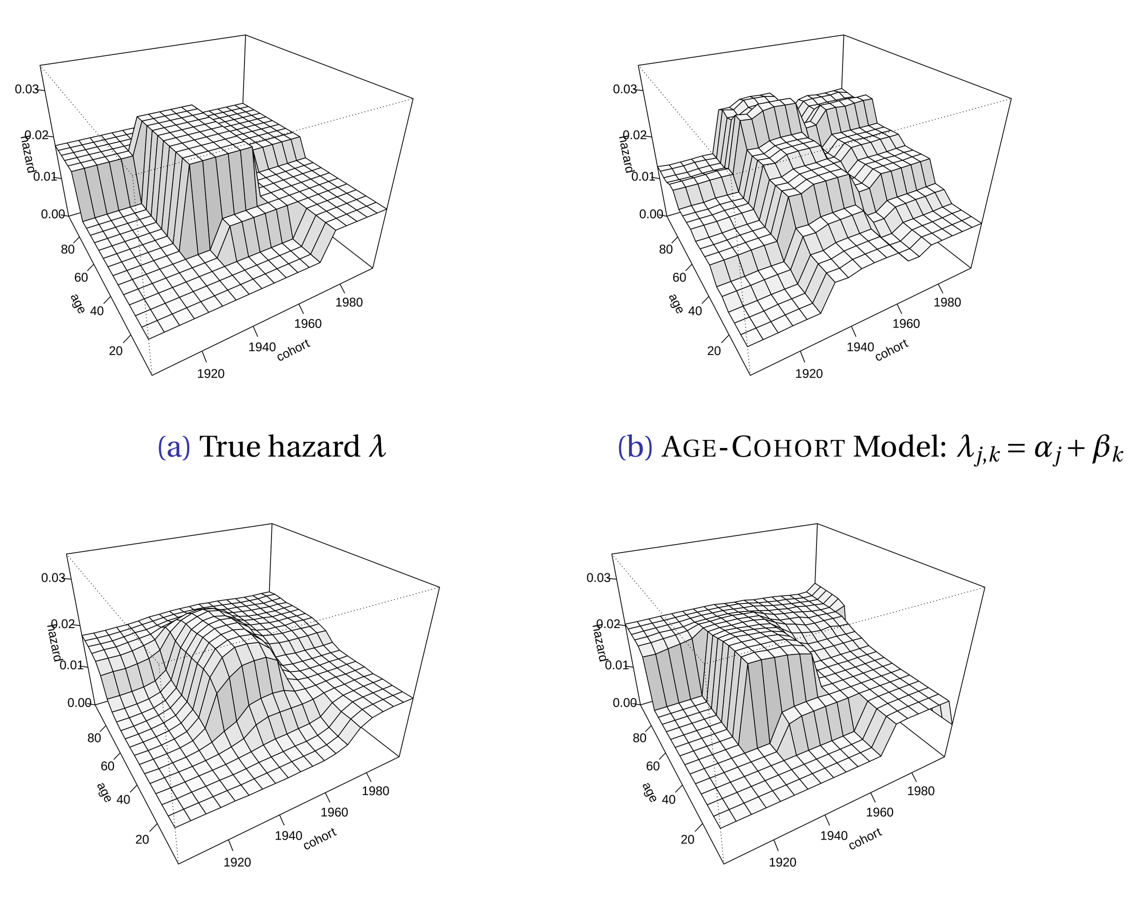
Step 1:
Representation of
$v_{j,k}\left(\eta_{j+1,k} - \eta_{j,k}\right)^2$
et $w_{j,k}\left(\eta_{j,k+1} - \eta_{j,k}\right)^2$

