PROD. TYPE: COM
PP: 1-13 (col.fig.: nil)

COMSTA3194 MODC+

ED: KK
PAGN: LAK – SCAN: Mouli

ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Computational Statistics & Data Analysis ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

# Additive two-way hazards model with varying coefficients

Göran Kauermann*, Pavel Khomski

*Department of Economics and Business Administration, P.O. Box 100131, University of Bielefeld, Bielefeld, Germany*

Received 9 June 2005; received in revised form 6 December 2005; accepted 13 December 2005

**Abstract**

The paper considers smooth modelling of hazard functions, where dynamics is modelled in both, duration time and calendar time. The model is specified with time dynamic covariate effects to replace restrictive assumptions of proportional hazards. Additivity of the time effects is assumed which allows for simple estimation in a backfitting style. Penalized splines are employed, which provide the welcome benefit of linking smoothing with mixed models. The model is applied to unemployment data taken from the German socioeconomic panel. The hazard function, here the chance for finding reemployment, varies with duration as well as calendar time.


*Keywords:* Generalized linear mixed models; Penalized spline smoothing; Survival time models; Two-way hazard models; Varying coefficient models



## 1. Introduction

Survival time or duration time studies primarily focus on one time scale only, namely the time to an event of interest. This is justifiable, if the individuals or subjects under investigation enter the study all at the same time point or at least in a short time interval. This implies that all subjects are under the same risk, modified by available covariates. In some data situations, however, the risk does not only change with survival time, expressed in the hazard function, but also with calendar time. This is particularly the case if the duration time is long and subjects enter the study at different timepoints. The example we consider in this paper are duration times of unemployment with data taken from the German socioeconomic panel (see www.diw.de). Subjects enter the state of unemployment at different timepoints ranging over the years 1983–2000. Clearly, with this wide time range and the economic dynamics, the success rate of finding a new job has to be modelled to depend on calendar time. Hazard function models which incorporate both, duration as well as calendar time, are known under the phrase two-way hazard models. A graphical representation of the survival data is available by a Lexis diagram, as shown in Fig. 1 for our data at hand (see also Keiding, 1990; Francis and Pritchard, 1998). The event of interest is defined as full time reemployment, while any other termination of unemployment (retirement, retraining, half time job, etc.) is taken as censored information. Unemployment spells longer than 36 months are truncated and taken as censored.

Approaches to model the hazard function of duration time data in both, duration as well as calendar time trace back to Cox (1972) and Cox and Farewell (1979), see also Anderson (1991). Recently, Efron (2002) shows that two-way hazard models can be fitted either by focussing on calendar time, or on duration time or on both at the same time, the

* Corresponding author. Tel.: +49 0521 106 4879; fax: +49 0521 106 89004.
*E-mail address:* gkauermann@wiwi.uni-bielefeld.de (G. Kauermann).

0167-9473/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.csda.2005.12.009

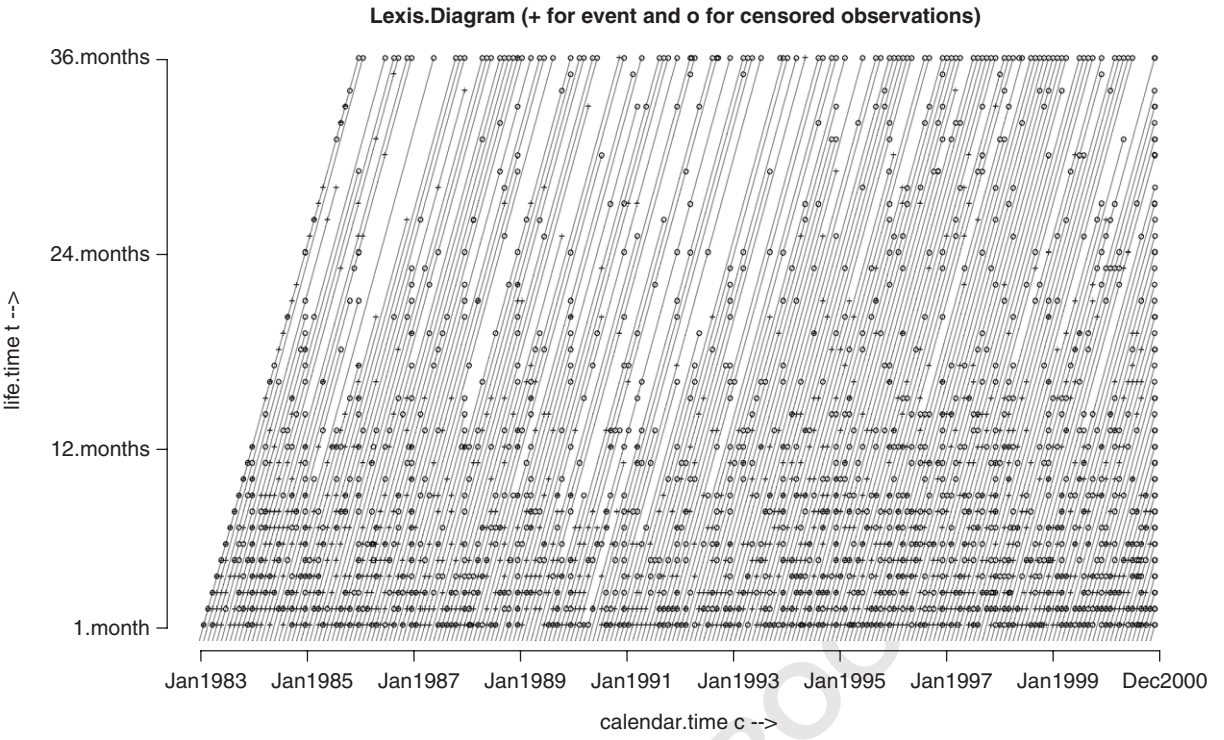


Fig. 1. Lexis diagram for unemployment data. Shown are the time individuals spend in unemployment. Observations are censored at 36 months.

latter using a Poisson-type model. Our approach uses the Poisson approach as starting point, but instead of fitting a parametric or semiparametric model, like in Efron (2002), we choose a nonparametric or more suitably called smooth model with both time scales entering the model in a smooth functional form. Additionally, we allow covariate effects to vary also with both, duration as well as calendar time. This leads to a complex varying coefficient model as introduced in Hastie and Tibshirani (1993). For interpretational reasons, and in order to keep the numerical effort feasible, we assume additivity for our hazard function. This means, on a log scale, the functional effects of duration and calendar time decompose additively. This leads to a generalized additive model in the style of Hastie and Tibshirani (1990) and allows with the backfitting principle a simple way of calculating the estimates.

As smoothing technique we employ penalized spline fitting as introduced as P-spline smoothing in Eilers and Marx (1996) (see also O'Sullivan, 1988). Spline-based approaches with penalized fitting in the context of survival time models have been suggested before, with early references given by Zucker and Karr (1990), Gray (1992, 1994). Kooperberg et al. (1995) provide a general approach with flexible, low-dimensional splines while Fan et al. (1997) or Cai and Sun (2003) use local techniques for smooth estimation. Recently Cai et al. (2002) propose P-spline smoothing for hazard modelling, which is further extended in Kauermann (2005) towards nonproportional hazard models. In all of the above cited papers the nonparametric structure is either over duration time or over some other exogenous metrically scaled covariate. The nonparametric inclusion of calendar time besides of duration time as proposed in this paper is new to our knowledge.

Whenever smoothing techniques are applied for fitting, there is a tuning parameter, commonly called bandwidth or smoothing parameter, to be chosen adequately. This should be done data driven based on some optimality criterion (see e.g. Hastie and Tibshirani, 1990). If P-spline smoothing is applied, the penalized estimation is found to be equivalent to estimation and prediction in linear mixed models, as has been demonstrated in Wand (2003). This link has been further exploited in Ruppert et al. (2003) and Kauermann (2004). The same idea is also used in this paper, building up a connection between P-spline fitting of a two-way hazard model and a generalized linear mixed model (GLMM). Smoothing parameter selection then corresponds to multivariate variance component estimation in a GLMM. This has the important advantage that multidimensional smoothing parameter selection can be easily carried out without

complicated grid searching. This is an essential point for our model, since the number of smoothing parameters to be chosen is 2 for the baseline and for each covariate under investigation. Moreover, software developed for estimation of GLMMs can be used for fitting our model. For details and listing of code we refer to Ngo and Wand (2004) who demonstrate the use of Splus, R and SAS. Their results are readily extendable to the model fitted here. We also refer to Therneau et al. (2003) or Brezger et al. (2005) for software developments in a similar direction.

