

Regularized hazard estimation for age-period-cohort analysis

Grégory Nuel, Olivier Bouaziz, Vivien Goepp

▶ **To cite this version:**

Grégory Nuel, Olivier Bouaziz, Vivien Goepp. Regularized hazard estimation for age-period-cohort analysis. 2018. hal-01714624

HAL Id: hal-01714624
https://hal.science/hal-01714624

Preprint submitted on 13 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Regularized hazard estimation for age-period-cohort analysis

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

[1]*MAP5 (Department of mathematics, 45, rue des Saints-Pères, 75006 Paris)*
[2]*LPMA (Department of mathematics, 4, Place Jussieu, 75005 Paris)*

December 2017

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 data are usually reported in the form of registers which contain the number of observed cases and the number of individuals at risk to contract the disease. These data 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 the E3N cohort data of breast cancer.

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 in the form of registers which contain the number of observed cases and the number of individuals at risk to contract the disease. 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 epidemiological and demographic studies, it is therefore very important to take into account the variability of the age, the cohort effect (date of birth) and the period effect (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].

The standard approach in age-period-cohort analysis is to fit the effects of age, period and cohort as factor variables in a regression model where the output is the logarithm of the hazard rate. One can use the full model which includes all three effects as factor variables. However, this induces an identifiability problem due to the relationship: period = age + cohort. One solution is to use instead a reduced model where only two of the age, period and cohort effects are modeled. These models are identifiable and since they have fewer parameters than there are variables, they are regularizing. But they are also restrictive because they assume that one effect does not depend on the other effect's value. Different constraints have then been proposed to make the age-period-cohort model identifiable. However, when no *a priori* knowledge is assumed on the data, these proposed constraints are absolutely arbitrary [Heuer, 1997, p 162]. Osmond and Gardner [1982] proposed to compute each

submodel and use a weighting procedure to measure which has the best goodness of fit, in order to choose the best constraint to add to the model. On the other hand, without additional constraints, the age, period, and cohort effects are identifiable only up to a linear trend. Therefore, Holford [1983] proposed to directly estimate the linear trends of each effect, a procedure leading to results that are difficult to interpret. The package `apc` [Kuang et al., 2008, Nielsen et al., 2015] offers to estimate the second order derivatives of the effects . Carstensen [2007] provides a detailed discussion of the identifiability problem of the age-period-cohort model. Namely, the author offers to first fit one submodel (say age-cohort) and then to fit the period effect over the residuals of the first model. Fitting two models sequentially is different from fitting the three effects at once, but the result is believed to be close to the maximum likelihood estimate. Moreover, this method offers a simple and convenient solution to the nonidentifiability problem. The function `apc.fit` from the R package `Epi` [Carstensen et al., 2017, Plummer and Carstensen, 2011] implements this method and provides estimates of the age, period, and cohort effects. Each effect is smoothed using one-dimensional splines. Finally, it is important to stress that both modeling approaches (models with two effects and the full model) assume a simplistic effect of age, period and cohort as no interaction terms are allowed. In other words, these models assume a same effect of the age for every cohort and period, a same effect of cohort for every age and period and a same effect of period for every age and cohort.

Another approach consists in considering the hazard rate as a function of age and either period or cohort and to estimate this function in a non-parametric setting. No specific structure is assumed on the effect of age, period and cohort on the hazard rate, however non-parametric approaches will suffer from overparametrization for moderate sample sizes. 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 McKeague and Utikal [1990] where the cumulative hazard is smoothed using a kernel function. Non-parametric methods for survival analysis over the Lexis diagram have been initiated by Beran [1981]. A spline method was proposed by Heuer [1997], estimating the age and period effects using restricted cubic splines (natural splines). Finally, Ogata and Katsura [1988] proposed a penalized likelihood estimator. The likelihood is penalized using the integral of the squared second order derivatives and the proximity between spline regression and penalized likelihood is explicited. The authors make use of the Bayesian interpretation of the penalized likelihood to propose a Bayes procedure to adjust the values of the penalty constants. See [Keiding, 1990] for a thorough discussion of methods for inference of the hazard for age-period-cohort analysis.

In this article, we propose a new model for age-period-cohort analysis. As the aforementioned 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 effect of age, period and cohort. All effects can be retrieved using the relationship: period $=$ age $+$ cohort. We emphasize that there is no need to directly model the three effects (and as a consequence add supplementary constraints on the model), since all effects can be retrieved from the original modelization. Interestingly, our model can be seen as a direct extension of the standard age-period-cohort models since interactions terms between the three effects of age, period and cohort are allowed in our modeling approach. Finally, in order to take into account the issue of overfitting, a general penalization procedure is introduced. The penalty can be specified either as a L_2 penalty or as an approximated L_0 penalty. The latter uses an approximation of the L_0 norm which was first introduced by Frommlet and Nuel [2016] and which is computationally tractable. We show that in our case, the likelihood maximization is made computationally feasible even for small age and cohort step sizes using the inversion of bandmatrices. Both approaches are of interest depending of the objective of the study: the L_2 penalty provides a smooth estimate of the hazard rate while the L_0 penalty provides a segmented estimate, leading to a parsimonious representation of the hazard rate.

In Section 1, after presenting a review of some of the existing models for age-period-cohort analysis our two new models are introduced. The two different regularization methods are then presented in Section 2: the L_2 penalty, which leads to a smoothed hazard, and the L_0 penalty, which leads to

a segmented hazard estimate. In Section 3, the penalty term selection problem is discussed and in Section 4, we explain how to construct confidence intervals for the L_0 penalty type estimator using a bootstrap procedure. Finally, the performance of our models is illustrated through a simulation study in Section 5 and on the breast cancer data from the MGEN French registry in Section 6.

1 Modeling strategy in the age-period-cohort setting

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 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 $c_K = d_K = \infty$. These points of discretization are often evenly spaced and, in the case of registered data, they are imposed by the data. On a sample of size n, 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. 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:

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

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 [Carstensen, 2007, for a discussion on the "Poisson" model]. Minimizing ℓ_n yields an explicit maximum likelihood estimator $\widehat{\lambda}^{\text{mle}}_{j,k} = O_{j,k}/R_{j,k}$ which is well known in the age-period-cohort literature. However, for moderate sample sizes this estimator will suffer from overfitting, especially in places of the age-cohort plane where few events are recorded. In the following, we first discuss some existing models and then we present two new approaches to analyze these types of age-period-cohort data.

1.1 Existing Age-Cohort and Age-Period-Cohort models

In the classical literature of age-period-cohort analysis, the focus is on modeling a specific effect of the age, period and cohort on the hazard rate. To this end, mainly two different parametrizations have

been used, leading to two different models. In the following, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{J+K-1})$ represent the vectors of the age, cohort, and period effects respectively.

1. In the AGE-COHORT model, it is assumed that

$$\log \lambda_{j,k} = \mu + \alpha_j + \beta_k, \tag{2}$$

   with the convention $\alpha_1 = \beta_1 = 0$. Note that this model contains only $J + K - 1$ parameters instead of the initial JK values of $\boldsymbol{\gamma}$. As a consequence, this model is highly regularizing and it also assumes a strong *a priori* on the structure of the hazard. Inference is made by using the Newton-Raphson algorithm to minimize $\ell_n(\boldsymbol{\lambda})$. See Clayton and Schifflers [1987a] for more information on these types of modeling approaches.

2. In the AGE-PERIOD-COHORT model, it is assumed that

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

   where $\boldsymbol{\gamma}$ is the vector of parameters estimating the *period effect*. Since age + cohort = period, this model is not identifiable as such, and the age, period and cohort effects are only identified up to a linear trend. See Clayton and Schifflers [1987b] for more information of this model. Discussions over the choice of the *best* constraints to add to make this model identifiable can be found for instance in Carstensen [2007].

Other approaches have been proposed. In particular regularized non-parametric estimators have been considered such as splines [Eilers and Marx, 1996, pp. 82-83] or kernel estimators [McKeague and Utikal, 1990]. In [Eilers and Marx, 1996], spline smoothing is done by projecting the hazard rate on a family of B-spline functions. Instead of choosing the number and positions of spline knots, the authors propose to use a relatively large number of knots and to penalize on the differences between coefficients of adjacent B-splines.

