Fahrmeir, Klinger:

A nonparametric multiplicative hazard model for event history analysis

Sonderforschungsbereich 386, Paper 12 (1995)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

A nonparametric multiplicative hazard model for event history analysis

LUDWIG FAHRMEIR[1]
AND ARTUR KLINGER[2]
Institut für Statistik, Universität München, Germany

SUMMARY

A major issue in exploring and analyzing life history data with multiple states and events is the development and availability of flexible methods that allow simultaneous incorporation and estimation of baseline hazards, detection and modelling of nonlinear functional forms of covariates and time-varying effects, and the possibility to include time-dependent covariates. In this paper we consider a nonparametric multiplicative hazard model that takes into account these aspects. Embedded in the counting process approach, estimation is based on penalized likelihoods and splines. The methods are illustrated by two real data applications, one to a more conventional survival data set with two absorbing states, and one to more complex sleep-electroencephalography data with multiple recurrent states of sleep.

Keywords: Counting processes; Life history data; Nonlinear effects; Penalized likelihood; Sleep patterns; Smoothing; Transition intensities; Varying coefficients.

[1]email: fahrmeir@stat.uni-muenchen.de

[2]email: artur@stat.uni-muenchen.de

1. Introduction

Cox's proportional hazard model is generally used as the standard tool for survival data analysis in studies where the effect of risk factors or covariates on the time until occurrence of a certain event is of prime interest. The hazard rate is written in semiparametric multiplicative form

$$\alpha(t; z_1, \ldots, z_p) = \alpha_0(t) \exp(\beta_1 z_1 + \ldots + \beta_p z_p),$$

where the baseline hazard rate $\alpha_0(t)$ is left unspecified and is estimated separately if necessary. Through the choice of a parametric exponential risk function for the second factor, covariates $z_1, \ldots, z_p$, which may also be time-dependent, act multiplicatively on the hazard rate.

In a number of applications there is a need for extending and further developing this basic model with respect to several aspects, such as allowing more flexible functional forms for covariates, inclusion of time-varying effects, thereby dropping the proportional hazards assumption, simultaneous estimation of baseline hazards and covariate effects and incorporation of unobserved heterogeneity. Increased flexibility becomes even more important in applications to more complex event history data as considered in this paper. To illustrate the methods by a simple example, we use a data set on survival with malignant melanoma, presented and analyzed in Andersen, Borgan, Gill and Keiding (1993). Patients were followed after operation, and survival times were recorded distinguishing 'death due to malignant melanoma' and 'death due to other causes'. Thus, transition rates α_1 and α_2 for these absorbing states and the influence of covariates like sex, thickness of tumor etc. are of interest.

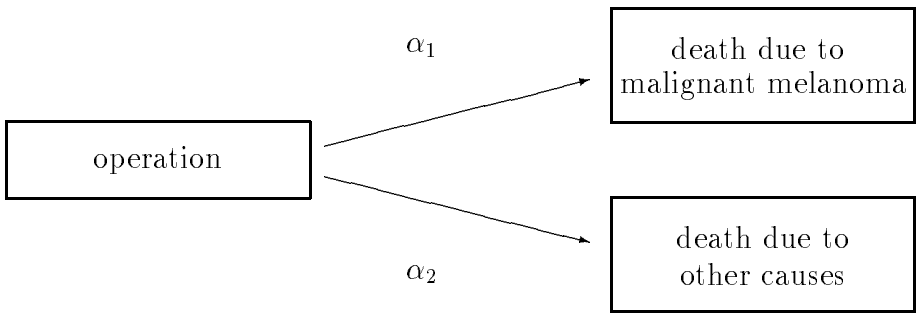

Figure 1: Transitions after operation for malignant melanoma

Our main application deals with sleep-electroencephalography (EEG) data, recording various stages of sleep. The data are described in more detail in Section 2. Here we consider only the three recurrent states rapid eye movement (REM) sleep, non rapid eye movement (NREM) sleep and AWAKE, following the diagram of Figure 2:

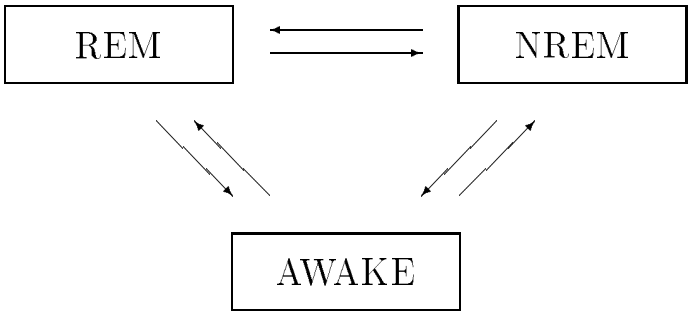

Figure 2: Transitions between sleep stages

Various extensions of the basic Cox model, mostly for survival data, have been developed with regard to the aspects mentioned above. First, the simple parametric form of the exponential risk function may not be suitable. Using local or penalized partial likelihood approaches, Hastie and Tibshirani (1986, 1990) and O'Sullivan (1988) model and estimate the effect of continuous covariates nonparametrically, replacing $\beta_j z_j$ by a smooth function $f_j(z_j)$. This approach is further extended in Hastie and Tibshirani (1993) to allow for varying effects of the form $\beta_j(x)z_j$, where $\beta_j(x)$ is a smooth function of some

covariate x. Viewing time t as a covariate, time-varying effects $\beta_j(t)z_j$, e.g., with $\beta_j(t)$ as the effect of a certain therapy z_j varying over time, are obtained as an important special case. A related multiplicative model for time-varying effects is also studied in Zucker and Karr (1990). A nonparametric additive model, incorporating time-varying effects, was introduced by Aalen (1980), further developed in Aalen (1989, 1993) and is described in some detail in Andersen et al. (1993, Ch. VII.4.) A general nonparametric regression model for survival data, without assuming additive or multiplicative hazards, is considered in Mc Keague and Utikal (1990) and in Keiding (1990), but dimensionality, i.e., the number of covariates included, becomes more critical here, and, as general with nonparametric models for complex data structures, more experience with applications is needed to gain insight into required sample sizes.

Time-varying effects can also be nicely dealt with in the Bayesian nonparametric framework of state space or dynamic models and Kalman filtering, see Gamerman (1991) for a dynamic version of the piecewise exponential model and Fahrmeir (1994), Fahrmeir and Wagenpfeil (1995) for dynamic discrete time survival and competing risk models. A related but somewhat different approach is proposed in Arjas and Liu (1995), using MCMC techniques like the Gibbs sampler for inference.

In this paper, we propose a nonparametric multiplicative model that takes the aspects discussed above into account and allows simultaneous incorporation and flexible estimation of baseline hazards and covariate effects for survival data and more complex event history data. Time t is essentially treated in the same way as other covariates or further time scales, including it as $\exp\{\beta_0(t)\}$, $\beta_0(t) = \log\{\alpha_0(t)\}$, in the predictor of the exponential risk function. The baseline effect, as well as continuous covariates and varying effects, is modelled by continuous or discrete-time smoothing splines, and a penalized likelihood approach is used to obtain smooth estimates. In certain circumstances, e.g. in the presence of several time scales, individual unobserved heterogeneity or frailty can be modelled by individual-specific effects, as

in the sleep data example. The degree of smoothness can be chosen subjectively, but data driven methods for choice of smoothing parameters are also discussed.

Section 2 describes the sleep study and the data set used in our main application in more detail. Section 3 introduces the model and the resulting penalized likelihood. Section 4 provides details on estimation. Section 5 contains analyses of the examples, in particular our main application to the sleep study.

2. Example: Sleep-EEG Data

Most sleep studies focus on sleep structure, characterized by recurrent alternations of electroencephalographic (EEG) patterns, and its relation to nocturnal hormonal secretion or to psychiatric diseases like depression. Sleep-EEG data are recordings of nocturnal sleep rhythm, usually classified in several stages such as awake, rapid eye movement (REM) and states of non-rapid eye movement (NREM) sleep. The sleep-EEG data in our example are part of a larger study at the Max-Planck-Institut für Psychiatrie in Munich. Sleep stages during one night, from 8 pm till 7 am next morning, are recorded every 30 seconds for a homogeneous group of 30 patients. In addition to REM stage and four NREM stages 1,2,3,4, indicating depth of sleep, the data include the stages AWAKE and, only for some patients, PAUSE (no recording during PAUSE). Figure 3 shows sleep-EEG data for two patients. In addition, secretion of several hormones is measured every 10, 20 or 30 minutes. Figure 4 contains corresponding recordings of cortisol plasma concentration for the same two patients. Figure 4 is typical for most patients of the study group: There is a low during the first hours of sleep followed by a marked increase in early morning. It is much more difficult for the human eye to detect typical patterns in sleep-EEG recordings, and some kind of smoothing and synchronization seems appropriate.

