Regularized Bidimensional Estimation of the Hazard Rate

Vivien Goepp, Jean-Christophe Thalabard, Grégory Nuel, Olivier Bouaziz

Regularized Bidimensional Estimation of the Hazard Rate

Vivien Goepp[1], Jean-Christophe Thalabard[1], Grégory Nuel[2], and Olivier Bouaziz[1]

[1]*MAP5 (CNRS UMR 8145, 45, rue des Saints-Pères, 75006 Paris)*
[2]*LPSM (CNRS UMR 8001, 4, Place Jussieu, 75005 Paris)*

May 2018

Abstract

In epidemiological or demographic studies, with variable age at onset, a typical quantity of interest is the incidence of a disease (for example the cancer incidence). In these studies, the individuals are usually highly heterogeneous in terms of dates of birth (the cohort) and with respect to the calendar time (the period) and appropriate estimation methods are needed. In this article a new estimation method is presented which extends classical age-period-cohort analysis by allowing interactions between age, period and cohort effects. In order to take into account possible overfitting issues, a penalty is introduced on the likelihood of the model. This penalty can be designed either to smooth the hazard rate or to enforce consecutive values of the hazards to be equal, leading to a parsimonious representation of the hazard rate. The method is evaluated on simulated data and applied on breast cancer survival data from the SEER program.

Keywords Survival Analysis, Penalized Likelihood, Piecewise Constant Hazard, Age-Period-Cohort Analysis, Adaptive Ridge Procedure

Introduction

In epidemiological or demographic studies, with variable age at onset, a typical quantity of interest is the incidence or the hazard rate of a disease (for example the cancer incidence). In these studies, individuals are recruited and followed-up during a long period of time,

usually from birth. The data are then reported either in the form of registers, which contain the number of observed cases and the number of individuals at risk to contract the disease, or in the form of the observed time for each individual. These types of studies are of great interest for the statistician, especially when the event of interest will tend to occur at late ages, such as in cancer studies. However, these data are usually highly heterogeneous in terms of dates of birth and with respect to the calendar time. In such cases, it is therefore very important to take into account the variability of the age, the cohort (date of birth) and the period (the calendar time) in the hazard rate estimation. This is usually done using age-period-cohort estimation methods (see Yang and Land, 2013, and citations therein).

In age-period-cohort analysis, the effects of age, period and cohort are fit as factor variables in a regression model where the output is the logarithm of the hazard rate. However, this induces an identifiability problem due to the relationship: period = age + cohort. There have been several solutions proposed to this problem. Osmond and Gardner (1982) proposed to compute each submodel (age-cohort, age-period, and period-cohort) and use a weighting procedure to combine the three models. Different constraints have also been proposed to make the age-period-cohort model identifiable. However, as noticed by Heuer (1997, p 162), the obtained estimates highly depend on the choice of the constraints. Holford (1983) proposed to directly estimate the linear trends of each effect. This procedure leads to results that are difficult to interpret. See Carstensen (2007) for a detailed discussion of the identifiability problem of the age-period-cohort model. More recently, Kuang et al. (2008) proposed to estimate the second order derivatives of the three effects. This model is implemented in the package `apc` Nielsen (2015). Finally, Carstensen (2007) proposed to first fit one submodel (say age-cohort) and then to fit the period effect over the residuals of the first model. This model is implemented in the `R` package `Epi` (Carstensen et al., 2017), Plummer and Carstensen (2011).

All these approaches can be viewed as parametric models, where the parameters are the age, period, and cohort vector parameters. As such they are also restrictive because they do not allow for interactions between the three effects, that is they assume that one effect does not depend on the other effect's value. A different approach consists in considering the hazard rate as a function of age and either period or cohort and to estimate this bi-dimensional function in a non-parametric setting. No specific structure of the hazard rate is assumed. However, for moderate sample sizes, non-parametric approaches are prone to overparametrization. As a consequence, regularized methods have been proposed in order to avoid overfitting in this non-parametric context. A kernel-type estimator was proposed by Beran (1981) and McKeague and Utikal (1990) where the cumulative hazard is smoothed using a kernel function. See Keiding (1990) for a thorough discussion of methods for hazard inference in age-period-cohort analysis. More recently, Currie and Kirkby (2009) proposed a spline estimation procedure to infer the hazard rate as a function of two variables. The authors use a generalized linear model using B-splines and overfitting is dealt with using a penalization over the differences of adjacent splines' coefficients.

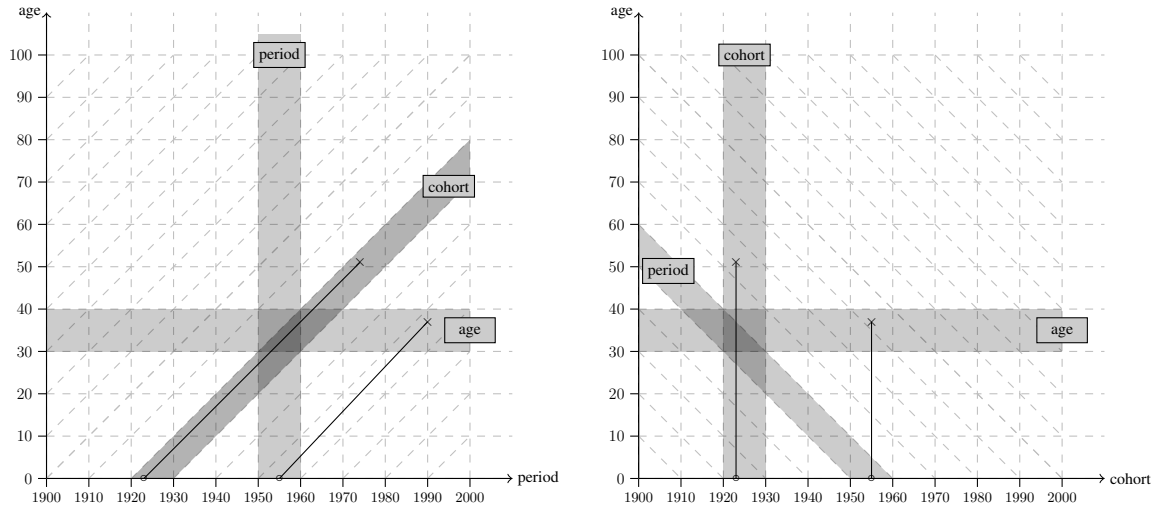
In this article, we propose a new non-parametric method for bi-dimensional hazard rate estimation. As the previous non-parametric approaches, this model considers the estimation of the hazard rate with respect to two variables, i.e. either age-cohort, age-period, or period-cohort, without assuming any specific structure on the hazard rate. Inference is made in two dimensions, but through the linear relationship period = age + cohort, the hazard rate can be represented as a function of any two of the three variables. Finally, in order to take into account the issue of overfitting, we use the L_0 penalization procedure introduced by Rippe et al. (2012), Frommlet and Nuel (2016), and Bouaziz and Nuel (2017). This penalty offers a segmentation of the hazard rate into constant areas. It makes use of an approximation of the L_0 norm which is computationally tractable. The novelty of this method lies in the parsimonious representation of the bi-dimensional hazard rate into segmented areas. In particular, the method can efficiently exhibit cohort, age or period effects, that is, specific changes of the hazard rate due to the date of birth, the age or the calendar time. Our approach also allows L_2 norm penalization, which will induce a smoothed estimate of the hazard in a similar way as the aforementioned non-parametric methods.