Knorr [1984] used a more straightforward smoothing method, sometimes referred to as graduation: the authors first estimate the maximum likelihood estimator $\widehat{\boldsymbol{\lambda}}^{\text{mle}}$ as the parameter that minimizes $\ell_n(\boldsymbol{\lambda})$, defined in (1), then they smooth it using a Whittaker-Henderson-type penalization :

$$\widehat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \sum_{j,k} W_{j,k}(\lambda_{j,k} - \widehat{\lambda}^{\text{mle}}_{j,k})^2 + \sum_{j,k} (\Delta^2 \lambda_{j,k})^2,$$

where $W_{j,k}$ are scalar weights that represent the importance given to the (j,k)-th hazard rate value and Δ^2 represents the second order difference operator. In this equation, the first term represents the estimate's goodness of fit to the data and the second term (the square of the derivative) represents its smoothness, such that the estimator achieves a compromise between model fit and smoothness of the estimate. The author presents this method in the general multidimensional setting.

On the other hand, in McKeague and Utikal [1990], the cumulative hazard is estimated and then smoothed using a kernel function. With this method, the type of kernel and the bandwidth need to be chosen by the user.

In the following, we present our model, which provides a different regularization approach than the spline or kernel methods.

1.2 The Age-Cohort-Interaction (ACI) Model

In the two previous models, no interactions between age, period and cohort are allowed. For instance, Model (i) imposes the age effect to be the same for every cohort and the cohort effect to be the same for every age. Model (ii) also captures the period effect but at the cost of introducing a cumbersome constraint on the model parameters.

In order to capture interaction effects and to avoid complex constraints on the model parameters, we define the new AGE-COHORT-INTERACTION (ACI) model as

$$\log \lambda_{j,k} = \mu + \alpha_j + \beta_k + \delta_{j,k}, \tag{3}$$

where $\delta_{1,k} = \delta_{j,1} = \alpha_1 = \beta_1 = 0$ such that there are JK freely varying parameters (1 parameter for μ, $J-1$ parameters for the αs, $K-1$ parameters for the βs and $(J-1)(K-1)$ parameters for the δs). The age and cohort effects are fitted with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as in the age-cohort model (2) but the hazard is allowed to deviate from this model using an extra parameter $\boldsymbol{\delta} = (\delta_{1,1}, \ldots, \delta_{J,K})$, called the *interaction* term. This model can be seen as a direct extension of Models (i) and (ii). In particular, even though only two effects (age and cohort) are modeled, any other combination of effects can be retrieved using the simple relation: age = period + cohort.

In order to avoid overfitting in the case where n is not large enough as compared to the number of parameters JK, a penalization strategy is considered for the estimation procedure. First, we rewrite the model in the following form:

$$\log \lambda_{j,k} = \eta_{j,k}, \tag{4}$$

where the $\eta_{j,k}$s are JK freely varying parameters and the model parameter is denoted $\boldsymbol{\eta} = (\eta_{1,1}, \ldots, \eta_{J,K})$. The correspondance between Equations (3) and (4) is made through the following relations:

$$\begin{aligned}
\mu &= \eta_{1,1} \\
\alpha_j &= \eta_{j,1} - \mu \\
\beta_k &= \eta_{1,k} - \mu \\
\delta_{j,k} &= \eta_{j,k} - \mu - \alpha_j - \beta_k.
\end{aligned}$$

Penalization of the ACI model will then be performed on the differences between adjacent values of the log-hazard. Using the parametrization (4), this amounts to penalize the $\eta_{j,k}$s in the following way:

$$\ell_n^\kappa(\boldsymbol{\eta}) = \ell_n(\boldsymbol{\eta}) + \frac{\kappa}{2} \sum_{j=1}^{J-1} \sum_{k=1}^{K} v_{j,k} \left(\eta_{j+1,k} - \eta_{j,k}\right)^2 + \frac{\kappa}{2} \sum_{j=1}^{J} \sum_{k=1}^{K-1} w_{j,k} \left(\eta_{j,k+1} - \eta_{j,k}\right)^2, \tag{5}$$

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 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 *a priori*. As such, it does not answer the question of quantifying the age, period and cohort effects but rather provides an accurate estimation of the hazard rate which takes into account potential overfitting. Two choices for the weights $\boldsymbol{v}$ and $\boldsymbol{w}$ will be presented in the next section: one will lead to a smooth hazard rate and the other to a segmented hazard rate. 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 score vector 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 the Hessian matrix. For $1 \leq j, j' \leq J$ and $1 \leq k, k' \leq K$, we have

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

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

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

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 $\mathcal{O}(J^3 K^3)$. Inversion of the Hessian matrix is performed with the C++ implemented function `bandsolve` (see `https://github.com/Monneret/bandsolve`).

Other types of penalization strategies that directly use the paramatrization (3) could have been considered. In particular, it could be of interest to introduce a similar penalty on the $\delta_{j,k}$ instead of directly penalizing the log-hazard. These types of penalties will not be studied in the present paper.

Algorithm 1 Newton-Raphson Procedure with constant weights

```
function NEWTON-RAPHSON(O, R, κ, v, w)
    η ← 0
    while not converge do
        η^new ← η − I_n^κ(η, v, w)^{-1} U_n^κ(η, v, w)
        η ← η^new
    end while
    return η
end function
```

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 penalization of the first order differences.

1. 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 1.1. 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 and we will use the abbreviation "L_2^{Δ} model" to emphasize that the penalization is performed on the differences of the log-hazard.

2. Following the work from Frommlet and Nuel [2016] an adaptive ridge procedure is performed when the weights are updated at each iteration of the Newton-Raphson algorithm. For example, in the penalized maximum likelihood model, at the m-th iteration of the Newton-Raphson algorithm the weights are computed from the following formulas:

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

where ε_v and ε_w are constants negligible compared to 1 (in practice one typically chooses $\varepsilon_v = \varepsilon_w = 10^{-6}$). We iterate between minimizing ℓ_n^{κ} for fixed weights and reevaluating 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 $w_{j,k}^{(m)}(\eta_{j,k+1}^{(m)} - \eta_{j,k}^{(m)})^2 \simeq \|\eta_{j,k+1}^{(m)} - \eta_{j,k}^{(m)}\|_0$. 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, this algorithm will be used to select the piecewise constant areas but we use the unpenalized maximum likelihood estimate to infer the value of the hazard over each constant area. More precisely, when the adaptive ridge algorithm converges, $v_{j,k}\left(\eta_{j+1,k} - \eta_{j,k}\right)^2$ are very close to 1 if the adjacent values $\eta_{j+1,k}$ and $\eta_{j,k}$ have been estimated to have different values and to 0 if they have been estimated to have the same value – and similarly for $w_{j,k}\left(\eta_{j,k+1} - \eta_{j,k}\right)^2$. We typically use a threshold of $\varepsilon = 10^{-8}$, so that values smaller than ε are set to 0 and values larger than $1-\varepsilon$ are set to 1. Then one creates the graph whose vertices are the JK discretization cells and whose edges are the connexion between adjacent cells that have a difference close to 0. Each connex component of this graph is a different area over which the hazard has been estimated to be constant. The extraction of connex 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 $\forall [c_{j-1}, c_j) \times [d_{k-1}, d_k) \in r, \eta_{j,k} = \eta^{(r)}$. The values of $\eta^{(r)}$ are not estimated using the results of the adaptive ridge algorithm, but by unpenalized maximum likelihood estimation: $\widehat{\eta}^{(r)} = O^{(r)}/R^{(r)}$ 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 method is illustrated through the toy-example of Figure 2 and the adaptive ridge procedure is summarized in Algorithm 2. This estimation method will be called L_0 regularized estimation and will shorten into L_0^{Δ} model. See also [Bouaziz and Nuel, 2017] for an implementation of this adaptive ridge procedure in a similar context of hazard rate estimation.

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

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 $(\boldsymbol{\eta}, \boldsymbol{v}^{\text{new}}, \boldsymbol{w}^{\text{new}})$
  $\boldsymbol{\eta}^{\text{new}} \leftarrow \boldsymbol{O}^{\text{new}} / \boldsymbol{R}^{\text{new}}$
  return $\boldsymbol{\eta}^{\text{new}}$
end function

variability of the hazard rate estimate. For the L_0^{Δ} 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 5 on simulated data.

