New Methods for Detecting and Modelling Heterogeneity in Survival Responses

Olivier Bouaziz
joint work with Arthur Carcano, Vivien Goepp, Grégory Nuel

MAP5 (CNRS 8145), Université Paris Descartes, Sorbonne Paris Cité
LPSM (CNRS 8001), UPMC, Sorbonne Universités.

Séminaire de statistique du LPSM

1. Background in time to event analysis

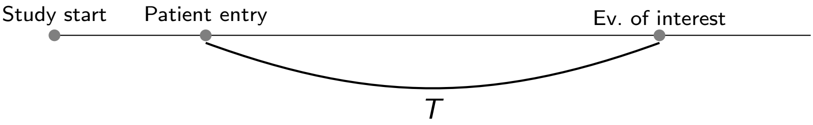
2. A change-point model for detecting heterogeneity in ordered survival responses

3. Regularized hazard estimation for age-period-cohort analysis

Outline

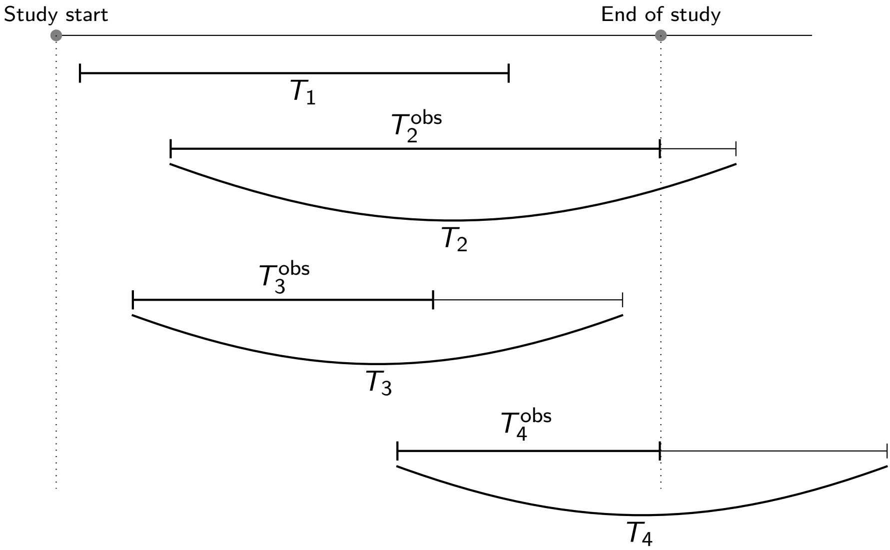
Background in time to event analysis

- We study a positive continuous time to event variable T.
- T represents the time difference between event of interest and patient entry.



- Examples : time to relapse of Leukemia patients, time to onset of cancer, time to death ...

Background in time to event analysis : right censoring



The hazard rate

- Observations :
$$\begin{cases} T_i^{\text{obs}} = T_i \wedge C_i \\ \Delta_i = I(T_i \leq C_i) \end{cases}$$

- Independent censoring : $T \perp\!\!\!\perp C$

- A key relation :

$$\lambda(t) := \lim_{\triangle t \to 0} \frac{\mathbb{P}[t \leq T < t + \triangle t | T \geq t]}{\triangle t}$$
$$= \lim_{\triangle t \to 0} \frac{\mathbb{P}[t \leq T^{\text{obs}} < t + \triangle t, \Delta = 1 | T^{\text{obs}} \geq t]}{\triangle t}.$$

Many estimators (Nelson Aalen, Kaplan-Meier, ...) are based on this relation.

Likelihood and the Cox model

- The likelihood of the observed data is equal to :

$$\prod_{i=1}^{n} f(T_i^{\text{obs}})^{\Delta_i} S(T_i^{\text{obs}})^{1-\Delta_i} = \prod_{i=1}^{n} \lambda(T_i^{\text{obs}})^{\Delta_i} \exp\left(-\int_0^{T_i^{\text{obs}}} \lambda(t)dt\right),$$

where f is the density of T and $S(t) = \mathbb{P}[T > t]$.

- Regression modelling : let $\boldsymbol{X} \in \mathbb{R}^d$ be a covariate.

$$\lambda(t|\boldsymbol{X}_i) = \lambda_0(t)\exp(\boldsymbol{X}_i\beta) \qquad \text{(Cox Model)}$$

For a binary covariate,

$$\frac{\lambda(t|\boldsymbol{X}_i = 1)}{\lambda(t|\boldsymbol{X}_i = 0)} = \exp(\beta).$$

Outline

The Steno memorial hospital dataset

- Cohort dataset of 2 709 Danish diabetic patients collected between 1933 and 1981 from *Andersen et al., 1993.*
- The variable of interest is the time from diabetes onset until death (in years).
- 74% of right censoring due to emigration or end of study (December, 31st 1984).
- Left truncation due to delayed entry into the study.
- Gender and calendar year of diabetes onset (range : 1933 – 1972) were also collected for each patient.