Our model is introduced in Section 1. The regularization method is then presented in Section 2. In Section 3, the penalty term selection problem is discussed. Finally, the performance of our model is assessed through a simulation study in Section 4 and illustrations on the SEER cancer dataset is provided in Section 5.

1 Modeling strategy

In the age-period-cohort setting, the date of birth (the cohort) U of each individual is available and the variable of interest is a time-to-event variable of this individual denoted T. The data are subject to right-censoring and they are represented as tabulated data over the J cohort intervals and the K age intervals $[c_0, c_1), [c_1, c_2), \ldots, [c_{J-1}, c_J)$ and $[d_0, d_1), [d_1, d_2), \ldots, [d_{K-1}, d_K)$ respectively, with the convention $c_0 = d_0 = 0$ and $c_K = d_K = \infty$. On a sample of n individuals, the available data can then be rewritten in terms of the exhaustive statistics $\boldsymbol{O} = (O_{1,1}, \ldots, O_{J,K})$, $\boldsymbol{R} = (R_{1,1}, \ldots, R_{J,K})$, where for $j = 1, \ldots, J$, $k = 1, \ldots, K$, $O_{j,k}$ represents the number of observed events that occurred in the j-th cohort interval $[c_{j-1}, c_j)$ and k-th age interval $[d_{k-1}, d_k)$ and $R_{j,k}$ represents the total times individuals were at risk in this j-th cohort and k-th age interval. In the case of register data, the discretization $(c_j), (d_k)$ is imposed by the data and the available data is directly $\boldsymbol{R}$ and $\boldsymbol{O}$, which are often called the *cases* and *person-years*, respectively. See for instance Carstensen (2007) for an example of such data. The aim is to use the available data to provide an estimator of the hazard rate, defined in the age-cohort setting as:

$$\lambda(t|u) = \lim_{dt \to 0} \frac{1}{dt} \mathrm{P}(t < T < t + dt | T > t, U = u),$$



(a) Lexis diagram: Age-Period diagram

(b) Age-Cohort diagram

Figure 1: Diagrams representing the lives of individuals: in the age-period plane (a) – called Lexis diagram – and in the age-cohort plane (b). Solid lines represent lives of individuals until occurrence of the event of interest. The same age, cohort, and period intervals are displayed in light gray. The intersection of two intervals forms a parallelogram and the intersection of three intervals forms a triangle.

in the situation where $\lambda(t,u)$ is assumed to be piecewise constant. That is, we assume that

$$\lambda(t|u) = \sum_{j=1}^{J}\sum_{k=1}^{K} \lambda_{j,k} \mathbb{1}_{[c_{j-1},c_j)\times[d_{k-1},d_k)}(t,u),$$

and inference is made over the $J \times K$ dimension parameter $\boldsymbol{\lambda} = (\lambda_{1,1}, \ldots, \lambda_{J,K})$. Note that the hazard can be equivalently defined as a function of age and period or as a function of period and cohort where the period is defined as the calendar time, that is: period = cohort + age. For illustration, the change of coordinates between the age-period and age-cohort diagrams is represented in Figure 1. In our models, the hazard will be considered as a function of solely age and cohort since the influence of any of the two elements of age, period or cohort can be retrieved using this reparametrization.

Following Aalen et al. (2008, p. 224) the negative log-likelihood takes the form

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

The authors also noticed that this log-likelihood is equivalent to a log-likelihood arising from a Poisson model. However, note that no distribution assumptions are made on the data and in particular the $O_{j,k}$ are not assumed to be Poisson distributed (see Carstensen, 2007, for a discussion on the "Poisson" model). Minimizing ℓ_n yields an explicit maximum likelihood estimator $\widehat{\lambda}^{\text{mle}}_{j,k} = O_{j,k}/R_{j,k}$. However, for moderate sample sizes this estimator is overfitted, especially in places of the age-cohort plane where few events are recorded. To remedy this problem we propose in the following to penalize the differences between adjacent values of the hazard in the log-likelihood.

For computation convenience, we first reparametrize the model: $\eta_{j,k} = \log \lambda_{j,k}$, for $1 \leq j \leq J$ and $1 \leq k \leq K$. The estimate is obtained by minimizing the penalized function

$$\ell_n^{\kappa}(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) = \ell_n(\boldsymbol{\eta}) + \frac{\kappa}{2}\sum_{j=1}^{J-1}\sum_{k=1}^{K} v_{j,k}\left(\eta_{j+1,k} - \eta_{j,k}\right)^2 + \frac{\kappa}{2}\sum_{j=1}^{J}\sum_{k=1}^{K-1} w_{j,k}\left(\eta_{j,k+1} - \eta_{j,k}\right)^2, \tag{2}$$

where $\ell_n(\boldsymbol{\eta})$ was defined in (1), κ is a penalty constant used as a tuning parameter, and $\boldsymbol{v} = (v_{1,1}, \ldots, v_{J-1,K})$, $\boldsymbol{w} = (w_{1,1}, \ldots, w_{J,K-1})$ are constant positive weights of respective dimensions $(J-1)K$ and $J(K-1)$. Note that the case $\kappa = 0$ corresponds to the maximum likelihood estimation and the case $\kappa = \infty$ corresponds to a hazard uniformly constant over the age and cohort intervals. The parameter κ needs to be chosen in an appropriate way in order to obtain a compromise between these two extreme situations.

This model does not attempt to estimate the age, period and cohort effect as parameter vectors. Instead, it performs a regularized estimation of $\boldsymbol{\lambda}$ that has no age-period-cohort-type structure. Two choices for the weights $\boldsymbol{v}$ and $\boldsymbol{w}$ can be made: one will lead to a

smooth hazard rate and the other to a segmented hazard rate. This will be discussed in the next section. The choice of the optimal value for κ is addressed in Section 3.

Minimization of ℓ_n^κ is performed using the Newton-Raphson algorithm (see Algorithm 1). Let $U_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) = \partial \ell_n^\kappa / \partial \boldsymbol{\eta}$ be the gradient vector of the negative log-likelihood and $I_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) = \partial U_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) / \partial \boldsymbol{\eta}^T$ be its Hessian matrix.

For $1 \leq j, j' \leq J$ and $1 \leq k, k' \leq K$, we have

$$\frac{\partial \ell_n}{\partial \eta_{j,k}}(\boldsymbol{\eta}) = \exp(\eta_{j,k}) R_{j,k} - O_{j,k}, \qquad \frac{\partial^2 \ell_n(\boldsymbol{\eta})}{\partial \eta_{j',k'} \partial \eta_{j,k}} = \mathbb{1}_{j=j',k=k'} \exp(\eta_{j,k}) R_{j,k}, \quad \text{and}$$

$$\begin{aligned}\frac{\partial \ell_n^\kappa}{\partial \eta_{j,k}}(\boldsymbol{\eta}) = \frac{\partial \ell_n(\boldsymbol{\eta})}{\partial \eta_{j,k}} &+ \kappa \left[-v_{j,k} (\eta_{j+1,k} - \eta_{j,k}) + v_{j-1,k} (\eta_{j,k} - \eta_{j-1,k}) \right] \\ &+ \kappa \left[-w_{j,k} (\eta_{j,k+1} - \eta_{j,k}) + w_{j,k-1} (\eta_{j,k} - \eta_{j,k-1}) \right],\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ell_n^\kappa(\boldsymbol{\eta})}{\partial \eta_{j',k'} \partial \eta_{j,k}} = \frac{\partial^2 \ell_n}{\partial \eta_{j',k'} \partial \eta_{j,k}}(\boldsymbol{\eta}) &+ \kappa \left[\mathbb{1}_{j=j',k=k'} (v_{j',k'} + v_{j'-1,k'} + w_{j',k'} + w_{j',k'-1}) \right. \\ &- v_{j',k'} \mathbb{1}_{j=j'+1,k=k'} - v_{j'-1,k'} \mathbb{1}_{j=j'-1,k=k'} \\ &\left. - w_{j',k'} \mathbb{1}_{j=j',k=k'+1} - w_{j',k'-1} \mathbb{1}_{j=j',k=k'-1} \right].\end{aligned}$$