We recall that $\boldsymbol{R}$ and $\boldsymbol{O}$ are the exhaustive statistics (i.e. the data) 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 the right-hand side of the approximation

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

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 et al., 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:

$$\mathrm{BIC}(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m \log n \tag{6}$$

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. 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 $\mathrm{P}(\mathcal{M}_m \in \mathcal{M}_{[q_m]}) = \binom{JK}{q_m}^s$ for some constant $s \in [0, 1]$. Thus

$$\pi(M_m) = \binom{JK}{q_m}^{1-s}$$

and

$$\mathrm{EBIC}_s(m) = 2\ell_n(\widehat{\boldsymbol{\eta}}_m) + q_m \log n - 2 \log \binom{JK}{q_m}^{1-s}. \tag{7}$$

Note that for $s = 0$, $\mathrm{P}(\mathcal{M}_m \in \mathcal{M}_{[q_m]}) = 1$, that is the $\mathrm{EBIC}_{s=0}$ 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 $\mathrm{EBIC}_{s=0}$ will be able to select this model more easily. Namely, when the true model's dimension is very small the $\mathrm{EBIC}_{s=0}$ will tend to choose very sparse models. In the following we will only consider the case $s = 0$ for the EBIC_0 criterion.

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^Δ estimation only, since the L_2^Δ does not perform a model selection. An alternative to performing model selection is to use the L-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. This score evaluates how well the model learned on one part of the data explains the other 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^Δ estimation and the L_2^Δ 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, L will be set to 10.

4 An Ensemble method for the adaptive ridge estimator

The adaptive ridge regularization provides a piecewise constant estimate of the hazard. While this might be useful in practice in order to obtain a parsimonious estimate, visualization of plots of the adaptive ridge hazard estimate can be hard to interpret, especially if the number of segmented areas is large. On the opposite, the ridge procedure provides nice plots of the hazard but the information contained in these plots can be hard to summarize. An alternative method consists in providing a smooth estimation of the hazard rate which also includes some few segmented areas corresponding to areas of the hazard which have very different values of the rest of the plot (typically areas which are highly peaked at some localized regions or areas with a very different level of hazard). This new type of estimation can be implemented using a resampling strategy of the adaptive ridge estimator. We compute bootstrapped samples of the original data and calculate the adaptive ridge estimator for each of these samples. Then, taking the median of all the bootstrapped estimators will lead to a new estimator, called the ensemble adaptive ridge estimator. This estimator represents a relaxed

version of the adaptive ridge segmentation. Moreover, the bootstrapped estimators can easily be used to compute pointwise confidence intervals: for example, removing the 2.5% largest and 2.5% smallest estimates values at a given time point will provide a 95% confidence interval at this point. Interestingly, these confidence intervals also take into account uncertainty about the choice of the segmentation since each bootstrapped estimate is allowed to have a different segmentation chosen by the adaptive ridge procedure than the original estimate. This ensemble method can be used directly for estimation of the hazard rate or of other related functions such as the survival function. This technique will be illustrated in Section 6 on real data.

5 Simulation study

5.1 Data simulated from a smooth hazard model

In this section, data are simulated using the true hazard presented in Figure 3 (a) and (b). To be more specific, we set $J = 10$ equally spaced age intervals and $K = 10$ equally spaced cohort intervals. The age intervals are defined as $[0, 10), \ldots, [90, 100]$ and the cohorts intervals are defined as $[1900, 1910), \ldots, [1990, 2000]$. The true hazard is generated using the age-cohort-interaction model (3) with $\mu = \log(10^{-2})$ and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are arithmetic sequences such that $\alpha_2 = 0$, $\alpha_J = 2.5$, $\beta_2 = 0$, and $\beta_K = 0.3$. The interaction term $\delta_{j,k}$ is defined as 10 times the Gaussian function with mean $(45, 1945)$ 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. It also features a peak around $(45, 1945)$ which we will try to recover using our estimation methods.

This true hazard will be referred to as smooth in opposition to the next simulation scenario (Section 5.2), where the hazard is piecewise constant. 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.

In the following, we first simulate 100 datasets of sample size equal to 4000 which corresponds to an average of 9.6 observed events per 5×5 age-cohort square unit. Estimation is performed using the standard age-cohort model presented in Section 1.1 and compared with our new penalized estimator presented in Section 1.2. The adaptive ridge procedure is performed using the different AIC, BIC, EBIC_0 and cross-validation criteria to select the penalty constant. For the ridge procedure, the penalty constant is chosen using the cross validation criterion. Perspectives plots of the median hazard estimations over the 100 datasets are presented in Figures 3 (c)-(d).

Figure 3 (c) represents the estimated hazard using the age-cohort model (2). The resulting estimate is smooth and relatively accurate except for the area of cohort 1945 and age 45 where a peak occurs for the true hazard rate. Since the age-cohort model has an outer product structure (no interaction terms between age and cohort is allowed), the hazard singularity cannot be detected using this model. Figure 3 (d) represents the estimated hazard using the L_2^{Δ} estimation method. The resulting estimator shrinks all adjacents values of the hazard and is consquently biased for high age values. However, the central bump is accurately estimated, which is not the case with the AGE-COHORT model.

The adaptive ridge estimations for our ACI model are displayed in figures 4 (a)-(d). They provide a comparable regularization effect with more flexibility on the model, since the singularity is also accurately estimated. Among the different methods implemented to choose the regularization parameter, the AIC seems to provide the best fit, followed by the BIC. The EBIC_0 (see Figure 4 (c))

turns out to be too penalizing for a hazard rate with strong variations as it is the case here. Interestingly, adaptive ridge segmentation with cross validation (see Figure 4 (d)) seems to be as regularizing as the EBIC_0. As a conclusion, the segmented hazard estimate with AIC selection can be considered a better overall estimate than the smooth estimate, since it performs almost as well in a case where the true hazard is smooth. More sparse criteria than the AIC perform slightly worse in this case, but not significantly so.

We then looked at the performance of the previous different estimation methods in terms of mean squared error. We considered samples of size 4000 where each sample was replicated 100 times and the Mean Squared Error (MSE) between the true model and the estimations was used as a criterion for estimation performance. The true hazard used for the simulations is the same as before (see Figures 3(a)-(b)). The ridge estimator (L_2^Δ with CV in the figure) and the adaptive ridge estimator with EBIC_0 seem to provide very similar results. The three other methods are significantly less performant in terms of MSE, all three yielding similar results. Other samples sizes were considered. The performance of our methods were similar and are therefore not shown here.

Simulations were also performed with a higher level of censoring. The resulting estimations showed a similar trend with a slightly worse performance of the estimation.

5.2 Data simulated from a piecewise constant hazard model

The data are now simulated using a piecewise constant hazard with four constant areas over the age-cohort square $[0, 100] \times [1900, 2000]$. The shape of the true hazard is displayed in Figure 6 (a) and (b). As previously, the median of the hazard over 100 simulations with sample size 4000 are computed for each estimation method. The results are shown in Figures 6 (c)-(d) for the AGE-COHORT model and L_2^Δ model and in Figures 7 (a)-(d) for the adaptive ridge estimation methods. The AGE-COHORT model performs as poorly as is expected from the fact that the true hazard is quite far from having an age-cohort outer product structure. The L_2^Δ model provides a smoothed estimate of the true hazard, which in this case makes him perform poorly. Interestingly, amongst the L_0^Δ models, the most parsimonious estimate is provided by the cross-validation, followed by EBIC_0, BIC and AIC. The regularizing power of EBIC_0 and, to some extent, BIC allows the estimated hazard to be regularized for high values of the age, whereas this region is poorly estimated by AIC. The AIC is the only criterion to correctly estimate the central bump.

Finally, mean squared errors are compared as in the previous section. Simulations were replicated 100 times for samples of sizes 100, 400, 1000 and 4000 using the same true hazard rate. The results in terms of MSE are presented for sample size 4000, on the log scale, in Figure 8. Since the true hazard is a piecewise constant hazard with only four different regions, the most performant criteria are the ones providing the most sparse estimation. In particular, EBIC_0 is seen to perform almost as well as CV. As in the previous simulation, the EBIC_0 displays many outlying values of the MSE. We suspect that this comes from the fact that the criterion grants importance to models with small or large dimensions, which can make it sensible to sampling outliers.