- Classical survival analysis except that we want to take into account a possible cohort effect due to the wide range of year of diabetes onset.

Illustration of the cohort effect

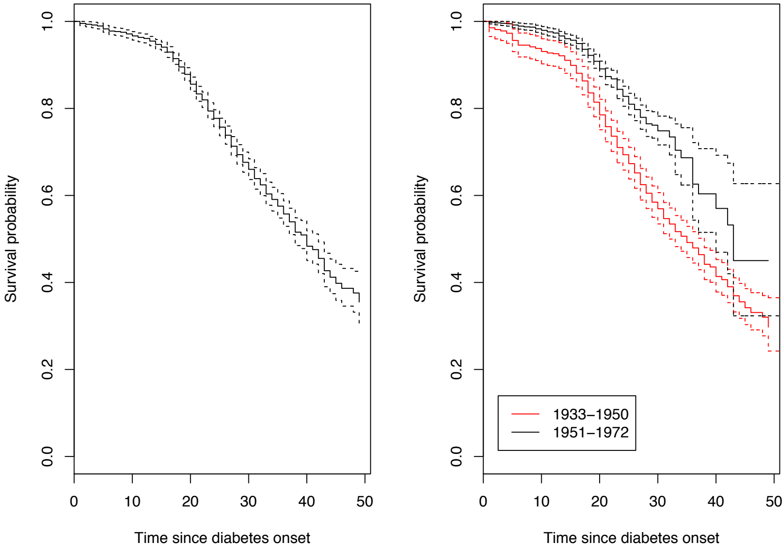


Illustration of the cohort effect

Heterogeneity in the survival time distribution according to year of diabetes onset !

Classical solutions are :

- Divide the dataset in arbitrary segments.
- Regression model (Cox for instance) adjusted with respect to year of diabetes onset.
- Age-period-cohort model.

We propose a different approach : deal with the cohort effect as an unsupervised clustering problem. We propose an iterative algorithm which :

- Automatically find the segments locations.
- Compute *a posteriori* probabilities of breakpoints.
- Estimate survival quantities in each segment.

The model

- Suppose there are K segments and let $R_1, \ldots, R_n$ be the segment indexes of each individual. For example, $n = 10$ and $R_{1:10} = 1112222333$ means 2 breakpoints occur in positions 3 and 7.
- The model is :

$$\lambda(t|\boldsymbol{X}_i, R_i = k) = \lim_{\triangle t \to 0} \frac{\mathbb{P}(t \leq T_i < t + \triangle t | T_i \geq t, \boldsymbol{X}_i, R_i = k)}{\triangle t}$$
$$= \lambda_k(t) \exp(\boldsymbol{X}_i \boldsymbol{\beta}_k)$$

The goal is :

- Estimate the *a posteriori* probability of a breakpoint, $\mathbb{P}(R_i = k, R_{i+1} = k + 1 | \text{data})$.
- Estimate the λ_ks and $\boldsymbol{\beta}_k$s.

The EM algorithm

Introduce data $= (T^{\text{obs}}_{1:n}, \Delta_{1:n}, \boldsymbol{X}_{1:n})$ and $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_K, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$.

- (E-step) Compute :

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = \int_{R_{1:n}} \mathbb{P}(R_{1:n}|\text{data}; \boldsymbol{\theta}_{\text{old}}) \log \mathbb{P}(\text{data}|R_{1:n}; \boldsymbol{\theta}) dR_{1:n}$$
$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{P}(R_i = k|\text{data}; \boldsymbol{\theta}_{\text{old}}) \log \mathbb{P}(\text{data}_i|R_i = k; \boldsymbol{\theta}),$$

  where $\boldsymbol{\theta}_{\text{old}}$ represents the previous update of the parameter.

- (M-step) Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}})$ with respect to $\boldsymbol{\theta}$.

Computation of the emission probability

- The contribution of the ith individual to the likelihood conditionally to its segment index is :

$$\log \mathbb{P}(T_i^{\text{obs}}, \Delta_i, \boldsymbol{X}_i | R_i = k; \boldsymbol{\theta})$$
$$= \Delta_i \left\{ \log \left(\lambda_k(T_i^{\text{obs}}) \right) + \boldsymbol{X}_i \boldsymbol{\beta}_k \right\} - \int_0^{T_i^{\text{obs}}} \lambda_k(t) \exp(\boldsymbol{X}_i \boldsymbol{\beta}_k) dt.$$

- Take λ_k as an Exponential, Weibull, Piecewise-Constant-Hazard or nonparametric baseline hazard.

