L_0 Regularization for the estimation of piecewise constant hazard

Joint work with O. Bouaziz[1] and G. Nuel[2]

Vivien Goepp [1]

[1]: MAP5 (CNRS 8145), Université Paris Descartes
[2]: LPMA (CNRS 7599), Université Pierre et Marie Curie

July 4, 2017

What we want to do

Classic setting: $\lambda(t) = \lim_{dt \to 0} \frac{1}{dt} \mathbb{P}(t < T < t + dt | T > t)$

What we want to do

Classic setting: $\lambda(t) = \lim_{dt\to 0} \frac{1}{dt}\mathbb{P}(t < T < t + dt | T > t)$

With cohort effect:

$$\lambda_c(t) = \lim_{dt\to 0} \frac{1}{dt}\mathbb{P}(t < T < t + dt | T > t, \text{cohort} = c)$$

What we want to do

Classic setting: $\lambda(t) = \lim_{dt \to 0} \frac{1}{dt} \mathbb{P}(t < T < t + dt | T > t)$

With cohort effect:

$$\lambda_c(t) = \lim_{dt \to 0} \frac{1}{dt} \mathbb{P}(t < T < t + dt | T > t, \text{cohort} = c)$$

Problem: with a discretization $(t_0, \ldots, t_J)$ and $(c_0, \ldots, c_K)$, estimate

$$\lambda_{j,k} = \lambda_c(t) \quad \text{such that} \quad t \in [t_{j-1}, t_j], c \in [c_{k-1}, c_k]$$

What we want to do

Classic setting: $\lambda(t) = \lim_{dt \to 0} \frac{1}{dt} \mathbb{P}(t < T < t + dt | T > t)$

With cohort effect:

$$\lambda_c(t) = \lim_{dt \to 0} \frac{1}{dt} \mathbb{P}(t < T < t + dt | T > t, \text{cohort} = c)$$

Problem: with a discretization $(t_0, \ldots, t_J)$ and $(c_0, \ldots, c_K)$, estimate

$$\lambda_{j,k} = \lambda_c(t) \quad \text{such that} \quad t \in [t_{j-1}, t_j], c \in [c_{k-1}, c_k]$$

Notation:

cohort	=	date of birth
period	=	date of event
age	=	age of patient at event

Age-Period Diagram

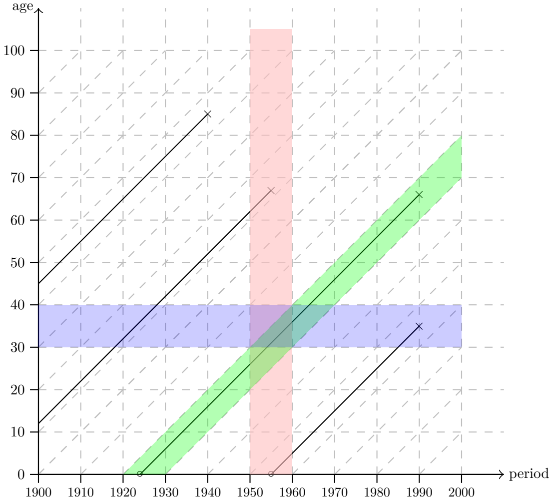


Figure 1: Lexis diagram: age and period

Age-Cohort Diagram

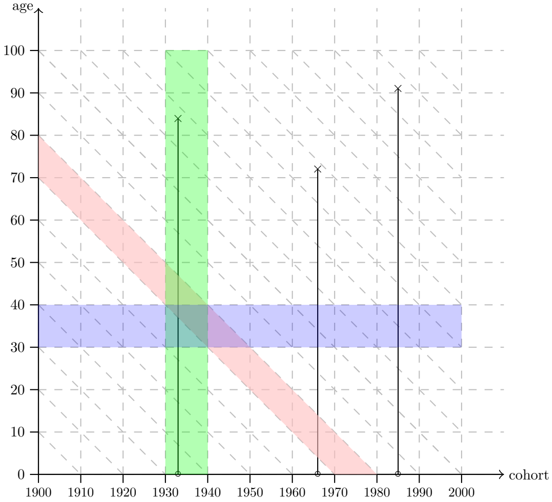


Figure 2: Age-Cohort diagram

Existing models

Age-Cohort model: $\lambda_{j,k} = \exp\Big(\mu + \underbrace{\alpha_j}_{\text{age effet}} + \underbrace{\beta_k}_{\text{cohort effect}}\Big)$:

