Regularized hazard estimation for age-period-cohort analysis

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

Statistical Methods for Post Genomic Data, January 11-12 2018, Montpellier, France

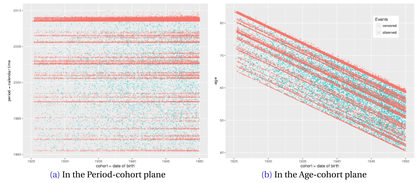
Presentation of the problem

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

- Epigemiological study focused on the link between cancer and nutrition (part of EPIC).
- Population: ~ 100000 women (mostly teachers).
- Medical data gathered every 2-3 years using questionnaires (9 questionnaires in total). Diet, physical exercice and medical treatments are monitored.
- Blood and saliva samples are also gathered for som participants (not used here).
- The event of interest is the **occurrence of breast cancer**
- These occurrences are spread over the period [1990, 2010]

Objective: estimate the hazard rate of breast cancer occurrence

Representation of the data:



(a) In the Period-cohort plane (b) In the Age-cohort plane

Right-Censoring

The age of cancer occurrence is observed for only 7% of the individuals.

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

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

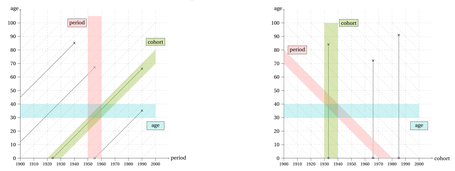
where C is a censoring r.v. independent from T.

- We observe $\Delta_i = 1_{T_i = Y_i}$.
- We infer the instantenous hazard rate

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

Age-Period-Cohort analysis

The hazard depends also on the date of birth



period = calendar time cohort = date of birth

The hazard is a function of **age**, **period**, and **cohort**.

Existing Models in age-period-cohort Analysis

In the literature: we infer α, β, and γ, parameters of the age, cohort and period effetcs.

- In the AGE-COHORT Model, it is assumed that

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

  - $J + K - 1$ parameters for JK variables: regularizing
  - Strong *a priori* on λ.
- In the AGE-PERIOD-COHORT Model, it is assumed that

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

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

New Approach: Penalized Likelihood

Reparametrization:

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

The unpenalized negative log-likelihood ℓ_n takes the form

$$\ell_n(\boldsymbol{\eta}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \exp(\eta_{j,k}) R_{j,k} - \eta_{j,k} O_{j,k},$$

where

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

The MLE is explicit:

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

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

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

- $\boldsymbol{v}$ et $\boldsymbol{w}$ are weights
- pen is a penalty constant

Types of regularization

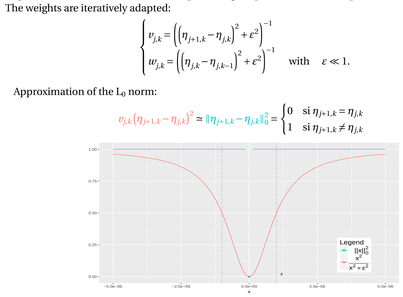
- Ridge Regularisation – L_2 norm with $\boldsymbol{v} = \boldsymbol{w} = \mathbf{1}$ → Smoothed estimation
- Regularisation L_0 with the iterative **Adaptive Ridge** [3] procedure → Segmented estimation

The weights are iteratively adapted:

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

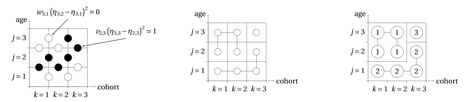
Approximation of the L_0 norm:

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



Principle of Model Selection using the L_0 norm

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



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

Step 2: Create corresponding graph

Step 3: Each connex component yields a constant area

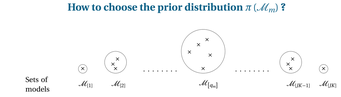
3. On each area : $\boldsymbol{\eta}$ is estimated by unpenalized maximum likelihood.