This paper is organized as follows. In Section 2 we describe the model and how this is fitted. In particular, a penalized backfitting is presented and linked to GLMM. Section 3 provides simulations and the data example. An outlook is given in Section 4.

## 2. Nonparametric two-way hazard model

### 2.1. P-spline estimation of hazard function

Assume we collect survival data having survival time $t$ and calendar time $c$ as quantities of interest. For each individual, we have the independent data pairs $(t_i, c_i, \delta_i, x_i)$ with $t_i$ the observed survival time, $\delta_i$ as censoring indicator and $c_i$ denoting the timepoint of failure. With $x_i$ we denote a set of covariates. For simplicity of presentation, we omit covariates $x_i$ for the moment and present our approach first for pure hazard function modelling. Based on the data we have $b_i = c_i - t_i$ as timepoint of "birth" and the hazard function is modelled as

$$h(t, b) = \exp\{\alpha_0(t, b)\}, \tag{1}$$

with $\alpha_0(.)$ as some smooth but otherwise unspecified function. Assuming additivity, we decompose $\alpha_0(.)$ to $\alpha_0(t, b) = \alpha_{t0}(t) + \alpha_{b0}(b)$. Based on hazard function (1) the log likelihood results to

$$\sum_{i=1}^{n}\left[\delta_i\{\alpha_{t0}(t_i) + \alpha_{b0}(b_i)\} - \exp\{\alpha_{b0}(b_i)\}\int_0^{t_i}\exp\{\alpha_{t0}(u)\}\,du\right]. \tag{2}$$

In classical proportional hazard models one now replaces the cumulated hazard function, that is the last component in (2), by a step function with steps at the observed timepoints and step heights as unknown parameters. This leads to Breslow's (1972) estimate and justifies the partial likelihood introduced by Cox (1972) as principle likelihood. We go a similar route, but replace the integral by a trapezia approximation with trapezoids constructed over the observed failure time points. Let therefore $0 = k_0, k_1, \ldots, k_K$ be the observed failure times and define $J_i$ as the index defined such that $k_{J_i}$ is the smallest knot larger than or equal to $t_i$, that is $k_{J_i-1} < t_i \leqslant k_{J_i}$. We then approximate

$$\begin{aligned}\int_0^{t_i}\exp\{\alpha_{t0}(u)\,du\} \approx\ & \exp\{\alpha_{t0}(0)\}(k_1 - k_0)/2 \\ & + \sum_{j=1}^{J_i-1}\exp\left\{\alpha_{t0}\left(k_j\right)\right\}\left(k_{j+1} - k_{j-1}\right)/2 \\ & + \exp\left\{\alpha_{t0}\left(k_{J_i}\right)\right\}\left(t_i - k_{J_i-1}\right)/2 \\ =\ & \sum_{j=0}^{J_i}\exp\left\{\alpha_{t0}\left(k_j\right) + o_j^{(i)}\right\},\end{aligned} \tag{3}$$

where $o_0^{(i)} = \log\{(k_1 - k_0)/2\}$, $o_j^{(i)} = \log\left\{\left(k_{j+1} - k_{j-1}\right)/2\right\}$ for $1 \leqslant j \leqslant J_i - 1$ and $o_{J_i}^{(i)} = \log\left(\left(t_i - k_{J_i-1}\right)/2\right)$. Since usually no information is available prior to the first event, we can also set the early hazard to zero and start the trapezoid integration at the first event time, that is, at $k_1 = \min(t_i, \delta_i = 1)$. The number $K$ of trapezoids used in approximation (3) clearly has an influence on the correctness of the fit, but if knots $k_l$ are placed at every observed failure time point, we take saturated information of our data. This is in particular a reasonable strategy, if survival times are clustered or measured on a discrete grid, like in our example where duration times are given in a months. Inserting (3) in (2) provides an approximation for the likelihood which should now be maximized with respect to both $\alpha_{t0}(.)$ and $\alpha_{b0}(.)$.

For fitting we replace the two functions by high-dimensional parametric curves, which are subsequently estimated in a penalized manner. This means we replace $\alpha_{t0}(t)$ by

$$\alpha_{t0}(t) = Z_{t0}(t)\beta_{t0} + B_{t0}(t)u_{t0},$$

with $Z_{t0}(t)$ as a low-dimensional basis in $t$. In the application below we set $Z_{t0}(t) = (1, t)$. This means we allow for a linear shape of the baseline hazard in an unpenalized form. In contrast, $B_{t0}(t)$ is a high-dimensional spline basis built e.g. from truncated polynomials. In our example we use truncated lines, i.e. $B_{t0}(t) = \left(\left(t - \tau_{t_1}\right)_+, \ldots, \left(t - \tau_{t_m}\right)_+\right)$ with knots $\tau_{t_1} < \cdots < \tau_{t_m}$ covering the range of observed failure times and $(t)_+ = t$ for $t > 0$ and zero otherwise. The number of knots $m$ is chosen in a lush and generous way, but apparently $m$ should be less than or equal to the number of observed failure time points. A more thorough investigation of how knots should be chosen in general is found in Ruppert (2002). In our example we worked with 15–20 dimensional basis. In the same way we fit $\alpha_{b0}(\cdot)$ by setting

$$\alpha_{b0}(b) = Z_{b0}(b)\beta_{b0} + B_{b0}(b)u_{b0}.$$

Again $Z_b(.)$ is low dimensional, where we set $Z_b(b) = b$. This means we allow for a linear change in an unpenalized form. Note that we attach the intercept to matrix $Z_{t0}(\cdot)$ and leave therewith $\alpha_{b0}(\cdot)$ without intercept. The spline basis $B_{b0}(.)$ is again chosen high dimensional. We employ truncated linear lines in the example below, that is, $B_{b0}(b) = \left(\left(b - \tau_{b_1}\right)_+, \ldots, \left(b - \tau_{b_q}\right)_+\right)$. Again, dimension $q$ should be chosen generously and knots $\tau_{b_l}, l = 1, \ldots, q$, should cover the range of observed birth dates. In our example, we use like above a 15–20 dimensional basis of truncated lines. The log likelihood (2) is now approximated by

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\left[\delta_{ij}\left(Z_{ij0}\beta_0 + B_{ij0}u_0\right) - \exp\left\{Z_{ij0}\beta_0 + B_{ij0}u_0 + o_j^{(i)}\right\}\right], \tag{4}$$

where $Z_{ij0} = \left(Z_{t0}\left(t_{ij}\right), Z_{b0}\left(b_i\right)\right)$, where $B_{ij0} = \left(B_{t0}\left(t_{ij}\right), B_{b0}\left(b_i\right)\right)$ and $t_{ij} = k_j$ for $j = 1, \ldots, J_i - 1$ and $t_{iJ_i} = t_i$. Coefficients are stacked to $\beta_0 = \left(\beta_{00}, \beta_{t0}, \beta_{b0}\right)$ and $u_0 = (u_{t0}, u_{b0})$ and the censoring indicator $\delta_{ij}$ takes values 1 if both, $j = J_i$ and $\delta_i = 1$, while $\delta_{ij}$ equals 0 otherwise. Note that (4) is the likelihood for Poisson variables $\delta_{ij}$ with intensity $\lambda(t, b) = \exp\left(\alpha_0\left(t, b\right)\right)$ and given offset $o_j^{(i)}$. Model (2) is now extended to incorporate smooth covariate effects. Let therefore $X_i = (1, x_i) = \left(x_{i0}, x_{i1}, \ldots, x_{ip}\right)$ denote the design matrix built from the intercept and $p$ covariates. The hazard function is modelled additively as

$$h\left(t, b, x_i\right) = \exp\left\{X_i\left[\boldsymbol{\alpha}_t(t) + \boldsymbol{\alpha}_b(b)\right]\right\},$$

