Tutz, Binder:

Flexible Modelling of Discrete Failure Time Including Time-Varying Smooth Effects

Sonderforschungsbereich 386, Paper 294 (2002)

Flexible Modelling of Discrete Failure Time Including Time-Varying Smooth Effects

Gerhard Tutz
Institut für Statistik, Ludwig-Maximilians-Universität München,
Akademiestr. 1, D–80799 München, Germany

Harald Binder
Klinik und Poliklinik für Psychiatrie und Psychotherapie, Universität Regensburg,
Universitätsstr. 84, D–93042 Regensburg, Germany

Abstract:

Discrete survival models have been extended in several ways. More flexible models are obtained by including time-varying coefficients and covariates which determine the hazard rate in an additive but not further specified form. In this paper a general model is considered which comprises both types of covariate effects. An additional extension is the incorporation of smooth interaction between time and covariates. Thus in the linear predictor smooth effects of covariates which may vary across time are allowed. It is shown how simple duration models produce artefacts which may be avoided by flexible models. For the general model which includes parametric terms, time-varying coefficients in parametric terms and time-varying smooth effects estimation procedures are derived which are based on the regularized expansion of smooth effects in basis functions.

Key words: Discrete survival model, varying-coefficients, additive modelling, time-varying smooth effects

1 Introduction

Statistical modelling of survival data has some tradition of trying to avoid mis- or overinterpretation of estimated effects or estimated hazard functions. Much of this effort has been dedicated to the incorporation of frailties (e.g. Vaupel et al., 1979 or more recently Keiding et al., 1997, Henderson & Oman, 1999). In the following we will focus on alternatives which are connected to more flexible models, e.g. the incorporation of time dependent effects. Simple models as well as flexible models may produce artifacts but it will be shown that adequate flexible models are helpful to avoid artifacts.

Since in applications duration or survival time is often measured on a discrete scale we will consider flexible models of discrete survival and first introduce some basic concepts. The discrete scale may be due to grouping if failure is known to occur within certain intervals $[a_0, a_1), \ldots, [a_{q-1}, a_q), [a_q, \infty)$ where a_q often denotes the final follow up. Alternatively, time T may be measured directly on a discrete scale, e.g. in days, weeks or months. The corresponding discrete hazard function is given as the conditional probability

$$\lambda(t|x) = P(T = t|T \geq t, x)$$

where $T = t$ denotes failure at time t (in interval $[a_{t-1}, a_t)$ if grouping is assumed). Models for discrete time have the common form

$$\lambda(t|x_i) = h(\beta_{0t} + x_{it}'\beta) \tag{1}$$

where $h : \mathbb{R} \to [0, 1]$ is a fixed response function. A model of type (1) may be derived from the (continuous time) proportional hazards model yielding the link function $h(u) = 1 - \exp(-\exp(u))$ (e.g. Kalbfleisch & Prentice, 1980). Cox (1972) considered the logistic link $h(u) = \exp(u)/(1 + \exp(u))$ which yields quite similar results for short intervals (see also Arjas & Haara, 1987 and Ryu, 1994). The analogue to the unspecified baseline hazard which is familiar from Cox's proportional hazard model is found in the parameters $\beta_{01}, \ldots, \beta_{0q}$ which determine the discrete baseline hazard.

In model (1) the parameterization of the effect of covariates is rather restrictive. A first step to more flexible modelling is to replace the linear predictor $\eta_{it} = \beta_{0t} + x_i'\beta$ by

$$\eta_{it} = \beta_{0t} + x_{it}'\beta_t \tag{2}$$

yielding

$$\lambda(t|\eta_{it}) = h(\eta_{it})$$

where for simplicity the predictor is given in the condition. The time dependence of β_t allows the effects of covariates to vary across time. For example, in treatment studies the impact of the treatment may vanish after some time. Models of this type may be seen as varying-coefficient models (Hastie & Tibshirani, 1993) where the effect of covariates is modified by an additional variable, the so-called effect modifier. In the present case the effect-modifier is time. Estimation procedures for varying coefficients models that are based

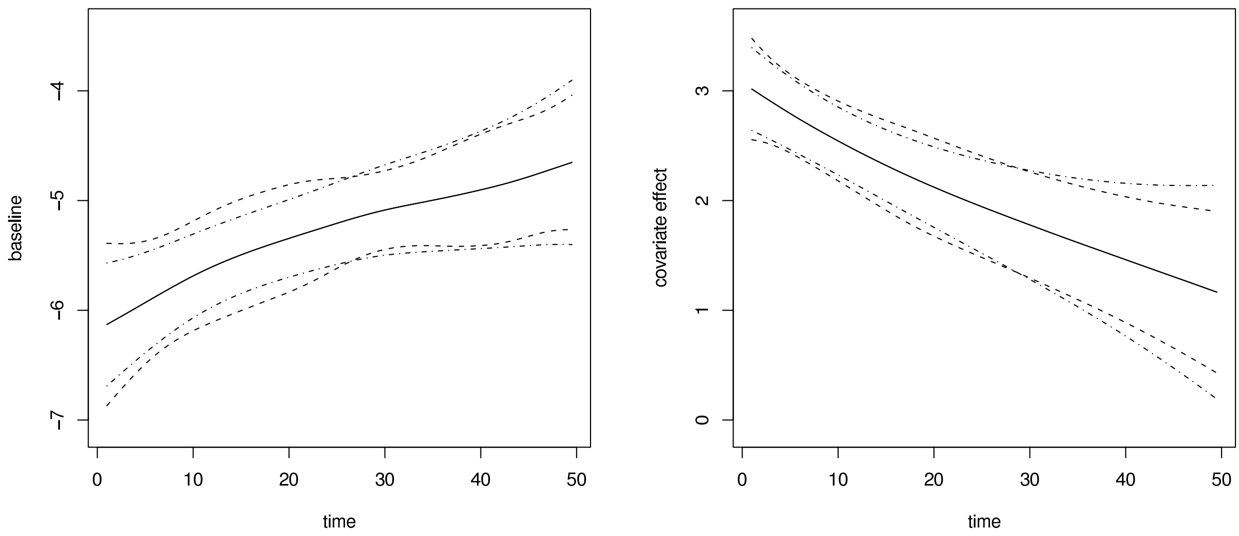

Figure 1: Mean estimate of β_{0t} (left panel) and β_t (scaled by covariate mean) (right panel) when fitting the model $\eta_{it} = \beta_{0t} + x_i\beta_t$ to data generated from $\eta_{it} = x_i^2$ (mean of estimates: solid line, empirical confidence bands: dashed lines, mean of estimated confidence bands: dash-dot lines).

on penalization are given in Hastie & Tibshirani (1993). Alternatively localizing techniques may be used. The general framework of localizing approaches is extensively treated in Fan & Gijbels (1996). For varying coefficients modelling, also in the context of survival modelling, see Kauermann & Tutz (2000, 2001). Bayesian modelling approaches are found in Fahrmeir & Wagenpfeil (1996).

Models with time-varying coefficients are more flexible by reflecting the variation of effect strength across time. In particular if the covariate is fixed at time $t = 0$ one has to expect vanishing effects (see Verweij & van Houwelingen, 1995 for an example). However, time-varying effects may also be due to the misspecification of the linkage between response and covariate. This may be seen from a simple example where the underlying predictor has the form $\eta_{it} = x_i^2$ with x_i being measured at time $t = 0$. Figure 1 shows the result if the varying coefficients model $\eta_{it} = \beta_{0t} + x_i\beta_t$ is fitted. It is seen that the estimates $\hat{\beta}_{0t}$, $\hat{\beta}_t$ are strongly misleading. While $\hat{\beta}_{0t}$ is an increasing function, $\hat{\beta}_t$ is decreasing. Since the linear term $x_i\beta_t$ tries to fit the nonlinear effect x_i^2 the effect $\hat{\beta}_t$ is larger in the beginning where all the subjects are in the study but decreases if the subjects with high risk are no longer in the study. Similar effects have been described for the contious time models by Abrahamowicz et al. (2002). Figure 2 shows similar effects for the underyling model $\eta_{it} = (1 - x_i)^2$. In both cases an analysis allowing for varying coefficients yields misleading effects which are strongly supported by confidence intervals which suggest that the effects are indeed decreasing or increasing.