As a consequence, the Hessian matrix can be written

$$I_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w}) = \frac{\partial^2 \ell_n(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} + \kappa B(\boldsymbol{\eta}),$$

where $B(\boldsymbol{\eta})$ is a band matrix of bandwidth equal to $\min(J, K) - 1$. Thus the Hessian matrix has the same structure as $B(\boldsymbol{\eta})$ and the calculation of $I_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w})^{-1} U_n^\kappa(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w})$ has a $\mathcal{O}(\min(J, K) JK)$ complexity instead of $\mathcal{O}(J^3 K^3)$. Fast inversion of the Hessian matrix is done using Cholesky decomposition as implemented in `Rcpp` in the package `bandsolve` [1].

2 Choice of the regularization parameters v and w

In this section, two different expressions of the weights $\boldsymbol{v}$ and $\boldsymbol{w}$ are proposed which correspond to two different types of regularization of the hazard rate. The first one yields a smooth estimate. The second one uses an iterated adaptation of the weights to approximate an L_0 norm penalization of the first order differences.

[1] http://github.com/Monneret/bandsolve

Algorithm 1 Newton-Raphson Procedure with constant weights

function NEWTON-RAPHSON($\boldsymbol{O}, \boldsymbol{R}, \kappa, \boldsymbol{v}, \boldsymbol{w}$)
  $\boldsymbol{\eta} \leftarrow \mathbf{0}$
  while not converge **do**
    $\boldsymbol{\eta}^{\text{new}} \leftarrow \boldsymbol{\eta} - I_n^{\kappa}(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w})^{-1} U_n^{\kappa}(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{w})$
    $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta}^{\text{new}}$
  end while
  return $\boldsymbol{\eta}$
end function

2.1 L_2 Norm Regularization

A ridge-type penalization is performed when setting $\boldsymbol{v} = \boldsymbol{w} = \mathbf{1}$. In this case the penalization corresponds to the square of the first-order differences of $\boldsymbol{\delta}$. In the penalized estimation model, this choice of weights yields a globally smooth estimator of the hazard rate. Note that our penalized maximum likelihood model will yield similar results as the spline method of Ogata and Katsura (1988) presented in Section 1. In our method the penalization is performed over the first order differences of the parameter while in the spline method it is performed over the second order differences. This means that for arbitrarily large values of the penalty constant, the regularized hazard will be a constant function instead of a linear function. This model will be referred to as L_2 regularized estimation or smooth estimation.

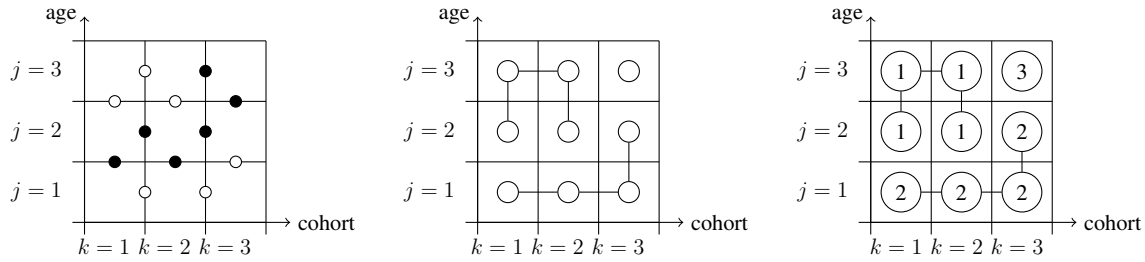
Finally, one notes that Equation 2 allows for some flexibility in the regularization. Indeed, manually setting the weights $\boldsymbol{v}$ and $\boldsymbol{w}$ will allow to tune the importance of the regularization between different regions of the plane and between the two variables.

2.2 Approximate L_0 Norm Regularization

Following the work from Rippe et al. (2012), Frommlet and Nuel (2016), and Bouaziz and Nuel (2017) an adaptive ridge procedure is performed when the weights are updated at each iteration of the Newton-Raphson algorithm. At the m-th iteration of the Newton-Raphson algorithm the weights are computed from the following formulas:

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

where ε_v and ε_w are constants negligible compared to 1 (in practice one typically chooses $\varepsilon_v = \varepsilon_w = 10^{-5}$). We iterate between minimizing ℓ_n^{κ} for fixed weights and reevaluat-



(a) Weighted differences (b) Corresponding graph (c) Connected components

Figure 2: Representation of the method used to select the constant areas for the adaptive ridge procedure. In this example, $J = K = 3$. In Panel (a), the circles represent the values of the differences $v_{j,k} \left(\eta_{j+1,k} - \eta_{j,k}\right)^2$ and $w_{j,k} \left(\eta_{j,k+1} - \eta_{j,k}\right)^2$: empty circles correspond to the value 0 and filled circles correspond to the value 1. Panel (b) represents the graph that is generated from these values. Adjacent nodes whose difference is null are connected by a vertice. Panel (c) represents the last step, where the connected components of the graph are extracted. Each connected component corresponds to one constant area. The numbering is arbitrary.

ing the weights such that at the m-th step, $v_{j,k}^{(m)}(\eta_{j+1,k}^{(m)} - \eta_{j,k}^{(m)})^2 \simeq \|\eta_{j+1,k}^{(m)} - \eta_{j,k}^{(m)}\|_0$ and $w_{j,k}^{(m)}(\eta_{j,k+1}^{(m)} - \eta_{j,k}^{(m)})^2 \simeq \|\eta_{j,k+1}^{(m)} - \eta_{j,k}^{(m)}\|_0$, where $\|.\|_0$ denotes the L_0 norm – i.e. $\|u\|_0 = 0$ if $u = 0$ and $\|u\|_0 = 1$ otherwise. In other words, this adaptive ridge procedure approximates the L_0 norm regularization over the differences of $\eta_{j,k}$ and yields a segmentation of $\eta_{j,k}$ into piecewise constant areas. As with other classical penalized methods (e.g. LASSO, ridge) and as pointed out in Frommlet and Nuel (2016), the adaptive ridge penalization scheme induces a shrinkage bias. Therefore, after segmentation of the $\eta_{i,j}$s, the hazard rate is estimated on each constant area using the unpenalized maximum likelihood estimator. More precisely, at convergence of the adaptive ridge algorithm, $v_{j,k} \left(\eta_{j+1,k} - \eta_{j,k}\right)^2$ will be approximately equal to 0 if $|\eta_{j+1,k} - \eta_{j,k}|$ is smaller than ε_v and approximately equal to 1 if $|\eta_{j+1,k} - \eta_{j,k}|$ is greater than ε_v – and similarly for $w_{j,k} \left(\eta_{j,k+1} - \eta_{j,k}\right)^2$. Then one creates the graph whose vertices are the JK discretization cells and whose edges are the connexions between adjacent cells that have differences close to 0. Each connected component of this graph is a different area over which the hazard has been estimated to be constant. The extraction of connected components from the graph is done using the package `igraph` (Csardi and Nepusz, 2006). The log-hazard $\eta^{(r)}$ of the r-th constant area is such that $\forall [c_{j-1}, c_j) \times [d_{k-1}, d_k) \in r, \eta_{j,k} = \eta^{(r)}$. Finally, the values of $\eta^{(r)}$ are not estimated using the results of the adaptive ridge algorithm, but by unpenalized maximum likelihood estimation: $\widehat{\eta}^{(r)} = \log\left(O^{(r)}/R^{(r)}\right)$ where $O^{(r)}$ is the number of events in the r-th constant area and $R^{(r)}$ is the time at risk in the r-th constant area.