where $\boldsymbol{\alpha}_t(t) = \left(\alpha_{t0}(t), \alpha_{t1}(t), \ldots, \alpha_{tp}(t)\right)^{\mathrm{T}}$ contains the smooth baseline and time dynamic covariate effects and analogous decomposition for $\boldsymbol{\alpha}_b(b) = \left(\alpha_{b0}(b), \alpha_{b1}(b), \ldots, \alpha_{bp}(b)\right)^{\mathrm{T}}$ compensating for calendar time effects. As before, $\alpha_{t0}(\cdot)$ and $\alpha_{b0}(\cdot)$ represent baseline smooth duration time and calendar time effects, respectively, while $\alpha_{tl}(t)$ mirrors the covariate effect of the $l$th covariate which varies with duration time, $l = 1, \ldots, p$. Accordingly, $\alpha_{bl}(\cdot)$ expresses smooth dynamics with calendar time. Like above we replace the smooth components for estimation by spline functions. This means we set

$$\alpha_{tl}(t) = Z_{tl}(t)\beta_{tl} + B_{tl}(t)u_{tl}.$$

Like above we assume a linear structure for $Z_{tl}(t)=(1, t)$ and let basis $B_{tl}(.)$ be constructed from truncated polynomials, in its most simple form resulting as truncated linear lines $B_{tl}(t) = \left(\left(t - \tau_{t_1}\right)_+, \ldots, \left(t - \tau_{t_m}\right)_+\right)$. Similarly, we replace $\alpha_{bl}(b)$ by $\alpha_{bl}(b) = Z_{bl}(b)\beta_{bl} + B_{bl}(b)u_{bl}$ where $Z_{bl}(.)$ is low dimensional and does not include the intercept, since this is contained in $Z_{tl}(.)$. In our application we chose $Z_{bl}(.) = b$ so that $\alpha_{bl}(b) = b\beta_{bl} + B_{bl}(b)u_{bl}$. If now $u_{bl}$ is set to zero, a linear trend in time results, that is covariate effects vary linearly with calendar time. The complete model leads now to the log likelihood

$$l(\boldsymbol{\beta}, \boldsymbol{u}) = \sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\left\{X_i\left[\mathbf{W}_t\left(t_{ij}\right)\boldsymbol{\theta}_t + \mathbf{W}_b\left(b_i\right)\boldsymbol{\theta}_b\right]\right\} - \exp\left\{X_i\left[\mathbf{W}_t\left(t_{ij}\right)\boldsymbol{\theta}_t + \mathbf{W}_b\left(b_i\right)\boldsymbol{\theta}_b\right] + o_j^{(i)}\right\}, \tag{5}$$

where $\mathbf{W}_t(t)$ is a block diagonal matrix with row matrices $\left(Z_{tl}\left(t_{ij}\right), B_{tl}\left(t_{ij}\right)\right)$ on its diagonal, $l = 0, 1, \ldots, p$. Accordingly $\mathbf{W}_b(b)$ is block diagonal with rows $\left(Z_{bl}\left(b\right), B_{bl}\left(b\right)\right)$, $l = 1, \ldots, p$, on the diagonal. Parameter vector $\theta_t$ decomposes to elements $\boldsymbol{\theta}_t = \left(\beta_{t0}, u_{t0}, \beta_{t1}, u_{t1}, \ldots, \beta_{tp}, u_{tp}\right)^{\mathrm{T}}$ and $\theta_b = \left(\beta_{b0}, u_{b0}, \beta_{b1}, u_{b1}, \ldots, \beta_{bp}, u_{bp}\right)^{\mathrm{T}}$. It is worth pointing out that in principle trapezoid integration is not necessary if the integral in (2) can be calculated analytically. This would be the case when replacing $\alpha_t(\cdot)$ by simple splines. Analytic integration has however to be bought for awkward and clumsy implementation and in fact the nice link to the Poisson model is lost. Since this link will be of importance later on we do not pursue analytic integration further on.

Direct maximization of (4) would provide unsatisfactory results $\beta_{bp}$ since the dimension of the basis matrices $B_t(.)$ and $B_b(.)$ is large, which is necessary to capture the unknown underlying functional structure. The idea is now to penalize spline coefficients $u$ so that smooth, unwiggled estimates result for $\boldsymbol{\alpha}(.)$. This is achieved by penalizing coefficients $u_{tl}$ with $\frac{1}{2}\lambda_{tl}u_{tl}\mathbf{D}_{tl}u_{tl}$, where $\mathbf{D}_{tl}$ is a penalty matrix chosen appropriately to spline basis $B_{tl}(.)$. For truncated polynomials a suitable choice for $\mathbf{D}_{tl}$ is the identity matrix, as suggested in Ruppert et al. (2003). After some calculation it can be shown that this choice is similar to the difference based proposal by Eilers and Marx (1996). The parameter $\lambda_{tl}$ steers the amount of penalty and is therewith playing the role of a smoothing parameter. Apparently, $\lambda_{tl}$ should be chosen adequately, which is discussed in the next section. Extending the idea to the remaining coefficients $u_{bl}$, $l = 0, 1, \ldots, p$, we get the penalized likelihood written in matrix notation as

$$l_p\left(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\lambda}_t, \boldsymbol{\lambda}_b\right) = l(\boldsymbol{\beta}, \mathbf{u}) - \tfrac{1}{2}\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{\Lambda}_t\mathbf{D}_t\boldsymbol{\theta}_t - \tfrac{1}{2}\boldsymbol{\theta}_b^{\mathrm{T}}\boldsymbol{\Lambda}_b\mathbf{D}_b\boldsymbol{\theta}_b, \tag{6}$$

where $\mathbf{D}_t = \mathrm{diag}\left(0, D_{tl}\right)$ and $\mathbf{D}_b = \mathrm{diag}\left(0, D_{bl}\right)$ and $\boldsymbol{\lambda} = \left(\lambda_{tl}, \lambda_{bl}\right), l = 1, \ldots, p$. Accordingly, $\boldsymbol{\Lambda}_t$ is a diagonal matrix containing the penalty parameters $\lambda_{tl}$ in matching dimension to $\theta_{tl}, l = 0, 1, \ldots, p$. In the same way we construct $\boldsymbol{\Lambda}_b$.

### 2.2. Penalized backfitting

In principle, estimation of (6) can be carried out in a straightforward manner by differentiating (6) with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$. This may however be numerically expensive, in particular if the number of covariates is large. We therefore propose a backfitting routine as alternative. Assume first that $\boldsymbol{\alpha}_t(\cdot) = \mathbf{W}_t(\cdot)\boldsymbol{\theta}_t$ is given and estimation is supposed to be carried out over $\boldsymbol{\alpha}_b(\cdot) = \mathbf{W}_b(\cdot)\boldsymbol{\theta}_b$ only. Then, the penalized likelihood equals

$$l_{pb}(\boldsymbol{\theta}_b, \boldsymbol{\lambda}_b) = \sum_{i=1}^{n}\delta_i X_i\mathbf{W}_b\left(b_i\right)\boldsymbol{\theta}_b - \exp\left\{X_i\mathbf{W}_b\left(b_i\right)\boldsymbol{\theta}_b + o_{bi}\right\} - \frac{1}{2}\boldsymbol{\theta}_b^{\mathrm{T}}\boldsymbol{\Lambda}_b\mathbf{D}_b\boldsymbol{\theta}_b, \tag{7}$$

with $o_{bi} = \log\sum_{j=1}^{J_i}\exp\left\{X_i\mathbf{W}_t\left(t_{ij}\right)\boldsymbol{\theta}_t + o_j^{(i)}\right\}$. Note that (7) equals a simple penalized likelihood for the $n$ Poisson data $\delta_i, i = 1, \ldots, n$. Differentiation with respect to $\boldsymbol{\theta}_b$ provides the estimating equation

$$0 = s_{pb}\left(\boldsymbol{\theta}_b, \boldsymbol{\lambda}_b\right) = \sum_{i=1}^{n}\mathbf{W}_b^{\mathrm{T}}\left(b_i\right)X_i^{\mathrm{T}}\left\{\delta_i - \exp\left(X_i\mathbf{W}_b\left(b_i\right)\boldsymbol{\theta}_b + o_{bi}\right)\right\} - \boldsymbol{\Lambda}_b\mathbf{D}_b\boldsymbol{\theta}_b, \tag{8}$$

and the Fisher-type matrix with respect to $\boldsymbol{\theta}_b$ is defined through

$$\mathbf{I}_{pb}\left(\boldsymbol{\theta}_b, \boldsymbol{\lambda}_b\right) = \sum_{i=1}^{n}\mathbf{W}_b^{\mathrm{T}}\left(b_i\right)X_i^{\mathrm{T}}X_i\mathbf{W}_b(b_i)\exp\left(X_i\mathbf{W}_b\left(b_i\right)\boldsymbol{\theta}_b + o_{bi}\right) + \boldsymbol{\Lambda}_b\mathbf{D}_b. \tag{9}$$