Computation of posterior segment distributions

- Let $\eta_i(k) = \mathbb{P}(R_i = k+1 | R_{i-1} = k)$ be a prior distribution (for instance uniform prior distribution). Under the constraint $R_n = K$, the model is a constrained Hidden Markov Model. We have :

$$\mathbb{P}(R_i = k | \text{data}; \boldsymbol{\theta}) \propto F_i(k; \boldsymbol{\theta}) B_i(k; \boldsymbol{\theta}),$$

  where
  - $F_i(k; \boldsymbol{\theta}) = \mathbb{P}(\text{data}_{1:i}, R_i = k; \boldsymbol{\theta})$ is the forward quantity.
  - $B_i(k; \boldsymbol{\theta}) = \mathbb{P}(\text{data}_{(i+1):n}, R_n = K | R_i = k; \boldsymbol{\theta})$ is the backward quantity.

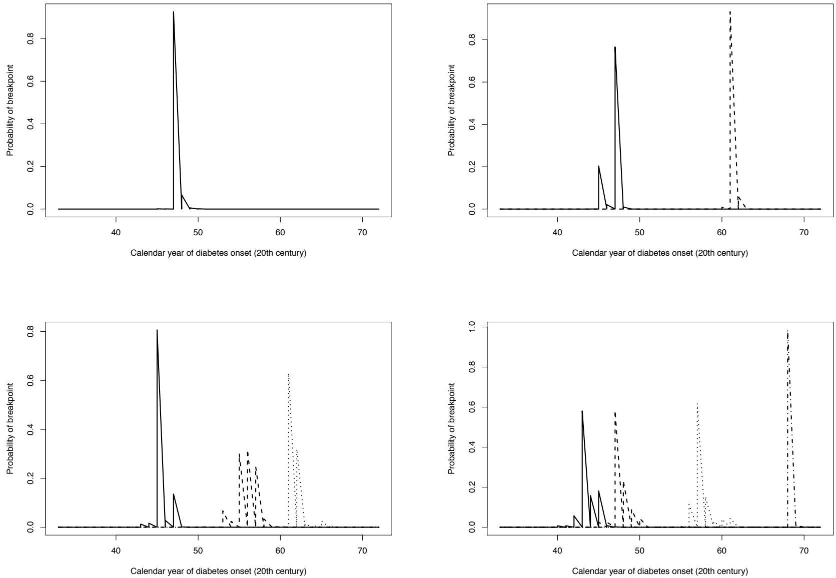
- The *posterior* probability of a breakpoint occurring at position i is :

$$\mathbb{P}(R_i = k, R_{i+1} = k+1 | \text{data}; \boldsymbol{\theta}) \\ \propto F_i(k; \boldsymbol{\theta}) \eta_{i+1}(k) e_{i+1}(k+1; \boldsymbol{\theta}) B_{i+1}(k+1; \boldsymbol{\theta}).$$

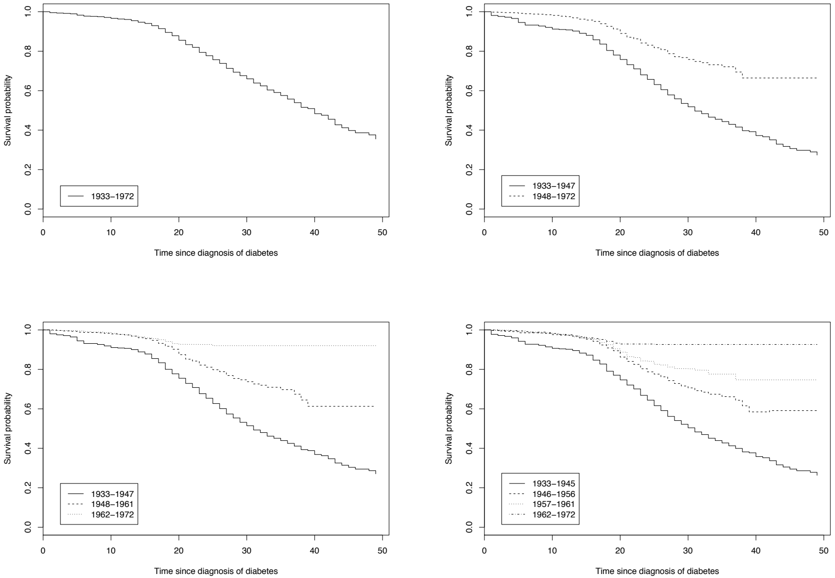
The Steno memorial hospital dataset (exp. baseline, one covariate : gender)

