CEM
Center for Empirical Macroeconomics

Working Paper No. 56

Penalised Spline Smoothing in Multivariable Survival Models with Varying Coefficients

by

Göran Kauermann

University of Bielefeld
Department of Economics
Center for Empirical Macroeconomics
P.O. Box 100 131

33501 Bielefeld, Germany

Penalised Spline Smoothing in Multivariable Survival Models with Varying Coefficients

Göran Kauermann*
Universität Bielefeld

28th July 2003

Abstract

The paper discusses penalised spline (P-spline) smoothing for hazard regression of multivariable survival data. Non-proportional hazard functions are fitted in a numerically handy manner by employing Poisson regression which results from numerical integration of the cumulative hazard function. Multivariate smoothing parameters are selected by utilizing the connection between P-spline smoothing and Generalised Linear Mixed Models. A hybrid routine is suggested which combines the Mixed Model idea with a classical Akaike information criteria. The model is evaluated with simulations and applied to data on the success and failure of newly founded companies.

KEYWORDS: Generalized Linear Mixed Model, P-Spline Smoothing, Survival Model, Varying Coefficient Model,

*University Bielefeld, Postfach 300131, 33501 Bielefeld, Germany. E-mail: gkauermann@wiwi.uni-bielefeld.de

1 Introduction

Modeling of survival data is largely dominated by the proportional hazard (PH) model introduced by Cox (1972). Even though the PH model appeals by simple numerical fitting based on the partial likelihood, the proportional hazard assumption often restricts the model in applications since it means that covariate effects remain constant over survival time. This assumption has been under major investigation and numerous papers suggest extensions and testing procedures, see for instance O'Sullivan (1988), O'Quigley & Pessione (1989), Hastie & Tibshirani (1990), Gray (1994), Hess (1994) or Abrahamowicz, MacKenzie & Esdaile (1996). For a general overview of estimation and tests in proportional hazard models we also refer to Lin & Wei (1991) or Sasieni (1999). Allowing covariate effects to be dynamic in time leads to a varying coefficient model as generally introduced by Hastie & Tibshirani (1993). Here, constant covariate effects are replaced by smooth but unknown functions. Smooth estimation can then be carried out using e.g. Spline fitting, as in Hastie & Tibshirani (1993), see also Kooperberg, Stone & Troung (1995) or by applying local techniques, see e.g. Fan, Gijbels & King (1997) or Cai & Sun (2003).

Smooth estimation in survival models is usually based on the partial likelihood function. There are, however, two points of criticism which should be raised against the use of the partial likelihood in the context of smoothing. First, in the simple case that covariate effects are in fact constant over time, that is if the PH assumption holds, the cumulative (integrated) hazard function in the likelihood function factorizes to the cumulative baseline hazard multiplied by the covariate effects. If the baseline hazard is then estimated by the empirical survivor function, the resulting profile likelihood for the parameters is equivalent to the partial likelihood suggested by Cox. This justification of the partial likelihood is due to Breslow (1972) (see also

Cox, 1975 or Wong, 1986). However, if covariate effects do vary with time, that is if the PH assumption is violated, such factorization of the cumulative hazard does not exist and consequently, the partial likelihood does not have any justifcation as profile likelihood function. Secondly, in partial likelihood estimation the baseline hazard is treated as nuisance component and not explicitly estimated. In applications, however, knowledge about the baseline hazard can be of interest, in particular if smooth, nonparametric regression is pursued. For this reason it seems worthwhile to work directly with the likelihood function. This approach is pursued in this paper in order to fit a smooth, non-proportional hazard model. The integrated hazard function in the likelihood is thereby approximated using numerical integration based on a trapezoid approximation. This in turn leads to a simple likelihood functions which resembles a Poisson model.

As smoothing technique we employ penalized spline fitting (P-spline). The approach was originally introduced by O'Sullivan (1986), but the procedure finally achieved general recognition with the paper by Eilers & Marx (1996). A comprehensive overview about the current state of the art is found in Ruppert, Wand & Carroll (2003). P-spline smoothing in survival models has been studied in Cai, Hyndman & Wand (2002) for baseline hazard smoothing. The underlying idea of P-spline smoothing is to fit a smooth curve by using a high dimensional basis. But instead of simple parametric fitting a penalized version is pursued to provide a smooth fit. The approach resembles standard spline smoothing as discussed e.g. in Wahba (1978), or in its generalized form in Green & Silverman (1994). The major difference is that for spline smoothing the dimension of the corresponding spline basis grows with the sample size. In contrast, for P-spline smoothing a finite dimensional basis is used, where the dimension is chosen in a rich and generous manner. The approach

is numerically very handy. It also has has strong links to Linear Mixed Models (see Wand, 2003) and to penalized quasi likelihood (PQL) estimation in Generalized Linear Mixed Models (GLMM), as discussed in Breslow & Clayton (1993). The connection becomes obvious if the penalty is rewritten as *a priori* distribution on the coefficients of the basis. In fact, the smoothing parameters steering the amount of penalisation is then playing the role of the *a priori* variance in the resulting Generalized Linear Mixed Model. We utilize the link for smoothing parameter estimation. It will be demonstrated that the PQL approach is numerically simple but fails to estimate reasonable smoothing parameters in low intensity hazard models. Alternatively an EM based procedure as suggested in Booth & Hobert (1999) could be used for the price of increased numerical effort. We suggest a hybrid approach based on the numerically attractive PQL estimates combined with an Akaike criterion.