This estimation method will be called L_0 regularized estimation or segmented estima-

tion. This method is illustrated through the toy-example of Figure 2 and the adaptive ridge procedure is summarized in Algorithm 2. In practice, the stopping criterion for the adaptive ridge algorithm is when the absolute difference between successive values of the weighted differences is smaller than a predefined value – we use 10^{-8} in our implementation.

Algorithm 2 Adaptive Ridge Procedure

function ADAPTIVE-RIDGE($\boldsymbol{O}, \boldsymbol{R}, \kappa$)
  $\boldsymbol{\eta} \leftarrow \mathbf{0}$
  $\boldsymbol{v} \leftarrow \mathbf{1}$
  $\boldsymbol{w} \leftarrow \mathbf{1}$
  while not converge **do**
    $\boldsymbol{\eta}^{\text{new}} \leftarrow$ NEWTON-RAPHSON($\boldsymbol{O}, \boldsymbol{R}, \kappa, \boldsymbol{v}, \boldsymbol{w}$)
    $v_{j,k}^{\text{new}} \leftarrow \left(\left(\eta_{j+1,k}^{\text{new}} - \eta_{j,k}^{\text{new}} \right)^2 + \varepsilon_v^2 \right)^{-1}$
    $w_{j,k}^{\text{new}} \leftarrow \left(\left(\eta_{j,k}^{\text{new}} - \eta_{j,k-1}^{\text{new}} \right)^2 + \varepsilon_w^2 \right)^{-1}$
    $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta}^{\text{new}}$
  end while
  Compute $(O^{\text{new}}, R^{\text{new}})$ for selected $(\boldsymbol{\eta}, \boldsymbol{v}^{\text{new}}, \boldsymbol{w}^{\text{new}})$
  $\boldsymbol{\eta}^{\text{new}} \leftarrow \log\left(\boldsymbol{O}^{\text{new}} / \boldsymbol{R}^{\text{new}}\right)$
  return $\boldsymbol{\eta}^{\text{new}}$
end function

3 Choice of the penalty constant κ

In practice, the hazard rate needs to be estimated for a set of penalty constants and the choice of κ is determined as the penalty that provides the best compromise between model fit and reduced variability of the hazard rate estimate. For the L_0 regularization model, different values of the penalty constant lead to different segmentations of the $\eta_{j,k}$. As a consequence, the problem of choosing the optimal penalty constant can be rephrased as the problem of choosing the optimal model among a set of models $\mathcal{M}_1, \ldots, \mathcal{M}_M$, where each of these models corresponds to a different segmentation of the $\eta_{j,k}$ and M is the maximum number of different models. In this section we propose different methods to select the optimal model. Comparison of the efficiency of the different methods will be analyzed in Section 4 on simulated data.

We recall that $\boldsymbol{R}$ and $\boldsymbol{O}$ are the exhaustive statistics and $\boldsymbol{\eta}$ is the parameter to be estimated in our two models. Bayesian criteria attempt to maximize the posterior probability $\mathrm{P}(\mathcal{M}_m|\boldsymbol{R}, \boldsymbol{O}) \propto \mathrm{P}(\boldsymbol{R}, \boldsymbol{O}|\mathcal{M}_m)\pi(\mathcal{M}_m)$, where $\mathrm{P}(\boldsymbol{R}, \boldsymbol{O}|\mathcal{M}_m)$ is the integrated likelihood

and $\pi(\mathcal{M}_m)$ is the prior distribution on the model. This problem is equivalent to minimizing $-2\log \mathrm{P}(\mathcal{M}_m|\boldsymbol{R},\boldsymbol{O})$. By integration

$$\mathrm{P}(\boldsymbol{R},\boldsymbol{O}|\mathcal{M}_m) = \int_{\boldsymbol{\eta}} \mathrm{P}(\boldsymbol{R},\boldsymbol{O}|\mathcal{M}_m,\boldsymbol{\eta})\pi(\boldsymbol{\eta})d\boldsymbol{\eta},$$

where $\mathrm{P}(\boldsymbol{R},\boldsymbol{O}|\mathcal{M}_m,\boldsymbol{\eta})$ is the likelihood and $\pi(\boldsymbol{\eta})$ is the prior distribution of the parameter, which is taken constant in the following. Thus Bayesian criteria are defined as

$$-2\log\left(\mathrm{P}(\mathcal{M}_m|\boldsymbol{R},\boldsymbol{O}\right) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m\log n - 2\log\pi(\mathcal{M}_m) + \mathcal{O}_{\mathrm{P}}(1),$$

where q_m is the dimension of the model $\mathcal{M}_m$ i.e., the number of constant areas selected by the adaptive ridge algorithm.

The BIC (Schwarz, 1978) corresponds to the Bayesian criterion obtained when one neglects the term $\pi(\mathcal{M}_m)$, which is equivalent to having a uniform prior on the model:

$$\mathrm{BIC}(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m\log n. \tag{3}$$

As explained by Żak-Szatkowska and Bogdan (2011), a uniform prior on the model is equivalent to a binomial prior on the model dimension $\mathcal{B}(JK, 1/2)$. When the true model's dimension is much smaller than the maximum possible dimension JK, the BIC tends to give too much importance to models of dimensions around $JK/2$, which will result in underpenalized estimators. To this effect, Chen and Chen (2008) have developed an extended Bayesian information criterion called EBIC_0 (or EBIC for short). One can write $\pi(\mathcal{M}_m) = \mathrm{P}(\mathcal{M}_m|\mathcal{M}_m \in \mathcal{M}_{[q_m]})\mathrm{P}(\mathcal{M}_m \in \mathcal{M}_{[q_m]})$ where $\mathcal{M}_{[q_m]}$ is the set of models of dimension q_m. The EBIC_0 criterion is defined by setting $\mathrm{P}(\mathcal{M}_m|\mathcal{M}_m \in \mathcal{M}_{[q_m]}) = 1/\binom{JK}{q_m}$ and $\mathrm{P}(\mathcal{M}_m \in \mathcal{M}_{[q_m]}) = 1$. Thus

$$\pi(M_m) = \binom{JK}{q_m}$$

and

$$\mathrm{EBIC}_0\left(m\right) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m\log n + 2\log\binom{JK}{q_m}. \tag{4}$$

Note that the EBIC_0 assigns the same *a priori* probability to all models of same dimension. Therefore, when the true model's dimension is not close to $JK/2$ the EBIC_0 will be able to select this model more easily. Namely, when the true model's dimension is very small the EBIC_0 will tend to choose very sparse models.

The last criterion that will be used is the Akaike Information Criterion (Akaike, 1998), or AIC, defined as $\mathrm{AIC}(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + 2q_m$. This criterion is known for performing better than the BIC in terms of mean squared error, however the BIC will tend to select sparser models than the AIC.