A different approach to achieve more flexibility is to weaken the assumption of linearity by

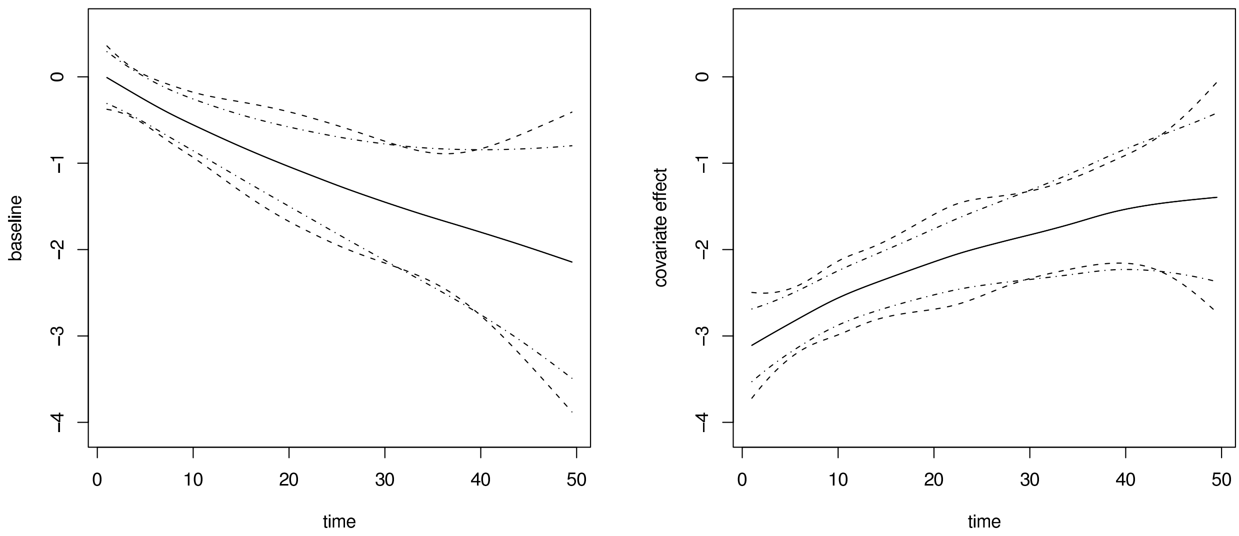

Figure 2: Estimate of β_{0t} (left panel) and β_t (scaled by covariate mean) (right panel) when fitting the model $\eta_{it} = \beta_{0t} + x_i\beta_t$ to data generated from $\eta_{it} = (1 - x_i)^2$ (mean of estimates: solid line, empirical confidence bands: dashed lines, mean of estimated confidence bands: dash-dot lines).

using an additive predictor

$$\eta_{it} = \beta_{0t} + \sum_{j=1}^{S} f_{(j)}(x_{itj})$$

where $f_{(j)}, j = 1, \ldots, S$ are unspecified smooth functions of covariates. Additive modelling has been investigated e.g. by Hastie & Tibshirani (1990) by use of backfitting estimators, Linton & Härdle (1996) by integration techniques, whereas Marx & Eilers (1998) and Aerts et al. (1999) use low-rank smoothers. An application of the latter approach to discrete survival modelling is found in Eilers (1998). One of the problems of the additive form is that it is time-independent. So the effect may not vary across time. If effects are vanishing across time the effect will be estimated with strong bias. In Figure 3 the underlying model is given by $\eta_{it} = \gamma_t\alpha(x_i)$ where α is a smooth function (Figure 5) and γ_t is an attenuation parameter which decreases over time. The fitted model $\eta_{it} = \alpha(x_i)$ assumes that there is no damping. It is seen that the effect of x_i is hardly recognizable from the estimate (with smoothing parameters optimized with respect to the Akaike Information Criterion) shown in Figure 3. The estimated function is a rather flat function and the only feature it shares with the true function is the decrease for low values of the covariate. The increase for higher values does not show up in the estimate.

In the following we will consider a model which allows for flexibility in several respects. The effects of covariates may change across time where the effects may be linear as in the parametric varying-coefficients model but the effects may also be represented by smooth functions which show varying intensities across time. The additive model and the parametric varying-coefficient model are limiting cases of this interaction between covariates and time. In Section 2 and 3 the model and estimation procedures are outlined. In Section 4 and 5

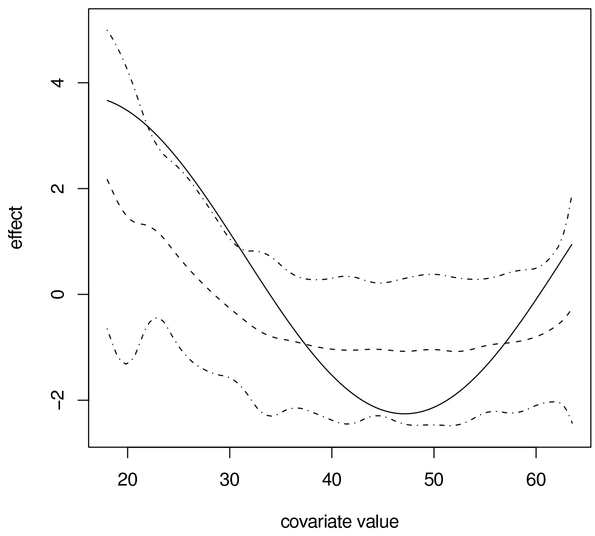

Figure 3: Estimate of the smooth function when fitting the model $\eta_{it} = \alpha(x_i)$ to data generated from $\eta_{it} = \gamma_t \alpha(x_i)$ (true function: solid line, mean of the estimated functions: dashed line, empirical confidence bands: dash-dot line)

simulation results and an application to duration time of patients in a psychiatric hospital are given.

2 The model

The model to be considered in the following is the discrete time survival model (1) with the generalized predictor for individual i at time t given by

$$\eta_{it} = \beta_{0t} + x'_{it}\beta_t + \sum_{j=1}^{S} \gamma_{tj} f_{(j)}(z_{itj}) \tag{3}$$

where $x'_{it} = (x_{it1}, \ldots, x_{itp})$, $z'_{it} = (z_{it1}, \ldots, z_{itm})$ are covariates and β_{0t}, $\beta'_t = (\beta_{t1}, \ldots, \beta_{tp})$, γ_{tj}, $t = 1, \ldots, T$, $j = 1, \ldots, S$, are unknown parameters. The essential components as far as the modelling of covariates is concerned are the linear varying-coefficients components and the additive component. The varying-coefficients component

$$\eta_{V,it} = x'_{it}\beta_t$$

allows that the effects of the x-covariates vary across time. In order to reduce the number of parameters to be estimated it has to be assumed that β_t varies at least in parts smoothly over time. By considering β_t as a function in time the varying-coefficients component becomes a semiparametric term reflecting the smooth interaction between the x-covariates and time. Of course in the limiting case one has $\beta_{tk} = \beta_{\cdot k}, t = 1, \ldots, T$, and therefore the case of a fixed effect is enclosed. Thus at least parts of $\eta_{V,it}$ may have simple linear parametric form.

The additive component

$$\eta_{A,it} = \sum_{j=1}^{S} \gamma_{tj} f_{(j)}(z_{itj})$$

reduces for $\gamma_{tj} = 1$ to a simple additive term where the effect of the z-variables has the unspecified functional form $f_{(j)}$. Thus for metrically scaled variables the restrictive linear form is avoided. However, for $\gamma_{tj} = 1$ the possibility of an interaction with time is excluded. If a covariate like age in a therapy study is distinctly nonlinear it has to be assumed that the effect does not change across time. In $\eta_{A,it}$ the interaction between a covariate and time is captured by introducing the parameters γ_{tj}. These parameters modify the effect of covariate z_{itj} which is given by $f_{(j)}(z_{itj})$ in the multiplicative form $\gamma_{tj} f_{(j)}(z_{itj})$. Therefore the effect of z_{itj}, nonlinear as it may be, is modified by time. If for example γ_{tj} decreases, i.e. $\gamma_{tj} > \gamma_{t+1,j}$, the γ-parameters act as damping constants which reduce the effect of z_{itj} across time. Model (3) gains flexibility in the additive part, however in exchange for the cost of introducing multiplicative components which makes straightforward use of established estimation procedures impossible. In the general form of $\eta_{A,it}$ the parameters γ_{tj} and the functions $f_{(j)}$ are not identifiable. Thus the restriction $\gamma_{1j} = 1$, $j = 1, \ldots, S$, is made.