	No bp	One bp 1948	Two bp 1948, 62	Three bp 1946, 57, 62	Four bp 1944, 48, 58, 69
$\hat{\lambda}_1$	0.012	0.022	0.023	0.023	0.024
$\hat{\lambda}_2$		0.006	0.008	0.011	0.015
$\hat{\lambda}_3$			0.003	0.006	0.009
$\hat{\lambda}_4$				0.003	0.004
$\hat{\lambda}_5$					0.001
$e^{\hat{\beta}_1}$	1.32	1.29	1.29	1.29	1.25
$e^{\hat{\beta}_2}$		1.61	1.60	1.41	1.43
$e^{\hat{\beta}_3}$			1.44	1.80	1.50
$e^{\hat{\beta}_4}$				1.46	1.66
$e^{\hat{\beta}_5}$					0.90
BIC	7426.405	7214.413	7179.012	7187.442	7194.631

Marginal distributions of the breakpoints



Weighted Kaplan-Meier estimators



Confidence intervals

A bootstrap procedure is implemented to obtain 95% confidence intervals.
In the two breakpoints model (with covariate gender) :

- 1933 – 1947

$$\hat{\lambda} = 0.023[0.020; 0.027] \quad \exp(\hat{\beta}) = 1.29[1.06; 1.55]$$

- 1948 – 1961

$$\hat{\lambda} = 0.008[0.007; 0.012] \quad \exp(\hat{\beta}) = 1.60[1.12; 2.09]$$

- 1962 – 1972

$$\hat{\lambda} = 0.003[0.001; 0.005] \quad \exp(\hat{\beta}) = 1.44[0.90; 3.40]$$

This procedure takes into account uncertainty about breakpoints location !

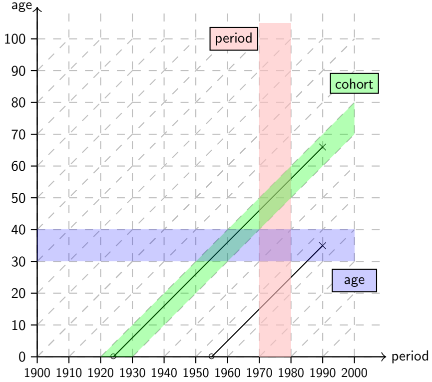
Summary

- Breakpoint locations are detected with high probability.
- The BIC criterion is very performant to find the number of segments.
  - Also in the null case of no breakpoints.
- Very accurate estimations on each segment.
  - Bootstrap procedure allows to compute valid confidence intervals.
- Estimation performance is not very sensitive to the choice of baseline.
  - Piecewise constant hazard gives a good compromise for accurate estimates and performant breakpoints detection.
- Ties can be handled through the prior distribution of breakpoints.

A change-point model for detecting heterogeneity in ordered survival responses. O. Bouaziz and G. Nuel. **Statistical Methods in Medical Research** (2017)

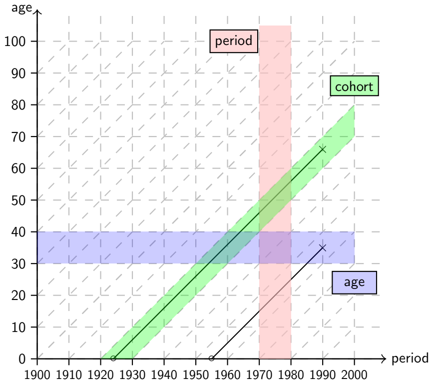
Outline

The Lexis diagram

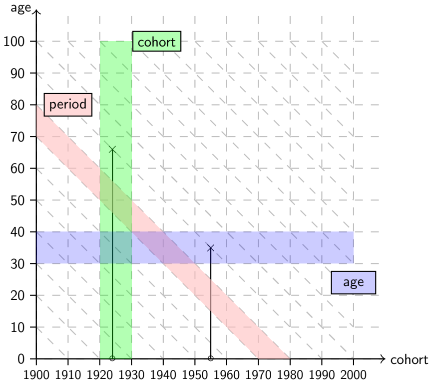


Age-Period Lexis diagram

The Lexis diagram



Age-Period Lexis diagram



Age-Cohort Lexis diagram

Key relation : cohort+age=period

The age-period-cohort approach

- Discretization of the hazard rate into $J \times K$ intervals :

$$\lambda(\text{age}, \text{cohort}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \lambda_{j,k} I(c_{j-1} \leq \text{age} < c_j, d_{k-1} \leq \text{cohort} < d_k)$$

- Decompose $\lambda_{j,k}$ through :
  - α_j : the age effect
  - β_k : the cohort effect
  - γ_{j+k-1} : the period effect

- The classical approches try to estimate α_j, β_k (and γ_{j+k-1}).

Existing models

1. The AGE-COHORT model :

$$\log \lambda_{j,k} = \mu + \alpha_j + \beta_k \qquad \text{(with } \alpha_1 = \beta_1 = 0\text{)}.$$

   - $J + K - 1$ parameters to estimate instead of $J \times K$.
   - But no interactions are allowed !

