Interaction Effect in Age-Period-Cohort Analysis

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

[1]MAP5, (UMR 8145), Université Paris Descartes [2]LPSM, (UMR 8001), Sorbonne Université

Statistical Methods for Post Genomic Data, 31^{st} Jan–1^{st} Feb 2019, Barcelona

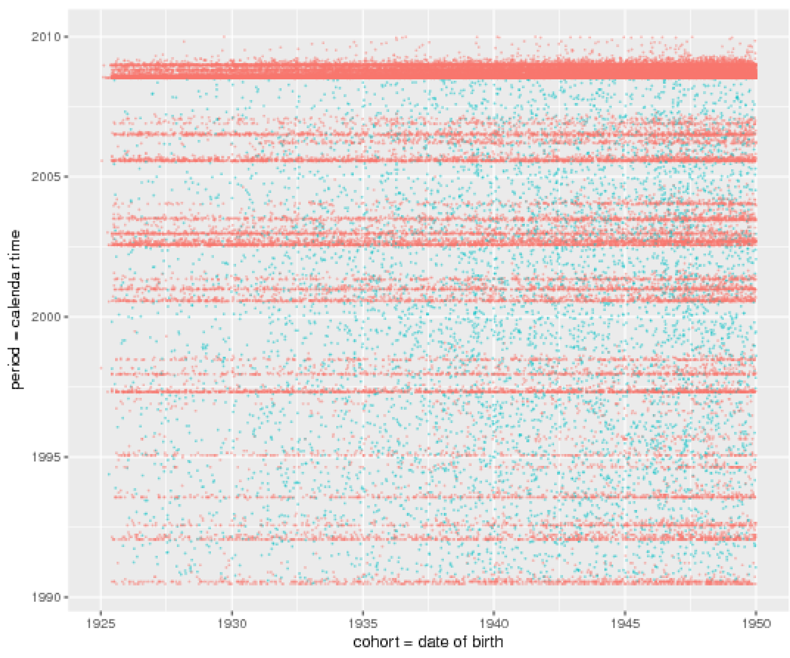
Motivating Example

The **E3N Cohort Study** [1]:

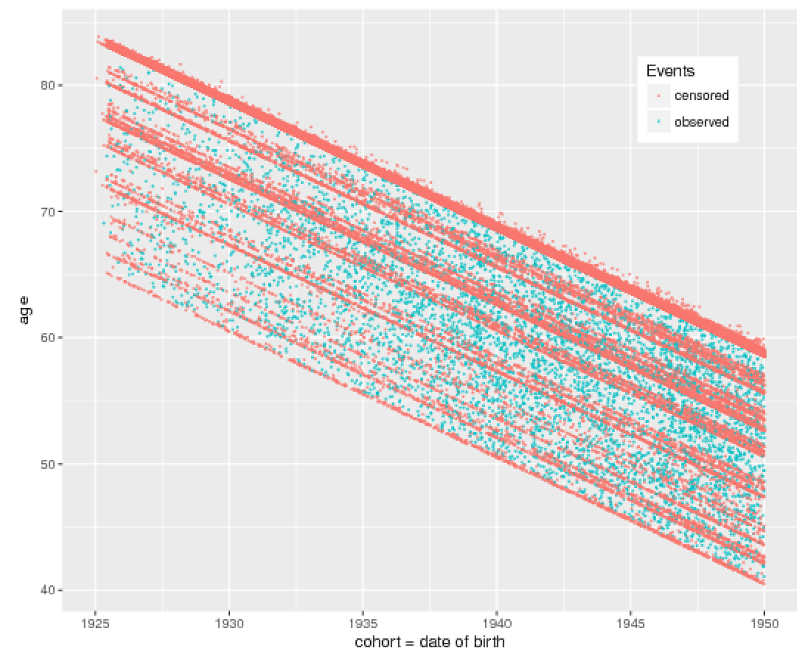
- Epidemiological study on the link between cancer and nutrition (EPIC).
- Population: $\sim$ **100000 women**.
- Medical data gathered every 2-3 years using questionnaires.
- Blood and saliva samples are also gathered for some participants (not used here).
- The event of interest is the **occurrence of breast cancer**
- These occurrences are spread over the period $[1990, 2010]$

Goal: estimate the incidence of breast cancer as a function of age and cohort

Data in the Period-cohort plane



Data in the Age-cohort plane



Right-censoring

- T_i is the age of cancer onset, U_i is the date of birth.
- We do not observe (T_i) but

$$Y_i = \min(T_i, C_i) \quad \text{and} \quad \Delta_i = 1_{T_i = Y_i}$$

where C is a censoring r.v. independent from (U, T).

- We infer the bivariate hazard rate

$$\lambda(t|u) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T \leq t + dt | T \geq t, U = u)}{dt}$$

Existing Models in Age-Period-Cohort Analysis

In the literature: we infer $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, parameters of the age, cohort and period effetcs.