In the following estimation is based on low-rank smoothers. We will consider direct fitting procedures by using penalized basis functions. The most prominent version of low-rank smoothers are penalized spines, also known as P-splines, as introduced by Eilers & Marx (1996). That means the unknown flexible parts of the model are represented as sums of basis functions. The unknown parts are the unknown functions $f_{(1)} \ldots, f_{(S)}$ and the parameters $\beta_{0t}, \beta_{tk}, \gamma_{tj}, k = 1, \ldots, p, j = 1, \ldots, S$, which are considered as functions in time, i.e. $\beta_0(t), \beta_k(t), \gamma_j(t)$. All these functions will be specified as sums of basis functions. Let m represent any of these functions, then one uses the form

$$m(x) = \sum_{s=1}^{M} \alpha_s B_s(x)$$

where $B_s(\cdot)$ are basic functions connected to knots κ_s. Of course basis functions as well as the number and location of knots may depend on the function which is fitted. The truncated power series used by Ruppert & Carroll (1999), Ruppert (2000) have the form

$$m(x) = \alpha_0 + \alpha_1 x + \alpha_2 x + \ldots \alpha_k x^k + \sum_s \alpha_{k+j}(x - \kappa_s)_+^k$$

where $(u)_+^k = u^k I(u \geq 0)$. In the following mostly B-splines will be used as basis functions. Eilers & Marx (1996) give many reasons for prefering B-splines. We found them to be numerically more stable than the truncated power series. It should be noted that parametric terms may be considered as special cases. If $M = 1, B_s(x) = x$ is chosen, then $m(x)$ has the form $m(x) = \alpha_1 x$ and is therefore simply parametric.

In order to have not to distinguish between the different terms in the predictor (3) in the following a simpler form is used. Appropriate choice of basis functions yields for the general predictor (3) the condensed additive form

$$\eta_{it} = \sum_{j=1}^{q} \gamma_{tj} \alpha_j' w_{itj} \tag{4}$$

where $w_{itj}' = (w_{itj1}, \ldots w_{itjP_j})$ is the vector of covariates, $\alpha_j' = (\alpha_{j1}, \ldots \alpha_{jP_j})$ and γ_{tj}, $t = 1, \ldots, T$, $j = 1, \ldots, q$, are unknown parameters. The observable variables w_{itj} are composed from the covariates x_{it} and z_{itj} and the knots. In order to show that (4) comprises (3) it is sufficient to demonstrate that the additive term $\gamma_{tj}\alpha_j' w_{itj}$ may take the form of an expansion in basis function of any of the terms used in (3). The choice $\gamma_{tj} = 1, w_{itjs} = x_{itj} B_s(t), s = 1, \ldots, M_j$, yields

$$\gamma_{tj}\alpha_j' w_{itj} = \sum_s x_{itj} \alpha_{js} B_s(t)$$

which represents the component $x_{itk}\beta_{tk}$ from $x_{it}'\beta_t$ where α_{js} are the weights and B_s are the basis functions used in the expansion of the varying coefficient β_{tk}. Smoothly varying coefficient β_{0t} corresponds to the special case where $x_{itk} = 1$. If just one knot is used and $B_1(t) = 1$ one obtains a non-varying parameter. The additive term $\gamma_{tj} f_{(j)}(z_{itj})$ in (3) follows directly from the expansion $f_{(j)}(z_{itj}) = \sum_s \alpha_{js} B_s(z_{itj})$. If no interaction between time and a function of a covariate is present one sets $\gamma_{tj} = 1$. In the following we will use the more compact form (4) of the predictor.

3 Estimation and inference

3.1 Penalized maximum likelihood estimates

Let the observations be given by (t_i, δ_i, x_i), $i = 1, \ldots, n$, where time $t_i = \min\{T_i, C_i\}$ is the minimum of failure time T_i and censoring time C_i with the indicator variable determined by

$$\delta_i = \begin{cases} 1 & \text{if } T_i < C_i \\ 0 & \text{if } T_i \geq C_i. \end{cases}$$

Assuming random censoring, the likelihood contribution of observation i is given by

$$L_i = c_i \lambda(t_i|\eta_{it_i})^{\delta_i} \prod_{j=1}^{t_i-1} (1 - \lambda(j|\eta_{ij})) \tag{5}$$

where the constant $c_i = P(C_i > t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}$ is considered as noninformative and in the following will be dropped. In (5) without loss of generality it is implicitely assumed that censoring takes place at the beginning of interval $[a_{t_i-1}, a_{t_i})$. Based on (5) it is easy to show that the log-likelihood function is given by

$$l = \sum_{t=1}^{q} \sum_{i \in R_t} y_{it} \log \lambda(t|\eta_{it}) + (1 - y_{it}) \log(1 - \lambda(t|\eta_{it})) \tag{6}$$

where y_{it} contains the transition of individual i from interval $[a_{t-1}, a_t)$ to $[a_t, a_{t+1})$ in the form

$$y_{it} = \begin{cases} 1 & \text{individual fails in } [a_{t-1}, a_t) \\ 0 & \text{individual survives in } [a_{t-1}, a_t). \end{cases}$$

The index set $R_t = \{i | s \leq t_i - (1 - \delta_i)\}$ represents the risk set, i.e. the individuals that are still under risk in interval $[a_{t-1}, a_t)$. For a derivation see e.g. Fahrmeir & Tutz (2001). It should be noted that for the derivation of (6) the time dependence of coefficients has no influence. From (6) it is easily seen that the likelihood is equivalent to the likelihood of a pseudo binary response model $P(y_{is} = 1|x_i) = h(\eta_{it})$ for binary observations $y_{11}, \ldots, y_{1,t_1-(1-\delta_1)}, y_{21}, \ldots, y_{n,t_n-(1-\delta_n)}$.

The basic concept of P-splines is to use the expansion in basis functions used in the construction of the predictor (4) together with maximization of the penalized likelihood. The penalty term penalizes the differences of the coefficient of adjacent basis functions where the ordering of the basis functions $B_1(\cdot), \ldots B_s(\cdot)$ is given through the ordering of the corresponding knots $\kappa_1 < \ldots < \kappa_s$. Following the suggestions of Eilers & Marx (1996) we will use many equally spaced knots and steer the smoothers by smoothing parameters. Ruppert (2000) investigated the number of knots and found that it has little influence for monotone functions. For non-monotone functions a number of knots given as $\min\{n/4, 35\}$ seems to perform well.

Assuming the predictor $\eta_{it} = \sum_j \gamma_{tj} \alpha_j' w_{itj}$ we use the penalized log-likelihood

$$l_p = l - \sum_j K_{1,j} + K_{2,j}.$$

The first penalty smoothes the parameters γ_{tj} across time and has the form

$$K_{1,j} = \frac{\tilde{\gamma}_j}{2} \sum_{t=2}^{T} (\gamma_{tj} - \gamma_{t-1,j})^2$$

with smoothing parameter $\tilde{\gamma}_j$. For reasons of identifiability $\gamma_{1j} = 1$ is fixed. The second penalty smoothes the weights $\alpha_{1j}, \ldots, \alpha_{jM_j}$ in the form

$$K_{2,j} = \frac{\tilde{\alpha}_j}{2} \sum_{s=1}^{M_j} (\alpha_{js} - \alpha_{j,s-1})^2$$

where $\tilde{\alpha}_j$ is the smoothing parameter. While $K_{1,j}$ is a penalty which ascertains smoothness across response categories, $K_{2,j}$ smoothes the effect of the covariate. We restrict the presentation to first order differences. Of course one could also consider higher order differences which in connection with higher order P-splines have nice properties in the limiting case where smoothing is maximal, i.e. $\tilde{\alpha}_j \to \infty$. Matrix forms of $K_{1,j}, K_{2,j}$ are given by

$$K_{1,j} = \frac{\tilde{\gamma}_j}{2} \{e_1' e_1 + 2e_1' P_\gamma \gamma_j + \gamma_{\cdot j}' P_\gamma' P_\gamma \gamma_{\cdot j}\}, \quad K_{2,j} = \frac{\tilde{\alpha}_j}{2} \alpha_j' P_\alpha' P_\alpha \alpha_j$$

where $\gamma'_{\cdot j} = (\gamma_{2j}, \ldots, \gamma_{Tj}), e_1^T = (-1, 0, 0, \ldots, 0)$, and P_γ, P_α are matrices with entries $0, 1, -1$ (for details see Appendix).