- Regularizing ($J + K$ parameters instead of $J \times K$)
- Strong *a priori* on the data

Existing models

Age-Cohort model: $\lambda_{j,k} = \exp\Big(\mu + \underbrace{\alpha_j}_{\text{age effet}} + \underbrace{\beta_k}_{\text{cohort effect}}\Big)$:

- Regularizing ($J + K$ parameters instead of $J \times K$)
- Strong *a priori* on the data

Age-Period-Cohort model: $\lambda_{j,k} = \exp\Big(\mu + \alpha_j + \beta_k + \underbrace{\gamma_{j+k-1}}_{\text{period effect}}\Big)$

- Not identifiable (period = age + cohort)
- Strong *a priori* on the data

Existing models

Age-Cohort model: $\lambda_{j,k} = \exp\Big(\mu + \underbrace{\alpha_j}_{\text{age effet}} + \underbrace{\beta_k}_{\text{cohort effect}}\Big)$:

- Regularizing ($J + K$ parameters instead of $J \times K$)
- Strong *a priori* on the data

Age-Period-Cohort model: $\lambda_{j,k} = \exp\Big(\mu + \alpha_j + \beta_k + \underbrace{\gamma_{j+k-1}}_{\text{period effect}}\Big)$

- Not identifiable (period = age + cohort)
- Strong *a priori* on the data

We want a regularizing model without any *a priori* of age, cohort or period effet.

Maximum Likelihood Estimation

Parametrization: $\lambda_{j,k} = \exp(\eta_{j,k})$.

Negative log-likelihood:[1]

$$\ell(\boldsymbol{\eta}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \exp(\eta_{j,k}) R_j^k - \eta_{j,k} O_j^k. \tag{1}$$

Explicit estimator:

$$\eta_{j,k} = \log\left(\frac{O_j^k}{R_j^k}\right).$$

O_j^k: number of events in rectangle (j,k)

R_j^k: total time at risk in rectangle (j,k)

[1] Aalen, Borgan, and Gjessing, *Survival and Event History Analysis*, p 224.

Likelihood Penalization

Penalization:

$$\ell^{\text{pen}}(\boldsymbol{\eta}) = \ell(\boldsymbol{\eta}) \quad + \quad \frac{\text{pen}}{2} \sum_{j=1}^{J-1} \sum_{k=1}^{K} v_{j,k} \left(\eta_{j+1,k} - \eta_{j,k}\right)^2$$
$$+ \quad \frac{\text{pen}}{2} \sum_{j=1}^{J} \sum_{k=1}^{K-1} w_{j,k} \left(\eta_{j,k+1} - \eta_{j,k}\right)^2$$

Likelihood Penalization

Penalization:

$$\ell^{\text{pen}}(\boldsymbol{\eta}) = \ell(\boldsymbol{\eta}) \quad + \quad \frac{\text{pen}}{2} \sum_{j=1}^{J-1} \sum_{k=1}^{K} v_{j,k} \left(\eta_{j+1,k} - \eta_{j,k}\right)^2$$
$$\quad + \quad \frac{\text{pen}}{2} \sum_{j=1}^{J} \sum_{k=1}^{K-1} w_{j,k} \left(\eta_{j,k+1} - \eta_{j,k}\right)^2$$

- For Ridge regularization: $\boldsymbol{v} = \boldsymbol{w} = 1$
- For L_0 norm regularization, the weights are adapted iteratively:

$$v_{j,k}^{(m)} = \frac{1}{\left(\eta_{j+1,k}^{(m)} - \eta_{j,k}^{(m)}\right)^2 + \varepsilon^2} \quad w_{j,k}^{(m)} = \frac{1}{\left(\eta_{j,k+1}^{(m)} - \eta_{j,k}^{(m)}\right)^2 + \varepsilon^2}$$

Principle of Adaptive Ridge

At convergence:[2]

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

This approximates the L_0 norm with a **convex** *and* **differentiable** function.

Remark: This work extends the univariate case[3].

[2] Frommlet and Nuel, "An Adaptive Ridge Procedure for L0 Regularization."

[3] Bouaziz and Nuel, "L0 Regularisation for the Estimation of Piecewise Constant Hazard Rates in Survival Analysis."