Solving (8) gives the first step in the backfitting procedure. Exchanging the role of $\boldsymbol{\alpha}_t(\cdot)$ and $\boldsymbol{\alpha}_b(\cdot)$ leads to the second backfitting step. We now consider $\boldsymbol{\alpha}_b(\cdot) = \mathbf{W}_b(\cdot)\boldsymbol{\theta}_b$ as given leading to the penalized likelihood

$$l_{pt}(\boldsymbol{\theta}_t, \boldsymbol{\lambda}_t) = \sum_{i=1}^{n}\sum_{j=1}^{J_i}\left[\delta_{ij}X_i\mathbf{W}_t\left(t_{ij}\right)\boldsymbol{\theta}_t - \exp\left\{X_i\mathbf{W}_t\left(t_{ij}\right)\boldsymbol{\theta}_t + o_{tij}\right\}\right] - \frac{1}{2}\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{\Lambda}_t\mathbf{D}_t\boldsymbol{\theta}_t, \tag{10}$$

with $o_{tij} = X_i \mathbf{W}_b(b_i)\,\boldsymbol{\theta}_b + o_j^{(i)}$. In this case, (10) is the penalized likelihood for Poisson data $\delta_{ij}$ with offset $o_{tij}$, $j = 1, \ldots, J_i$, $i = 1, \ldots, n$. Estimating equations are found via

$$0 = s_{pt}(\boldsymbol{\theta}_t, \boldsymbol{\lambda}_t) = \sum_{i=1}^{n} \sum_{j=1}^{J_i} \mathbf{W}_t^{\mathrm{T}}(t_{ij})\, X_i^{\mathrm{T}} \left\{ \delta_{ij} - \exp\left( X_i \mathbf{W}_t(t_{ij})\,\boldsymbol{\theta}_t + o_{tij} \right) \right\} - \boldsymbol{\Lambda}_t \mathbf{D}_t \boldsymbol{\theta}_t, \tag{11}$$

with Fisher matrix

$$\mathbf{I}_{pt}(\boldsymbol{\theta}_t, \lambda_t) = \sum_{i=1}^{n} \sum_{j=1}^{J_i} \mathbf{W}_t^{\mathrm{T}}(t_{ij})\, \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_i \mathbf{W}_t(t_{ij}) \exp\left( \mathbf{X}_i \mathbf{W}_t(t_{ij})\,\theta_t + o_{tij} \right) + \boldsymbol{\Lambda}_t \mathbf{D}_t.$$

The backfitting algorithm results now by solving (8) and (11) gradually. In particular, the procedure is as follows

(a) Start with an initial estimate for $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_b$, obtained e.g. by setting $\mathbf{u}_t \equiv 0$ and $\mathbf{u}_b \equiv 0$ and estimating $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_b$ parametrically using a generalized linear model fitted to data $\delta_{ij}$. We denote the resulting estimates by $\hat{\boldsymbol{\theta}}_t^{(0)}$ and $\hat{\boldsymbol{\theta}}_b^{(0)}$.

(b) In the $r$th loop of the algorithm we update the estimate for $\boldsymbol{\theta}_b$ by the one step Fisher scoring

$$\hat{\boldsymbol{\theta}}_b^{(r)} = \hat{\boldsymbol{\theta}}_b^{(r-1)} + \mathbf{I}_{pb}^{-1}\left( \hat{\boldsymbol{\theta}}_b^{(r-1)}, \boldsymbol{\lambda}_b \right) s_{pb}\left( \hat{\boldsymbol{\theta}}_b^{(r-1)}, \boldsymbol{\lambda}_b \right),$$

with offset $o_{bi}$ in (8) calculated from $\boldsymbol{\theta}_t^{(r-1)}$.

(c) Exchanging the roles of $t$ and $b$ we update $\boldsymbol{\theta}_t^{(r-1)}$ by

$$\hat{\boldsymbol{\theta}}_t^{(r)} = \hat{\boldsymbol{\theta}}_t^{(r-1)} + \mathbf{I}_{pt}^{-1}\left( \hat{\boldsymbol{\theta}}_t^{(r-1)}, \boldsymbol{\lambda}_t \right) s_{pt}\left( \hat{\boldsymbol{\theta}}_t^{(r-1)}, \boldsymbol{\lambda}_t \right),$$

where offset $o_{tij}$ in (11) is now calculated from $\boldsymbol{\theta}_b^{(r)}$.

(d) Iterating between (b) and (c) leads to the backfitting routine.

For the final estimate we can calculate the variance of our fitted smooth curve in the following way. With standard asymptotic arguments, assuming the number of individuals to grow but leaving the number of knots to be finite, we obtain the asymptotic behavior for $\hat{\boldsymbol{\theta}}$ as solution of (8) and (11) as

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \sim N\left( 0, I_p(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1} I_p(\boldsymbol{\theta}, \boldsymbol{\lambda} = \mathbf{0}) I_p^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \right), \tag{12}$$

with $\mathbf{I}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is the joint Fisher matrix defined through

$$\mathbf{I}_p(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} \sum_{j=1}^{J_i} \mathbf{W}_{ij}^{\mathrm{T}} X_i^{\mathrm{T}} X_i \mathbf{W}_{ij} \exp\left( X_i \mathbf{W}_{ij} \boldsymbol{\theta} + o_j^{(i)} \right) + \mathrm{diag}\left( \boldsymbol{\Lambda}_t \mathbf{D}_t, \boldsymbol{\Lambda}_b \mathbf{D}_b \right),$$

where $\mathbf{W}_{ij} = \left( \mathbf{W}_t(t_{ij}), \mathbf{W}_b(b_i) \right)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_t, \boldsymbol{\theta}_b)$. Note that the variance in (12) can be rewritten as

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \mathbf{I}_p^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \mathbf{I}_p^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda})\, \mathrm{diag}\left( \boldsymbol{\Lambda}_t \mathbf{D}_t, \boldsymbol{\Lambda}_b \mathbf{D}_b \right) \mathbf{I}_p^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda}).$$

Note that since we are interested in variance estimates for estimates $\hat{\alpha}_{tl}(.)$ or $\hat{\alpha}_{bl}(.)$ we are only interested in the two block diagonals of the Fisher matrix referring to elements in $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_b$. It is now a simple step to obtain pointwise confidence intervals for $\hat{\alpha}_{tl}(t)$ via the variance estimate

$$\mathrm{Var}\left( \hat{\alpha}_{tl}(t) \right) = (Z_{tl}(t), B_{tl}(t))\, \mathrm{Var}\left( \left( \hat{\beta}_{tl}, \hat{u}_{tl} \right)^{\mathrm{T}} \right) \left( Z_{t,l}^{\mathrm{T}}(t), B_{tl}^{\mathrm{T}}(t) \right)^{\mathrm{T}}. \tag{13}$$

In the same way we get the variance for $\alpha_b(\cdot)$.