- In the AGE-COHORT model,

$$\log \lambda_{\mathbf{j,k}} = \alpha_{\mathbf{j}} + \beta_{\mathbf{k}}$$

  - $J + K - 1$ parameters for JK variables: regularizing
  - Strong *a prior* on λ.

- In the AGE-PERIOD-COHORT model,

$$\log \lambda_{\mathbf{j,k}} = \alpha_{\mathbf{j}} + \beta_{\mathbf{k}} + \gamma_{\mathbf{j+k-1}}$$

  - Regularizing
  - Strong *a priori* on λ
  - Non identifiable

The Age-Cohort-Interaction model

- We introduce the AGE-COHORT-INTERACTION model:

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

where $\delta_{\mathbf{j,k}}$ is the **interaction** between age and cohort.

- Estimation by penalized likelihood:

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

Aim: enforce $\delta_{j,k} \simeq 0$ except where relevant.

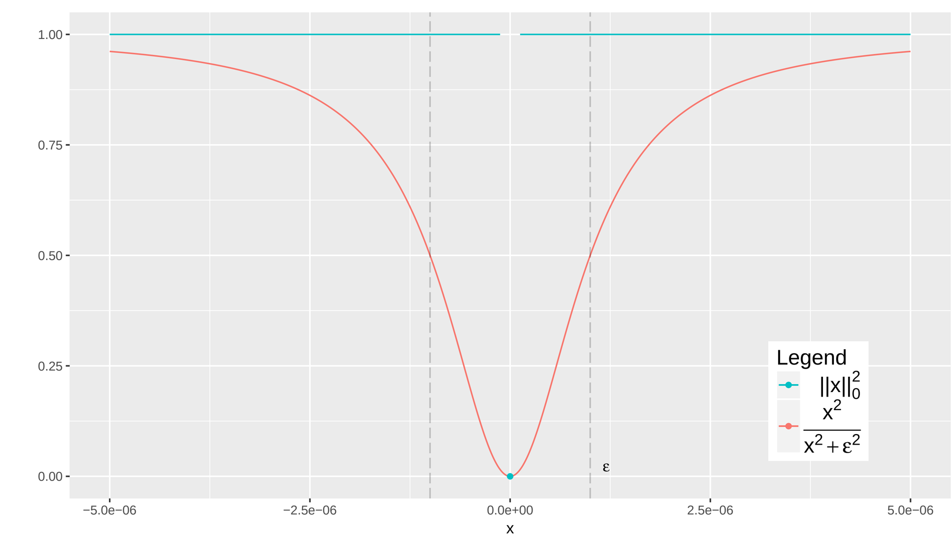
Regularization using L_0 penalization

- We enforce $\boldsymbol{\delta}$ to be **piecewise constant**.
- Fused L_0 regularization with the iterative **Adaptive Ridge** [3] procedure.
- The estimation is iterative:

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

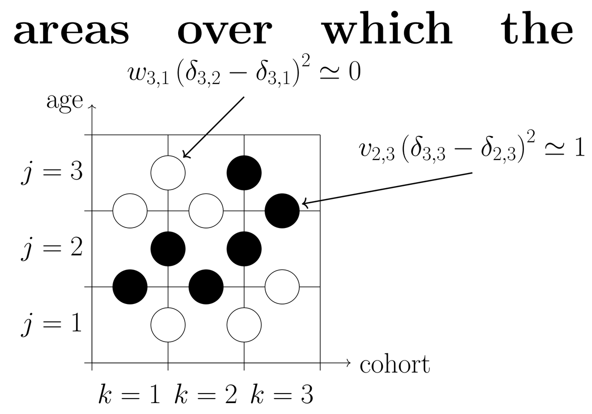
Principle of the approximation:

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

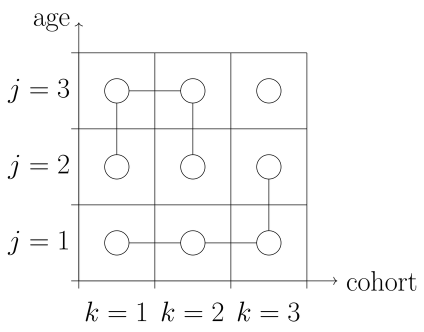


Model Selection with L_0 norm

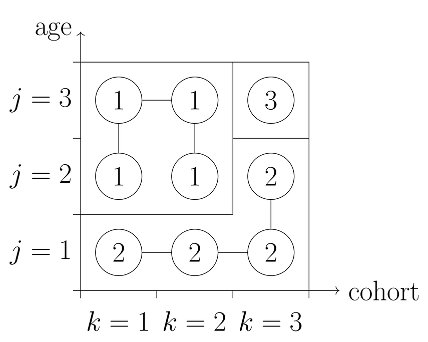
1. We alternate until convergence:
   - Minimize $\ell_n^{\text{pen}}(\boldsymbol{\theta})$ for fixed $\mathbf{v}$ and $\mathbf{w}$.
   - Adapt $\mathbf{v}$ and $\mathbf{w}$ using $\boldsymbol{\theta}$.
2. The weighted differences of $\boldsymbol{\eta}$ are used to **select areas over which the hazard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