Step 2:
Create corresponding graph
Extract connex components

3. On each area : $\boldsymbol{\eta}$ is estimated by unpenalized maximum likelihood.

Simulation: with a Piecewise Constant Hazard

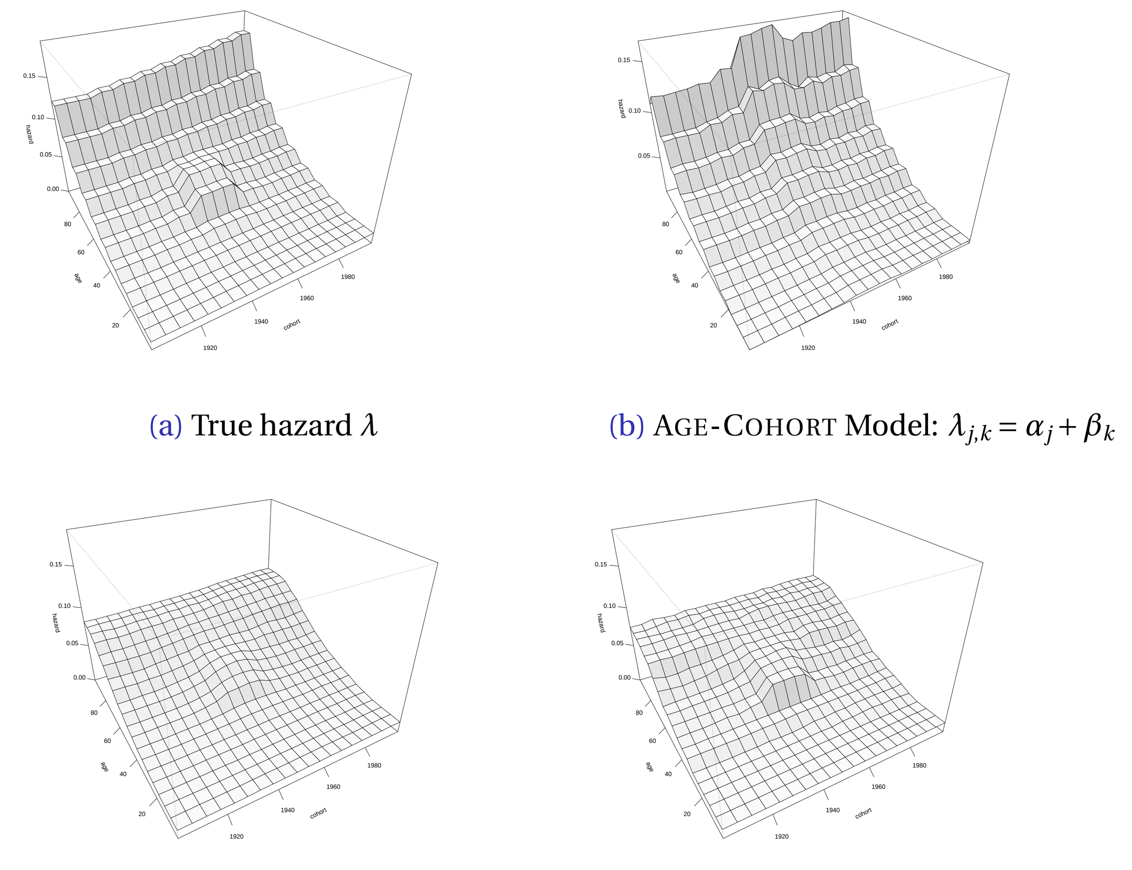
- 4000 **events** are generated using the true hazard
- The hazard is estimated using different methods
- We represent the **median estimate** of 500 such replications