2. The AGE-PERIOD-COHORT model :

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

   - Identifiability issues :
     - Estimate second order differences.
     - Add arbitrary constraints.
   - Still no interactions allowed !

Our approach : penalizing the maximum likelihood estimator

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

The log-likelihood is equal to :

$$\ell_n(\boldsymbol{\lambda}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \{O_{j,k} \log(\lambda_{j,k}) - \lambda_{j,k} R_{j,k}\}$$

The maximum likelihood estimator is :

$$\lambda_{j,k}^{\text{mle}} = \frac{O_{j,k}}{R_{j,k}}$$

Our approach : penalizing the maximum likelihood estimator

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

The log-likelihood is equal to :

$$\ell_n(\boldsymbol{\lambda}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \{O_{j,k} \log(\lambda_{j,k}) - \lambda_{j,k} R_{j,k}\}$$

The maximum likelihood estimator is :

$$\lambda_{j,k}^{\text{mle}} = \frac{O_{j,k}}{R_{j,k}}$$

Overfitting issues : $J \times K$ parameters need to be estimated !

Our approach : penalizing the maximum likelihood estimator

Set $\log \lambda_{j,k} = \eta_{j,k}$ Estimation of $\boldsymbol{\eta}$ is achieved through penalized

log-likelihood :

$$\ell_n^{\text{pen}}(\boldsymbol{\eta}) = \underbrace{\ell_n(\boldsymbol{\eta})}_{\text{log-likelihood}}$$

Our approach : penalizing the maximum likelihood estimator

Set $\log \lambda_{j,k} = \eta_{j,k}$ Estimation of $\boldsymbol{\eta}$ is achieved through penalized

log-likelihood :

$$\ell_n^{\text{pen}}(\boldsymbol{\eta}) = \underbrace{\ell_n(\boldsymbol{\eta})}_{\text{log-likelihood}} - \underbrace{\frac{\text{pen}}{2}\left\{ \sum_{j,k} v_{j,k}\left(\eta_{j+1,k} - \eta_{j,k}\right)^2 + w_{j,k}\left(\eta_{j,k+1} - \eta_{j,k}\right)^2 \right\}}_{\text{regularization term}},$$

- $\boldsymbol{v}$ and $\boldsymbol{w}$ represent weights
- pen is a penalty term

Two types of regularization

1. L_2 regularization (Ridge) with $\boldsymbol{v} = \boldsymbol{w} = \mathbf{1}$
2. L_0 regularization with the adaptive ridge procedure.
   Iterative updates of the weights :

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

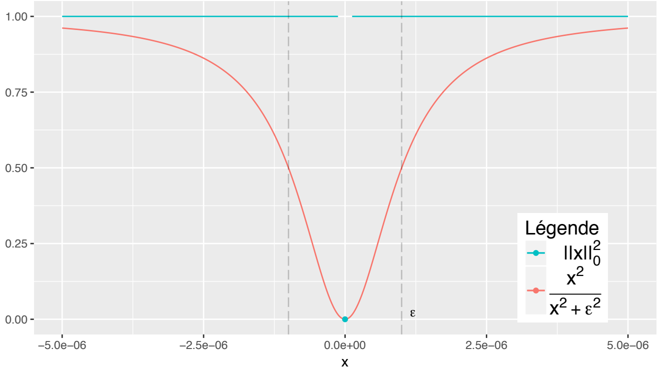
   with $\varepsilon \ll 1$.

F. Frommlet and G. Nuel, *An Adaptive Ridge Procedure for L0 Regularization*. **PlosOne** (2016).

L_0 norm approximation

When $\varepsilon \ll 1$:

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



The *Adaptive Ridge* procedure

procedure ADAPTIVE-RIDGE($\boldsymbol{O}, \boldsymbol{R}$, pen)

$(\boldsymbol{\eta},\boldsymbol{v},\boldsymbol{w}) \leftarrow (\mathbf{0},\mathbf{1},\mathbf{1})$

while not converge **do**

$\boldsymbol{\eta}^{\text{new}} \leftarrow \text{NEWTON-RAPHSON}(\boldsymbol{O}, \boldsymbol{R}, \text{pen}, \boldsymbol{v}, \boldsymbol{w})$

$$v_{j,k}^{\text{new}} \leftarrow \left(\left(\eta_{j+1,k}^{\text{new}} - \eta_{j,k}^{\text{new}} \right)^2 + \varepsilon^2 \right)^{-1}$$

$$w_{j,k}^{\text{new}} \leftarrow \left(\left(\eta_{j,k}^{\text{new}} - \eta_{j,k-1}^{\text{new}} \right)^2 + \varepsilon^2 \right)^{-1}$$