The estimation equations are obtained by considering the penalized score function. Collecting the parameters in

$$\delta' = (\gamma'_{\cdot 1}, \ldots, \gamma'_{\cdot q}, \alpha'_1, \ldots, \alpha'_q)$$

one obtains the score function

$$s_p(\delta) = \partial l_p / \partial \delta = (\partial l_p / \partial \gamma'_{\cdot 1}, \ldots, \partial l_p / \partial \alpha'_q)'$$

with components

$$s_{p,\alpha_j} = \frac{\partial l_p}{\partial \alpha_j} = \sum_{t=1}^{T} \sum_{i \in R_t} \gamma_{tj} w_{itj} \frac{\partial h(\eta_{it})}{\partial \eta} (y_{it} - \mu_{it}) / \sigma_{it}^2 - \tilde{\alpha}_j P'_\alpha P_\alpha \alpha_j$$

$$s_{p,\gamma_{tj}} = \frac{\partial l_p}{\partial \gamma_{tj}} = \sum_{i \in R_t} \alpha'_j w_{itj} \frac{\partial h(\eta_{it})}{\partial \eta} (y_{it} - \mu_{it}) / \sigma_{it}^2 - \tilde{\gamma}_j \{[e_1^T P_\gamma]_t + [P'_\gamma P_\gamma]_t \gamma_j\}$$

where $\mu_{it} = h(\eta_{it}), \sigma_{it}^2 = h(\eta_{it})(1 - h(\eta_{it}))$ and $[\quad]_t$ denotes the component t of a vector or in the case of a matrix the tth row. The estimation equation $s_p(\delta) = 0$ cannot be solved in the usual way by using (penalized) Fisher-scoring because the incorporation of the damping constants γ_{tj} makes the predictor non-linear. We suggest a stepwise procedure which uses estimation procedures from generalized linear models.

In the *first step* the parameters γ_{tj} are considered as fixed. For given parameters γ_{tj} the equations

$$\frac{\partial l_p}{\partial \alpha_j} = 0, j = 1, \ldots, q, \tag{7}$$

correspond to the fitting of a generalized linear model with linear predictor $\eta_{it} = \sum \alpha'_j \tilde{w}_{itj}$ where $\tilde{w}_{itj} = \gamma_{tj} w_{itj}$ are the observed covariates. Equation (7) may be solved by iterative (penalized) Fisher-scoring where $\alpha' = (\alpha'_1, \ldots, \alpha'_q)$ is updated by

$$\alpha^{(s+1)} = \alpha^{(s)} + F_{p,\gamma}^{-1}(\alpha^{(s)}) s_{p,\gamma}(\alpha^{(s)})$$

where $s_{p,\gamma}(\alpha)' = (s'_{p,\alpha_1}, \ldots, s'_{p,\alpha_p})$ is the score function evaluated at α and the corresponding penalized Fisher matrix evaluated at α is given by

$$F_{p,\gamma}(\alpha) = \sum_{t=1}^{T} \sum_{i \in R_t} \gamma_{tj}^2 w_{itj} w'_{itj} \frac{\partial h^2(\eta_{it}(\alpha))}{\partial \eta} / \sigma_{it}^2(\alpha) + P_{\tilde{\alpha}} \tag{8}$$

where $P_{\tilde{\alpha}}$ is given as block diagonal matrix $P_{\tilde{\alpha}} = Diag(\tilde{\alpha}_1 P'_\alpha P_\alpha, \ldots, \tilde{\alpha}_q P'_\alpha P_\alpha)$. The index γ is used because of the dependence on the given parameters γ_{tj}.

The *second step* uses the estimate of α which results from the first step. If the parameters $\alpha_1, \ldots, \alpha_q$ are given the equations

$$\frac{\partial l_p}{\partial \gamma_{tj}} = 0, t = 1, \ldots, T, j = 1, \ldots, q \tag{9}$$

correspond to the fitting a generalized linear model with predictor $\eta_{it} = \sum_{j=1}^{q} \gamma_{tj} \tilde{w}_{itj}$ where $\tilde{w}_{itj} = \alpha_j' w_{itj}$ is the linear predictor. Let the parameters be rearranged by $\gamma' = (\gamma_{\cdot 1}', \ldots, \gamma_{\cdot p}')$ with $\gamma_{\cdot j}' = (\gamma_{2j}, \ldots, \gamma_{Tj})$. The Fisher scoring takes the form

$$\gamma^{(s+1)} = \gamma^{(s)} + F_{\alpha}^{-1}(\gamma^{(s)}) s_{\alpha}(\gamma^{(s)})$$

where $s_{\alpha}(\gamma)' = (s_{\gamma_{11}}, \ldots, s_{\gamma_{Tq}})$ is the score vector evaluated at γ and

$$F_{\alpha}(\gamma) = \sum_{t=1}^{T} \sum_{i \in R_t} v_{it} v_{it}' \frac{\partial h^2(\eta_{it})}{\partial \eta} / \sigma_{it}^2 + P_{\tilde{\gamma}}$$

is the corresponding Fisher matrix where $v_{it}' = (0_q', \ldots, \tilde{w}_{it}', \ldots, 0_q')$, $\tilde{w}_{it} = (\alpha_1' w_{it1}, \ldots, \alpha_q' w_{itq})$ and 0_q is a q-dimensional vector containing zeros, i.e. $0_q' = (0, \ldots, 0)$. The penalty is given as the block diagonal matrix $P = Diag(\tilde{\gamma}_1 P_{\gamma}' P_{\gamma}, \ldots, \tilde{\gamma}_q P_{\gamma}' P_{\gamma})$. The Fisher matrix F_{α} depends on the parameters α which are contained in v_{it}.

Penalized maximum estimates are computed by solving (7) with starting values γ_{tj}. Then (9) is solved by considering the obtained estimates $\hat{\alpha}_j$ as known. The resulting estimates $\hat{\gamma}_{tj}$ are then used as known to solve (7). The iteration of these cycles is ended if both estimates do no longer change.

3.2 Inference

Let us first consider the case without multiplicative terms, i.e. $\gamma_{tj} = 1, t = 2, \ldots, T, j = 1, \ldots, q$. Then the estimation is reduced to the first step where (7) is solved. As estimator for the covariance Marx & Eilers (1998) propose for this case the sandwich matrix

$$\hat{\text{cov}}(\hat{\alpha}) = F_{p,1}^{-1}(\hat{\alpha}) F_1(\hat{\alpha}) F_{p,1}^{-1}(\hat{\alpha}) \tag{10}$$

where $F_{p,1}$ is the special form of (8) where $\gamma_{tj} = 1$, i.e.

$$F_{p,1} = \sum_{t=1}^{T} \sum_{i \in R_t} w_{itj} w_{itj}' \frac{\partial h^2(\eta_{it})}{\partial \eta} / \sigma_{it}^2 + P_{\tilde{\alpha}}$$

and F_1 is equivalent to $F_{p,1}$ without penalization term, i.e. $\tilde{\alpha}_j = 0$ yielding $P_{\tilde{\alpha}}$ as a matrix containing only zeros. If one assumes that the number of knots is large enough to represent the underlying smooth structure of influential terms the covariance $\hat{\text{cov}}(\hat{\alpha})$ is an approximation based on asymptotic arguments when $n \to \infty, \tilde{\alpha}_j / n \to 0$ (see also Tutz & Scholz, 2000).

If interaction between time and a function of covariates is in the model in the form of parameters γ_{tj} the penalized Fisher matrices used in the two estimation steps are of limited use since they assume that either γ or α is known. Then the covariance is based on the total penalized Fisher matrix $F_p = \partial l_p / \partial\delta\partial\delta'$. The approximation

$$\text{cov}(\hat{\delta}) = F_p^{-1}(\hat{\delta})F(\hat{\delta})F_p^{-1}(\hat{\delta})$$

holds if it can be assumed that the chosen knots and basis functions represent the underlying functions. This condition is weak when many basis functions, say 30, are used with equidistant knots. In addition non-informative censoring is assumed which makes it possible to treat the observations as asymptotically uncorrelated responses y_{it}. A short sketch of the approximation and the total Fisher matrix is given in the Appendix.

4 Simulation

4.1 Method

In a small simulation study the gain in efficiency by adequate modelling and the performance of the basis function approach is investigated. The two concepts introduced, time-varying effects of variables and time-varying effects of smoothly transformed variables, are evaluated separately. In the first case one compares

$$\eta_{it} = \beta_{0t} + x_{it}'\beta \tag{11}$$

to

$$\eta_{it} = \beta_{0t} + x_{it}'\beta_t. \tag{12}$$

The simulated model is given by

$$\eta_{it} = \beta_{0t} + x_{it}'b\left[\left(1 - \frac{1}{1+\exp(-c\cdot(t-1))}\right)\cdot 2\right] \tag{13}$$

where b determines the covariate effect at $t = 1$ and c determines how fast the effect decreases over time. Therefore b and c form a time dependent parameter β_t. Figure 4 shows the attenuation of β_t for several values of c. For simplicity $\beta_{0t} = \beta_0$ is chosen.