Note that Bayesian criteria and the AIC can only be used for the L_0 regularized estimation only, since the L_2 model does not perform a model selection. An alternative to performing model selection is to use the K-fold cross validation. With this method, the data are split at random into L parts. The estimated parameter obtained when the l-th part is left out is noted $\widehat{\eta}^{-l}(\kappa)$ and the cross-validated score is defined as

$$\mathrm{CV}(\kappa) = \sum_{l=1}^{L} \ell_n^{\kappa,l}(\widehat{\boldsymbol{\eta}}^{-l}),$$

where $\ell_n^{\kappa,l}$ is the negative log-likelihood evaluated on the l-th part of the data. The optimal penalty constant is obtained by minimizing $\mathrm{CV}(\kappa)$ with respect to κ. The L-fold cross validation method can be used for both the L_0 regularized estimation and the L_2 regularized estimation. However, this method is numerically time consuming as the estimator has to be computed L times while Bayesian criteria or the AIC provide direct methods to perform model selection from the original estimator. In the simulation studies and data analysis, we set $L = 10$.

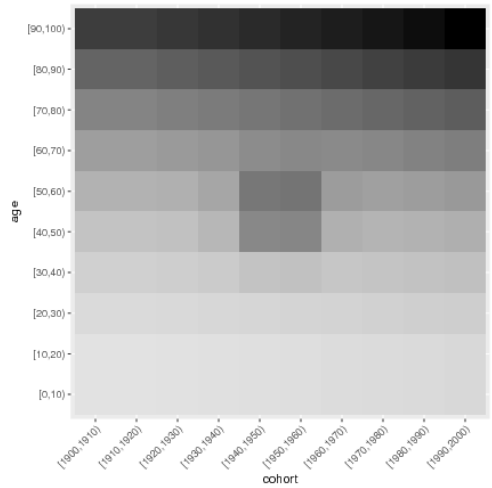
4 Simulation study

4.1 Simulation designs

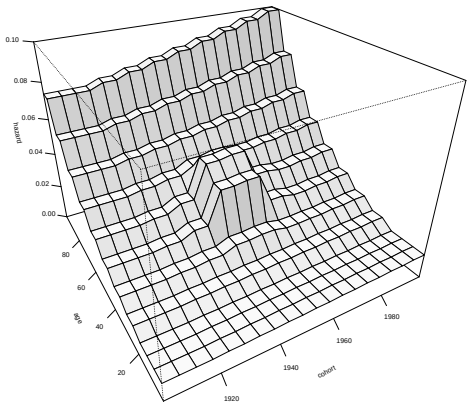
In this section, our segmented estimation method with L_0 norm is compared with the AGE-COHORT model and with the smoothed hazard estimate with the L_2 norm. The different criteria for model selection are also compared with each other. We present two simulation designs. In the first one, the true hazard rate is generated from a smooth age-cohort model which includes an interaction term on a small region of the age-cohort plane. In the second case, the true hazard rate is a piecewise constant function with four heterogeneous areas. The two true hazards are displayed in Figure 3, both in greyscale and in perspective plot.

The simulation design is as follows. We set $J = 10$ equally spaced age intervals and $K = 10$ equally spaced cohort intervals. The age intervals are defined as $[0, 10)$, ..., $[90, 100]$ and the cohorts intervals are defined as $[1900, 1910)$,..., $[1990, 2000]$. In order to simulate a dataset, the cohorts are first sampled on $K = 10$ cohort group intervals of 10 years length ranging from 1900 to 2000. Censoring is then simulated as a uniform distribution over the age interval $[75, 100]$ for all cohorts such that all observed events are comprised in the age interval $[0, 100]$. Since in practice one does not know the appropriate discretization in advance, a different discretization was used for the estimation procedure : the age and cohort intervals were defined as 5-year length intervals instead of 10 for the true hazard. As a result, a total of 20×20 parameters need to be estimated.

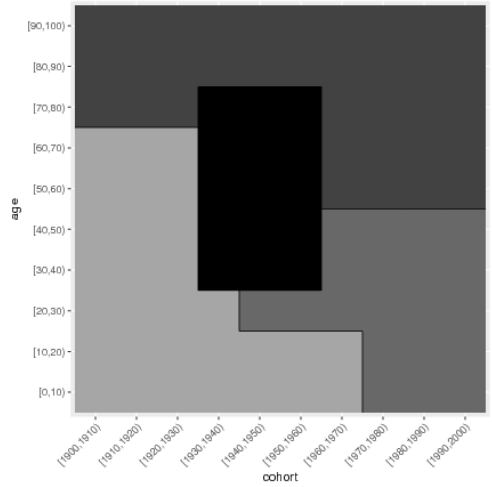
For each of the two designs, we simulated data of sample sizes 100, 400, 1000, 4000, and 10000. For each sample size, the simulation and estimation were replicated 500 times.



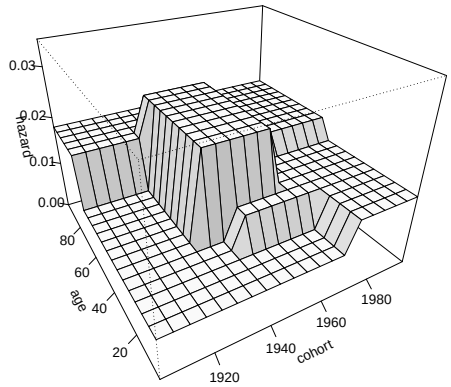
(a) Smooth true hazard – heatmap



(b) Smooth true hazard – perspective



(c) Piecewise constant true hazard – heatmap



(d) Piecewise constant true hazard – perspective

Figure 3: True hazard of the two simulation designs: smooth hazard in heatmap (a) and perspective plot (b) and piecewise constant hazard in heatmap (c) and perspective plot (d).

	L_0 method				L_2 method
Sample size	**AIC**	**BIC**	**EBIC**	**CV**	**CV**
100	412.90	401.50	4.60	4.50	1.00
400	269.60	225.50	37.90	7.10	1.00
1000	168.00	111.50	4.30	3.60	1.00
4000	79.40	24.90	5.00	3.40	1.00
10000	26.10	5.90	4.60	2.40	1.00

(a) Smooth true hazard

	L_0 method				L_2 method
Sample size	**AIC**	**BIC**	**EBIC**	**CV**	**CV**
100	429.90	428.90	1.30	1.30	1.00
400	81.70	64.30	3.00	2.40	1.00
1000	34.20	16.80	3.80	3.50	1.00
4000	12.20	2.20	1.50	1.90	1.00
10000	6.70	0.80	0.60	0.80	1.00

(b) Piecewise constant true hazard

Table 1: Relative mean squared errors with respect to the cross-validated L_2 estimator, for different sample sizes and different estimation methods. Panel (a): smooth true hazard. Panel (b): piecewise constant true hazard.

Smooth true hazard The smooth true hazard (Figures 3a and 3b) is generated using the age-cohort model $\log \lambda_{j,k} = \mu + \alpha_j + \beta_k$ with an intercept $\mu = \log(10^{-2})$. The age effect vector $\boldsymbol{\alpha}$ and cohort effect vector $\boldsymbol{\beta}$ are arithmetic sequences such that $\alpha_2 = 0$, $\alpha_J = 2.5$, $\beta_2 = 0$, and $\beta_K = 0.3$. An interaction term is added to the hazard. It corresponds to a bump in the hazard located in the neighbourhood of the region of the age-cohort plane (45,1945). The bump is defined as 10 times the Gaussian density function with mean $(1945, 45)$ and with a diagonal variance-covariance matrix with diagonal equal to $(50, 50)$. This true hazard displays a sharp increase for high values of the age, which implies that few events will be recorded in this region. On average, 91 % of the events are observed in this simulation design.