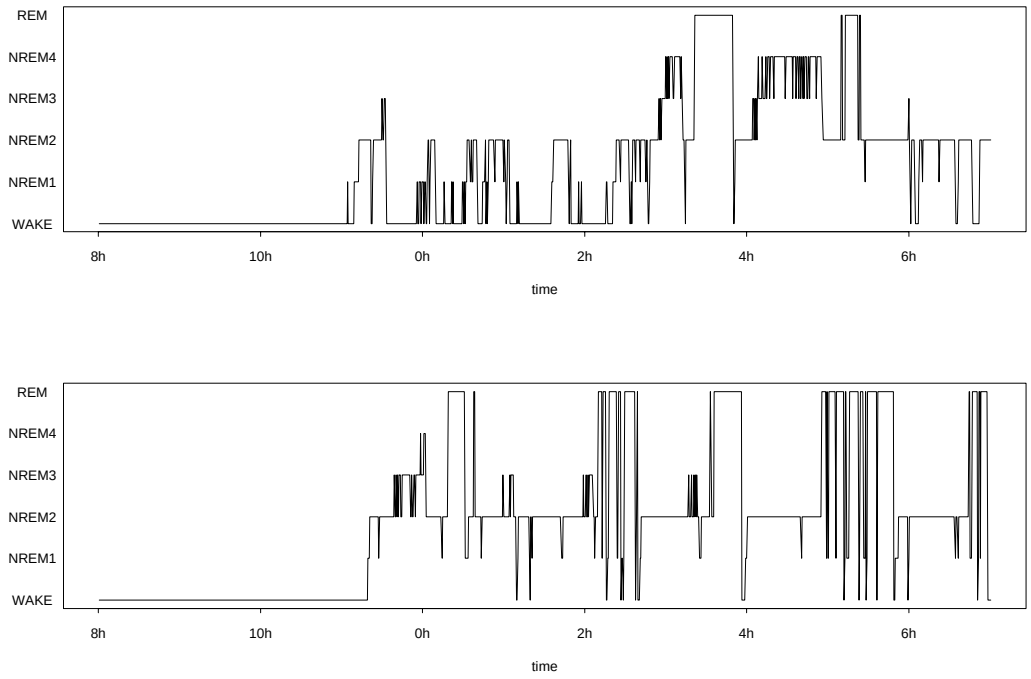

Figure 3: Individual sleep processes for two patients.

Previous statistical analyses of possible interrelation between hormonal secretion and sleep structure is mostly based on first constructing and extracting simpler characteristic variables from the original data and then applying more conventional methods like correlation and variance analysis. In Section 5, we will apply a specific nonparametric multiplicative model for transition intensities between sleep stages, providing some evidence on sleep structure and the effect of cortisol on it.

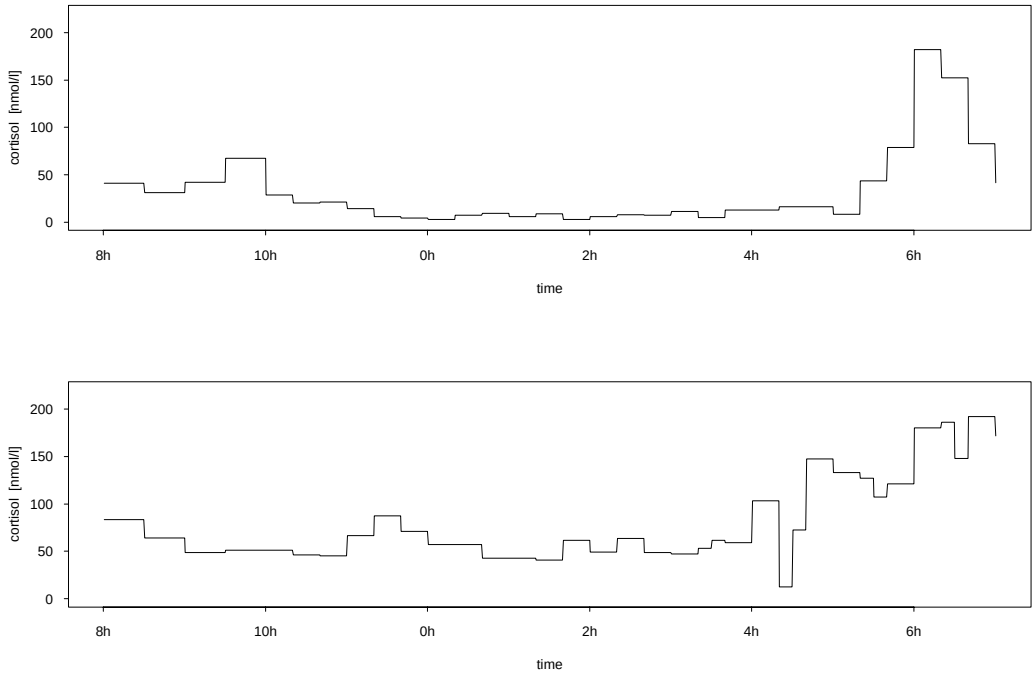

Figure 4: The two patients' nocturnal hormone secretion.

3. Models and likelihood

We first discuss the special but important case of time-independent covariates. Consider n individuals and let N_{hi}, $h = 1, \ldots, k$, $i = 1, \ldots, n$, denote the individual counting processes for events of type h, i.e. $N_{hi}(t)$ indicates the number of observed type h events experienced by the ith individual up to time t. We assume that individual intensity processes $\lambda_{hi}(t)$ exist and have multiplicative structure

$$\lambda_{hi}(t) = Y_{hi}(t)\alpha_{hi}(t; z_{hi}), \quad h = 1, \ldots, k, \quad i = 1, \ldots, n,$$

compare for example Andersen et al. (1993, Ch.VII) The predictable $0-1$ processes $Y_{hi}(t)$ indicate whether individual i is at risk for experiencing a type h event just before time t. The individual type h hazard or transition

rates $\alpha_{hi}(t; z_{hi})$ generally depend on time t and type-specific covariates or design vectors z_{hi}, often constructed from basic covariates.

3.1. Models

Hazard rates are related to predictor functions $\eta_{hi}(t; z_{hi})$ with additive structure by the exponential link

$$\alpha_{hi}(t; z_{hi}) = \exp\{\eta_{hi}(t; z_{hi})\}.$$

By the properties of the exponential function, hazard rates are nonnegative and have multiplicative structure. Before describing a general and flexible form for the predictors, we discuss some simpler examples. To simplify notation, we consider only two basic covariates x and w, where x is a continuous variable like tumor thickness and w is binary, indicating for example sex or treatment group. The simplest model is

$$\begin{aligned} \alpha_{hi}(t; x_i, w_i) &= \alpha_{h0}(t) \exp(\beta_{h1} x_i + \beta_{h2} w_i) \\ &= \exp\{\beta_{h0}(t) + \beta_{h1} x_i + \beta_{h2} w_i\}, \end{aligned}$$

i.e.

$$\eta_{hi} = \eta_{hi}(t; x_i, w_i) = \beta_{h0}(t) + \beta_{h1} x_i + \beta_{h2} w_i.$$

The predictor η_{hi} maintains the linear parametric form for the influence of the covariates as for the Cox model. The effect may be type-specific or common to some or all predictors, i.e. $\beta_{h1} = \beta_1$, $\beta_{h2} = \beta_2$. If a covariate is included only in one or some of the predictors,it becomes type-specific. Baseline effects $\beta_{h0}(t)$ are modelled nonparametrically by smoothing splines or 'smooth' piecewise constant functions over a fine grid $0 < a_0 < \ldots < a_{t-1} < a_t < \ldots < a_T = T$ of the observation period $[0, T]$. The gridpoints or knots $\{a_t\}$ can be determined by observed event times, or can be chosen subjectively, usually with small intervals $(a_{t-1}, a_t]$ in periods with many observations and larger intervals towards the end of the observation period, where data become sparse. Estimation is carried out simultaneously with

estimation of covariate effects β_{h1}, β_{h2}, using a penalized likelihood approach with penalty terms enforcing smoothness of the estimated baseline-effects, see further below.

More flexibility is obtained by dropping the simple linear parametric assumption for modelling covariate effects. If a certain functional form for the influence of x_{hi} cannot be specified in advance, $\beta_{h1}x_i$ can be replaced by a smooth function $\beta_{h1}(x)$ evaluated at x_i. Simultaneous estimation of $\beta_{h0}(t)$ and $\beta_{h1}(x)$ is carried out in analogy to generalized additive models (Hastie and Tibshirani, 1990).

Models with time-varying effects are obtained by assuming

$$\eta_{hi} = \beta_{h0}(t) + \beta_{h1}(t)x_i + \beta_{h2}(t)w_i,$$

where, for example, $\beta_{h2}(t)$ could be the effect of a certain therapy decreasing over time. In the more restricted context of survival data such time-varying coefficient models have recently gained much interest, and several proposals have been made for modelling and estimation. Note that for fixed t, this is a conventional linear predictor model. Nonparametric methods are based on penalized likelihood estimation (Zucker and Karr, 1990, Hastie and Tibshirani, 1993 , Klinger, 1993, on local likelihoods (Tutz, 1995) or on smoothing of appropriate residual plots (Grambsch and Therneau, 1994). Bayesian approaches are considered in Gamerman (1991), Fahrmeir (1994), in a discrete-time setting, and Arjas and Liu (1995).