### 2.3. Mixed model representation

We can rewrite the penalized estimation above as prediction in a GLMM by formulating the penalty as a priori distribution on the spline coefficients. This link has been worked out in Wand (2003) for normal response data, and has been explored further in Ruppert et al. (2003) (see also Kauermann, 2004). In particular, the connection between smoothing and mixed models is advantageous, since the smoothing parameter is playing the role of a random effect variance in the mixed model formulation. This in turn allows for maximum likelihood estimation and a simple way to obtain data-driven smoothing parameters. This approach is also employed here. Additionally, we use the backfitting idea to simplify the numerical effort. This means for fitting $\alpha_b(.)$ we assume that $\alpha_t(.)$ is given or fixed at its current estimate. Assuming spline coefficients as given, the first component in the penalized likelihood (7) equals the likelihood for the mixed model

$$\delta_i|\mathbf{u}_b, \alpha_t(\cdot) \sim \text{Poisson}\left(\exp\left\{X_i \mathbf{W}(b_i)\,\boldsymbol{\theta}_b + o_{bi}\right\}\right).$$

Assuming now that $\mathbf{u}_b$ is random induces the penalty as a priori distribution

$$u_{bl} \sim N\left(0, \lambda_{bl}^{-1} D_{bl}^{-}\right), \quad l = 0, \ldots, p, \tag{14}$$

with $D_{bl}^{-}$ as the (generalized) inverse of $D_{bl}$. Following Breslow and Clayton (1993) we integrate out $\mathbf{u}_b$ by Laplace approximation leading to the marginal likelihood, here conditioned on $\alpha_t(\cdot)$. It is not difficult to see that this marginal likelihood resembles (7) and maximizing this with respect to $\boldsymbol{\beta}_b$ and $\mathbf{u}_b$ leads to the estimate defined as solution of (8). The maximized approximate marginal likelihood, based on the Laplace approximation, results now to

$$\begin{aligned} l_{mb|t}(\boldsymbol{\lambda}_b) &\approx -\frac{1}{2}\log\left|\tilde{\boldsymbol{\Lambda}}_b^{-1}\tilde{\mathbf{D}}_b^{-}\right| + l_{pb}\left(\hat{\boldsymbol{\theta}}_b, \boldsymbol{\lambda}_b\right) - \frac{1}{2}\log\left|-\frac{\partial^2 l_{pb}\left(\hat{\boldsymbol{\theta}}_b, \boldsymbol{\lambda}_b\right)}{\partial \mathbf{u}_b \partial \mathbf{u}_b^{\mathrm{T}}}\right| \\ &= l_{pb}\left(\hat{\boldsymbol{\theta}}_b, \boldsymbol{\lambda}_b\right) - \frac{1}{2}\log\left|\tilde{\mathbf{I}}_{pb}\left(\hat{\boldsymbol{\theta}}_b, \boldsymbol{\lambda}_b\right)\tilde{\boldsymbol{\Lambda}}_b^{-1}\tilde{\mathbf{D}}_b^{-}\right|, \end{aligned} \tag{15}$$

where $\hat{\boldsymbol{\theta}}_b$ is the maximizer of $l_{pb}(\boldsymbol{\theta}_b, \boldsymbol{\lambda}_b)$ and $\tilde{\mathbf{D}}_b$ taken those columns and rows of $\mathbf{D}_b$ is the submatrix of $\mathbf{D}_b$ with nonzero diagonal elements. Note that these are the elements matching to coefficients $\mathbf{u}_b$. Accordingly $\tilde{\boldsymbol{\Lambda}}_b$ results as submatrix of $\boldsymbol{\Lambda}_b$ built from components $\boldsymbol{\lambda}_b$. In the same way $\tilde{\mathbf{I}}_{pb}(.)$ is the submatrix of the Fisher matrix given in (9) with elements corresponding to $\mathbf{u}_b$. Observing the structure of $\tilde{\boldsymbol{\Lambda}}_b$, we can now maximize (15) to obtain an estimate for the $l$th component of $\boldsymbol{\lambda}_b$. Differentiation yields

$$\frac{1}{\hat{\lambda}_{bl}} = \frac{\operatorname{tr}\left\{\left(\tilde{\mathbf{I}}_{pb}\left(\hat{\boldsymbol{\theta}}_b, \lambda_b\right)^{-1}\right)_{ll} D_{bl}\right\} + \hat{u}_{bl}^{\mathrm{T}} D_{bl} \hat{u}_{bl}}{m_{bl}}, \tag{16}$$

with $m_{bl}$ as dimension of spline basis $B_{bl}(\cdot)$ and subscript $ll$ indicating the $l$th block diagonal of the inverse Fisher matrix. Note that (16) is not an analytic solution by itself, since the right-hand side depends on $\boldsymbol{\lambda}_b$ as well, explicitly through $\tilde{\mathbf{I}}_{pb}(\cdot)$ as well as implicitly through $\hat{\boldsymbol{\theta}}_b$. However, (16) can be used in an interactive manner keeping the components on the right-hand side as fixed and updating $\hat{\boldsymbol{\lambda}}_{bl}$ on the left-hand side. It can be shown that this corresponds to a Newton-type algorithm as motivated in Krivobokova and Kauermann (2005).

In complete analogy we obtain a GLMM for $\alpha_t(\cdot)$, now keeping $\boldsymbol{\theta}_b$ as fixed. This means, with the backfitting idea we obtain the GLMM corresponding to (14) which is

$$\begin{aligned} &\delta_{ij}\left|\mathbf{u}_t, \alpha_b(\cdot) \sim \text{Poisson}\left(\exp\left\{X_i \mathbf{W}_t\left(t_{ij}\right)\boldsymbol{\theta}_t + o_{tij}\right\}\right),\right. \\ &u_{tl} \sim N\left(0, \lambda_{tl}^{-1} D_{tl}^{-}\right), \quad l = 1, \ldots, p. \end{aligned} \tag{17}$$

Applying Laplace approximation we obtain the marginal likelihood

$$l_{mt|b}(\lambda_t) \approx l_{pt}\left(\hat{\boldsymbol{\theta}}_t, \boldsymbol{\lambda}_t\right) - \tfrac{1}{2}\log\left|\tilde{\mathbf{I}}_{pt}\left(\hat{\boldsymbol{\theta}}_t, \boldsymbol{\lambda}_t\right)\tilde{\boldsymbol{\Lambda}}_t\tilde{\mathbf{D}}_t\right|,$$

with obvious definition for the tilde notation. Like above this suggests the approximate likelihood estimates

$$\frac{1}{\hat{\lambda}_{tl}} = \frac{\operatorname{tr}\left\{\left(\tilde{\mathbf{I}}_{pt}\left(\hat{\boldsymbol{\theta}}_t, \lambda_t\right)^{-1}\right)_{ll} D_{tl}\right\} + \hat{u}_{tl}^{\mathrm{T}} D_{tl}\hat{u}_{tl}}{m_{tl}}, \tag{18}$$

with $m_{tl}$ as dimension of $B_{tl}(\cdot)$. The backfitting algorithm from above can now be extended by updating the estimates for $\boldsymbol{\lambda}_t$ and $\boldsymbol{\lambda}_b$ in each step. This is achieved by supplementing (16) to step (b) and (16) to step (c).