In the second case one compares

$$\eta_{it} = \beta_{0t} + \sum_{j=1}^{S} f_{(j)}(z_{itj}) \tag{14}$$

to

$$\eta_{it} = \beta_{0t} + \sum_{j=1}^{S} \gamma_{tj} f_{(j)}(z_{itj})\,. \tag{15}$$

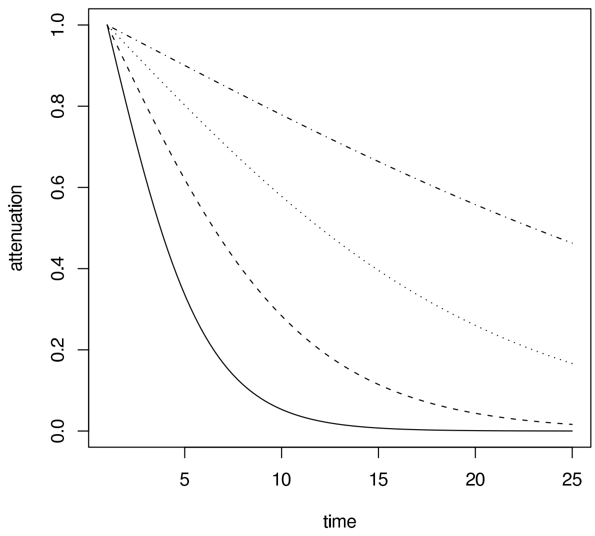

Figure 4: Attenuation component of the generating models (13) and (16) for different values of c (0.4: solid line, 0.2: dashed line, 0.1: dotted line, 0.05: dash-dot line).

The generating model (with $S = 1$) then is

$$\eta_{it} = \beta_{0t} + \left(\sin\left(\frac{x_{it}}{10}\right) b_1 + b_2\right) \left[\left(1 - \frac{1}{1 + \exp(-c \cdot (t-1))}\right) \cdot 2\right]. \tag{16}$$

The difference to the first model is that b is replaced by a nonlinear term. Figure 5 shows the nonlinear term for various values of b_1 and b_2 which are chosen such that it is centered for the range of x_{it} used.

The measure M used for comparison is squared error

$$\mathrm{SE} = \sum_{it} (p_{it} - \hat{p}_{it})^2$$

with $t = 1, \ldots, T$ and T fixed (For the results presented here $T = 15$) and p_{it} denoting the probability from model (13) or (16) transformed into the probability of a multinomial distribution with $\sum_t p_{it} = 1$. With $\lambda_{it} = \lambda(t|x_i)$ denoting the hazard function it is obtained by $p_{it} = \lambda_{it} \prod_{s=1}^{t-1} (1 - \lambda_{is})$. The estimate $\hat{p}_{it}$ from model (11), (12), (14) or (15) is given in the same way with λ_{it} being replaced by $\hat{\lambda}_{it} = h(\hat{\eta}_{it})$.

An alternative measure we used for comparison is Kullback-Leibler or entropy

$$\mathrm{E} = \sum_{it} p_{it} \log\left(\frac{p_{it}}{\hat{p}_{it}}\right).$$

The results from this measure did not differ much from squared error. For this reason squared error will be reported exclusively.

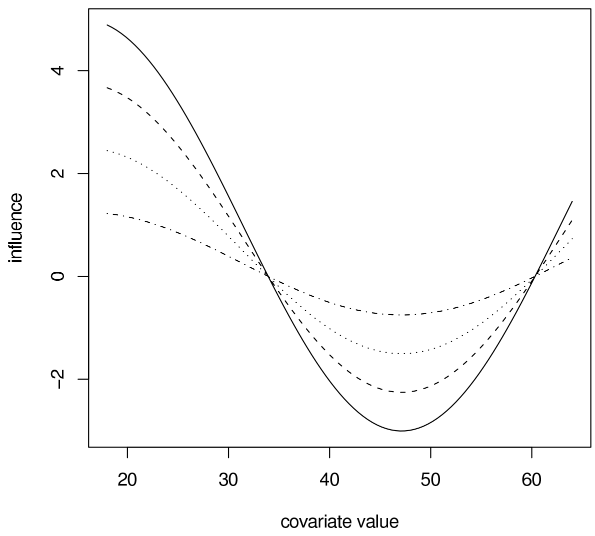

Figure 5: Smooth covariate component of the generating model (16) for various values of b_1/b_2 (4/0.9914: solid line, 3/0.7435: dashed line, 2/0.4957: dotted line, 1/0.2478: dash-dot line).

The improvement achieved by adequate modelling is evaluated by the ratio of measures

$$\mathrm{R} = \frac{\mathrm{M(flexible\ model, AIC\text{-}smoothing)}}{\mathrm{M(restricted\ model, AIC\text{-}smoothing)}} \tag{17}$$

where the flexible model is the adequate model for the underlying process, i.e. (12) for time-varying coefficients and (15) for attenuated smooth effects. The restricted models are (11) and (14). In order to obtain an adequate fit for each simulated data set the smoothing parameters are chosen from the data set once for the flexible and once for the restricted model. The algorithm used for searching the optimal smoothing parameters (i.e. the optimal penalties) in terms of the Akaike Information Criterion (AIC) is Powell's unconstrained optimization by quadratic optimization (UOBYQA) (Powell, 2002). It forms quadratic models by interpolation and minimizes with respect to these models. If the derivatives exist (albeit they don't have to be evaluated) an optimum will be found. The algorithm has been chosen because it requires no derivatives, which for the AIC are not obtained easily. In order to meet the constraint that the smoothing parameters may not be smaller than zero, optimization by the AIC is done for the the logarithm of the smoothing parameters. To evaluate the effect of the true structure behind the data the data generation has been varied systematically. The parameters b and c of the generating model (13) and b_1, b_2 and c of model (16) determine how strong the covariate influence and the variation of attenuation are. The baseline parameters β_{0t} are set constant over time, i.e. $\beta_{0t} = \beta_0$ for all t. Nevertheless all models fitted include a baseline that is allowed to vary over time. The value of β_0 does not seem to influence the results systematically and is chosen such that it yields a probability of 0.1 when no covariate influence is present. There also does not seem to be a systematic influence of the distribution of the covariate values and so a uniform distribution is chosen. The values are between -23 and 23 for the generating model (13) and between 18 and 64 for model (16). It is to be expected that the result of the comparisons will depend on the

Table 1: Median of the ratio 'SE(model with time-varying effects)/SE(model with fixed effects)' for different data generation parameters (using generating model (13)) with 100 comparisons for each combination.

		c				
		.1		.2		.4
		n=100	n=20	n=50	n=100	n=100
	.1	0.6605	1.001	0.7654	0.4706	0.3296
b	.2	0.2847	0.7911	0.4034	0.2026	0.1548
	.3	0.1788	0.6662	0.2889	0.1323	0.1365

number of subjects in the data set and so the simulation is done with 20, 50, 100 and 200 subjects. For each combination of parameters 100 simulations have been done.

4.2 Results

4.2.1 Time-varying effects of variables

Table 1 shows the improvement obtained by including the time-varying effects for different combinations of parameters of the generating varying-coefficients model (13). The values given are the medians of the ratio (17) for the squared error measure being computed for each simulated data set individually. Since the distribution of these ratios is not symmetric the mean values are larger than the median. The standard deviation varies from 0.0950 to 2.0815, but for most parameter combinations it is smaller than 0.4.

As a general tendency the median becomes smaller, i.e. the flexible model performs much better, when b gets larger. That means that when covariate influence is stronger flexible/time-varying modelling becomes more important. In a similar way the improvement by using the time-varying model becomes more distinct when the decrease across time is stronger. The relation between the number of subjects used for each data set and the medians is as one would expect. An increased number subjects increases the benefit of using a more flexible model. In general it is seen that misspecification by falsely assuming constant parameters may yield very bad estimates. The incorporation of varying coefficients reduces the mean squared error strongly. Even for moderate values (e.g. $b = 0.2$) the reduction of mean squared error is between 84% and 20%.