Piecewise constant true hazard The piecewise constant true hazard (Figures 3c and 3d) has four constant areas over the age-cohort square $[0, 100] \times [1900, 2000]$. On average, 71 % of the events are observed in this simulation design.

4.2 Performance of the estimation methods in terms of MSE

Our two estimation methods (L_0 and L_2 norm) are compared in terms of the Mean Squared Errors in each simulation scenario. The different selection methods for the penalty (AIC, BIC, EBIC and cross-validation) are also compared. The results are presented in Table 1. On the overall, the EBIC and cross-validated criteria outperform the AIC and the BIC for the two simulations scenarios. This is particularly true for small sizes where the AIC and the BIC behave very poorly. As expected, the L_2 norm estimator is the most performant of all estimators in the smooth true hazard scenario (Table 1a) and the L_0 method performs better in the piecewise constant hazard scenario (Table 1b) than in the smooth true hazard scenario. The L_2 norm estimator is also the most performant of all estimators in the piecewise constant hazard scenario except for very large sample sizes ($n = 10000$) where the BIC, EBIC and cross-validated criterion provide slightly better performances. Finally, in both scenarios, the EBIC always outperforms the AIC, the BIC and the cross-validated criterion. Different censoring rates were also studied which showed a degradation of the performances of the overall estimators as the percentage of censoring increased. The performance in terms of number of selected areas was also investigated. It showed that the EBIC and CV criterion perform better at selecting sparse models with few areas, while the AIC and BIC tend to overestimate the true number of areas. Indeed, for sample size 4000, the 80% inter-quantile range of the selected number of areas is $[3, 5]$ for the EBIC and $[1, 5]$ for the CV, whereas it is $[3, 13]$ and $[36, 72]$ for the BIC and AIC respectively. These experiments are not reported here.

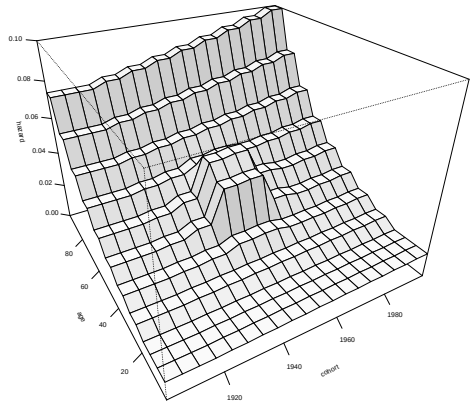
In conclusion, the simulation experiments suggest to use the EBIC among all different criteria for the L_0 norm estimator as it provides the best tradeoff between computation time and estimation performance.

4.3 Perspective plots of the estimation methods

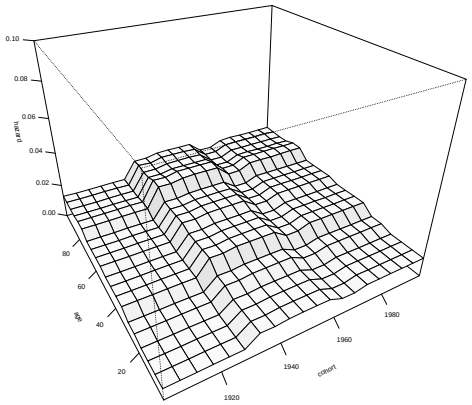
In this section the performance of our adaptive ridge (L_0 norm) and ridge (L_2 norm) estimates is assessed visually by comparison of the true hazard. The standard age-cohort model (Holford, 1983) has also been implemented. This model assumes that the hazard has the following expression:

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

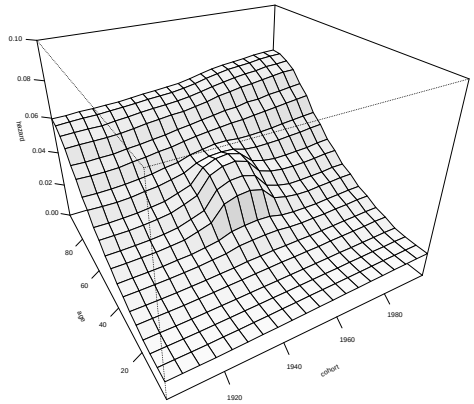
where μ is the intercept, $\boldsymbol{\alpha}$ is the age effect and $\boldsymbol{\beta}$ is the cohort effect. It should be noted that this model does not allow for interactions between age and cohort effects. Perspective plots of the median hazard estimations over 500 replications are presented in Figures 4 and 5 for the smooth and piecewise constant true hazard respectively. For the L_0 regularized estimate, the penalty constant is chosen using the EBIC.



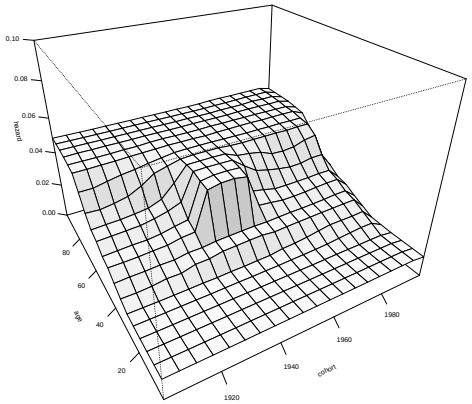
(a) True hazard



(b) Median of age-cohort estimates

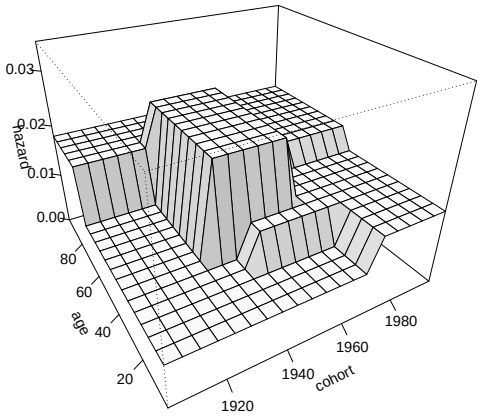


(c) Median of smooth estimates

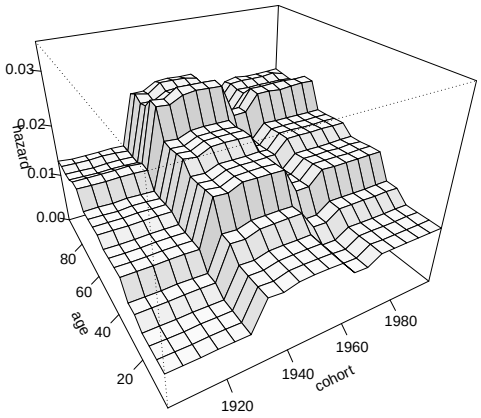


(d) Median of segmented estimates

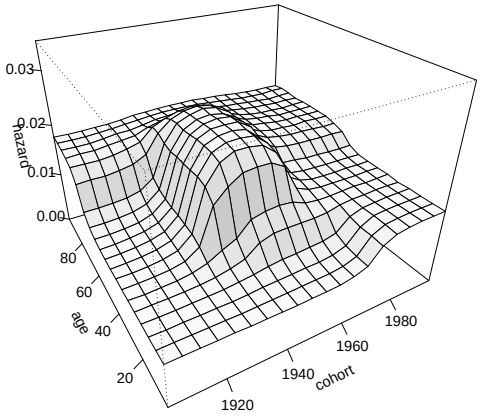
Figure 4: Smooth true hazard and corresponding estimates. The sample size is 4000 and the hazard estimates are medians taken over 500 simulations. The estimations are performed in the age-cohort plane and with different methods. Panel (a) represents the true hazard used to generate the data, Panel (b) represents the hazard estimated using the age-cohort model, Panel (c) represents the smoothed estimate, and Panel (d) represents the segmented estimate with the EBIC criterion.