## 3. Data example and simulations

### 3.1. Simulation

We run a small simulation study to explore the performance of our routine. We therefore simulate survival data, where we use discrete survival times as they occur in our data example. Based on the Poisson-type model we simulate data with hazard function $\exp\{\alpha_{t0}(t) + \alpha_{b0}(b) + x\,(\alpha_{t1}(t) + \alpha_{b1}(b))\}'$ with $x$ as a binary covariate with $P(x=1)=0.7$ and $P(x=0)=0.3$. The explicit functions are specified to $\alpha_{t0}(t) = -2.5 - t/30$, $\alpha_{b0}(b) = 3(b/50)^2 - 2.5b/50$, $\alpha_{t1}(t) = 0.5 + (t/25)^2$, $\alpha_{b1}(b) = -b/60$, where $t$ ranges from 1 to 30, with survival times beyond 30 taken as censored. The birth time is drawn uniform on [0, 50] and censoring is applied for survival times exceeding $t > 30$ and calender time exceeding 80, that is if $t + b > 80$. We simulate $n = 1000$ observations and fit the model with the backfitting procedure described above. In each simulation, the smoothing parameter is chosen data driven exploiting the link to generalized linear mixed models. Fig. 2 shows the true function and simulated coverage intervals showing pointwise, that is for each timepoint $t$ and $b$, respectively, the 5%, 50% and 95% quantile of the simulated estimates based on 200 simulations. These graphs suggest that in this example bias is not a serious problem. For each simulation we can calculate confidence bands based on (13) and corresponding formula of $\alpha_{bl}(\cdot)$. To assess the goodness of the variance estimate, we investigate the simulated coverage probability of estimated confidence intervals. To do so we check in each simulation whether the estimated confidence interval covers the true underlying function. Fig. 3 shows the simulated coverage probability, that is the proportion of simulation in which at a fixed point $t$ or $b$, respectively, the estimated confidence interval contains the true function. The nominal value is 95% based on plus/minus two times the standard deviation. Overall, the variance estimation seems acceptable, even though for $\alpha_{b0}(\cdot)$ there is undercoverage at for small values of $b$. We think that this is due to a small bias for small values of $b$ occurring due to the spline basis used. We worked with truncated lines while the function is quadratic. The functional shape of $\alpha_{b0}(\cdot)$ is however well captured so that we do feel not too discouraged by the poor performance of the variance estimates for $\hat{\alpha}_{b0}(\cdot)$. Note also that $\alpha_{b0}(0) = 0$ by construction and therewith $\operatorname{Var}\left(\hat{\alpha}_{b0}(0)\right) = 0$. We also experimented with different functions for $\alpha_{b0}(\cdot)$ as well as different splines. We observed that the coverage probability of $\alpha_{bl}(\cdot)$ can be improved if the spline used naturally captures the true shape of the function. This holds for $\alpha_{tl}(\cdot)$ as well, but by far weaker. Given the fact however that trends over calender time are usually less strong than trends over survival time, we still feel most comfortable with the truncated linear basis we used, even though, of course, this issue can be further disputed.

### 3.2. Unemployment data

We now analyze the unemployment data referred to in the introduction. Based on the German socio economic panel we consider unemployment spells from 4020 individuals who became unemployed between 1983 and 2000 and were domiciled in West Germany. Generally, for individuals in the panel with more spell of unemployment we randomly chose one of their spells and ignored the others. This guarantees independence of our observations. The empirical distribution of the beginning of unemployment is shown in Fig. 4. As covariates we consider

- $x_1$: foreigner (1 for Foreigner, 0 for German);

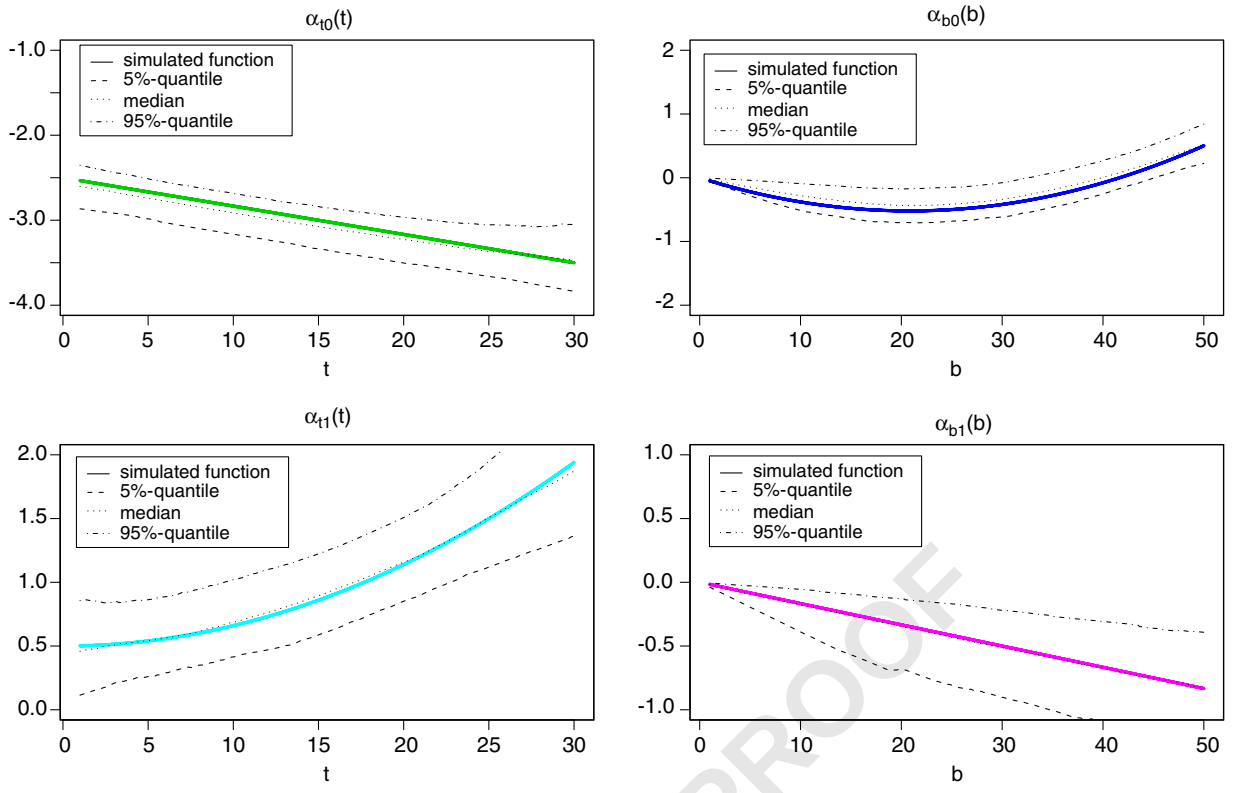


Fig. 2. Simulated confidence intervals with pointwise 5%, 50% and 95% quantiles based on 200 simulations. The function is shown as thick line.



- $x_2$: female (1 for female, 0 for male);
- $x_3$: age $\leqslant 25$ (1 if younger than 25, 0 otherwise when getting unemployed);
- $x_4$: age $\geqslant 50$ (1 if older than 50, 0 otherwise when getting unemployed);
- $x_5$: no education (1 if individual has no professional education, 0 otherwise);
- $x_6$: higher education (1 if person has university or comparable degree, 0 otherwise).

Based on coding of variables "no education" and "high education" the reference category are individuals with apprenticeship or comparable education but without university degree. Unemployment duration is censored at 36 months which counts as threshold to long-term unemployment. Moreover, as event we count the return to full time occupation while any other occupation (half-time, retirement, continuing education, etc.) is taken as censored. Our analysis pursues a macroeconomic viewpoint to investigate how conditions have changed with calendar time, that is with different economic circumstances. Figs. 5 and 6 show baseline and covariate effects for the data at hand. The upper left plot in Fig. 5 shows the baseline $\alpha_{b0}(t)$. Clearly, the chances of returning to professional life decrease with duration of unemployment. Moreover, over the years, the chances reduce even though this effect does not occur to be significant. The effect of nationality varies with duration of unemployment, as can be seen from the plot in the second row, first column in Fig. 5. In the first months of unemployment, Foreigners have lower probability of finding a new job. This effect vanishes however and changes signs, even though it does not show significant behavior later on. Moreover, there is no evidence that the effect of nationality changes with calendar time. Looking now at gender we see that females generally have less chances of finding a new job, regardless of their unemployment duration. Hence, gender has a proportional effect on the hazard. However, over the years, the negative effect of gender has reduced. Next we consider the effect of age. As can be seen, younger unemployed people have higher chances of finding a new job while individuals aged 50 or higher reduce their chances. For young unemployed workers, the positive age effect increased in the eighties but decreased and changed sign in the nineties. The latter effect is however not significant.