$(\boldsymbol{\eta},\boldsymbol{v},\boldsymbol{w}) \leftarrow (\boldsymbol{\eta}^{\text{new}},\boldsymbol{v}^{\text{new}},\boldsymbol{w}^{\text{new}})$

end while

end procedure

The *Adaptive Ridge* procedure

procedure $\textsc{Adaptive-Ridge}(\boldsymbol{O}, \boldsymbol{R}, \text{pen})$

  $(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) \leftarrow (\mathbf{0}, \mathbf{1}, \mathbf{1})$

  while not converge **do**

    $\boldsymbol{\eta}^{\text{new}} \leftarrow \textsc{Newton-Raphson}(\boldsymbol{O}, \boldsymbol{R}, \text{pen}, \boldsymbol{v}, \boldsymbol{w})$

    $v_{j,k}^{\text{new}} \leftarrow \left(\left(\eta_{j+1,k}^{\text{new}} - \eta_{j,k}^{\text{new}} \right)^2 + \varepsilon^2 \right)^{-1}$

    $w_{j,k}^{\text{new}} \leftarrow \left(\left(\eta_{j,k}^{\text{new}} - \eta_{j,k-1}^{\text{new}} \right)^2 + \varepsilon^2 \right)^{-1}$

    $(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) \leftarrow (\boldsymbol{\eta}^{\text{new}}, \boldsymbol{v}^{\text{new}}, \boldsymbol{w}^{\text{new}})$

  end while

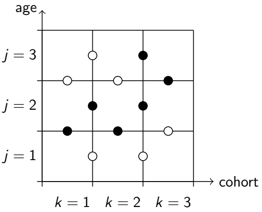
  Compute $(\boldsymbol{O}^{\text{sel}}, \boldsymbol{R}^{\text{sel}})$ **from** $(\boldsymbol{\eta}^{\text{new}}, \boldsymbol{v}^{\text{new}}, \boldsymbol{w}^{\text{new}})$

  $\exp(\boldsymbol{\eta}^{\text{mle}}) \leftarrow \boldsymbol{O}^{\text{sel}} / \boldsymbol{R}^{\text{sel}}$

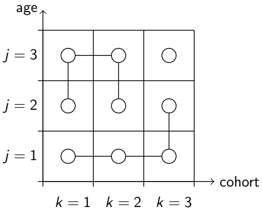
  return η^{mle}

end procedure

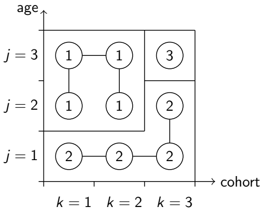
Model selection using the *Adaptive Ridge*



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



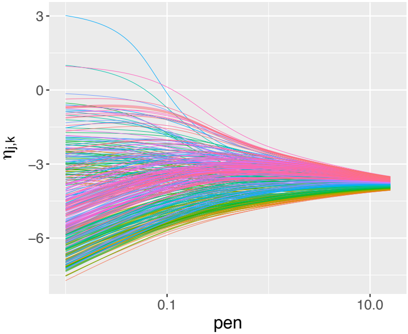
(b) Corresponding graph



(c) Segmentation through connected components

Comparison of the two regularization methods

$$\begin{aligned} \text{pen} \to 0 \quad &: \quad \widehat{\boldsymbol{\eta}} \to \widehat{\boldsymbol{\eta}}^{\text{mle}} \\ \text{pen} \to \infty \quad &: \quad \widehat{\boldsymbol{\eta}} \to \text{ constant} \end{aligned}$$

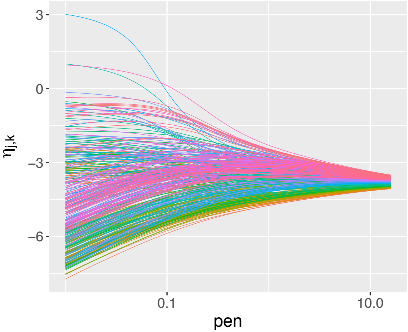


L_2 regularization

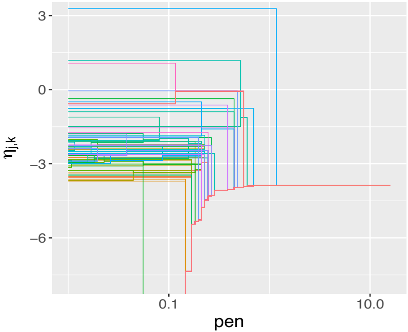
Comparison of the two regularization methods

pen $\to 0$: $\widehat{\boldsymbol{\eta}} \to \widehat{\boldsymbol{\eta}}^{\text{mle}}$

pen $\to \infty$: $\widehat{\boldsymbol{\eta}} \to$ constant



L_2 regularization



L_0 regularization

Model selection for the *Adaptive Ridge* estimator

Four different methods to perform model selection :

1. $\text{BIC}(m) = -2\ell_n(\widehat{\boldsymbol{\eta}}_m^{\text{mle}}) + q_m \log n$
2. $\text{EBIC}_0(m) = -2\ell_n(\widehat{\boldsymbol{\eta}}_m^{\text{mle}}) + q_m \log n + 2 \log \binom{JK}{q_m} \quad (*)$
3. $\text{AIC}(m) = -2\ell_n(\widehat{\boldsymbol{\eta}}_m^{\text{mle}}) + 2q_m$
4. K-fold Cross validation (CV),

with q_m the dimension of the model.