Bayesian Selection Criteria

- Problem: how to choose between m penalties $\text{pen}_1, \ldots, \text{pen}_m$?
- In other words: how to choose between M models $\mathscr{M}_1, \ldots, \mathscr{M}_M$ of dimensions $q_1, \ldots, q_M$?
- Solution: maximize $\mathbb{P}(\mathscr{M}_m | \boldsymbol{R}, \boldsymbol{O}) \propto \mathbb{P}(\boldsymbol{R}, \boldsymbol{O} | \mathscr{M}_m) \pi(\mathscr{M}_m)$.
- By approximation:

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

How to choose the prior distribution $\pi(\mathscr{M}_m)$?



$\mathscr{M}_{(q_m)}$ is the set of models of dimension q_m

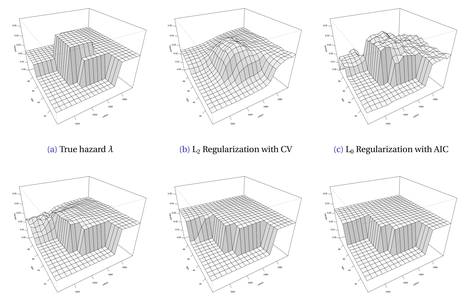
BIC: $\pi(\mathscr{M}_m) = 1$
All $\mathscr{M}_m$ are equipossible
All models are equipossible

EBIC$_0$: $\mathbb{P}\left(\mathscr{M}_m \in \mathscr{M}_{(q_m)}\right) = 1$
All $\mathscr{M}_{(q_m)}$ are equipossible
All model dimensions are equipossible [4]

We will compare three criteria: the **EBIC$_0$**, the **BIC**, the **AIC** and the cross-validation.

Simulation Results: with a piecewise constant hazard

- 4000 **data points** are generated using the true hazard
- The hazard is estimated using different methods
- We represent the **median estimate** of 100 such replications

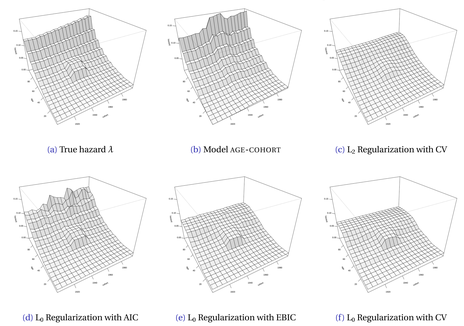


(a) True hazard λ (b) L_2 Regularization with CV (c) L_0 Regularization with AIC

(d) L_0 Regularization with BIC (e) L_0 Regularization with EBIC (f) L_0 Regularization with CV

Simulation Results: with a smooth hazard

Simulations are carried under the same conditions, but with a non-piecewise true hazard.

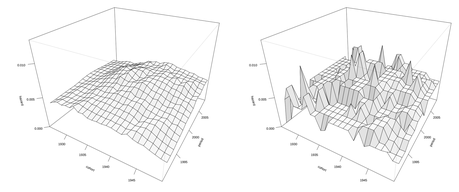


(a) True hazard λ (b) Model AGE-COHORT (c) L_2 Regularization with CV

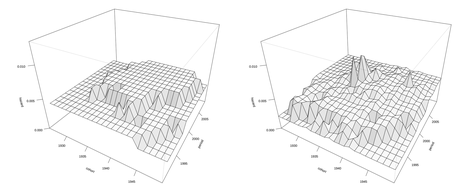
(d) L_0 Regularization with AIC (e) L_0 Regularization with EBIC (f) L_0 Regularization with CV

Application to real data: The E3N Cohort

The hazard is estimated using different methods:



(a) L_2 Regularization with CV (b) L_0 Regularization with AIC



(c) L_0 Regularization with EBIC (d) **Bootstrapped** L_0 Regularization with EBIC

Figure (d) is a obtained by ENSEMBLE method: the data is sampled 100 times, yielding 100 models. We represent the median of the corresponding hazards.

Conclusion & Perspectives

- The method allows a **segmented estimation** of the hazard
- EBIC$_0$ more efficient than AIC or BIC
- 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 & Ackownledgment

References

[1] E 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] E 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.

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.

[1] MAP5, UMR CNRS 8145, Paris, France; [2] LPMA, UMR CNRS 7599, Paris, France;

Correspondance to: vivien.goepp@parisdescartes.fr