6 Application to the French E3N cohort study

In this section we apply our estimation methods to the E3N dataset from the MGEN (Mutuelle Générale de l'Éducation Nationale) which is the French national health security for teachers. In this cohort women insured by the MGEN were followed through questionnaires in order to estimate breast cancer incidence. The cohort comprised $n = 98,995$ female participants who were born between 1925 and 1950. In the current dataset, the period starts in June 1990 and finishes in January 2010. Questionnaires were sent approximately every $2 - 3$ years. For more information about the E3N dataset, see Clavel-Chapelon and Group [2014]. The discretization was made on the year scale, such that there are $J = 26$ cohorts (from year of birth 1925 to 1950) and $K = 46$ different ages (from age

40 to 85). On average, there are 92 observed events on each 1×1 age-cohort square, but there are way fewer events for the older cohorts than for the most recent cohorts. Visualization of the observed breast cancers and censored events can be seen in the period-cohort plane or age-cohort plane in Figure 9. For the sake of simplicity, inference is made in the age-cohort plane, but visualization of the estimates can be made on any two-dimensional plane (age-cohort plane, age-period plane or period-cohort plane) by applying the same linear transformation that transforms Figure 9 (a) into Figure 9 (b). In the following, visualization is made in the period-cohort plane.

In Figure 10 estimation is performed using different models. The unpenalized maximum likelihood estimator $\widehat{\lambda}^{\text{mle}}$ is represented in Figure 10 (a) for the sake of comparison. This estimate clearly suffers from overfitting. Both the ridge and the adaptive ridge procedures are implemented. Following the results of the simulation section, the AIC and EBIC_0 criteria are chosen for the penalty term of the adaptive ridge procedure and cross validation is used for the ridge procedure. The estimates obtained with the adaptive ridge method yield a segmented hazard, which are visually hard to interpret but allow to identify breakpoint changes in the hazard rate. As explained in Section 4, compromise between the ridge and the piecewise constant estimators can be derived using an ensemble method. These estimators were also implemented and are also represented in Figures 10 (e)-(f).

Ridge regularized estimation appears as an appealing solution to overfitting as illustrated in Figure 10 (b). However, the information can be hard to summarize from this smoothed plot. On the other hand, the EBIC_0 criterion selects a very sparse segmentation of the data which provides a simple message: the risk of developing breast cancer is higher for women living between 1995 and 2005 who were born between 1935 and 1945 (see Figure 10 (d)).

The decrease of the hazard rate for large values of the cohort and small values of the period is also more clearly identifiable than in the two previous models. The AIC criterion (Figure 10 (c)) selects a less sparse model than EBIC_0, which in turn is more difficult to interpret even though this estimator should achieve the best bias and variance tradeoff as shown in the simulation section. For the AIC, the ensemble method provides a median hazard rate that is only partially smooth. Many peaks are present and the hazard rate is difficult to interpret. The ensemble method estimate with EBIC_0 is the most interesting one, since it seems to yield a simple yet complete visualization of the features of the hazard. Namely, it allows to identify some of the peaks present in Figure 10 (c) as well as the plateau present in Figure 10 (d). It also displays a somewhat important variation in the hazard for smaller values of the cohort.

Ensemble methods allow to provide empirical confidence intervals of the pointwise estimates of the hazard. Since the survival function can be computed from the hazard rate, confidence intervals can also be inferred for the survival function. As an illustration, we have computed the survival functions as a function of age only, and for different cohort intervals. Figure 11 represents two survival functions with cohort intervals $[1935, 1936)$ and $[1942, 1943)$ estimated using the ensemble method adaptive ridge estimates with AIC. The pointwise median value of the survival over all bootstrapped samples is represented with solid lines and the lower and upper bounds of the 95% confidence intervals are represented with dotted lines. The two survival curves are similar, with a slightly better survival for the cohort $[1935, 1936)$ where for instance the survival at age 65 is equal to 0.96 while it is equal to 0.94 for the other cohort at the same age. Results for other selection criteria have also been computed and showed similar results.

Conclusion

In this article, we have introduced a new estimation method to deal with age-period-cohort analysis. In this model no specific structure of the effects of age and cohort on the hazard rate is assumed and the likelihood is directly maximized 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 hazards 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 type 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 were the most performant, in terms of MSE . On the other hand, when the interest lies in obtaining nice visualization of hazard plots we advocate the use of ensemble methods for the adaptive ridge estimator which provide a good compromise between a parsimonious representation of the hazard and smooth estimation on some selected areas of the hazard. Among all proposed methods, the cross validation criterion was shown to provide the best fit of the hazard rate, but it is the most computationally intensive method of all. Among the other criteria, the $EBIC_0$ was shown to outperform the AIC and the BIC criteria in the context of segmentation, while if the true hazard is a smooth function then the AIC seemed to provide the best fit of the hazard rate. This brings us to recommend the L_0^{Δ} method with AIC as an alternative to the ridge regularized model when no piecewise constant areas in the data are expected and the L_0^{Δ} method with $EBIC_0$ piecewise when constant segmentation is desired. Overall, these different criteria have shown to provide interesting results when applied to the E3N Cohort breast cancer data. Notably, segmented hazard estimation has permitted to highlight interesting features in the cancer incidence: the AIC selection exhibits the presence of a precise hazard peak and the $EBIC_0$ selection exhibits two change-of-regime moments in the breast cancer hazard rates. Both of these informations could not be conveyed using standard age-period-cohort models, the unpenalized maximum likelihood estimator, existing smooth estimators such as splines, kernel estimators or using our ridge estimator.

The 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). This could yield a more adapted discretization that is non preferential with respect to the three effects (age, period and cohort). Another extension would be to consider other types of penalizations. In particular, penalizing the second order differences (instead of the first order) could generalize this current work in the sense that $\kappa = \infty$ would correspond to a linear hazard instead of a constant hazard.

Finally, a similar type of penalization could be implemented using the parametrisation (3), which seperates the effects of age, cohort, and the interaction term. With this parametrization, the penalty could be applied to the $\delta_{j,k}$s only using the same type of L_0^{Δ} or L_2^{Δ} penalty. This would yield a different model, where the case $\kappa = 0$ corresponds to the MLE and the case $\kappa = \infty$ corresponds to the AGE-COHORT model. A detailed analysis of this model is left for future work.

Acknowledgement 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.

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

References

Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

Rudolf Beran. Nonparametric regression with randomly censored survival data. Technical report, Technical Report, Univ. California, Berkeley, 1981.

Olivier Bouaziz and Gregory Nuel. L0 regularization for the estimation of piecewise constant hazard rates in survival analysis. *Applied Mathematics*, 8(3):377–394, 2017.

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

Bendix Carstensen, Martyn Plummer, Esa Laara, and Michael Hills. *Epi: A Package for Statistical Analysis in Epidemiology*, 2017. R package version 2.19.

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

Françoise Clavel-Chapelon and E3N Study Group. Cohort profile: the french e3n cohort study. *International journal of epidemiology*, 44(3):801–809, 2014.

D Clayton and E Schifflers. Models for temporal variation in cancer rates. i: age–period and age–cohort models. *Statistics in medicine*, 6(4):449–467, 1987a.

D Clayton and E Schifflers. Models for temporal variation in cancer rates. ii: age–period–cohort models. *Statistics in medicine*, 6(4):469–481, 1987b.

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

Florian Frommlet and Gregory Nuel. An adaptive ridge procedure for l 0 regularization. *PloS one*, 11(2):e0148620, 2016.

Carsten Heuer. Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics*, pages 161–177, 1997.

Theodore R Holford. The estimation of age, period and cohort effects for vital rates. *Biometrics*, pages 311–324, 1983.

Niels 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.

Frank E Knorr. Multidimensional whittaker-henderson graduation. *Transactions of the Society of Actuaries*, 36:213–240, 1984.

Di Kuang, Bent Nielsen, and JP Nielsen. Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, 2008.

Ian W McKeague and Klaus J Utikal. Identifying nonlinear covariate effects in semimartingale regression models. *Probability theory and related fields*, 87(1):1–25, 1990.

Bent Nielsen et al. apc: An r package for age-period-cohort analysis. *R Journal*, 7(2), 2015.

Yosihiko Ogata and Koichi Katsura. Likelihood analysis of spatial inhomogeneity for marked point patterns. *Annals of the Institute of Statistical Mathematics*, 40(1):29–39, 1988.