Figure 6 shows the estimated smooth function for β_t from the flexible model (12) for one of the parameter combinations in Table 1 based on 100 simulations. The mean estimate (dotted line) is quite close to the true function (solid line). The estimated pointwise confidence bands which are based on the sandwich matrix (mean $\pm$ 1.96 $\cdot$ standard error)(dash-dot line) are well comparable to the empirical confidence bands (dashed line) although they are slightly

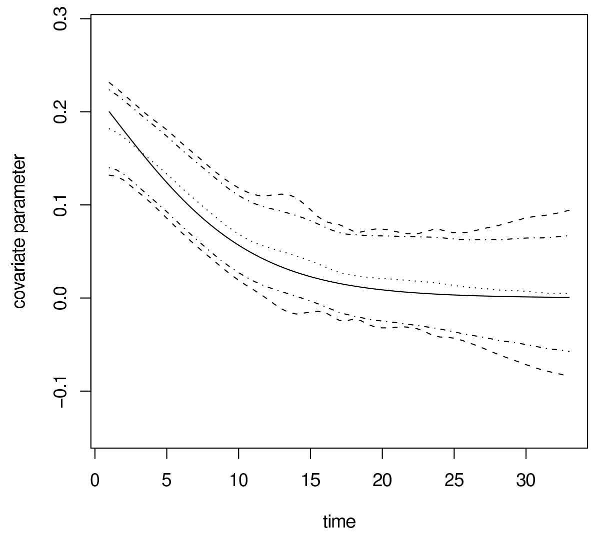

Figure 6: Estimation results for β_t based on the the varying coefficient model (13) with generating model parameters $b = 0.2$ and $c = 0.2$ (true function: solid line, mean of the estimated functions: dotted line, empirical confidence bands: dashed line, mean of the estimated confidence bands: dash-dot line)

smaller. If the number of subjects (here: 100) increases the difference between the empirical and estimated standard errors becomes even smaller.

4.2.2 Time-varying effects of smoothly transformed variables

Table 2 shows the median values for the comparison of a model with time-varying effects of smoothly transformed variables to a model that incorporates only smoothly transformed variables but no time-varying effects. The comparison has been done with different parameters for the generating model. The median of the ratio of measures (17) depends on the parameters that determine the strength of the covariate effect (here: b_1 and b_2). A stronger covariate effect (i.e. a larger value of b_1) is beneficial for the flexible model. Also when the decrease of attenuation over time becomes stronger (i.e. a larger value of c) the benefit of using the time-varying model increases. An increase in the number of subjects also increases the benefit of using the time-varying model. Even in the case of a rather flat function and weak time dependence the flexible model is rather stable, although it has no advantage over the time-independent model.

The left panel of Figure 7 shows the estimated smooth function (dotted lines) from the model with time-varying effects. The right panel shows the corresponding estimated $\hat{\gamma}_t$ values (circles). Albeit the true function (solid line) is within the confidence bands (empirical and estimated) the mean estimate is slightly larger than the true values. This effect has been found in all the simulations. It might be attributed to the penalization of variation in the

Table 2: Median of the ratio 'SE(model with attenuation of smooth effect of covariates)/SE(model with fixed effects of smooth effect of covariates)' for different data generation parameters based on generating model (16) for 100 comparisons for each combination.

		c			
		.2	.4		
		n=100	n=50	n=100	n=200
	1 / .2478	1.008	1.006	1.001	0.8057
b_1 / b_2	2 / .4957	0.9910	0.9982	0.5643	0.3594
	3 / .7435	0.7922	0.8587	0.5874	0.3648

estimation procedure and the fixed value of $\hat{\gamma}_1 = 1$, as this may lead to a function that is forced to decrease only slowly from the starting value. When working with increasing attenuation functions the estimates were consistently too small. This effect could explain why the estimate of the smooth function seems to be flatter than the true function (to be seen for example in the left panel of Figure 7): The combination of larger values of $\hat{\gamma}_t$ and smaller values of the estimated smooth function produces results similar to those from the true functions.

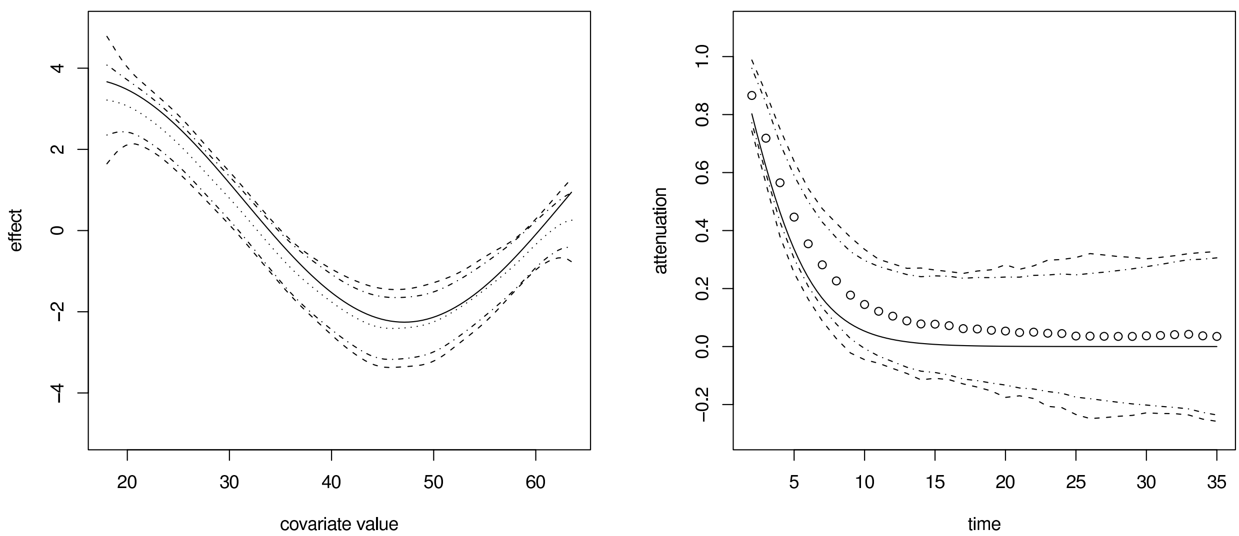

Figure 7: Estimates of the smooth function (left panel) and γ_t (right panel) with data from generating model (16) with parameters $b_1 = 3$, $b_2 = .7435$ and $c = .4$ (real function: solid line/circles, mean of the estimated functions: dotted line, empirical confidence bands: dashed line, mean of the estimated confidence bands: dash-dot line).

The mean estimated confidence bands (dash-dot lines) for the estimate of the smooth function (left panel of Figure 7) as well as for the γ_ts are smaller than the empirical confidence bands. As with the estimate of the time-varying parametric effects this difference decreases when more subjects are available.

5 Application to hospital data

The data presented here are from 1922 patients of a German psychiatric hospital with diagnosis "schizophrenic disorder of paranoid type". The response is time spent in hospital measured in days. The covariates available for smooth modelling are age at admission, calendar time between 1/1/1995 and 12/31/1999 (measured in days since 1/1/1995) and GAF (Global Assessment of Functioning) score at admission which is a physician's judgment of the patient's level of functioning. The other covariates are 0-1-coded variables and are given by: gender (MALE=1: male), education (EDU=1: above high school level), partner situation (PART=1: has a permanent partner), job situation (JOB=1: full/part time job at admission), first hospitalization (FIRST=1: first admission in a psychiatric hospital) and suicidal action (SUI=1: suicidal act previous to admission).

The predictor of the model for "probability of dismissal" has the form

$$\begin{aligned}\eta_{it} &= \beta_{0t} + f_T(\text{calendar time}) \\ &\quad + \text{MALE} \cdot \beta_{\text{MALE}} + \text{EDU} \cdot \beta_{\text{EDU}} + \text{PART} \cdot \beta_{\text{PART}} \\ &\quad + \text{JOB} \cdot \beta_{\text{JOB}} + \text{FIRST} \cdot \beta_{\text{FIRST}} + \text{SUI} \cdot \beta_{\text{SUI}} \\ &\quad + f_{\text{A}}(\text{age}) + \gamma_t \cdot f_{\text{G}}(\text{GAF score})\end{aligned}$$

where f_{T}, f_{A} and f_{G} are smooth functions. A specific feature of the model is the incorporation of two time scales, namely the time which patients stay in the hospital and the calendar time. The flexibility of the model allows to model both effects.

The predictor given is already the result of model selection. Starting from a general model the model has been selected based on the Akaike Information Criterion (AIC). Table 3 gives the AIC for selected models. In particular this criterion favors the inclusion of a time-dependent smooth term for "GAF score at admission" over a (time-dependent) parametric or a fixed smooth term.

Table 3: AIC for different models for "probability of dismissal" with optimal smoothing parameters ("..." refers to the binary components mentioned in the text, "ct" to calendar time, "age" to age at admission and "GAF" to GAF score at admission).