Algorithmic procedure

Algorithm: Newton-Raphson

Data: data, pen, $\boldsymbol{v}$, $\boldsymbol{w}$
Result: $\widehat{\boldsymbol{\eta}} = \arg\min \ell^{\text{pen}}(\boldsymbol{\eta})$
$\boldsymbol{\eta} \leftarrow \mathbf{0}$;
while *not converge* **do**

$\boldsymbol{\eta}_{\text{new}} \leftarrow \boldsymbol{\eta} - Hess^{-1}(\boldsymbol{\eta}, \text{data}, \text{pen}, \boldsymbol{v}, \boldsymbol{w}) S(\boldsymbol{\eta}, \text{data}, \text{pen}, \boldsymbol{v}, \boldsymbol{w})$;

$\boldsymbol{\eta} \leftarrow \boldsymbol{\eta}_{\text{new}}$;

end

Algorithmic procedure

Algorithm: Newton-Raphson

Data: data, pen, $\boldsymbol{v}$, $\boldsymbol{w}$
Result: $\widehat{\boldsymbol{\eta}} = \arg\min \ell^{\text{pen}}(\boldsymbol{\eta})$
$\boldsymbol{\eta} \leftarrow \mathbf{0}$;
while *not converge* **do**
  $\boldsymbol{\eta}_{\text{new}} \leftarrow \boldsymbol{\eta} - Hess^{-1}(\boldsymbol{\eta}, \text{data}, \text{pen}, \boldsymbol{v}, \boldsymbol{w}) S(\boldsymbol{\eta}, \text{data}, \text{pen}, \boldsymbol{v}, \boldsymbol{w})$;

  $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta}_{\text{new}}$;
end

The hessian is a band matrix:inversion in $\mathcal{O}(JK^2)$ instead of $\mathcal{O}(J^3K^3)$

Algorithmic procedure

Algorithm: Newton-Raphson

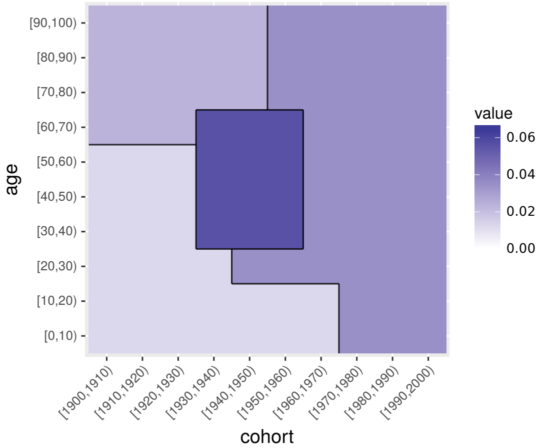
Data: data, pen, $\boldsymbol{v}$, $\boldsymbol{w}$
Result: $\widehat{\boldsymbol{\eta}} = \arg\min \ell^{\text{pen}}(\boldsymbol{\eta})$
$\boldsymbol{\eta} \leftarrow \mathbf{0}$;
while *not converge* **do**
  $\boldsymbol{\eta}_{\text{new}} \leftarrow \boldsymbol{\eta} - Hess^{-1}(\boldsymbol{\eta}, \text{data}, \text{pen}, \boldsymbol{v}, \boldsymbol{w}) S(\boldsymbol{\eta}, \text{data}, \text{pen}, \boldsymbol{v}, \boldsymbol{w})$;
  $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta}_{\text{new}}$;
end

The hessian is a band matrix:inversion in $\mathcal{O}(JK^2)$ instead of $\mathcal{O}(J^3K^3)$

Algorithm: Adaptive Ridge

Data: data, pen
Result: $\widehat{\boldsymbol{\eta}}^{\text{sel}}$
$\boldsymbol{v} \leftarrow \boldsymbol{w} \leftarrow \mathbf{1}$;
$\boldsymbol{\eta} \leftarrow \mathbf{0}$;
while *not converge* **do**
  $\boldsymbol{\eta}^{\text{new}} \leftarrow \text{NR}(\boldsymbol{\eta}, \text{data}, \text{pen}, \boldsymbol{v}, \boldsymbol{w})$;
  $v_{j,k} \leftarrow \frac{1}{\left(\eta^{\text{new}}_{j+1,k} - \eta^{\text{new}}_{j,k}\right)^2 + \varepsilon^2}$;
  $w_{j,k} \leftarrow \frac{1}{\left(\eta^{\text{new}}_{j,k+1} - \eta^{\text{new}}_{j,k}\right)^2 + \varepsilon^2}$;
  $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta}_{\text{new}}$;
end
region $\leftarrow$ selection$(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w})$;
$\widehat{\boldsymbol{\eta}}^{\text{sel}} \leftarrow \frac{O^{\text{region}}}{R^{\text{region}}}$;