Step 1: Weighted differences



Step 2: Corresponding graph

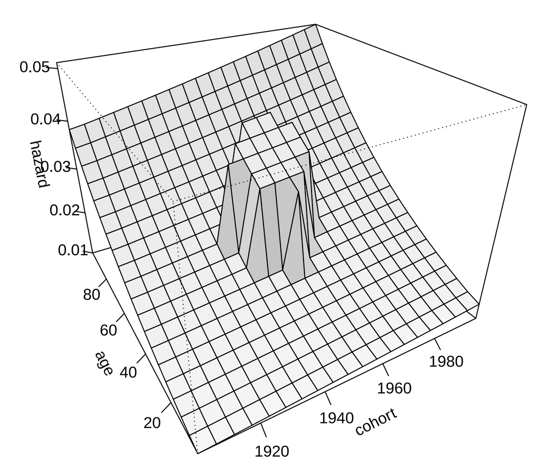


Step 3: Connected components

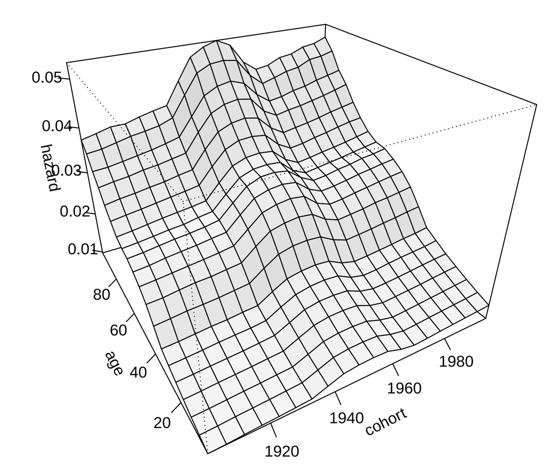
3. On each area : $\boldsymbol{\theta}$ is estimated by unpenalized maximum likelihood.

Simulation results

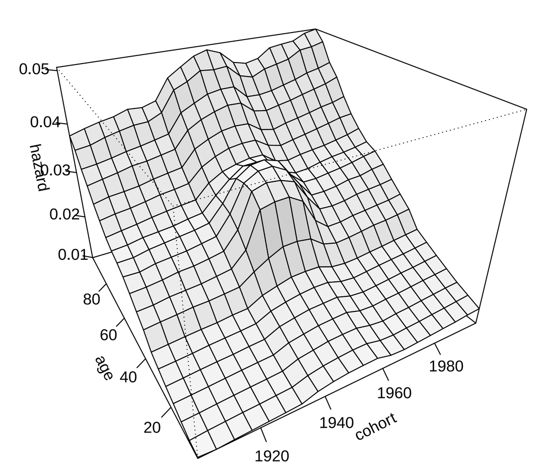
- 10000 **data points** are generated using the true hazard
- We represent the **median estimate** of 500 such replications



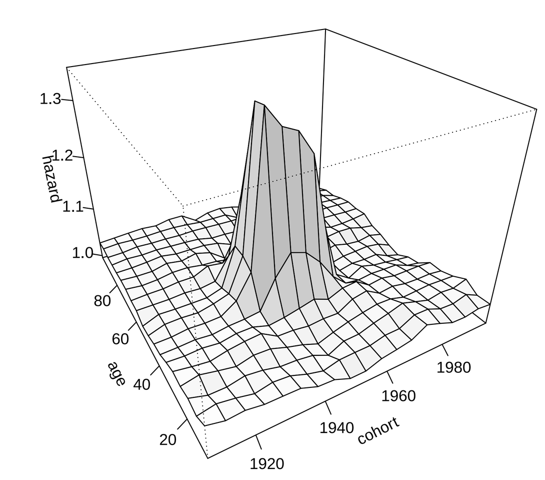
True $\lambda_{j,k}$



AC model



$\lambda_{j,k}$ (ACI model)



$\delta_{j,k}$ (ACI model)

Conclusion & References

Conclusion

- Joint estimation of the effects and their interaction
- More general model than APC
- Can use ensemble methods for better predictive performance

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, 2008.

Acknowledgment

This study was realized using data from the Inserm E3N cohort that has been established and is maintained with the financial support of the MGEN, the Gustave Roussy Institute and La Ligue contre le Cancer.

Correspondance to:
vivien.goepp@parisdescartes.fr

LaTeX TikZposter