As in Hastie and Tibshirani (1993), one may go a step further and consider varying coefficient models of the form

$$\eta_{hi} = \beta_{h0}(t) + \beta_{h1}(x_i) + \beta_{h2}(x_i)w_i + \beta_{h3}(t)w_i.$$

Here the smooth function β_{h2} may be viewed as an effect of w_i varying over the covariate x, or $\beta_{h2}(x_i)w_i$ is interpreted as an interaction term between the continuous covariate x and the binary covariate w.

In some cases of event history data it is also possible to include individual-

specific effects, common to some or all type-specific predictors, i.e.

$$\eta_{ih} = \gamma_i(t) + \text{other terms.}$$

We include such individual-specific effects in our model for analyzing the sleep data to separate individual-specific sleep intensities, that cannot be explained by covariates, from more systematic effects, e.g. the influence of cortisol. Subsection 3.4 provides a more detailed discussion on the incorporation of individual-specific effects.

A general form for all these models is

$$\eta_{hi}(t; z_{hi}) = \gamma_i(t) + \sum_{j=1}^{p} z_{hij}\beta_j(t) + \sum_{j=p+1}^{p+q} z_{hij}\beta_j(x_j), \tag{1}$$

where $x_1, \ldots, x_q$ are continuous covariates, and $z_{hi} = (z_{hij}, j = 1, \ldots, p+q)$ is a design vector, formed from basic covariates. By defining corresponding $0-1$ dummies in z_{hi}, the functions $\beta_j(t)$, $\beta_j(x_j)$ can be made type-specific or can be common to some or all predictors.

3.2. Likelihood and penalty function

Under appropriate assumptions on censoring or filtering mechanisms, e.g. noninformative right censoring, the corresponding likelihood has the form

$$\begin{aligned} l(\eta) &= \sum_{i=1}^{n}\sum_{h=1}^{k}\left[\int_0^T \log\{\alpha_{hi}(t; z_{hi})\}dN_{hi}(t) - \int_0^T \alpha_{hi}(t; z_{hi})Y_{hi}(t)dt\right] \quad (2)\\ &= \sum_{i=1}^{n}\sum_{h=1}^{k}\left[\int_0^T \eta_{hi}(t; z_{hi})dN_{hi}(t) - \int_0^T \exp\{\eta_{hi}(t; z_{hi})\}Y_{hi}(t)dt\right], \end{aligned}$$

see Andersen et al. (1993, Ch. III and VII). To obtain computationally tractable expressions for the likelihood, the predictors $\eta_{hi}(t; z_{hi})$ are considered – or approximated – as piecewise constant functions over the intervals $(a_{t-1}, a_t]$ of the chosen time-grid. This means that the smooth time-varying effects $\beta_j(t)$ are treated as piecewise constant functions over $(a_{t-1}, a_t]$, with value

$\beta_j(a_t)$, regardless whether continuous-time smoothing splines or discrete splines are used for modelling $\beta_j(t)$. For fixed z_{hi}, let now $\eta_{hi}(t) := \eta_{hi}(t; z_{hi})$ denote the value of η_{hi} over $(a_{t-1}, a_t]$. Then the log-likelihood (2) becomes

$$l(\eta) = \sum_{i,h} \sum_{t=1}^{T} \Big[\eta_{hi} \Delta N_{hi}(t) - Y^*_{hi}(t) \exp\{\eta_{hi}(t)\}\Big],$$

where $\Delta N_{hi}(t)$ indicates a type h event in $(a_{t-1}, a_t]$ for individual i and

$$Y^*_{hi}(t) = \int_{a_{t-1}}^{a_t} Y_i(t) dt$$

is the total amount of time of being at risk for a type h event in $(a_{t-1}, a_t]$. Defining the risk set $R_{th} = \{i : Y^*_{hi} > 0\}$, one obtains

$$l(\eta) = \sum_{t=1}^{T} \sum_{h=1}^{k} \sum_{i \in R_{th}} \Big[\eta_{hi}(t) \Delta N_{hi}(t) - Y^*_{hi}(t) \exp\{\eta_{hi}(t)\}\Big] \tag{3}$$

More details of computational evaluation of $l(\eta)$ are given in Section 4.

Smooth estimates of the functions β_j are obtained by maximizing a penalized log-likelihood

$$lp(\beta_1, \ldots, \beta_{p+q}) = l(\eta) - \sum_{j=1}^{p+q} \lambda_j J(\beta_j) \quad , \tag{4}$$

where $J(\beta_j)$ is a roughness penalty. The most popular smoother is a cubic smoothing spline, obtained with the integrated squared curvature

$$J(\beta_j) = \int \{\beta''_j(x)\}^2 dx \tag{5}$$

as roughness penalty. Alternatively we use discrete versions, replacing derivatives by differences. For example,

$$\sum_{s \geq 2} \frac{\Big\{\beta_j(x_s) - \beta_j(x_{s-1})\Big\}^2}{x_s - x_{s-1}} \quad , \tag{6}$$

$0 = x_0 < x_1 < \ldots < x_{S-1} < x_S$, corresponds to a discrete first order spline. For the special covariate x = time t, the knots x_s are given by the grid points a_s of the time axis. Using second differences leads to discrete second order splines. For equally spaced small intervals, the latter are more or less indistinguishable from cubic smoothing splines.

3.3. Time-dependent covariates

So far, discussion was restricted to time-independent covariates. Formally, time-dependent covariates are included in hazard rates and predictors by writing $z_{hi}(t)$ instead of z_{hi}. For so-called defined time-dependent covariates (Kalbfleisch and Prentice, 1980, p.123) the (conditional) likelihood remains the same, and inference is performed as if covariate paths had been fixed in advance. For truly random processes $z_{hi}(t)$, joint likelihoods for $\{N_{hi}(t), z_{hi}(t)\}$ and censoring processes have to be considered, in principle. Under appropriate assumptions, the log-likelihood $l(\eta)$ can be looked at as the relevant conditional log-likelihood. A thorough discussion of model specification in the presence of time-dependent covariates can be found in Andersen et al. (1993, Ch. III) and Arjas (1989). A fundamental assumption is that the $z_{hi}(t)$ are predictable, i.e. the covariate value at time t is already known just before t. For a continuously observed time-dependent covariate, not fixed in advance, its path has to be approximated by a discretized version. In our application to sleep data, where duration in certain states and cortisol concentration are included as covariates, these assumptions are fulfilled. To formulate the log-likelihood in analogy to the time-independent case, it is convenient to consider individual covariate-specific counting processes N_{hzi}, where z is an element of the discrete set E_h of possible covariate values $z_h(t)$. Then $N_{hzi}(t)$ is the number of type h events up to time t experienced by individual i under the covariate value z. For time-independent covariates $N_{hzi}(t)$ reduces to $N_{hi}(t)$. Assuming

$$\lambda_{hzi}\{t; z_{hi}(t) = z\} = Y_{hzi}(t) \exp[\eta_{hzi}\{t; z_{hi}(t) = z\}]$$

for the individual covariate-specific intensity processes, defining $Y^*_{hzi}(t)$ as the total amount of time in $(a_{t-1}, a_t]$ of individual i at risk for a type h event under covariate value z, one arrives at

$$l(\eta) = \sum_{t=1}^{T} \sum_{h=1}^{k} \sum_{i=1}^{n} \sum_{z \in E_h} \left[\eta_{hzi}(t) \Delta N_{hzi}(t) - Y^*_{hzi}(t) \exp\{\eta_{hzi}(t)\}\right] \quad .$$

To group over individuals i with $z_{hi}(t) = z$, we define

$$\eta_{hz}(t) = \eta_{hzi}\{t; z_{hi}(t) = z\}, \qquad Y^*_{hz}(t) = \sum_{i=1}^{n} Y^*_{hzi}(t)$$

and the risk set $R_{thz} = \{i : Y^*_{hzi}(t) > 0\}$. As resulting log-likelihood we have

$$l(\eta) = \sum_{t=1}^{T} \sum_{h=1}^{k} \sum_{z \in R_{thz}} \left[\eta_{hz}(t)\Delta N_{hz}(t) - Y^*_{hz}(t) \exp\{\eta_{hz}(t)\}\right] \quad , \tag{7}$$

in complete analogy to (3). Here, $\Delta N_{hz}(t)$ counts the number of type h events under covariate value z observed until time t, $Y^*_{hz}(t)$ is the total amount of time being at risk for a type h event during $(a_{t-1}, a_t]$ for all individuals with covariate value $z_{hi}(t) = z$, and R_{thz} is the corresponding risk set.