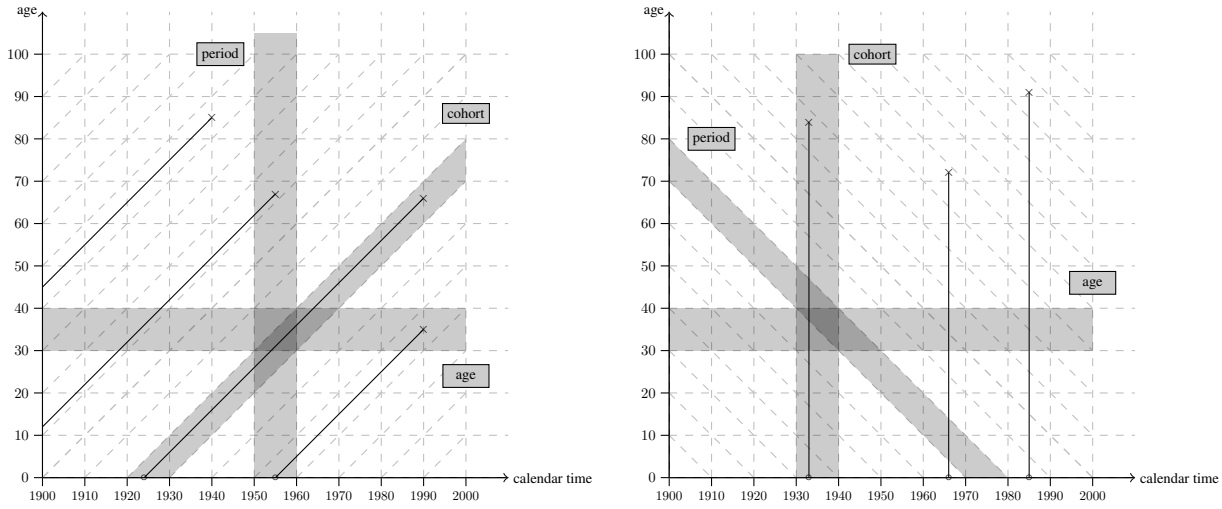
C Osmond and MJ Gardner. Age, period and cohort models applied to cancer mortality rates. *Statistics in Medicine*, 1(3):245–259, 1982.

Martyn Plummer and Bendix Carstensen. Lexis: An R class for epidemiological studies with long-term follow-up. *Journal of Statistical Software*, 38(5):1–12, 2011.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Yang Yang and Kenneth C Land. *Age-period-cohort analysis: New models, methods, and empirical applications*. Chapman & Hall/CRC Interdisciplinary Statistics, 2013.

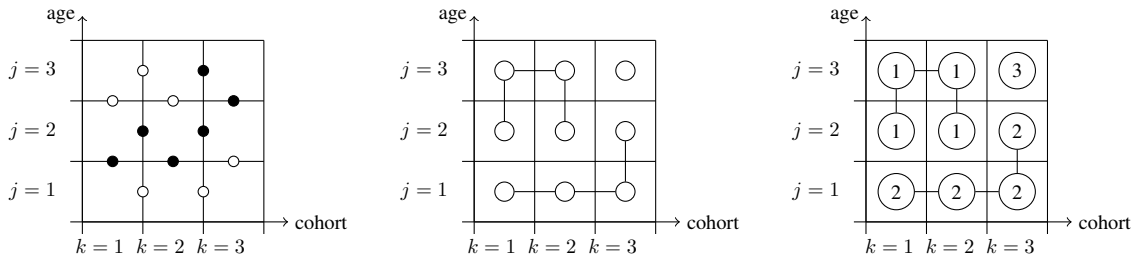
Małgorzata Żak-Szatkowska and Małgorzata Bogdan. Modified versions of the bayesian information criterion for sparse generalized linear models. *Computational Statistics & Data Analysis*, 55(11): 2908–2924, 2011.



(a) Lexis diagram: Age-Period iagram

(b) Age-Cohort diagram

Figure 1: Diagram 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 time of interest. The same age, cohort, and period intervals are displayed in light gray. The intersections of two intervals form either a rectangle or a parallelogram. The intersections of three intervals form a triangle.

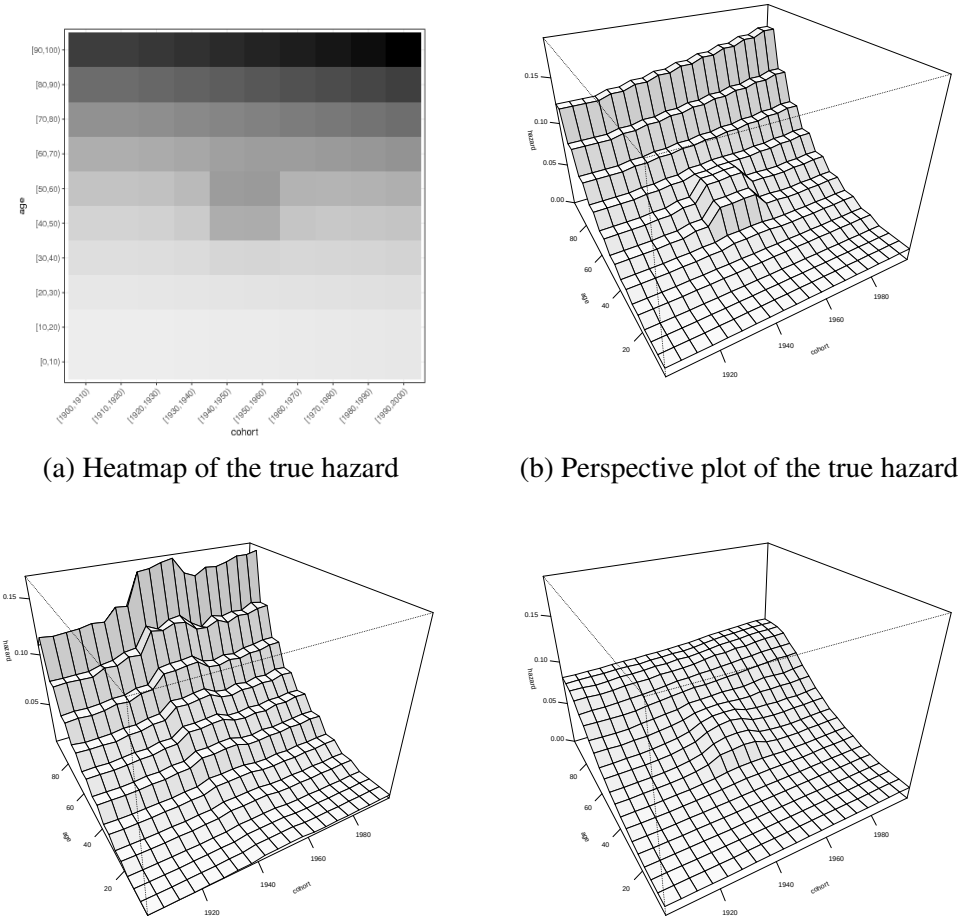


(a) Representation of $v_{j,k}\left(\eta_{j+1,k} - \eta_{j,k}\right)^2$ and $w_{j,k}\left(\eta_{j,k+1} - \eta_{j,k}\right)^2$

(b) Corresponding graph

(c) Segmentation into the graph's connex components

Figure 2: Representation of the method used to select the constant areas from the adaptive ridge procedure. In this example, $J = K = 3$. In the 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 connex components of the graph are extracted. Each connex component corresponds to one constant area. The numbering is arbitrary.