The paper is organized as follows. In Section 2 we first motivate the use of P-splines for fitting non-proportional hazard models. We demonstrate how integrals of the hazard function can be approximated by trapezoid integration, yielding a Poisson type model. We provide some asymptotic consideration and discuss practical adjustments of the fitting algorithm. In Section 3 we derive the link to GLMMs and discuss the estimation of the smoothing parameter. An application and simulations are provided in Section 4. A discussion finalizes the paper. Technical details are found in the Appendix.

2 Smooth Hazard Model

2.1 P-Spline Fitting

Let T_i denote the survival time of the ith individual or observational units and let C_i be the corresponding right censored time, $i = 1, \ldots, N$. We observe $Y_i = \min(T_i, C_i)$

and define the censoring indicator $\delta_i = 1$ if $T_i < C_i$ and $\delta_i = 0$ otherwise. With x_i we denote the p dimensional covariate vector for the i-th individual, which for simplicity of presentation is assumed to be time constant. The hazard function is then modeled as

$$h(t, x_i) \quad = \quad \lambda_0(t) \exp\{x_i^T \beta_x(t)\} \tag{1}$$

with $\lambda_0(t)$ as baseline hazard and $\beta_x(t)$ as vector of covariate effects varying smoothly with survival time t. For convenience we rewrite (1) as $h(t, x_i) = \exp\{z_i \beta(t)\}$ with $z_i^T = (1, x_i^T)$ and $\beta(t) = \{\log \lambda_0(t), \beta_x^T(t)\}^T$. The task is to estimate $\beta(t)$ smoothly by avoiding any stringent parametric assumptions. This is achieved by penalized spline regression.

For the sake of simplicity let us first consider smooth estimation of the baseline function $\beta_0(t) = \log \lambda_0(t)$. Let $B(t) = \{b_1(t), \ldots, b_q(t)\}$ be a high dimensional basis developed over the knots $t_1, \ldots, t_q$. Convenient choices are a B-spline basis (see de Boor, 1978) or truncated polynomials (see e.g. Wand, 2003). The dimension q of the basis is chosen lavish, such that the model bias $\beta_0(t) - B(t)\alpha_0^0$ is negligible, where $\alpha_0^0 = (\alpha_{01}^0, \ldots \alpha_{0q}^0)^T$ is the vector of "best" coefficients in the sense of having minimal Kullback-Leibler distance. More details are found later in the paper. Since q is supposed to be large, simple maximum likelihood estimation of α_0 would be highly variable and numerically unstable. Therefore, in order to achieve smoothness and numerical stability the penalty term $\lambda_0 \alpha_0^T D_0 \alpha_0$ is introduced, with D_0 as an appropriately chosen penalty matrix and λ_0 as a bandwidth parameter steering the amount of penalization. Possible choices for D_0 are differences based penalties as suggested in Eilers & Marx (1996) or taking D as identity matrix (see Wand, 2003). The latter choice is a reasonable candidate when working with truncated

polynomials.

In the same fashion we now fit all remaining components in the model. It is thereby tactically an advantage to extract the intercept from the smooth function. This means for estimation we decompose $\beta_l(t)$ to $\beta_{0l} + \tilde{B}(t)\alpha_l$, $l = 0, \ldots, p$, where β_{0l} is the constant part and $\tilde{B}(t)$ as a basis matrix containing no intercept. We define $\theta_l = (\beta_{0l}, \alpha_l^T)^T$ and using the Kronecker product we can jointly write $\beta(t) = \boldsymbol{W}(t)\boldsymbol{\theta}$ with $\boldsymbol{W(t)} = I_{p+1} \otimes \{1, \tilde{B}(t)\}$ and parameter vector $\boldsymbol{\theta}^T = (\theta_0^T, \ldots, \theta_p^T)$, where I_{p+1} is the $p + 1$ dimensional identity matrix. In principle the spline bases used for fitting $\beta_l(t)$ can differ among the separate components of $\beta(t)$ so that $\boldsymbol{W}(t)$ is of block diagonal form with different spline bases on its diagonal. For simplicity of presentation, however, we ignore this generalization here. To achieve a smooth fit the coefficients α_l are now jointly penalized which leads to the penalized likelihood function