3.4. Individual-specific effects and different time scales

Frailty concepts are incorporated into the framework of nonparametric multiplicative models by introducing individual-specific effects $\gamma_i(t)$. To illustrate this, let us consider a simple sleep-EEG model where we are mainly interested in the effect of high hormone concentration on the duration of the first REM phase. Besides duration of the first REM phase (d_i) we also make use of time since sleep onset (t) as second time scale. We suppose that characteristics of individual sleep processes do also depend on unobserved covariates such as personal habits. Because time since first entry in a sleep phase t is more appropriate to describe individual sleep processes, it is used as basic time, while d_i is included as discretized time-dependent covariate. Let $w_i(t) = 1$ if the hormone level is high at t and $w_i(t) = 0$ elsewhere. A multiplicative model with predictor

$$\eta_i = \gamma_i(t) + I(d_i > 0)\{\beta_0(d_i) + \beta_1(d_i)w_i(t)\} \tag{8}$$

for the process counting terminations of the first REM phase describes the patients individual sleep histories. By the smoothness restrictions imposed and the different time scales used, identifiability usually is guaranteed

if $\beta_0(d_i)$ is restricted to have mean 0, see the next section for details. In model (8) individual intensities depend on multiplicative individual specific components $\exp\{\gamma_i(t)\}$ characterizing the patients propensity to change sleep states or frailty. The effect $\beta_0(d_i)$ can be interpreted as a baseline effect and indicates whether an 'ideal' patient has high or low propensity to terminate the first REM phase after spending d_i minutes in this state. The coefficient of interest $\exp\{\beta_1(d_i)\}$, can be seen as interaction of d_i and $z_i(t)$ and thus explains relations between REM duration and high concentration of hormones. This concept is only based on exact description of individual histories and no additional assumptions about frailty parameters are made.

Basically, individual-specific effects can be introduced when the model assumption decomposes individual counting processes N_{hi} into two or more type- or covariate-specific counting processes N_{hzi}. This decomposition can be made by considering different or recurrent events, different time scales or time-dependent covariates. The whole approach can be transferred to the wide area of clinical studies, for example, by introducing a time scale t as the patients age and considering duration d_i as time since disease onset or operation. However, the basic time scale, age or calendar time, should be chosen such that censoring processes and stochastic covariates are predictable given the history in t.

4. Estimation

In this section we first derive the backfitting algorithm for estimating the functions $\beta_j(t)$ and $\beta_j(x_j)$. Introducing appropriate matrix notation, this can be formulated in terms of familiar generalized linear or additive modelling framework. Furthermore, we outline computation of confidence bands and selection of smoothing parameters. Although discussion here focusses on hazard models, extensions to other types of varying coefficient models are immediate.

4.1. The estimation procedure

Suppose n_{ht} distinct values $z \in E_h$ of the covariate vector contributing to transition h were observed during time interval $(a_{t-1}, a_t]$. Now let y_{ht} be a $n_{ht} \times 1$ vector containing the counts of type h events under each covariate value z and define $\exp(\eta_{ht})$ as the corresponding vector of componentwise exponential predictor evaluations. The experienced total time under risk for this event, $Y^*_{hz}(t)$, is stored in the same order in a diagonal matrix $Q_{ht} = \text{diag}\{Y^*_{hz}(t)\}$. Rewriting the penalized log-likelihood criteria of Section 3.2, the vector of point evaluations $\beta_j = \{\beta_j(x_1), \ldots, \beta_j(x_S)\}', x_1 < \ldots < x_S$, for each function $\beta_j(t), \beta_j(x_j)$, is then estimated by maximizing

$$lp(\beta_1, \ldots, \beta_{p+q}) = \sum_{h=1}^{k} \sum_{t=1}^{T} \{y'_{ht}\eta_{ht} - 1'_{n_{ht}} Q_{ht} \exp(\eta_{ht})\} - \sum_{j=1}^{p+q} \lambda_j J(\beta_j), \qquad (9)$$

where $1'_{n_{ht}} = (1, \ldots, 1)'$, and $J(\beta_j)$ is one of the roughness penalties described in Section 3.2. Note that the assumption of piecewise constant hazards may reduce the length of the vectors drastically, since grouping can be done within each time interval and for each transition type separately. Hence complex models for large datasets are becoming computationally feasible within this framework.

It is well-known that the roughness penalty derived from the integrated squared curvature can be written as a quadratic form of the vector of point evaluations, $J(\beta_j) = \beta'_j K_j \beta_j$, and the uniquely minimizing functions are natural cubic splines. See e.g. Green and Silverman (1994, Ch. 2) for details. Clearly, discrete penalties can be written in the same form and the penalty matrices K_j have simple band structure. For example for the discrete first order spline penalty (6) we have

$$K_j = \begin{pmatrix} \frac{1}{x_2 - x_1} & \frac{-1}{x_2 - x_1} & \cdots & 0 \\ \frac{-1}{x_2 - x_1} & \frac{1}{x_2 - x_1} + \frac{1}{x_3 - x_2} & \frac{-1}{x_3 - x_2} & \cdots \\ \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots \\ 0 & \cdots & \frac{-1}{x_S - x_{S-1}} & \frac{1}{x_S - x_{S-1}} \end{pmatrix}.$$

Assuming the space of continuous functions with continuous derivatives in intervals (x_{s-1}, x_s), the first order penalty (6) is equivalent to a continuous penalty

$$J(\beta_j) = \sum_{s=2}^{S} \int_{x_{s-1}}^{x_s} \{\beta_j'(u)\}^2 du.$$

The unique minimizer in this function space is a polygon with knots in $x_1, \ldots, x_S$. Furthermore we introduce a transition specific response vector as $y_h = (y_{h1}', \ldots, y_{hT}')'$. To write the design in matrix notation, suppose during interval $(a_{t-1}, a_t]$ the pairs $(z_1, x_1), \ldots, (z_{n_{ht}}, x_{n_{ht}})$ are the observed values of covariates z_{hj} and x_{hj}, where x_{hj} is a metrical variable. Let the pairs be arranged in same order as y_{ht}. Then we define design matrices $Z_{hj} = \{Z_{hj}'(1), \ldots, Z_{hj}'(T)\}'$ with blocks given by

$$Z_{hj}(t) = \begin{pmatrix} 0 & \cdots & 0 & z_1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & z_{n_{ht}} & \cdots & 0 \end{pmatrix}$$

$$\uparrow$$

$$t\text{-th column}$$

for a time-varying effect $\beta_j(t) z_{hj}$, respectively

$$x_1\text{-th column}$$

$$\downarrow$$

$$Z_{hj}(t) = \begin{pmatrix} 0 & \cdots & z_1 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & z_{n_{ht}} & \cdots \end{pmatrix}$$

$$\uparrow$$

$$x_{n_{ht}}\text{-th column}$$

for an effect $\beta_j(x_j) z_{hj}$ with effect modifier x_j. By the definition of the design matrices it is easy to see that the column vectors of each Z_{hj} are orthogonal. Since there is only one element in each row of Z_{hj}, the design matrix can efficiently be stored in two vectors. Now we can write the transition specific

predictor $\eta_h = (\eta'_{h1}, \ldots, \eta'_{hT})'$ as $\eta_h = Z_{h1}\beta_1 + \ldots + Z_{h,p+q}\beta_{p+q}$, where Z_{hj} is a matrix of zeros if z_j doesn't contribute to a type h event.

Using the notations above and equating the derivatives of (9) to zero, yields the $p+q$ generalized score equations

$$u_j(\beta) = \partial lp(\beta)/\partial\beta_j = \sum_{h=1}^{k} Z'_{hj} s(\eta_h) - \lambda_j K_j \beta_j = 0, \tag{10}$$

where

$$s(\eta_h) = y_h - Q_h \exp(\eta_h)$$

is the log-likelihood score vector with Q_h =diag$(Q_{h1}, \ldots, Q_{hT})$.

It follows from Whaba (1990, Ch. 1 and 10) and Whaba, Wang, Gu, Klein and Klein (1994) that the solution of (10) exists and is unique in a broad class of penalized likelihood schemes as soon as an embedded parametric model, obeying $J(\beta_1) = \ldots = J(\beta_{p+q}) = 0$, has a unique solution. For first order penalties as in (6), this embedded parametric model is defined by constant functions $\beta_j \equiv \beta_j(t)$, $\beta_j(x) \equiv \beta_j$, and for second order penalties by linear functions of t or x. If the sample provides no information about a certain point evaluation $\beta_j(x_j)$, the unique maximizer of the penalized log-likelihood is the linear or polynomial interpolant at this point. This happens for example when all covariate values z_{hj} for this effect are zero within one time interval. Hence the dimension of the function space containing the solution can be smaller than the number of point evaluations. Now consider the solution for a model with predictor $\eta_h = \beta_1(t) + \beta_2(x) + \beta_3(t)w + \beta_4(x)w$. Since the embedded parametric model $\eta_h = \beta_1 + \beta_2 + (\beta_3 + \beta_4)w$ contains constant terms for the intercept and for the effect of w twice, the solution is not unique. One way to overcome this phenomenon called concurvity (Buja, Hastie and Tibshirani, 1989), i.e. collinearity in function spaces , is to choose a reference value or reference interval x_R and write the predictor as

$$\eta_h = \beta_1(t) + I(x \notin x_R)\beta_2(x) + \beta_3(t)w + I(x \notin x_R)\beta_4(x)w.$$

Technically the rows corresponding to x_R are omitted in the design matrices and the point evaluations $\beta_2(x_R)$ resp. $\beta_4(x_R)$ are inter- or extrapolated. This

is similar to dummy coding of a covariate with possible categories $x_1, \ldots, x_S$ and reduces the dimension of the function space by one. An alternative solution to concurvity introduced by Buja, Hastie and Tibshirani (1989) is discussed below.