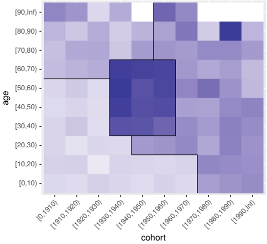
Results on simulated data



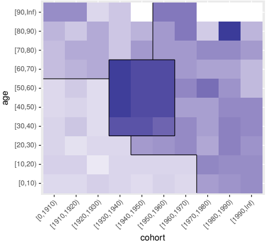
True piecewise constant hazard

Simulating 5000 points with censoring $\sim \mathcal{U}([75, 100])$

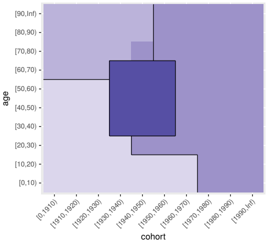
Estimated Hazard



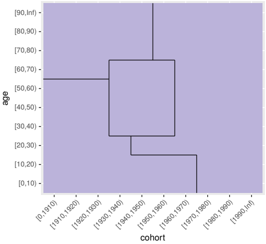
pen $= 0.001$



pen $= 0.1$



pen $= 1$



pen $= 100$

Real data application: The E3N Cohort[4]

Sample size: 91992 (women)

Cohort $\in [1925, 1950]$

Period $\in [1990, 2010]$

The event of interest is the outbreak of breast cancer.

Percentage of censored events: 93%

Cohort	$[1925, 1930)$	$[1930, 1935)$	$[1935, 1940)$	$[1940, 1945)$	$[1945, 1950]$
Number of cancers	374	808	1322	1604	1785
Total time at risk	538402.1	852763.2	1184437	1422128	1829870

Summary of E3N Data

[4] Clavel-Chapelon, "Cohort Profile of the French E3n Cohort Study."

Visualization of the E3N Data

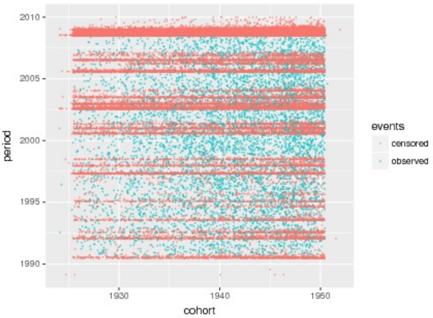


Figure 3: In the period-cohort plane

Visualization of the E3N Data

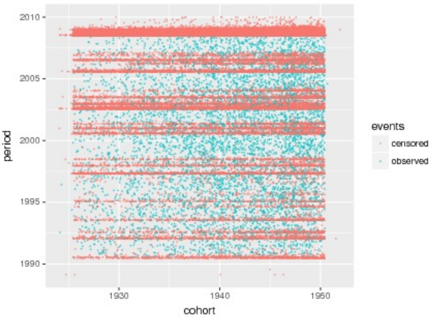


Figure 3: In the period-cohort plane

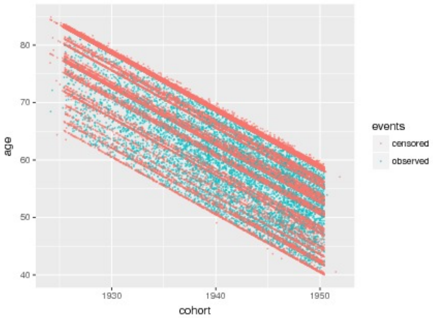
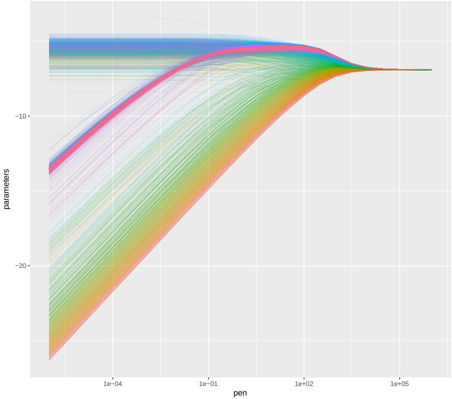