$(*)$ J. Chen and Z. Chen, *Extended Bayesian information criteria for model selection with large model spaces*, **Biometrika**, 2008.

Simulated data

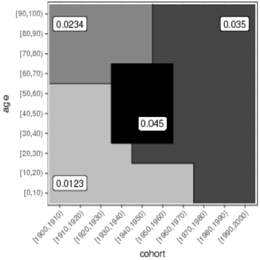
Two scenarios ($n = 4\,000$, 15% of censoring) :

- Piecewise constant hazard
- Smooth hazard.

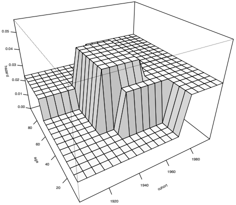
Comparison of estimators :

- Age-Cohort model : $\log \lambda_{j,k} = \mu + \alpha_j + \beta_k$
- L_2 regularization with CV criterion
- L_0 regularization with AIC, BIC, $EBIC_0$ and CV criterions.

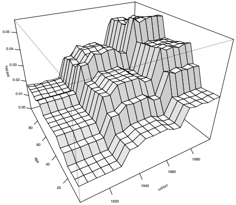
Simulations : piecewise constant hazard scenario



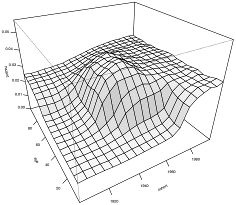
Truth



Truth

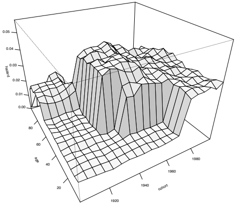


AGE-COHORT model

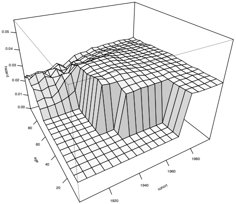


L_2 CV

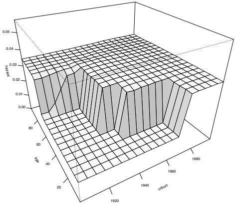
Simulations : piecewise constant hazard scenario