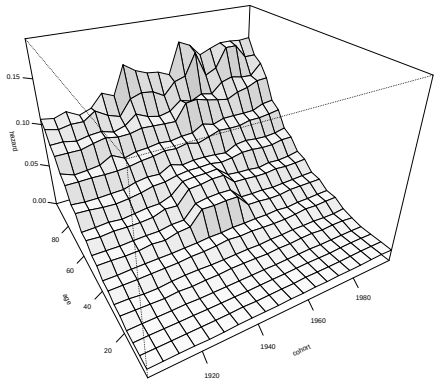
(a) Heatmap of the true hazard

(b) Perspective plot of the true hazard

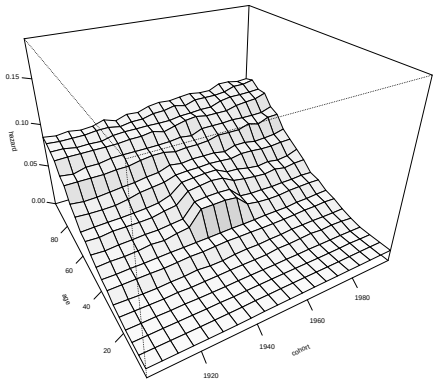
(c) AGE-COHORT Model

(d) L_2^{Δ} estimate with CV

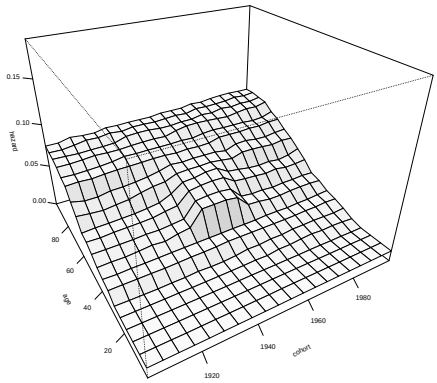
Figure 3: Smooth true hazard and estimations using the AGE-COHORT model and the ridge penalization estimation for 100 simulations. The estimations are performed in the age-cohort plane and with different methods. Figure (a) is a heatmap of the true hazard used to generate the data. Figure (b) represents the same hazard in a perspective plot. Figure (c) represents the hazard estimated using the AGE-COHORT model 2. Figure (d) represents the estimated hazard with ridge penalization estimation.



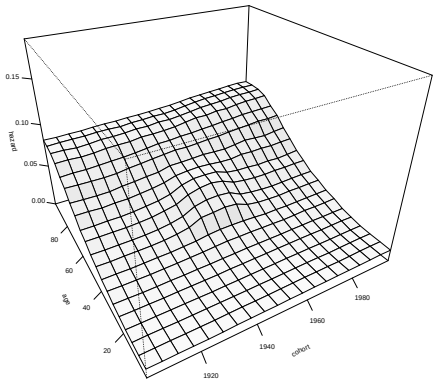
(a) L_0^{Δ} estimate with AIC



(b) L_0^{Δ} estimate with BIC



(c) L_0^{Δ} estimate with EBIC



(d) L_0^{Δ} estimate with CV

Figure 4: Median value of hazard estimates over 100 simulations and with a smooth true hazard. The estimations are performed in the age-cohort plane with the L_0^{Δ} method using different criteria. Figure (a) represents the AIC criterion, Figure (b) represents the BIC criterion, Figure (c) represents the $EBIC_0$ criterion, and Figure (d) represent the estimate selected using cross-validation.

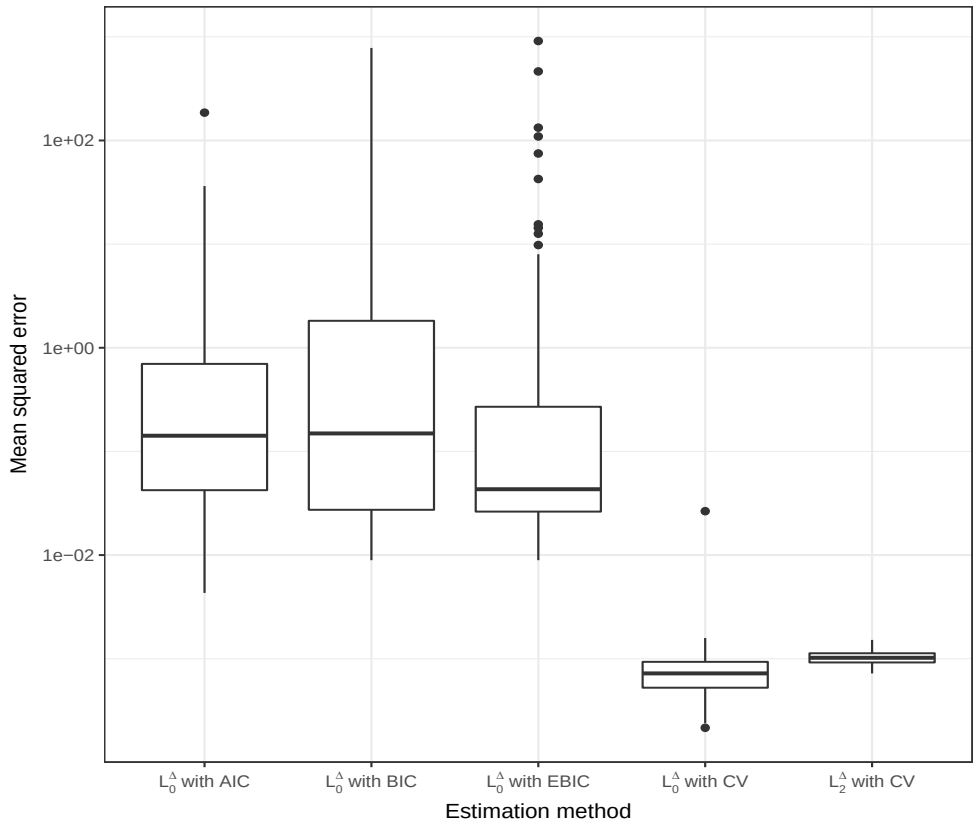
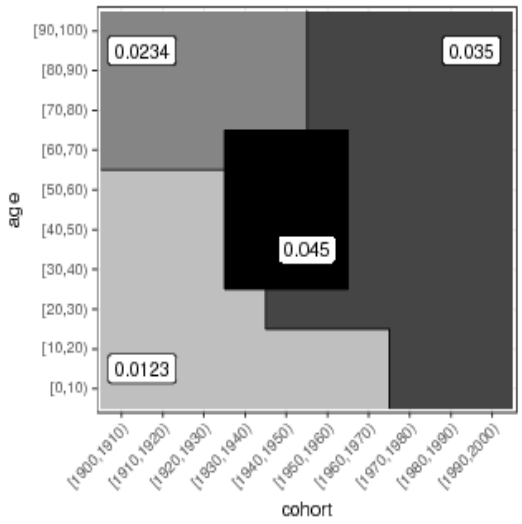
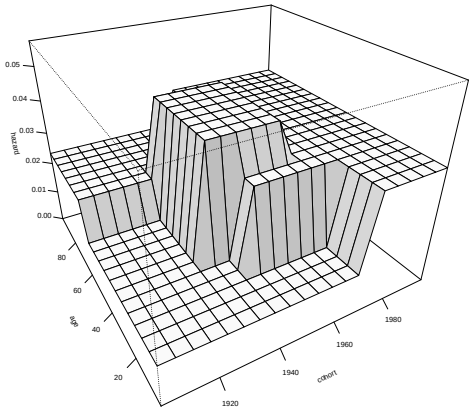


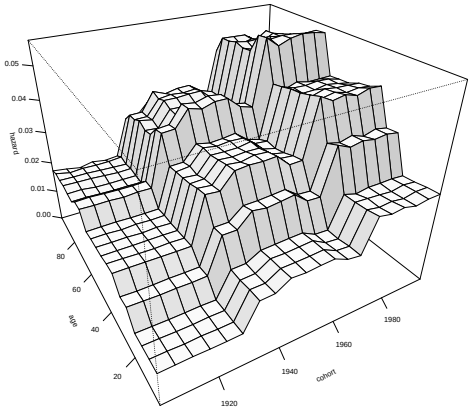
Figure 5: Mean squared errors of hazard estimation methods, where the true hazard is a smooth function of age and cohort. The boxplot represents 100 simulations over the same true hazard.