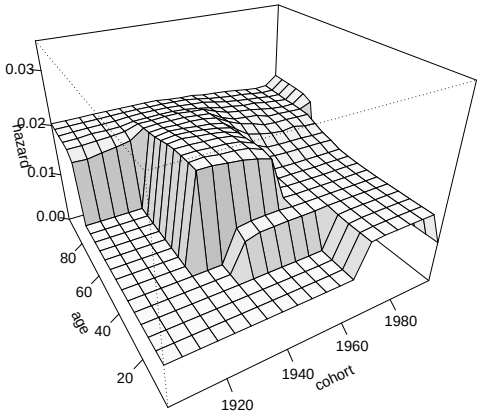
(a) True hazard



(b) Median of age-cohort estimates



(c) Median of smooth estimates



(d) Median of segmented estimates

Figure 5: Piecewise constant true hazard and corresponding estimates. The sample size is 4000 and the hazard estimates are medians taken over 500 simulations. The estimations are performed in the age-cohort plane and with different methods. Panel (a) represents the true hazard used to generate the data, Panel (b) represents the hazard estimated using the age-cohort model, Panel (c) represents the smoothed estimate, and Panel (d) represents the segmented estimate with the EBIC criterion.

In Figure 4, it is seen that the age-cohort model is not able to estimate the central bump in the hazard. On the contrary, the smoothed estimate accurately recovers the shape of the true hazard except for the high values of age where few events are observed. Interestingly, it is seen that our segmentation method provides similar results as the smoothing technique even though the true hazard is not piecewise constant.

The results in Figure 5 yield similar conclusions. The age-cohort model behaves very poorly due to its constrained structure while the ridge and adaptive estimates provide satisfactory results. In particular the shape of the true hazard is correctly captured by the adaptive ridge on the majority of replicated samples.

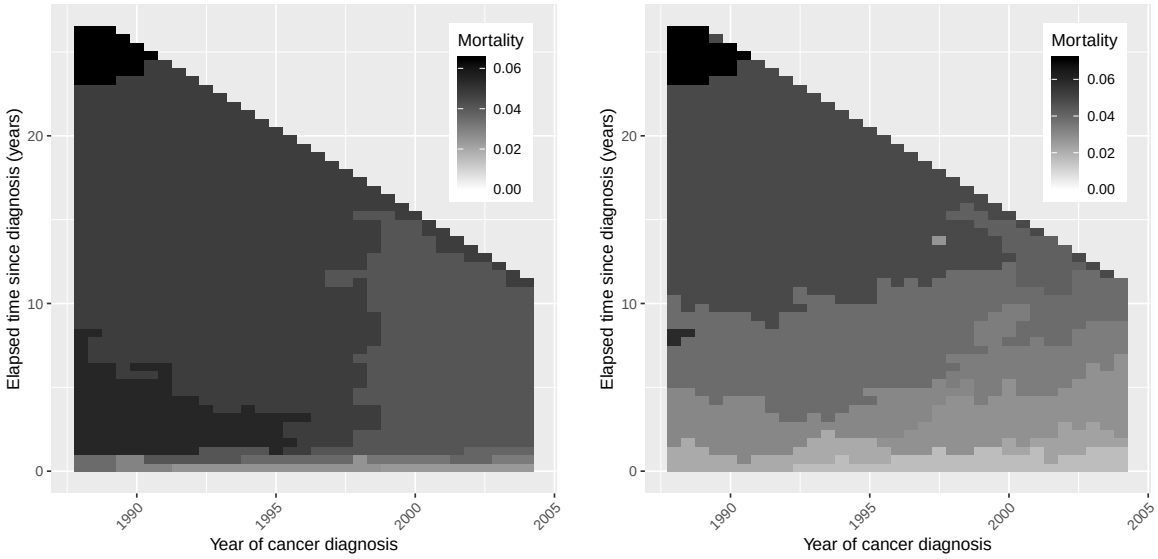
5 Real data application

Our method is applied to data of survival times after diagnosis of breast cancer. The dataset is provided by the Surveillance, Epidemiology, and End Results (SEER) Program from the US National Cancer Institute (NCI). SEER collects medical data of cancers (including stage of cancer at diagnosis and the type of tumor) and follow-up data of patients in the form of a registry. Around 28 percent of the US population is covered by the program. The registry started in February 1973 and the available current dataset includes follow-up data until January 2015. We refer to the website `https://seer.cancer.gov/` for information about the SEER Program and its publicly available cancer data.

In this study the duration of interest T is the time from breast cancer diagnosis to death in years, the variable U is the date of diagnosis (in years) and the period is the calendar time (in years). Patients continuously entered the study between 1973 and 2015 and right-censoring occurred for patients that were still alive at the end of follow-up or for those that were lost to follow-up.

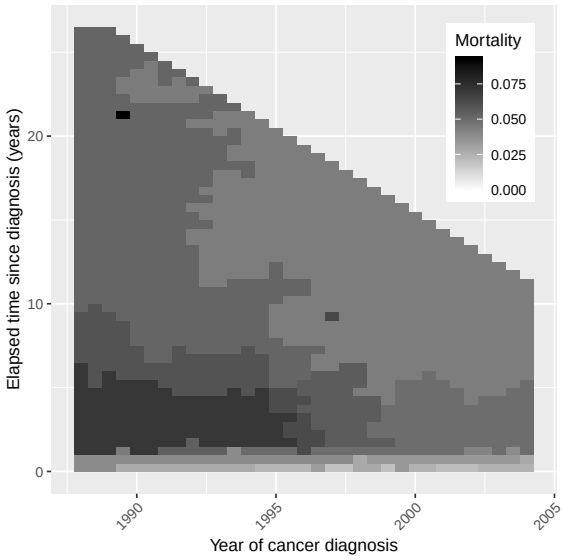
The breast cancer data was extracted using the package `SEERaBomb`. For the sake of comparison, the subsample of malignant, non-bilateral breast tumor cancers was extracted from the dataset, such that the data comprises $1,265,277$ women with 60 percent of censored individuals. Times from diagnosis to last day of follow-up vary between 0 and 41 years, and the dates of cancer diagnosis U_i vary between 1973 and 2015. Death from another cause than cancer is available in the dataset and is accounted for as right-censoring.