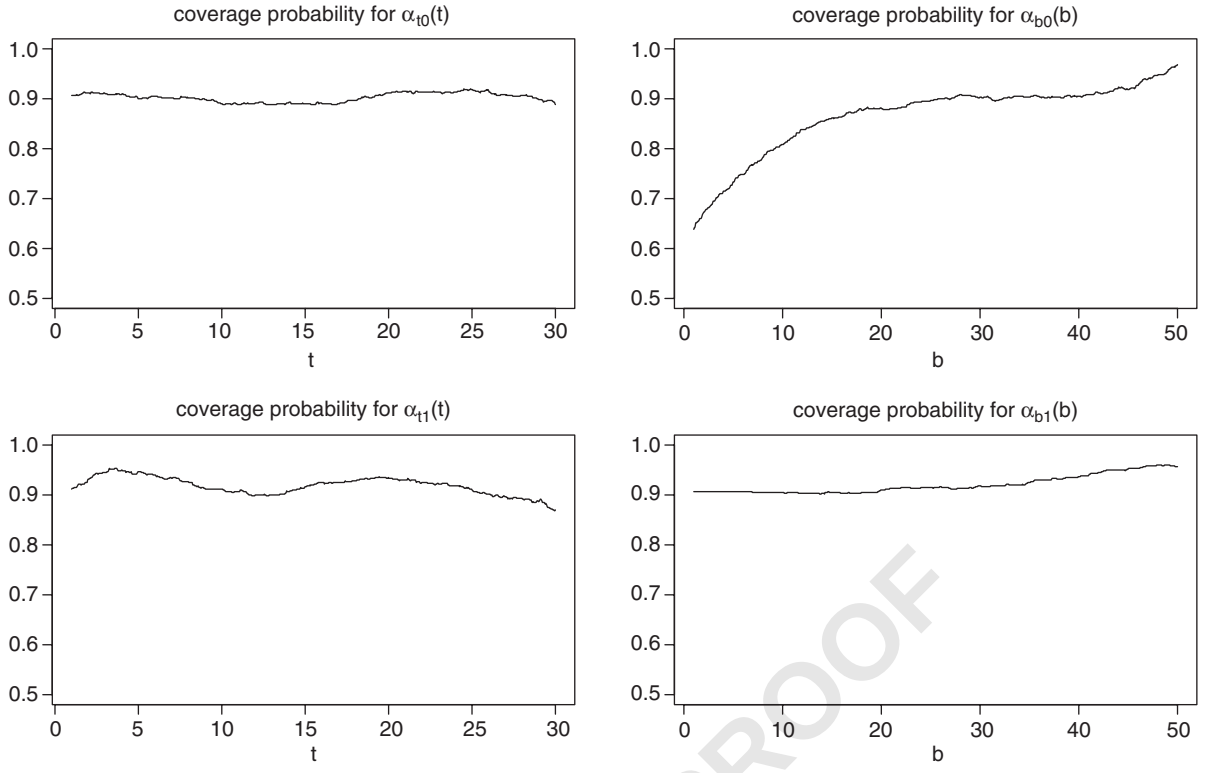


Fig. 3. Simulated coverage probability.

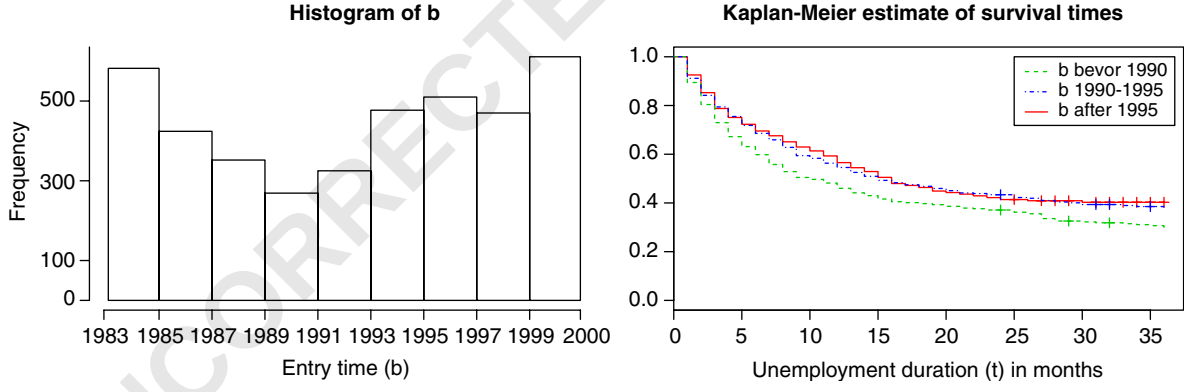


Fig. 4. Histogram for start of unemployment and Kaplan–Meier estimates for duration of unemployment grouped in three time intervals.



Moreover, for older individuals the chance for reemployment decreases over the years but in no significant manner. Finally, education has a positive effect on finding a new job regardless of the duration of unemployment. This can be shown from the bottom left plots in Fig. 6. A significant variation with calendar time was not observed for the education effects. Overall, it seems necessary to allow the covariate effects to vary with calender time as well as duration time of unemployment.

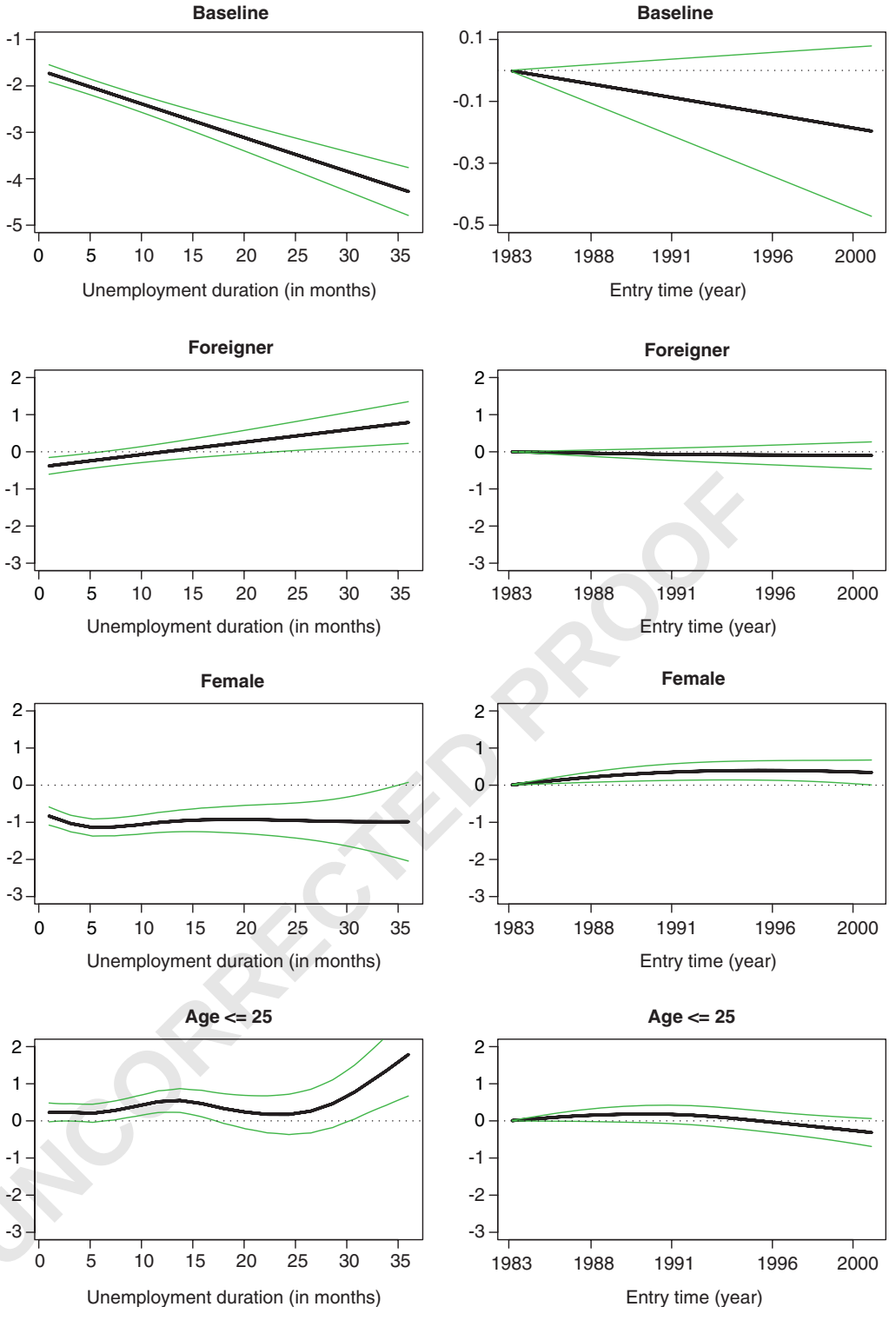




Fig. 5. Functional shape of $\alpha_{tl}(\cdot)$ (left-hand side) and $\alpha_{bl}(\cdot)$ (right-hand side) for unemployment data.