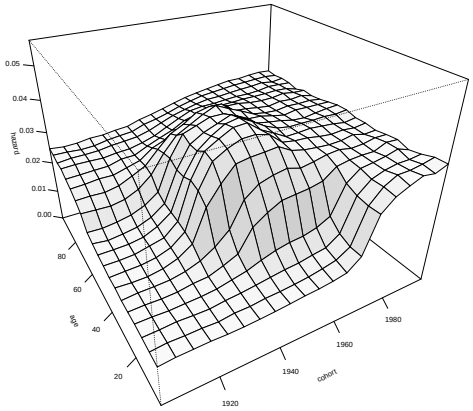
(a) Heatmap of the true hazard



(b) Perspective plot of the true hazard

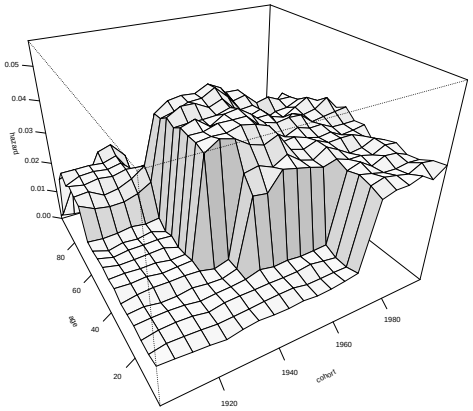


(c) AGE-COHORT Model

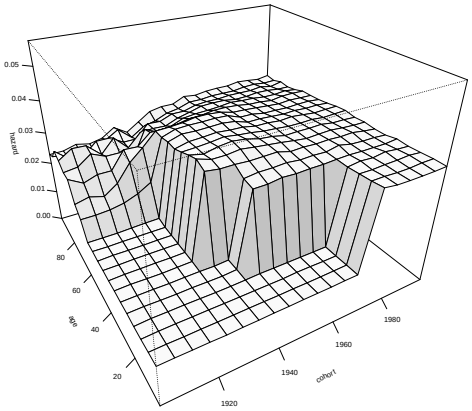


(d) L_2^{Δ} estimate with CV

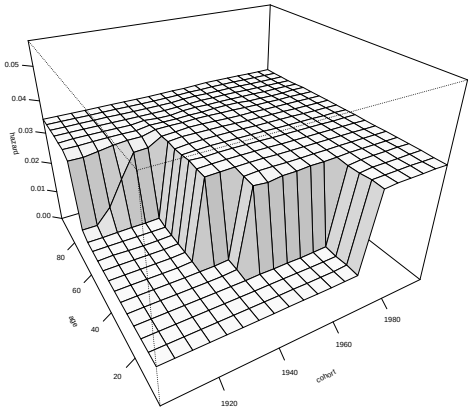
Figure 6: Piecewise constant true hazard and estimations using the AGE-COHORT model and the ridge penalization estimation for 100 simulations. The estimations are performed in the age-cohort plane and with different methods. Figure (a) is a heatmap of the true hazard used to generate the data, on which the hazard values are annotated on each constant area. Figure (b) represents the same hazard in a perspective plot. Figure (c) represents the hazard estimated using the AGE-COHORT model 2. Figure (d) represents the estimated hazard with ridge penalization estimation.



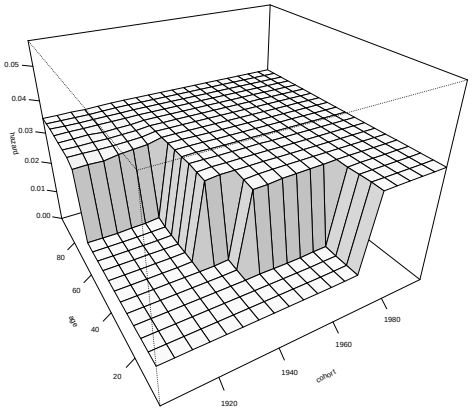
(a) L_0^{Δ} estimate with AIC



(b) L_0^{Δ} estimate with BIC



(c) L_0^{Δ} estimate with EBIC_0



(d) L_0^{Δ} estimate with CV

Figure 7: Median value of hazard estimates over 100 simulations and with a piecewise constant true hazard. The estimations are performed in the age-cohort plane with the L_0^{Δ} method using different criteria. Figure (a) represents the AIC criterion, Figure (b) represents the BIC criterion, Figure (c) represents the EBIC_0 criterion, and Figure (d) represent the estimate selected using cross-validation.

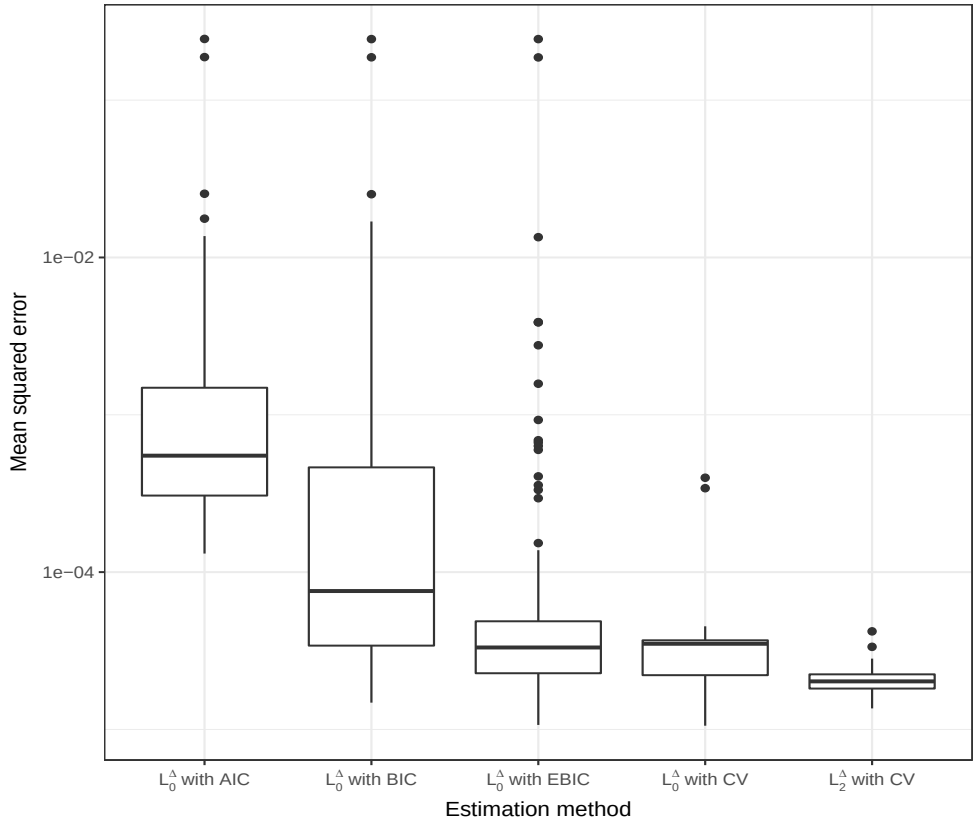
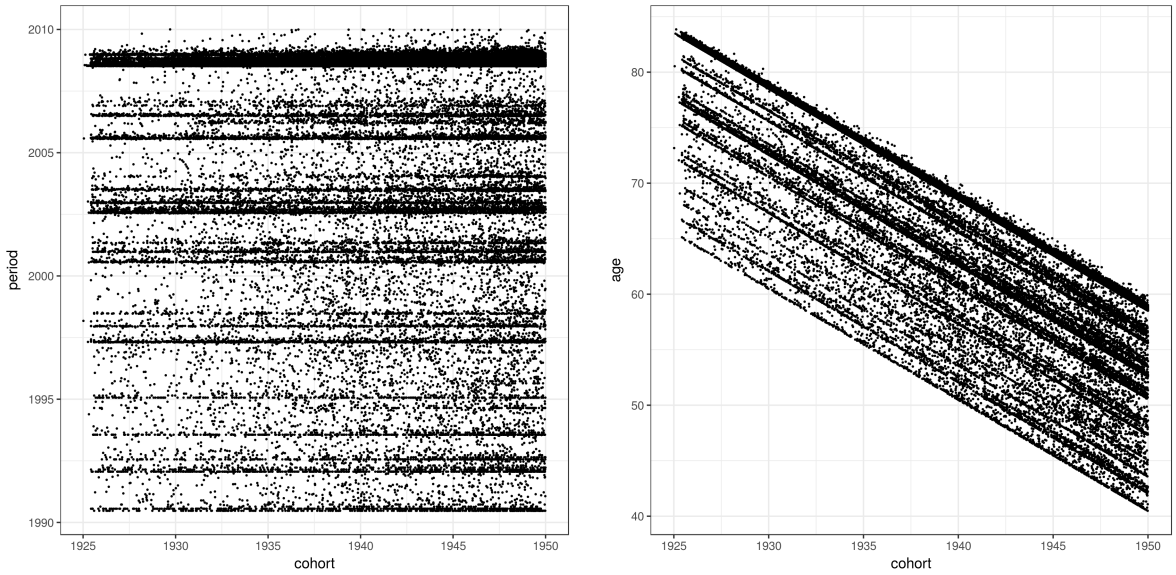


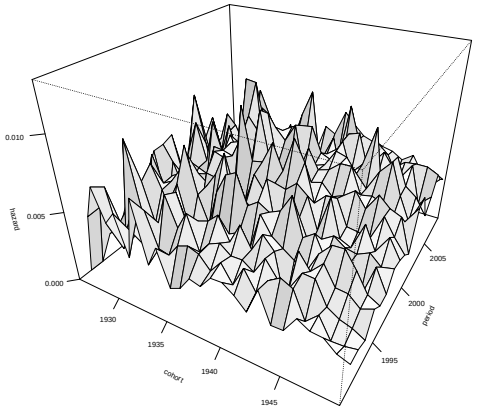
Figure 8: Mean squared errors of hazard estimation methods, with a piecewise constant true hazard and for different criteria. The boxplot represents 100 simulations over the same true hazard with sample size 4000.