System (10) is solved iteratively by a Fisher scoring procedure with inner Gauss-Seidel loops or the equivalent local scoring procedure (Hastie and Tibshirani, 1990, Ch. 6, 1993). With $\beta = (\beta_1', \ldots, \beta_{p+q}')'$ the matrix of negative expected second derivatives of the penalized log-likelihoods is given by

$$H(\beta) = -\partial^2 lp(\beta)/\partial\beta\partial\beta' = \begin{pmatrix} \sum Z_{h1}' F(\eta_h) Z_{h1} + \lambda_1 K_1 & \cdots & \sum Z_{h,p+q}' F(\eta_h) Z_{h1} \\ \sum Z_{h1}' F(\eta_h) Z_{h2} & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ \sum Z_{h1}' F(\eta_h) Z_{h,p+q} & \cdots & \sum Z_{h,p+q}' F(\eta_h) Z_{hp+q} + \lambda_{p+q} K_{p+q} \end{pmatrix}, \tag{11}$$

where

$$F(\eta_h) = -\partial^2 l(\eta)/\partial\eta_h\partial\eta_h' = Q_h \exp(\eta_h), \quad F(\eta) = \mathrm{diag}\{F(\eta_1), \ldots, F(\eta_k)\}$$

is the usual Fisher information matrix for η. With first derivative vector $u(\beta) = \{u_1'(\beta), \ldots, u_{p+q}'(\beta)\}'$, Fisher scoring iterations have the common form

$$H(\beta^{(o)})(\beta^{(n)} - \beta^{(o)}) = u(\beta^{(o)}),$$

where $\beta^{(o)}$ denotes results from the previous loop whereas $\beta^{(n)}$ is the actual coefficient vector. Using working observations

$$\tilde{y}^{(o)} = \eta^{(o)} + F^{-1}(\eta^{(o)}) s(\eta^{(o)}), \qquad s(\eta^{(o)}) = \{s(\eta_1^{(o)}, \ldots, s(\eta_k^{(o)})\}$$

for a current coefficient vector, the Fisher scoring algorithm can be transformed to

$$H(\beta^{(o)})\beta^{(n)} = Z' F(\eta^{(o)}) \tilde{y}^{(o)}. \tag{12}$$

Thus in each iteration normal equations for a penalized least squares problem with design matrix

$$Z = \begin{pmatrix} Z_{11} & \cdots & Z_{1,p+q} \\ \vdots & & \vdots \\ Z_{k1} & \cdots & Z_{k,p+q} \end{pmatrix}$$

have to be solved. This iteratively penalized least squares estimation stops at convergence of β, i.e. $\hat{\beta} = \beta^{(n)} \approx \beta^{(o)}$. Due to the special structure of (11) a backfitting algorithm of Gauss-Seidel type can efficiently solve the normal equations. Working out each block row of the normal equations (12) results in

$$\left\{ \sum_{h=1}^{k} Z'_{hj} F(\eta^{(o)}) Z_{hj} + \lambda_j K_j \right\} \beta_j^{(n)} = \sum_{h=1}^{k} Z'_{hj} F(\eta^{(o)}) \left\{ \tilde{y}^{(o)} - \sum_{l \neq j} Z_{hl} \beta_l^{(n)} \right\} \tag{13}$$

for each Gauss-Seidel iteration. Since $Z'_{hj} F(\eta^{(o)}) Z_{hj}$ are diagonal, only few modifications to standard fast smoothing-spline algorithms have to be done to solve (13), see Klinger (1993) and Fahrmeir, Gieger and Klinger (1995) for details. Backfitting cycles the smoothing or projection operators

$$S_j = \left\{ \sum_{h=1}^{k} Z'_{hj} F(\eta^{(o)}) Z_{hj} + \lambda_j K_j \right\}^{-1} \sum_{h=1}^{k} Z'_{hj} F(\eta^{(o)})$$

for $j = 1, \ldots, p+q, 1, \ldots, p+q, 1, \ldots$ on actual partial residuals

$$\tilde{y}^{(o)} - \sum_{l \neq j} Z_{hl} \beta_l^{(n)}$$

until $\beta_1^{(n)}, \ldots, \beta_{p+q}^{(n)}$ only change within a small given range. Thus the algorithm solves the system

$$\begin{array}{rcll} \beta_1^{(n)} & = S_1 & \left(\tilde{y}^{(o)} - \right. & 0 \qquad -Z_{h2}\beta_2^{(n)} \qquad \ldots - Z_{h,p+q}\beta_{p+q}^{(n)} \left.\right) \\ \vdots & = & \vdots & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ \beta_{p+q}^{(n)} & = S_{p+q} & \left(\tilde{y}^{(o)} - \right. & Z_{h1}\beta_1^{(n)} \quad \ldots - Z_{h,p+q-1}\beta_{p+q-1}^{(n)} \qquad\qquad -0 \left.\right) \end{array}$$

which is equivalent to the normal equations (12). From Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1993) it is known that for a

certain class of projection operators, including those proposed here, backfitting converges to any solution within the concurvity space, i.e. the space of all functions minimizing the corresponding penalized least squares problem.

To obtain unique results, the authors propose to apply centered smoothers where the average of β_j is substracted from β_j in each backfitting step. Thus the effects $\beta_j(t)$ or $\beta_j(x_j)$ are forced to have zero mean. When using centered smoothers, linear terms have to be included into the predictor and our concurvity example becomes

$$\eta_h = \gamma_1 + \gamma_2 w + \beta_1(t) + \beta_2(x) + \beta_3(t)w + \beta_4(x)w.$$

Estimation of the 'parametric' effects γ_1, γ_2 is incorporated into the backfitting algorithm by substituting S_j with an appropriate projection matrix $(X'X)^{-1}X, \quad X = (1, w)$, familiar from linear models. Alternatively, by centering only $\beta_3(t)$ and $\beta_4(x)$, the parameters γ_1 and γ_2 are automatically added to the 'baseline' effects $\beta_1(t)$ and $\beta_2(x)$. In our applications however, we found it more convenient to deal with concurvity by introducing a reference value as sketched above.

In presence of approximate concurvity backfitting tends to converge slowly and solutions may get unstable. An analysis of the embedded parametric model can help to detect this situation. Use of first order penalties instead of cubic smoothing splines when the slope is not very distinct may help to overcome instability due to approximate concurvity.

4.2. Confidence bands

Heuristic derivations of approximate confidence bands are usually based on appropriate first order expansions. As outlined by Gray (1992) in the context of survival data, more rigorous results may be obtained by assuming that the number of time intervals and different covariate values is held fixed as n increases. For a given vector λ_n of smoothing parameters, let $\beta(\lambda_n)$ denote a maximizer of the expected log-likelihood or, in case of uniqueness, equivalently a zero of the expected penalized score function $u\{\beta(\lambda_n)\}$.

Along similar lines as in asymptotic theory of maximum likelihood estimation in misspecified generalized linear models (e.g. Fahrmeir, 1990) it can be shown that $n^{1/2}\{\hat{\beta}_n - \beta(\lambda_n)\}$ is asymptotically normal with mean zero and covariance matrix

$$V = \lim n \quad H^{-1}\{\beta(\lambda_n)\}\mathrm{cov}\left[u\{\beta(\lambda_n)\}\right] H^{-1}\{\beta(\lambda_n)\}.$$

If the true model, characterized by β_0 say, coincides with the embedded model, i.e. the penalty terms are zero for β_0, then $\beta(\lambda_n) = \beta_0$. Generally however, $\beta(\lambda_n) \neq \beta_0$, but convergence $\beta(\lambda_n) \to \beta_0$ can be obtained by appropriate asymptotic rate of smoothing, e.g. assuming $\lambda_n = O(n^{1/2})$. Then it can be shown that $n^{1/2}\left\{\hat{\beta}_n - \beta(\lambda_n)\right\}$ is asymptotically normal with mean zero and covariance matrix $\lim n\hat{V}$, with the sandwich matrix

$$\hat{V} = \lim n \quad H^{-1}\{\beta(\lambda_n)\}Z'F(\hat{\eta})ZH^{-1}\{\beta(\lambda_n)\} \tag{14}$$

Pointwise confidence bands can be computed from the diagonal of $\hat{V}$. In practice, the quality of approximation will of course depend on the ratio of sample size versus numbers of parameters involved and the actual degree of smoothing. Yet we use (14) as a useful approximation.

Asymptotic analysis becomes much more complicated if the number of parameters increases with n, as for cubic smoothing splines. Consistency and convergence rate results for the Cox model are given in O'Sullivan (1993), but rigorous asymptotic distribution theory is still not available.

Since $H(\hat{\beta})$ is usually very big and unstructured, computation of $\hat{V}$ requires still a lot of time and memory. In principle this can be done by applying the backfitting algorithm to an appropriate set of vectors and solving the linear system $H(\hat{\beta})X = I$. However in our experience this is a very unstable procedure and thus we use direct inversion methods.

Based on Bayesian arguments Gu (1992) and Whaba, Wang, Gu, Klein and Klein (1994) give some evidence, that by imposing appropriate Gaussian smoothness priors for posterior mode estimation, leading to our penalized likelihood equations, the posterior distribution of $\hat{\beta}$ is approximate normal

with covariance matrix $H(\hat{\beta})^{-1}$. Hence pointwise confidence bands may also be computed from the diagonal of this matrix.