Model	AIC
$\eta_{it} = \beta_{0t} + \text{ct} \cdot \beta_{\text{T}} + \ldots + \text{age} \cdot \beta_{\text{A}} + \text{GAF} \cdot \beta_{\text{G}}$	15790.7
$\eta_{it} = \beta_{0t} + \text{ct} \cdot \beta_{\text{T}} + \ldots + \text{age} \cdot \beta_{\text{A}} + \gamma_t \cdot f_{\text{G}}(\text{GAF})$	15775.3
$\eta_{it} = \beta_{0t} + f_{\text{T}}(\text{ct}) + \ldots + \text{age} \cdot \beta_{\text{A}} + \gamma_t \cdot f_{\text{G}}(\text{GAF})$	15773.8
$\eta_{it} = \beta_{0t} + f_{\text{T}}(\text{ct}) + \ldots + f_{\text{A}}(\text{age}) + f_{\text{G}}(\text{GAF})$	15777.1
$\eta_{it} = \beta_{0t} + f_{\text{T}}(\text{ct}) + \ldots + f_{\text{A}}(\text{age}) + \text{GAF} \cdot \beta_{\text{G}t}$	15772.6
$\eta_{it} = \beta_{0t} + f_{\text{T}}(\text{ct}) + \ldots + f_{\text{A}}(\text{age}) + \gamma_t \cdot f_{\text{G}}(\text{GAF})$	15770.3

Table 4: Estimates of the parametric terms of the model for "probability of dismissal"

covariate	parameter estimate	standard deviation
gender (male)	-0.0525	0.0550
education	-0.0387	0.0716
partner situation	0.2619	0.0654
job situation	0.0428	0.0731
first hospitalization	0.0845	0.0680
suicidal action	-0.0828	0.1145

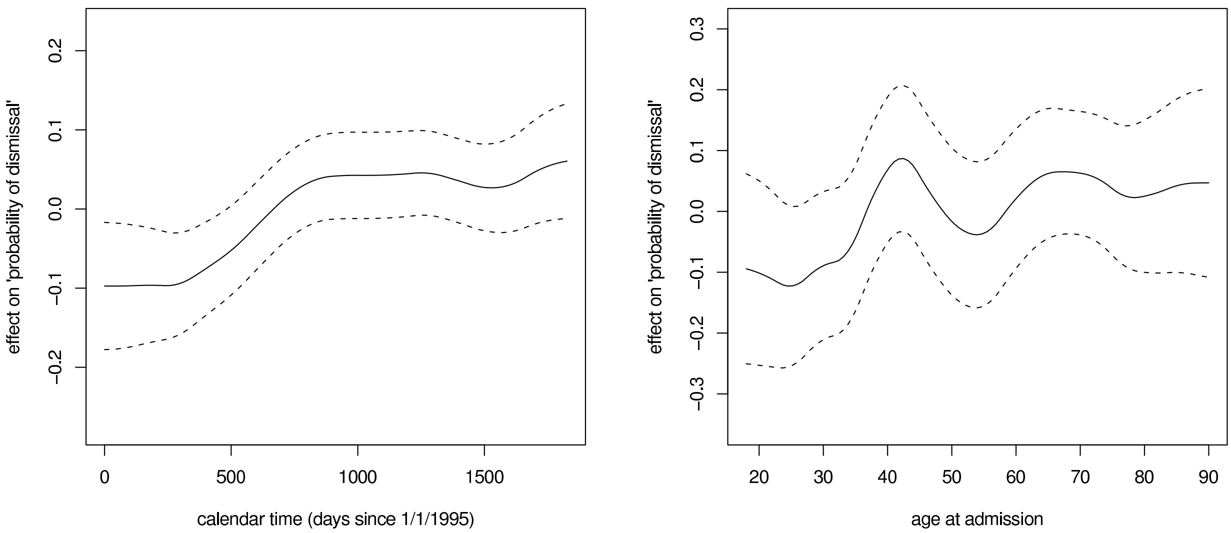

Figure 8: Estimated smooth effect of calendar time f_{T} (left panel) and smooth effect of age at admission f_{A} (right panel) of the model for "probability of dismissal" (estimates: solid lines; confidence bands: dashed lines).

Table 4 shows the parameter estimates of the parametric terms. The only variable that seems have influence is partner situation. The effect for this variable is rather strong. If the patient has a permanent partner the time spent in the hospital is strongly reduced. The rest of the variables has no significant effect. For schizophrenia as an endogenic disorder it is not to be expected that variables like job situation or gender have an influence on the healing process. The right panel of Figure 8 shows the smooth estimate of the effect of the variable age at admission. With increasing age the stay in the hospital is shortened. The effect of calendar time is shown in the left panel of Figure 8. It is seen that the time spent in hospital decreases almost continuously with calendar time. However it is not clear if the effect is more connected to the organization of the hospital or to new developments in the treatment of schizophrenia.

An interesting effect is that of the GAF score which is an assessment score at admission. Figure 9 shows the estimated effect f_G in the left panel and the corresponding time-varying

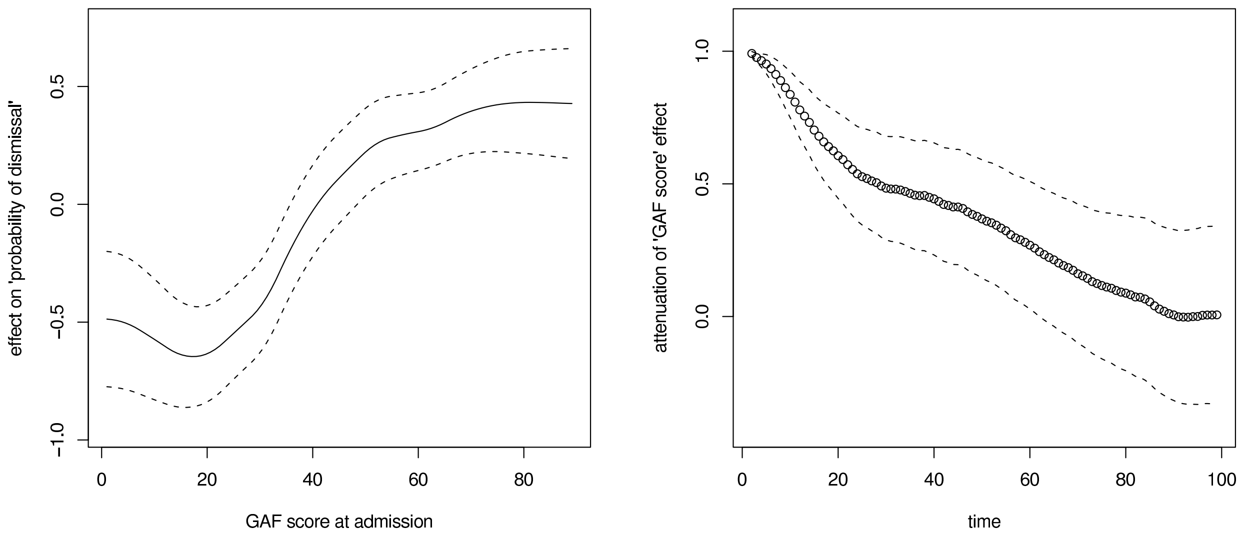

Figure 9: Estimated smooth effect of the covariate "GAF score at admission" on "probability of dismissal" (left panel) and its attenuation over time (right panel) (estimate: solid line/circles; confidence bands: dashed lines).

attenuation parameters γ_t in the right panel. The shape of the smooth function clearly indicates that a lower GAF score at admission (indicating a lower level of functioning) results in a lower probability of dismissal. This is reasonable as a patient in a worse condition can be expected to require longer treatment. The plateaus at scores under 30 and over 60 also deserve attention. They indicate that the essential difference is between low and high GAF score. Only for a small window between 30 and 50 points the effect is changing. Moreover, the curves clearly show that a linear effect with time-varying coefficients is not appropriate. This is also supported by the Akaike criterion which favors the time-dependent smooth component for the covariate "GAF score at admission" over a time-dependent parametric component. The estimate of γ_t (right panel of Figure 9) shows that the effect of the GAF score at admission vanishes over time. This is plausible as the condition of a patient is expected to change over time and so the predictive power of the initial score diminishes.