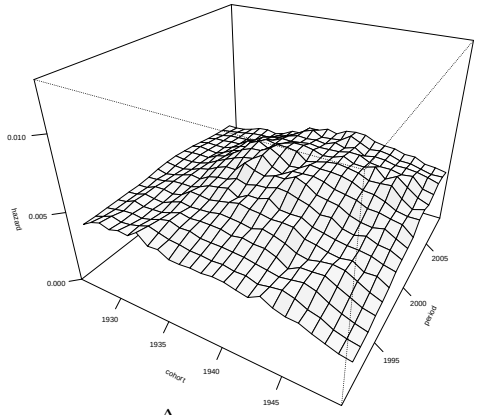
(a) In the period-cohort plane

(b) In the age-cohort plane

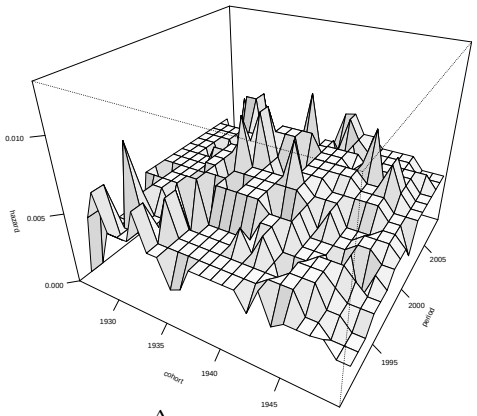
Figure 9: Visualization of the E3N breast cancer data. Each point is an individual's breast cancer occurrence or the last time the person has been checked on being cancer-free, whichever comes first. The data is represented (a) in the period-cohort plane and (b) in the age-cohort plane. Many censoring points are almost aligned because updates are made periodically, every two or three years.



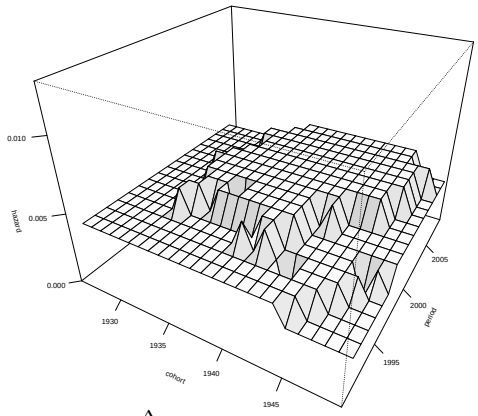
(a) Maximum likelihood estimate



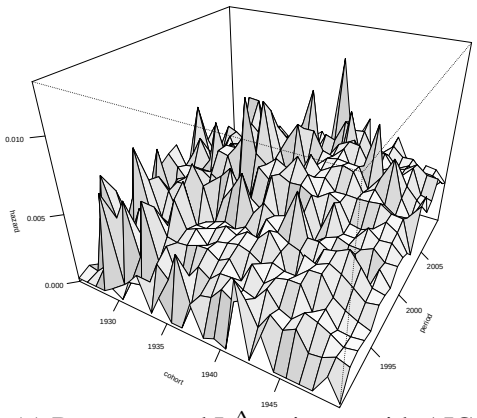
(b) L_2^{Δ} estimate with CV



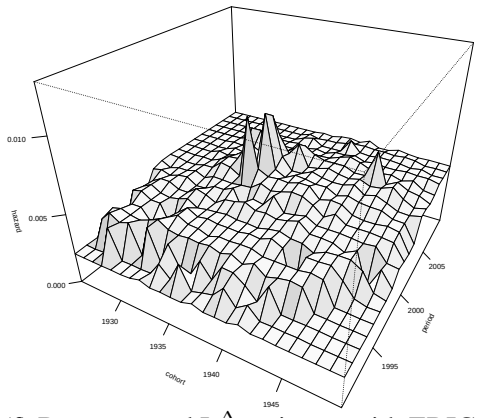
(c) L_0^{Δ} estimate with AIC



(d) L_0^{Δ} estimate with $EBIC_0$



(e) Bootstrapped L_0^{Δ} estimate with AIC



(f) Bootstrapped L_0^{Δ} estimate with $EBIC_0$

Figure 10: Hazard estimates of the breast cancer incidence in E3N data with different models. Inference is made in the age-cohort plane but the hazard is represented in the period-cohort plane. Figure (a) represents the unpenalized maximum likelihood estimate and Figure (b) represents the ridge regularized estimate. Figures (c) and (d) represent the adaptive ridge model with criteria AIC and $EBIC_0$. Finally, Figures (e) and (f) represent the same estimates using the ensemble method.

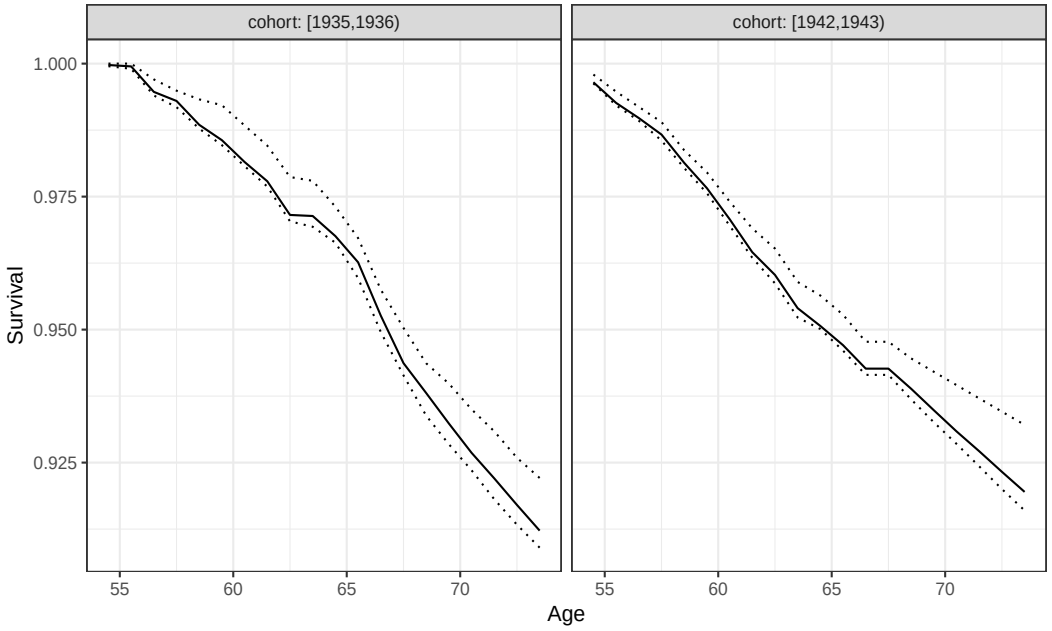


Figure 11: Survival functions of the breast cancer hazard rate in the E3N Cohort study from the ensemble method L_0^{Δ} estimate with the AIC criterion. The survival functions are represented as a function of age and for two cohort intervals : $[1935, 1936)$ and $[1942, 1943)$. Ensemble method provides a set of 100 survival estimates. The solid lines represent the median value of these estimates and the dotted lines represent the lower and upper bound of the 95% confidence intervals around the median value.