4.3. Selection of smoothing parameters

A common way to select smoothing parameters is to consider the traces of the matrices $\nu_j = \mathrm{tr}\,(Z_{1j}S_j + \ldots + Z_{kj}S_j)$ as 'effective number of parameters' or 'degrees of freedom' of a smooth as proposed by Hastie and Tibshirani (1990, Ch.2 and 6). Smoothing parameters $\lambda_1, \ldots, \lambda_{p+q}$ are then chosen according to a given number of parameters. Applying the penalties proposed in Section 3.2, λ_j tunes the degrees of freedom from 1, respectively 2, corresponding to the number of parameters for the embedded parametric model, up to the number of distinct time intervals or covariate values x_j or, more precisely, up to the dimension of the vector space spanned by the columns of $(Z_{1j}', \ldots, Z_{kj}')'$. By using deviance statistics or looking at appropriate residual plots, one can decide whether more or less smoothing is adequate and how much degrees of freedom to use.

Basically most criterions for automatic smoothing parameter selection, such as generalized cross validation (GCV) or Akaikes information criterion (AIC), require the trace $\nu = \mathrm{tr}\left\{F^{-T/2}(\hat{\eta})Z'H^{-1}(\hat{\beta})ZF^{-1/2}(\hat{\eta})\right\}$ of the hat-matrix. Since computation of $H^{-1}(\hat{\beta})$ is very demanding and the criterion has to be optimized over several parameters, smoothing parameter selection by exact optimization of one of those quantities is still too time consuming for practical use. One way to overcome this problem, is the proposal of Girard (1991), who studies GCV where a Monte-Carlo simulation based on the relation

$$\varepsilon \sim N(0, I) \Rightarrow E\,(\varepsilon' A \varepsilon) = \mathrm{tr}(A)$$

is used to approximate the required trace.

Alternatively one can use only the effective number of parameters ν_j which is cheaply calculated, and construct fast iterative algorithms for smoothing parameter selection. In principle, these procedures mimic a statistician

watching goodness of fit criterions and tuning smoothing parameters. One such algorithm designed for survival data and general varying coefficient models is described in Klinger (1993) and Dannegger, Klinger and Ulm (1995). There it was applied successfully to various data sets.

In more complex situations like the sleep EEG study, where individual-specific effects are included, further considerations are necessary. Heuristically, the degree of smoothing for individual-specific effects should not depend on the sample size whereas for other effects smoothness should decrease with increasing n. To ensure that the number of smoothing parameters does not increase with order $O(n)$, grouping of the λ_j e.g. those belonging to individual specific effects seems to be appropriate. However, still more experience is needed for such complex models.

5. Applications

5.1. Survival with malignant melanoma

We first illustrate the methods by an application to this survival data set which is described in detail and used in a number of examples in Andersen, Borgan, Gill and Keiding (1993). Survival time is measured in days after operation. There are 57 patients who died from melanoma within the observation period and 14 patients who died from other causes. The remaining 134 are censored. Covariates included are sex S (1 = male, 0 = female), tumor thickness X in mm and ulceration U (1 = present, 0 = absent). Let $\alpha_1(t;z)$ and $\alpha_2(t;z)$ denote the hazard rates for death from malanoma and death from other causes. We choose a multiplicative model $\alpha_1(t;z_1) = \exp\{\eta_1(t;z_1)\}$, $\alpha_2(t;z_2) = \exp\{\eta_2(t;z_2)\}$ with

$$\begin{aligned}\eta_1(t;z_1) &= \beta_1(t) + \beta_2(t) + \beta_3(t)S + \beta_4(t)S + \beta_5(t)U + I(X > 0.2)\beta_6(X)\\ \eta_2(t;z_2) &= \beta_1(t) + \beta_3(t)S \quad .\end{aligned}$$

Thus, for the hazard of dying from other causes, $\beta_1(t)$ is a global baseline effect and $\beta_3(t)$ is the global possibly time-varying effect of sex. The baseline

effect for dying from melanoma is modelled additively by $\beta_1(t)+\beta_2(t)$, and the time-varying effect of sex as $\beta_3(t)+\beta_4(t)$. Identifiability is guaranteed, since $\beta_1(t)$ and $\beta_3(t)$ appears in both predictors. The effect $\beta_5(t)$ of ulceration is also modelled as time-varying, and $\beta_6(X)$ is the effect of tumor-thickness X. Incorporation of the indicator function $I(X > 0.2)$ guarantuees uniqueness, compare Section 4. While $\beta_1(t)$, $\beta_2(t)$, $\beta_3(t)$, $\beta_5(t)$ and $\beta_6(t)$ are modelled by cubic splines, the additional effect $\beta_4(t)$ of sex in η_1 is modelled by a discrete first order spline. The reason is that $\beta_4(t)$ is near to zero for all t, causing instable estimation when using cubic splines due to near-concurvity, compare the remarks in Section 4. The effects are displayed in Figure 5, together with confidence bands obtained from the sandwich estimate $\hat{V}$. Smoothing parameters are selected by tuning degrees of freedom.

The global baseline effect $\beta_1(t)$ in Figure 5 (a) has bath-tub shape, in consistency with a simpler competing risks model in Andersen et al. (1994, p.495). It is modified for 'death from melanoma' by addition of the slightly bell-shaped effect $\beta_2(t)$. The global effect of sex is not clearly significant, but nearly constant and almost the same in both groups, since $\beta_4(t)$ is close to zero in Figures 5 (c) and (d). Thus, considering sex alone, a proportional hazards assumption seems plausible. On the contrary the effect of ulceration $\beta_5(t)$ is time-varying, violating a proportional hazards assumption. This is again in accordance with Andersen et al. (1994, p.550). Thickness has a nonlinear effect, increasing in logarithmic form up to about 5 mm, then becoming slightly decreasing, and increasing again for more then 10 mm. Note however, that the right tail is influenced by a small number of outlying observations.

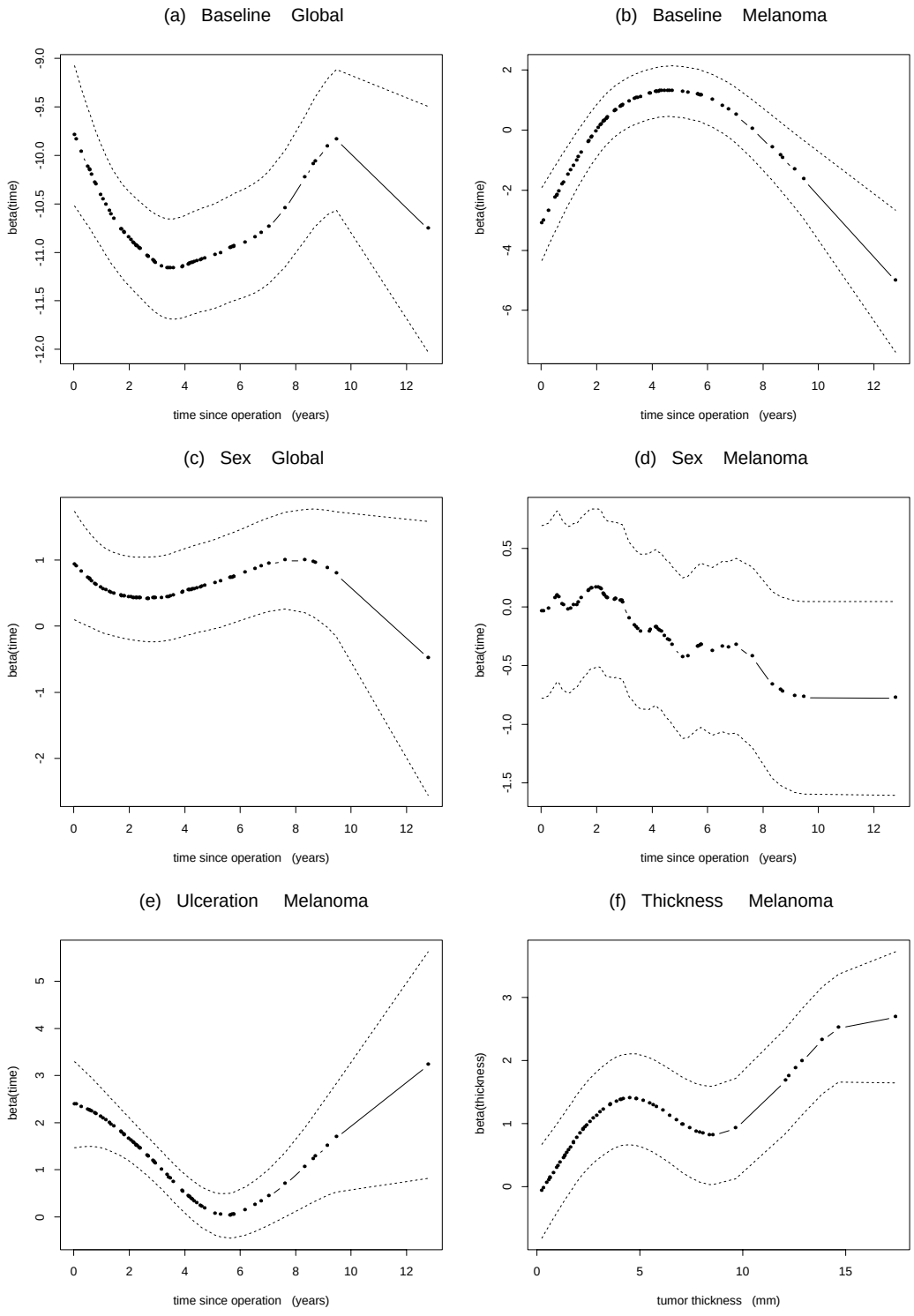

Figure 5: Estimated varying coefficients for the melanoma model with pointwise 1σ–confidence bands.