(a) True hazard λ (b) Age-Cohort Model: $\lambda_{j,k} = \alpha_j + \beta_k$

(c) L_2 Regularization (d) L_0 Regularization

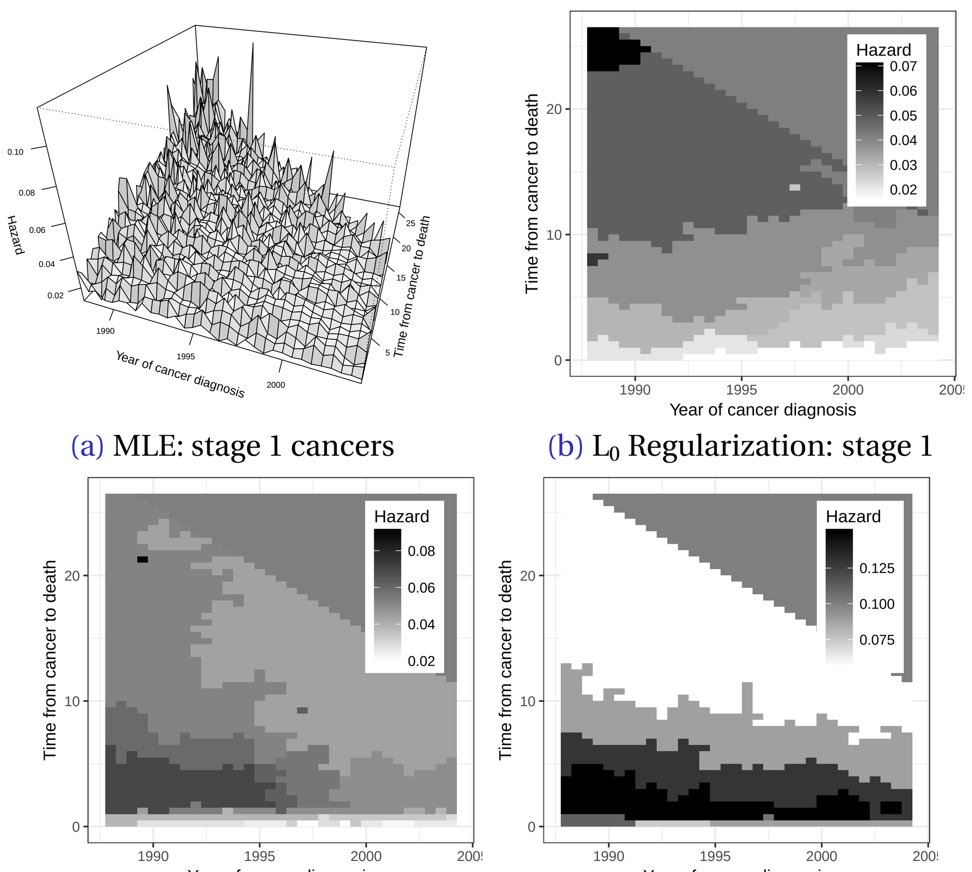
Simulation: with a Smooth Hazard



(a) True hazard λ (b) Age-Cohort Model: $\lambda_{j,k} = \alpha_j + \beta_k$

(c) L_2 Regularization (d) L_0 Regularization

Application to Real Data: Breast Cancer Mortality



(a) MLE: stage 1 cancers (b) L_0 Regularization: stage 1

(c) L_0 Regularization: stage 2 (d) L_0 Regularization: stage 3

Conclusion & Perspectives

- The method allows a **segmented estimation** of the hazard
- Inference is computationally tracatable
- The model can be extended:

$$\log \lambda_{j,k} = \underbrace{\alpha_j}_{\text{age effet}} + \underbrace{\beta_k}_{\text{cohort effect}} + \underbrace{\delta_{j,k}}_{\text{interaction term}}$$

with regularization over $\delta_{j,k}$.

References

References

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

[2] O. Bouaziz and G. Nuel, *L0 Regularization for the Estimation of Piecewise Constant Hazard Rates in Survival Analysis.* Applied Mathematics, 2017.

[3] F. Frommlet and G. Nuel, *An Adaptive Ridge Procedure for L0 Regularization.* PloS one, 2016.

[4] J. Chen and J. Chen, *Extended Bayesian information criteria for model selection with large model spaces.* Biometrika, 20008.

[1] MAP5, UMR CNRS 8145, Paris, France; [2] LPSM, UMR CNRS 8001, Paris, France;

Correspondance to: vivien.goepp@parisdescartes.fr