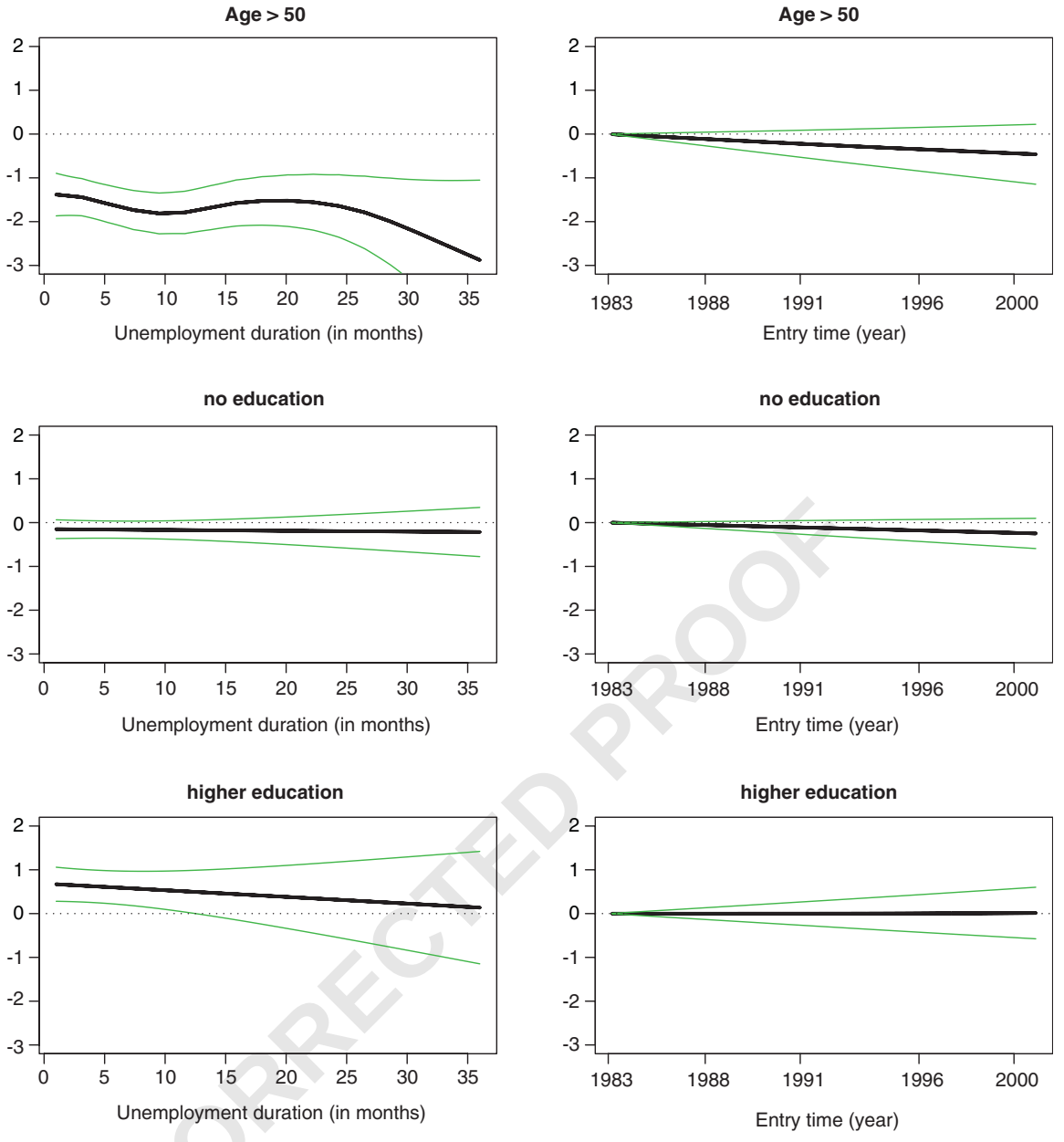


Fig. 6. Continuation of Fig. 5.

## 4. Discussion

The paper demonstrates how calendar time can be included in duration time modelling if the observed birth times have a wide range compared to the duration time. The nonparametric approach based on penalized splines easily allows to fit such data with flexible smooth structures in both, calender time and survival time. The delicate issue of smoothing parameter selection can be elegantly solved by linking penalized spline smoothing to mixed models. This shows advantageous given the number of smoothing parameters to be chosen. Fitting can numerically benefit from a backfitting idea so that the model is fitted easily. The modelling exercise can be extended or tackled in various alternative ways. First, one can include interaction effects both, between categorical covariates as well as between duration time

and calendar time. In the simplest case this can be done multiplicatively by including $x \cdot b$, for instance, in the model. Alternatively, one could include $b$ parametrically in the model and then check for interactions with duration time. Both approaches are somewhat ad hoc and a more coherent modelling framework seems worthwhile to be developed.

## References

Anderson, W., 1991. Continuous-time Markov Chains: An Applications-Oriented Approach. Springer, New York.
Breslow, N.E., 1972. Comment on "regression and life tables" by D.R. Cox. J. Roy. Statist. Soc. Ser. B 34, 216–217.
Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed model. J. Amer. Statist. Assoc. 88, 9–25.
Brezger, A., Kneib, T., Lang, S., 2005. J. Statist. Software 14 (11),
Cai, Z., Sun, Y., 2003. Local linear estimation for time-dependent coefficients in cox's regression models. Scand. J. Statist. 30, 93–111.
Cai, T., Hyndman, R., Wand, M., 2002. Mixed model-based hazard estimation. J. Comput. Graphical Statist. 11, 784–798.
Cox, D.R., 1972. Regression models and life tables (with discussion). J. Roy. Statist. Soc. Ser. B 34, 187–220.
Cox, D., Farewell, V., 1979. A note on multiple time scales in life testing. J. Roy. Statist. Soc. Ser. C 28, 73–75.
Efron, B., 2002. The two-way proportional hazards model. J. Roy. Statist. Soc. Ser. B 64, 899–909.
Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties. Statist. Sci. 11 (2), 89–121.
Fan, J., Gijbels, I., King, M., 1997. Local likelihood and local partial likelihood in hazard regression. Ann. Statist. 25, 1661–1690.
Francis, B., Pritchard, J., 1998. Bertin, lexis and the graphical representation of event histories. Bull. Comite Francais de Cartographie 156, 80–87.
Gray, R.J., 1992. Flexible methods for analyzing survival data using splines, which applications to breast cancer prognosis. J. Amer. Statist. Assoc. 87, 942–951.
Gray, R.J., 1994. Spline-based tests in survival analysis. Biometrics 50, 640–652.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, London.
Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. J. Roy. Statist. Soc. Ser. B 55, 757–796.
Kauermann, G., 2004. A note on smoothing parameter selection for penalised spline smoothing. J. Statist. Plann. Inference 127, 53–69.
Kauermann, G., 2005. Penalised spline fitting in multivariable survival models with varying coefficients. Comput. Statist. Data Anal. 49, 169–186.
Keiding, N., 1990. Statistical inference in the lexis diagram. Philos. Trans. Roy. Soc. London A 332, 487–509.
Kooperberg, C., Stone, C., Troung, Y., 1995. Hazard regression. J. Amer. Statist. Assoc. 90, 78–94.
Krivobokova, T., Kauermann, G., 2005. A short note on penalized spline smoothing with correlated errors. Technical report.
Ngo, L., Wand, M., 2004. Smoothing and mixed models software. J. Statist. Software 9 (1),
O'Sullivan, F., 1988. Nonparametric estimation of relative risk using splines and cross-validation. SIAM J. Sci. Statist. Comput. 9, 531–542.
Ruppert, D., 2002. Selecting the number of knots for penalized splines. J. Comput. Graphical Statist. 11, 735–757.
Ruppert, D., Wand, M., Carroll, R., 2003. Semiparametric Regression. Cambridge University Press, Cambridge.
Therneau, T.M., Grambsch, P.M., Pankratz, V.S., 2003. Penalized survival models and frailty. J. Comput. Graphical Statist. 12, 156–175.
Wand, M., 2003. Smoothing and mixed models. Comput. Statist. 18, 223–249.
Zucker, D., Karr, A., 1990. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. Ann. Statist. 18, 329–353.