5.2. Sleep-EEG Data

The following analysis illustrates the flexibility in modelling more complex event history data. It is a preliminary attempt to explore sleep patterns and its association with secretion of hormones with the proposed methods, and it does not provide a 'final' model but is only a first step towards more refined investigations in cooperation with clinical partners.

Since we are here mainly interested in the influence of cortisol on REM states, we consider only transitions between the three states AWAKE, REM and NREM , without further differentiating between different states of NREM sleep. For a few patients, an additional state PAUSE is recorded, where measurements are interrupted for some reason. If a patient is in state PAUSE for some time, its risk indicator is set to zero. We distinguish four types of events:

$h = 1$: transition from AWAKE to REM or NREM, $(A \to RN)$
$h = 2$: transition from REM to NREM, $(R \to N)$
$h = 3$: transition from NREM to REM, $(N \to R)$
$h = 4$: transition from REM or NREM to AWAKE. $(RN \to A)$

There are several time scales that might be considered, e.g. real time, that is time since beginning of recordings at 8.00 p.m., time since onset of sleep, and durations in sleep states. To simplify and to achieve some synchronisation, we consider time t since onset of sleep as the basic time scale and introduce duration in REM states in form of a time-dependent covariate $d_i = (t-$time of last entry into a REM state). For two patients, d_i is not well-defined because recordings were interrupted by the state PAUSE. For simplicity, d_i was taken as the time in REM since end of PAUSE. Concentration of plasma cortisol was dichotomized in 'high' and 'low' by introducing the time-dependent covariate $z_i(t)$ =('concentration of plasma cortisol in person i at time t'$> 100 \quad n\,mol/l$). Looking at individual sleep patterns, it seems that the general tendency of changing states is higher for some persons than for others and is varying during night. To separate such individual-specific intensities, that cannot be explained by covariates from more systematic

effects, we introduce individual-specific effects $\gamma_i(t)$ as a common baseline into all predictors $\eta_{hi} = \eta_{hi}\{t; d_i, z_i(t_i)\}, h = 1, \ldots, 4$. These considerations led to the following model

$$\begin{aligned}
\eta_{1i} &= \gamma_i(t_i) + \beta_1(t_i) && (A \to RN),\\
\eta_{2i} &= \gamma_i(t_i) + I(d_i > 0.5)\beta_2(d_i) + z_i(t_i)\beta_3(d_i) && (R \to N),\\
\eta_{3i} &= \gamma_i(t_i) + \beta_4(t_i) + z_i(t_i)\beta_5(t_i) && (N \to R),\\
\eta_{4i} &= \gamma_i(t_i) + \beta_6(t_i), && (RN \to A).
\end{aligned}$$

Thus, $\beta_1(t_i)$ is a (population-averaged) effect of falling asleep if one is awake at time t_i since onset of sleep, $\beta_2(d_i)$ is the effect of duration in REM state for a transition to NREM state, and $\beta_3(d_i)$ is an additional effect for high levels of cortisol at time t_i. Interpretation of the effects $\beta_4(t_i)$, $\beta_5(t_i)$ and $\beta_6(t_i)$ is quite analogous, for example $\beta_5(t_i)$ is the additional effect for transitions from NREM to REM in periods of high levels of cortisol. Effects β_1 to β_6 are all modelled by cubic splines, corresponding to the penalty (6). Individual-specific effects γ_i are modelled by discrete first order splines, corresponding to the penalty (7). They are more appropriate for modelling effects that remain more or less constant within longer periods of time, interrupted by shorter periods of high transition rates, as for example in Figure 5. For both time scales, an equidistant grid of knots is used, with 10 minute intervals for time t and 30 second intervals for duration d in REM state. The finer grid for duration d makes the time-dependent covariate d predictable and discrete-valued, so that the basic assumptions for time-dependent covariates are fulfilled.

The following figures show relative risk functions or intensities, i.e. the factors in the multiplictive models $\alpha_{hi} = \exp(\eta_{hi})$, for example the risk functions $\alpha_i(t_i) = \exp\{\gamma_i(t_i)\}$ and $\alpha_1(t_i) = \exp\{\beta_1(t_i)\}$ in $\alpha_{1i}(t_i) = \alpha_i(t_i)\alpha_1(t_i) = \exp\{\eta_{1i}(t_i)\}$.

Figure 6 shows sleep patterns and associated individual-specific relative sleep intensities $\alpha_i(t_i) = \exp\{\gamma_i(t_i)\}$ for the same two individuals already considered in Section 2. For both individuals, smoothed relative intensi-

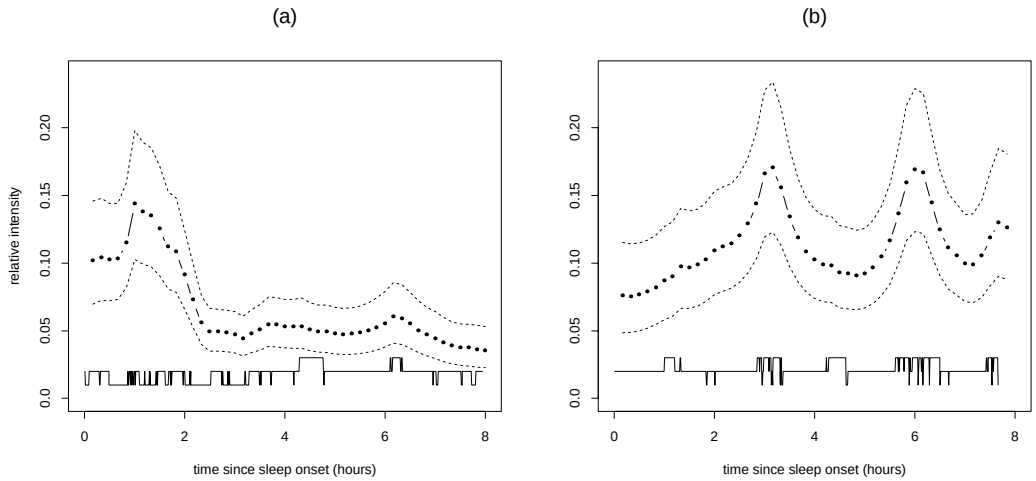

Figure 6: Individual–specific effects for two patients together with 1σ-confidence bands. The lower line indicates the states WAKE, NREM and REM.

ties reflect quite well phases of more 'restless' or more 'quiet' sleep. For example, the first individual experiences many transitions between NREM and AWAKE after one hour of sleep, and the peak in the relative intensitiy clearly indicates this. For the second individual, the two peaks and the smaller one towards the end of sleep reflect individual phases of more restless sleep. Figures 7 (a) and (b) show the relative intensities $\alpha_4(t_i)$ and $\alpha_5(t_i)$ corresponding to the main effect $\beta_4(t_i)$ for transitions from NREM to REM and the additional effect $\beta_5(t_i)$ for individuals with plasma concentration of cortisol over 100 $n\,mol/l$. The intensity $\alpha_4(t_i)$ supports well-known evidence: The probability for REM phases increases with time since onset of sleep, and $\alpha_5(t_i)$ clearly exhibits an additional effect in the early morning hours for individuals with higher level of cortisol concentration, thus providing evidence of the hypothesized association between REM phases and the level of cortisol concentration. The baseline intensity $\alpha_2(d_i)$ in Figure 7 (c) for transitions from REM to NREM is almost constant for about 30 minutes of REM sleep and increases slightly for longer REM sleep durations. For individuals with

high concentration of cortisol, transition intensities $\alpha_3(d_i)$ to NREM sleep are decreasing In Figure 7 (c). A possible interpretation is that longer duration in REM sleep becomes more likely for a patient who stays at a high cortisol level during the REM phase. Baseline intensities $\alpha_1(t_i)$ in Figure 7 (e) for transitions from SLEEP, i.e. REM or NREM, to AWAKE decrease rapidly at the beginning of sleep, remain at a constantly low level during most of the night, and increase in the morning, as to be expected. Baseline intensities $\alpha_6(t_i)$ for transitions from AWAKE to SLEEP, shown in Figure 7 (e), exhibit more variation during the night: The intensity for falling asleep has a distinct low about one hour after onset of sleep, that means if individuals are AWAKE at that time they have particular difficulty to fall asleep again. On the other side, the intensity for falling asleep again has a distinct maximum about the middle of the night. In the early morning hours, of course, there is a natural decrease for transitions from AWAKE to SLEEP, or in other words, it is difficult to fall asleep again after awakening in the morning.