Figure 4: In the age-cohort plane

Results: Ridge

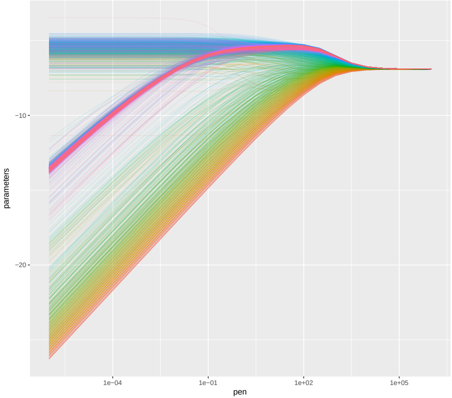
$J = K = 40$

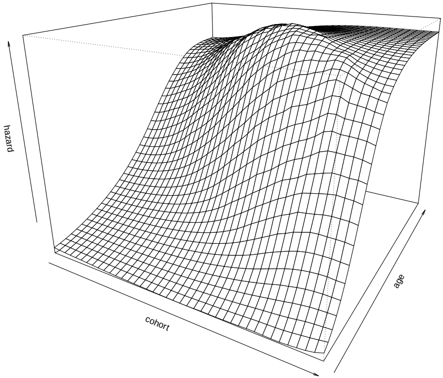


Regularization path

Results: Ridge

$J = K = 40$

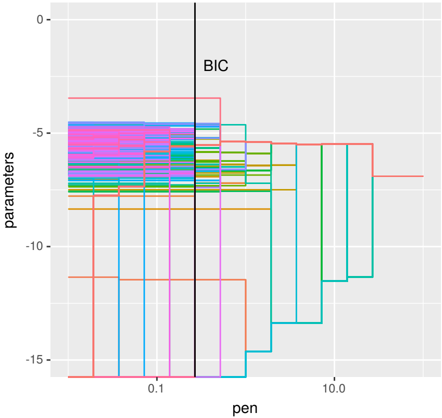




Regularization path

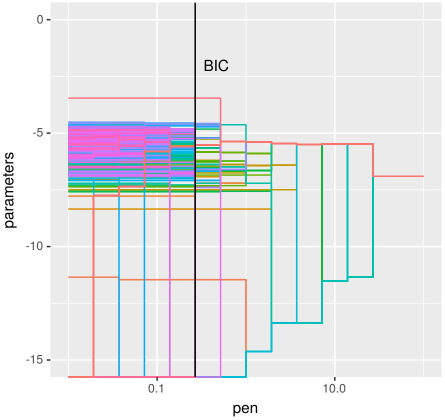
pen = 300

Results: Adaptive Ridge

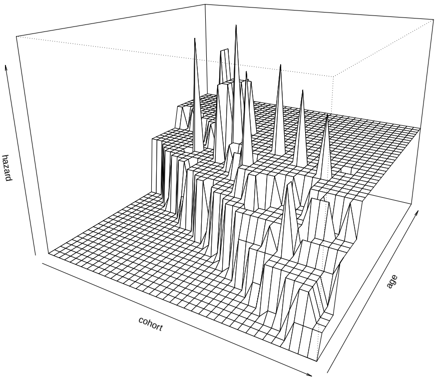


Regularization path

Results: Adaptive Ridge



Regularization path



BIC Selected Model (42 parameters)

Summary & Perspectives

What we have done:

- Regularized estimation of λ without APC-type *a priori*
- Ridge regularization with constant weights
- L_0 regularization with adapted weights

Summary & Perspectives

What we have done:

- Regularized estimation of λ without APC-type *a priori*
- Ridge regularization with constant weights
- L_0 regularization with adapted weights

Upcoming work

- K-fold cross-validation for Ridge regularization
- Confidence intervals using resampling

Thank you for your attention

Aalen, Odd, Ornulf Borgan, and Hakon Gjessing. *Survival and Event History Analysis: A Process Point of View*. Springer Science & Business Media, 2008.

Bouaziz, Olivier, and Grégory Nuel. "L0 Regularisation for the Estimation of Piecewise Constant Hazard Rates in Survival Analysis." *ArXiv Preprint ArXiv:1609.04595*, 2016.

Clavel-Chapelon, Françoise. "Cohort Profile of the French E3n Cohort Study." *International Journal of Epidemiology* 44, no. 3 (2014): 801–9.

Frommlet, Florian, and Grégory Nuel. "An Adaptive Ridge Procedure for L0 Regularization." *PLOS ONE* 11, no. 2 (February 2016): 1–23. doi:10.1371/journal.pone.0148620.

Nuel, Gregory, and Gilles Monneret. *Bandsolve: Fast Solver for Symetric Band Matrix*, 2016R package version 1.2.3.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2015. https://www.R-project.org/