The implementation of our adaptive ridge method aims at two goals. Firstly we aim at simultaneously detecting a cohort effect and an age effect, that is the evolution of the mortality with respect to the time elapsed since cancer diagnosis (age effect) and with respect to the date of diagnosis (cohort effect). Secondly, our method will provide estimation of the hazard rates on the resulting heterogeneous areas. The method is first applied on the whole sample of 1265277 individuals. In order to take into account the fact that mortality from cancer highly depends on the cancer stage, we also perform a stratified analysis with respect to the stage of cancer at diagnosis. For this purpose, we use the cancer stage classification

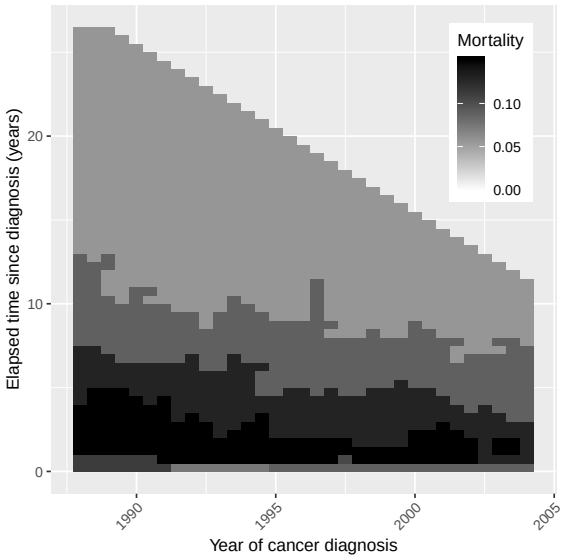


(a) All stages of cancer

(b) Stage 1



(c) Stage 2



(d) Stage 3

Figure 6: Estimated hazard of death after diagnosis of breast cancer for different stages of cancer. The estimate is obtained with the L_0 regularization. The upper right corner of every graph corresponds to the region where no data are available. Note that the scales are different between panels.

provided by the SEER data: we keep the patients with cancer stages 1, 2, and 3 at the time of diagnosis. This classification closely follows that of the American Joint Comitee on Cancer (AJCC), 3^{rd} Edition; the details are given at page 86 of the manual entitled Comparative Staging Guide for Cancer, available at `https://seer.cancer.gov`. The main difference between the two classifications is that the SEER Program classifies the cases where lymph node status cannot be assessed as if there was no regional lymph node metastasis.

The L_0 estimates for the whole sample and for each cancer stage are displayed in Figure 6. We see that the different stages of cancer at diagnosis have a great impact on the survival times. For Stage 1 cancers, the mortality is low between 0 and 4–5 years after diagnosis, and steadily increases afterwards. The date of diagnosis seems to have no impact on the mortality of Stage 1 cancers. On the other hand, Stage 2 cancers exhibit a strong effect of the date of diagnosis: around $1995 - 1997$, the mortality significantly decreases. This can correspond to an improvement of the treatment of breast cancer around that period in the United States. Finally, Stage 3 cancers display a very high hazard rate across all dates of diagnosis. This seems to indicate that the evolution in treatments of breast cancer had a significant impact on the survival times after diagnosis, but almost exclusively when cancers were diagnosed at Stage 2. Two additional analyses of the hazard rate with stratification with respect to age at diagnosis and estrogen receptor status were performed in the Supplementary Materials. The results suggest that the shift in mortality around year 1996 could correspond to the introduction of hormone-blocking therapy.

Conclusion

In this article, we have introduced a new estimation method to deal with age-period-cohort analysis. This model assumes no specific structure of the effects of age and cohort and the hazard rate is directly estimated without estimating the effects. In order to take into account possible overfitting issues, a penalty is used on the likelihood to enforce similar consecutive values of the hazard to be equal. Two different types of penalty terms were introduced. One leads to a ridge type regularization while the other leads to a L_0 regularization. Different selection methods of the penalty parameter were also introduced. To our knowledge, a segmented estimation model of this kind has never been introduced in this context.

Using simulated data, it has been shown that the cross validated ridge estimator and the $EBIC_0$ adaptive ridge estimator perform the best in terms of mean squared error. The cross validation criterion was shown to provide the best fit of the hazard rate, but its very high computationally cost makes it non-competitive. In this context, this modified BIC criterion comes out as a powerful tool to select the *best* bias-variance tradeoff.

The method was successfully applied to data of survival after breast cancer provided by the SEER program. The segmented estimate of the hazard rate displays important information about the shift in mortality after being diagnosed of breast cancer in the United States

in the mid-1990s.

Our method could be directly extended to a different discretization of the age-period-cohort plane, such as $1 \times 1 \times 1$-year triangles that are represented in dark gray in Figure 1 (see Section 3 of Carstensen, 2007, for an example of this discretization). Another extension would be to consider other types of penalizations. Instead of estimating a piecewise constant hazard, one could estimate a piecewise linear hazard by penalizing over second order differences of the hazard.

Acknowledgement The authors are thankful to the National Cancer Institute for providing U.S. mortality data on cancer.

Conflict of Interest The authors have declared no conflict of interest.

References

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

H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

R. Beran. Nonparametric Regression with Randomly Censored Survival Data. Technical report, Technical Report, University of California, Berkeley, 1981.

O. Bouaziz and G. Nuel. L0 Regularization for the Estimation of Piecewise Constant Hazard Rates in Survival Analysis. *Applied Mathematics*, 08(03):377–394, 2017.

B. Carstensen. Age–Period–Cohort Models for the Lexis Diagram. *Statistics in Medicine*, 26(15):3018–3045, 2007.

B. Carstensen, M. Plummer, E. Laara, and M. Hills. *Epi: A Package for Statistical Analysis in Epidemiology*. 2017.

J. Chen and Z. Chen. Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika*, 95(3):759–771, 2008.

G. Csardi and T. Nepusz. The igraph Software Package for Complex Network Research, 2006.

I. D. Currie and J. G. Kirkby. Smoothing Age-Period-Cohort Models with P -splines: A Mixed Model Approach. 2009.

F. Frommlet and G. Nuel. An Adaptive Ridge Procedure for L0 Regularization. *PLoS ONE*, 11(2):e0148620, 2016.

C. Heuer. Modeling of Time Trends and Interactions in Vital Rates Using Restricted Regression Splines. *Biometrics*, 53(1):161–177, 1997.

T. R. Holford. The Estimation of Age, Period and Cohort Effects for Vital Rates. *Biometrics*, 39(2):311–324, 1983.

N. Keiding. Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 332 (1627):487–509, 1990.

D. Kuang, B. Nielsen, and J. P. Nielsen. Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, 95(4):979–986, 2008.

I. W. McKeague and K. J. Utikal. Identifying Nonlinear Covariate Effects in Semimartingale Regression Models. *Probability Theory and Related Fields*, 87(1):1–25, 1990.

B. Nielsen. Apc: An R Package for Age-Period-Cohort Analysis. *The R Journal*, 7(2), 2015.

Y. Ogata and K. Katsura. Likelihood Analysis of Spatial in Homogeneity for Marked Point Patterns. *Annals of the Institute of Statistical Mathematics*, 40(1):29–39, 1988.

C. Osmond and M. J. Gardner. Age, Period and Cohort Models Applied to Cancer Mortality Rates. *Statistics in Medicine*, 1(3):245–259, 1982.

M. Plummer and B. Carstensen. Lexis: An R Class for Epidemiological Studies with Long-Term Follow-Up. *Journal of Statistical Software*, 38(5):1–12, 2011.

R. C. A. Rippe, J. J. Meulman, and P. H. C. Eilers. Visualization of Genomic Changes by Segmented Smoothing Using an L0 Penalty. *PLoS ONE*, 7(6):e38230, 2012.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

Y. Yang and K. C. Land. *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications*. Chapman & Hall/CRC Interdisciplinary Statistics, 2013.

M. Żak-Szatkowska and M. Bogdan. Modified Versions of the Bayesian Information Criterion for Sparse Generalized Linear Models. *Computational Statistics & Data Analysis*, 55(11):2908–2924, 2011.