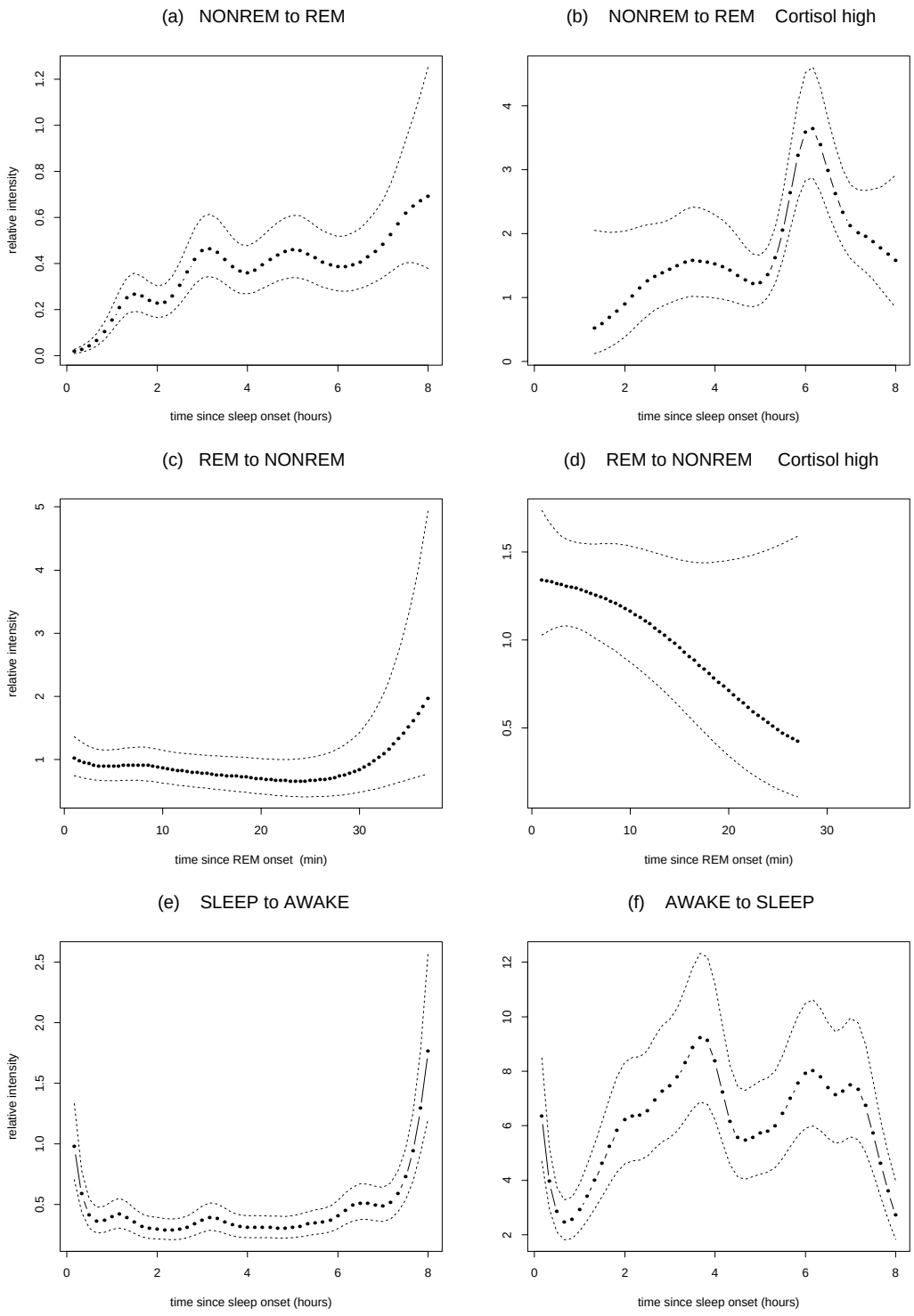

Figure 7: Estimated population averaged effects for the sleep study model together with 1σ-confidence bands.

6. Conclusions

As has been illustrated in the applications, the proposed multiplicative model family provides flexible tools for refined exploration and analysis of event history data and may therefore supplement existing methods. Although we focused here on continuous time, the appproach can also be transferred to the situation of discrete-time or grouped duration data.

There are some issues that have not been addressed to this paper. Model checking can be based on martingal residuals along the lines of Therneau, Grambsch and Fleming (1990). Computational efficiency might be greatly enhanced by special numerical techniques for inverting large sparse matrices, instead of using a backfitting algorithm. This would also be of particular value for data-driven choice of smoothing parameters. Also, reduced computation time will allow to conduct larger Monte Carlo studies to investigate finite sample properties of estimators and to support results or conjectures on asymptotic distributions.

References

Aalen, O. O. (1980). A model for non-parametric regression analysis of counting processes, *in* W. Klonecki, A. Kozek and J. Rosiński (eds), *Mathematical Statistics and Probability Theory*, Springer Lect. Notes Statist.

Aalen, O. O. (1989). A linear regression model for the analysis of life time, *Statist. Med.* **8**, 907–925.

Aalen, O. O. (1993). Further results on the non-parametric linear regression model in survival analysis, *Statist. Med.* **12**, 1569–1588.

Andersen, P., Borgan, O., Gill, R. and Keiding, N. (1993). *Statistical models based on counting processes*, Springer, New York.

Arjas, E. (1989). Survival models and martingale dynamics, *Scand. J. Statist.* **16**, 177–225.

Arjas, E. and Liu, L. (1995). Assessing the losses caused by an intervention: A hierarchical Bayesian approach, *Applied Statistics* **44**, 357–368.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models, *Ann. Statist.* **17**, 453–555.

Dannegger, F., Klinger, A. and Ulm, K. (1995). Identifikation of prognostic factors with censored data, *Discussion paper 11*, SFB 386, Munich.

Fahrmeir, L. (1990). Maximum likelihood estimation in misspecified generalized linear models, *Statistics* **21**, 487–502.

Fahrmeir, L. (1994). Dynamic modelling and penalized likelihood estimation for discrete time survival data, *Biometrika* **81**, 317–330.

Fahrmeir, L. and Wagenpfeil, S. (1995). Smoothing hazard functions and time-varying effects in discrete duration and competing risk models, *Discussion paper 7*, SFB 386, Munich.

Fahrmeir, L., Gieger, C. and Klinger, A. (1995). Additive, dynamic and multiplicative regression, *Allg. Stat. Archiv* **29**, 95–130.

Gamerman, D. (1991). Dynamic Bayesian models for survival data, *Appl. Statist.* **40**, 63–79.

Girard, D. (1991). Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization, *Ann. Statist.* **19**, 1950–1963.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazard tests and diagnostics based on weighted residuals, *Biometrika* **81**, 515–526.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognoses, *J. A. Statist. Assoc.* **87**, 942–951.

Green, P. and Silverman, B. (1994). *Nonparametric regression and generalized linear models*, Chapman and Hall, London.

Gu, C. (1992). Penalized likelihood regression: A Bayesian analysis, *Statistica Sinica* **2**, 255–264.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models: Some applications, *Proceedings 2nd International GLIM Conference, Lancaster*, Vol. 32, Springer Lect. Notes in Statistics, Berlin.

HASTIE, T. AND TIBSHIRANI, R. (1990). *Generalized additive models*, Chapman and Hall, London.

HASTIE, T. AND TIBSHIRANI, R. (1993). Varying-coefficient Models, *J.R. Statist. Soc. B* **55**, 757–796.

KALBFLEISCH, J. D. AND PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

KEIDING, N. (1990). Statistical inference in the Lexis diagram, *Phil. Trans. Roy. Soc. London A* **332**, 487–509.

KLINGER, A. (1993). *Spline–Glättung in zeitdiskreten Verweildauermodellen*, Diploma theses, Ludwig Maximilians Universität München, Institut für Statistik.

MCKEAGUE, I. W. AND UTIKAL, K. J. (1990). Inference for a nonlinear counting process regression model, *Ann. Statist.* **18**, 1172–1187.

O'SULLIVAN, F. (1988). Nonparametric estimation of relative risk using splines and cross–validation, *SIAM J. Sci. Stat. Comput.* **9**, 531–542.

O'SULLIVAN, F. (1993). Nonparametric estimation in the Cox model, *Ann. Statist.* **21**, 124–145.

THERNEAU, T., GRAMBSCH, P. AND FLEMING, T. (1990). Martingale–based residuals for survival models, *Biometrika* **77**, 147–160.

TUTZ, G. (1995). Dynamic modelling of discrete duration data: A local likelihood approach, *Forschungsbericht*, Fachbereich Informatik, Technische Universität Berlin.

WHABA, G. (1990). Spline Models for Observational Data, *CBMS-NSF Regional Conference Series in Applied Mathematics*, Vol. 59, SIAM, Philadelphia.

WAHBA, G., WANG, Y., GU, C., KLEIN, R. AND KLEIN, B. (1994). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy, *Technical Report 940*, Department of Statistics, University of Wisconsin, Madison. To appear, Ann. Statist, 1996.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach, *Ann. Statist.* **18**, 329–353.