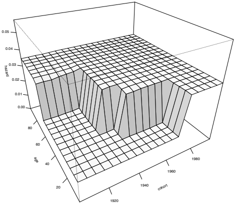
L_0 AIC



L_0 BIC

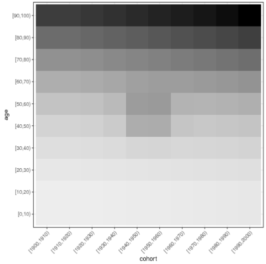


L_0 $EBIC_0$

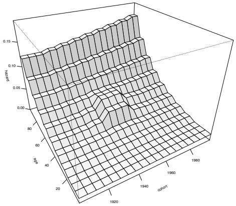


L_0 CV

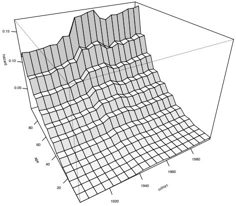
Simulations : smooth hazard



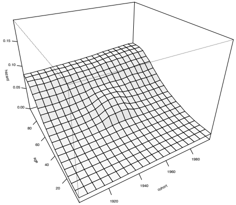
Truth



Truth

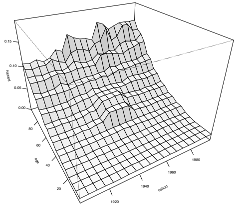


AGE-COHORT model

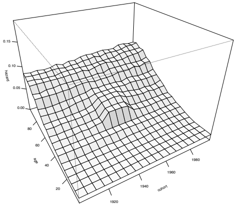


L_2 CV

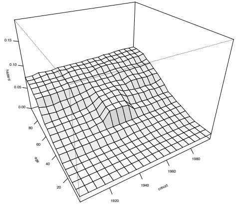
Simulations : smooth hazard



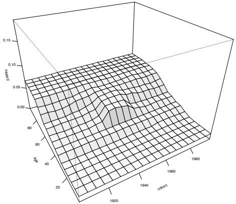
L_0 AIC



L_0 BIC



L_0 $EBIC_0$



L_0 CV

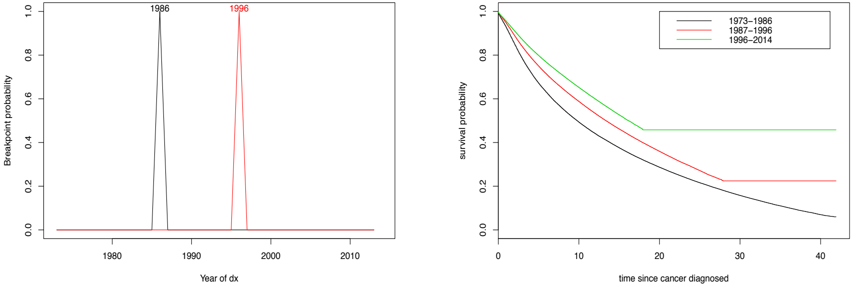
The SEER data

- Huge american registry dataset of breast cancer
- Primary, unilateral, malignant and invasive cancers
- 1.2 million of patients
- 60% of censoring
- The cancer diagnostics range from 1973 to 2014
- The variable of interest is the time from cancer diagnosis until death.

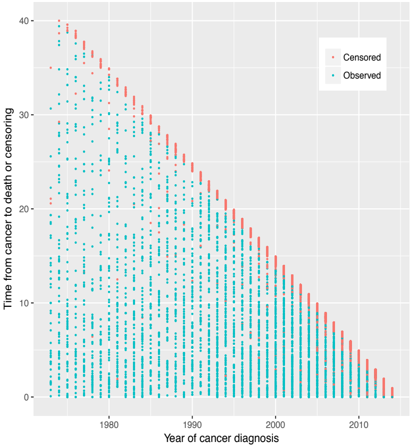
- **https ://seer.cancer.gov**

Application of the two methods to the SEER data

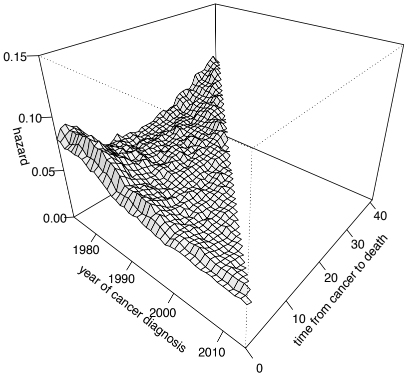
- The breakpoint model chooses 2 breakpoints with the BIC criterion.
- Piecewise constant hazard with no covariates.



Application of the two methods to the SEER data

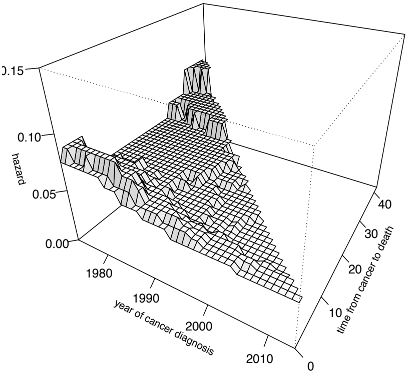


Observed data

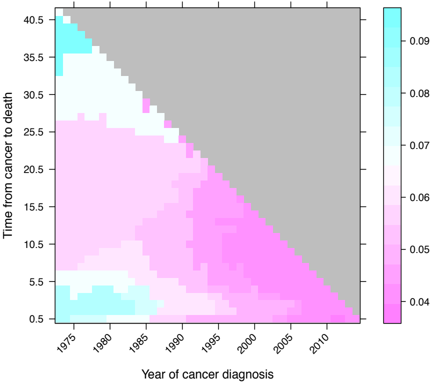


L_2 CV

Application of the two methods to the SEER data



L_0 $EBIC_0$



L_0 $EBIC_0$

Perspectives

- For the breakpoint model :
  - Generalization of the breakpoint model to a soft change of survival distribution between two dates.
  - Development of statistical tests for no breakpoint versus at least one breakpoint.
  - Extension of the method to a “multidimensional proximity space”.
- For the age-period-cohort model :
  - Penalization on second order differences.
  - Inclusion of an interaction term in the age-cohort model :

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

    with L0 regularization on the $\delta_{j,k}$'s.

Bibliography

[1] Olivier Bouaziz and Grégory Nuel. A change-point model for detecting heterogeneity in ordered survival responses. *Statistical methods in medical research*, 2016.

[2] Bendix Carstensen. Age–period–cohort models for the lexis diagram. *Statistics in medicine*, 26(15) :3018–3045, 2007.

[3] Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3) :759–771, 2008.

[4] Florian Frommlet and Grégory Nuel. An adaptive ridge procedure for l0 regularization. *PloS one*, 11(2), 2016.

[5] The Minh Luong, Yves Rozenholc, and Grégory Nuel. Fast estimation of posterior probabilities in change-point analysis through a constrained hidden markov model. *Computational Statistics and Data Analysis*, 68 :129–140, 2013.