Appendix

1. Penalization
   Consider with restriction $\gamma_1 = 1$ and $\gamma' = (\gamma_2, \ldots, \gamma_T)$ the penalty $K = \sum_{t=2}^{T}(\gamma_t - \gamma_{t-1})^2$ which has the form
   $$K = (e_1 + P_\gamma\gamma)^T(e_1 + P_\gamma\gamma)$$
   where
   $$e_1' = (-1, 0, \ldots, 0), \quad P_\gamma \quad = \quad \begin{pmatrix} 1 & 0 & 0 & & 0 \\ -1 & 1 & 0 & & 0 \\ 0 & -1 & 1 & & 0 \\ \vdots & & \ddots & -1 & 1 \end{pmatrix}$$
   One obtains
   $$K = e_1^T e_1 + 2e_1^T P_\gamma\gamma + \gamma' P_\gamma' P_\gamma\gamma.$$
   For $K = \sum_{s=1}^{M}(\alpha_{js} - \alpha_{j,s-1})^2$ one obtains the simpler form
   $$K = (P_\alpha\alpha)'(P_\alpha\alpha) = \alpha' P_\alpha' P_\alpha\alpha$$
   where $\alpha' = (\alpha_1, \ldots, \alpha_M)$ and the $(M-1) \times M$ matrix P_α is given by
   $$P_\alpha = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & & \\ & & & -1 & 1 \end{pmatrix}$$

2. Approximative covariance
   The information matrix that corresponds to the penalized likelihood is given by $F_p(\delta) = E(-\partial l_p/\partial\delta\partial\delta')$. Its components are given by
   $$\begin{aligned}
   E\left(-\frac{\partial l_p}{\partial\alpha_j\partial\alpha_j'}\right) &= \sum_{t=1}^{T}\sum_{i\in R_t}\gamma_{tj}^2 w_{itj}w_{itj}'\left(\frac{\partial h(\eta_{it})}{\partial\eta}\right)^2/\sigma_{it}^2 + \tilde{\alpha}_j P_\alpha' P_\alpha, \\
   E\left(-\frac{\partial l_p}{\partial\alpha_j\partial\alpha_l}\right) &= \sum_{t=1}^{T}\sum_{i\in R_t}\gamma_{tj}\gamma_{tl}w_{itj}w_{itl}'\left(\frac{\partial h(\eta_{it})}{\partial\eta}\right)^2/\sigma_{it}^2, l \neq j, \\
   E\left(-\frac{\partial l_p}{\partial\gamma_{tj}\partial\gamma_{tj}}\right) &= \sum_{i\in R_t}(\alpha_j' w_{it})^2\left(\frac{\partial h(\eta_{it})}{\partial\eta}\right)^2/\sigma_{it}^2 + \tilde{\gamma}_j[P_\gamma' P_\gamma]_{tt}, \\
   E\left(-\frac{\partial^2 l_p}{\partial\gamma_{tj}\partial\gamma_{\tilde{t}j}}\right) &= \tilde{\gamma}_j[P_\gamma' P_\gamma]_{t\tilde{t}} \qquad , \qquad \tilde{t} \neq t, \\
   E\left(-\frac{\partial^2 l_p}{\partial\alpha_j\partial\gamma_{tj}}\right) &= \sum_{i\in R_t}\gamma_{tj}w_{itj}w_{itj}'\alpha_j\left(\frac{\partial h(\eta_{it})}{\partial\eta}\right)^2/\sigma_{it}^2, \\
   E\left(-\frac{\partial^2 l_p}{\partial\alpha_j\partial\gamma_{tl}}\right) &= \sum_{i\in R_t}\gamma_{tj}w_{itl}w_{itj}'\alpha_l\left(\frac{\partial h(\eta_{it})}{\partial\eta}\right)^2/\sigma_{it}^2, l \neq j,
   \end{aligned}$$

where $[\ ]_{t\tilde{t}}$ denotes the elements $(t\tilde{t})$ of a matrix.

The components of the Fisher matrix without penalty $F(\delta)$ contain the first terms as given above, i.e. with $\tilde{\alpha}_j = \tilde{\gamma}_j = 0$. Thus one has $F_p(\delta) = F(\delta) + P$ where P is the penalty matrix depending on $\tilde{\alpha}_j, \tilde{\gamma}_j$. In the same way the score function $s = \partial l/\partial\delta$ is connected to the penalized score function $s(\delta) = \partial l/\partial\delta$ by $s_p(\delta) = s(\delta) + P\delta$. Common assumptions are $\partial l/\partial\delta = 0_p(n^{1/2}), \partial l/\partial\delta\partial\delta' = -F + 0_p(n^{1/2})$ with $F^{-1} = 0(n^{-1})$ and similar assumptions for higher order derivatives. Then the expansion

$$0 = s_p(\hat{\delta}) = s_p(\delta) + \frac{\partial s_p}{\partial\delta'}(\hat{\delta} - \delta) + \ldots$$

yields

$$\begin{aligned} \hat{\delta} - \delta &= \left(-\tfrac{\partial s_p}{\partial\delta'}\right)^{-1} s_p(\delta) + 0_p(n^{-1}) + 0_p(n^{-3/2}\tilde{\lambda}) \\ &= (F(\delta) + P)^{-1}\,(s(\delta) + P\delta) + 0_p(n^{-1}) + 0_p(n^{-3/2}\tilde{\lambda}) \end{aligned}$$

where $\tilde{\lambda} = \max\{\tilde{\lambda}_j, \tilde{\alpha}_j\}$. Based on this expansion the covariance may be shown to have the form

$$\begin{aligned} \mathrm{cov}(\hat{\delta}) &= (F(\delta) + P)^{-1}\mathrm{cov}(s(\delta))(F(\delta) + P)^{-1} + 0(n^{-2}) + 0(n^{-3}\lambda^2) \\ &= F_p(\delta)^{-1}F(\delta)F_p(\delta)^{-1} + 0(n^{-2}) + 0(n^{-3}\lambda^2). \end{aligned}$$

3. Akaike Information Criterion (AIC)
   Following Hastie & Tibshirani (1990) and Eilers & Marx (1996) the AIC is computed by

   $$\text{AIC} = \text{deviance} + 2 \cdot \text{(effective) dimensions of the vector of parameters } \delta$$

   wherein the dimension component is given by the trace of the smoother matrix. The smoother matrix is obtained by treating the Fisher matrix introduced above as if it would be from a *one step* iterative estimation procedure. The smoother matrix then is

   $$H = (F + P)^{-1}F.$$

Acknowledgment
We thank the German Science Foundation DFG for financial support.

References

Abrahamowicz, M., MacKenzie, T., Leffondre, K., du Berger, R. & Siemiatycki, J. (2002). Joint modelling of time-dependent and non-linear effects of continuous predictors in survival analysis. In *Proceedings of the 17th International Workshop on Statistical Modelling*.

Aerts, M., Claeskens, G. & Wand, M. (1999). Some theory for penalized spline generalized additive models. Preprint.

Arjas, E. & Haara, P. (1987). A logistic regression model for hazard: Asymptotic results. *Scand. J. Statist.* **14**, 1–18.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* **B 34**, 187–220.

Eilers, P. H. C. (1998). Hazard smoothing with B-splines. In *Proceedings of the 13th International Workshop on Statistical Modelling*.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer, 2nd ed.

Fahrmeir, L. & Wagenpfeil, S. (1996). Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *J. Amer. Statist. Assoc.* **91**, 1584–1594.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc.* **B 55**, 757–796.

Henderson, R. & Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *J. Roy. Statist. Soc.* **B 61**, 367–379.

Kalbfleisch, J. & Prentice, R. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Kauermann, G. & Tutz, G. (2000). Local likelihood estimation in varying-coefficient models including additive bias correction. *Journal of Nonparametric Statistics* **12**, 343–371.

Kauermann, G. & Tutz, G. (2001). Vanishing of risk factors for the success and survival of newly founded companies. Discussion Paper 242, SFB 368, LMU München.

Keiding, N., Andersen, P. K. & Klein, J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* **16**, 215–224.

Linton, O. B. & Härdle, W. (1996). Estimation of additive regression models with known links. *Biometrika* **83**, 529–540.

Marx, B. D. & Eilers, P. H. C. (1998). Direct generalized additive modelling with penalized likelihood. *Comp. Stat. & Data Analysis* **28**, 193–209.

Powell, M. J. D. (2002). UOBYQA: unconstrained optimization by quadratic approximation. *Math. Program.* **92**, 555–582.

Ruppert, D. (2000). Selecting the number of knots for penalized splines. Preprint.

Ruppert, D. & Carroll, R. J. (1999). Spatially-adaptive penalties for spline fitting. *Australian Journal of Statistics* **42**, 205–223.

Ryu, K. (1994). Group duration analysis of the proportional hazard model: Minimum chi-squared estimators and specification tests. *J. Amer. Statist. Assoc.* **42**, 1386–1397.

Tutz, G. & Scholz, T. (2000). Semiparametric modelling of multicategorical data. Discussion Paper 209, SFB 368, LMU München.

Vaupel, J. . W., Manton, K. G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454.

Verweij, P. J. M. & van Houwelingen, J. C. (1995). Time-dependent effects of fixed covariates in cox regression. *Biometrics* **51**, 1550–1556.