$$l^P(\boldsymbol{\theta}, \lambda) \;=\; \sum_{i=1}^{N} l_i(\boldsymbol{\theta}) - \frac{1}{2}\sum_{l=0}^{p} \lambda_l \alpha_l^T D_l \alpha_l \tag{2}$$

with $l_i(\boldsymbol{\theta}) = \delta_i\Big(z_i^T \boldsymbol{W}(Y_i)\boldsymbol{\theta}\Big) - \int_0^{Y_i} \exp\{z_i^T \boldsymbol{W}(t)\boldsymbol{\theta}\}\mathrm{d}t$ as likelihood contribution (see Cox & Oakes, 1984) and $\lambda = (\lambda_0, \ldots, \lambda_p)$ as component-wise smoothing parameters steering the amount of penalization for each component. For notational convenience the penalty component in (2) can be rewritten as $\boldsymbol{\theta}^T(\Lambda \boldsymbol{D})\boldsymbol{\theta}$ with $\boldsymbol{D}$ as block diagonal matrix built from matrices $\mathrm{diag}(0, D_l)$, $l = 0, \ldots, p$, where $\mathrm{diag}(0, D_l)$ is the $q + 1$ dimensional diagonal basis having D_l in the bottom right corner and 0 elsewhere. Bandwidth matrix $\boldsymbol{\Lambda}$ matches accordingly as a diagonal matrix with $(\lambda_0 \otimes 1_{q+1}^T, \ldots, \lambda_p \otimes 1_{q+1}^T)$ on the diagonal, with 1_q as q dimensional unit vector. Differentiating (2)with respect to $\boldsymbol{\theta}$ leads to the penalized score equation

$$\frac{\partial l^p(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta}} \;=\; \sum_{i=1}^{N} \boldsymbol{s}_i(\boldsymbol{\theta}) - \Lambda \boldsymbol{D}\boldsymbol{\theta} \;=\; 0 \tag{3}$$

with $\boldsymbol{s}_i(\boldsymbol{\theta}) = \delta_i \boldsymbol{W}^T(Y_i) z_i - \int_0^{Y_i} \boldsymbol{W}^T(t) z_i \exp\{z_i^T \boldsymbol{W}(t)\boldsymbol{\theta}\} dt$. Accordingly, the second order derivative results to

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \quad = \quad \sum_{i=1}^{N} \nabla \boldsymbol{s}_i(\boldsymbol{\theta}) - \Lambda \boldsymbol{D} \tag{4}$$

where $\nabla \boldsymbol{s}_i(\boldsymbol{\theta}) = -\int_0^{Y_i} \boldsymbol{W}^T(t) z z_i^T \boldsymbol{W}(t) \exp\{z_i^T \boldsymbol{W}(t)\boldsymbol{\theta}\} dt$.

2.2 Integration

The penalized score function (3) and its derivatives contain integrals based on the hazard function. Since no analytic solution is readily available numerical integration is employed. A computationally handy version is to approximate the integrals by trapezoids. Let therefore $0 = \tau_0 < \tau_1 < \ldots < \tau_K$ be a grid of points that span the range of the observed failure times, i.e. $\tau_1 = \min\{Y_i : \delta_i = 1\}$ and $\tau_K = \max\{Y_i : \delta_i = 1\}$. Index K_i is defined through $\tau_{K_i - 1} < Y_i \leq \tau_{K_i}$ and with $\boldsymbol{m}_i(t) = \boldsymbol{W}^T(t) z_i \exp\{z_i^T \boldsymbol{W}(t)\boldsymbol{\theta}\}$ we denote the integrand in (3). This is approximated by a polygon going through the points $\Big(\tau_k, \boldsymbol{m}_i(\tau_k)\Big)$, $k = 0, 1, \ldots, K_i - 1$ which leads to

$$\int_0^{Y_i} \boldsymbol{m}_i(t) dt \tag{5}$$

$$\begin{aligned} \approx \quad & d(K_i > 1) \sum_{k=1}^{K_i - 1} \frac{1}{2} (\tau_k - \tau_{k-1}) \Big\{ \boldsymbol{m}_i(\tau_k) + \boldsymbol{m}(\tau_{k-1}) \Big\} \\ & + \frac{1}{2} (Y_i - \tau_{K_i - 1}) \Big\{ \boldsymbol{m}_i(\tau_{K_i - 1}) + \boldsymbol{m}(\tau_{K_i}) \Big\} \end{aligned} \tag{6}$$

$$= \quad \frac{1}{2} \min(\tau_1, Y_i) \boldsymbol{m}_i(\tau_0) + \frac{1}{2} \sum_{k=1}^{K_i} \Big\{ \min(\tau_{k+1}, Y_i) - \min(\tau_{k-1}, Y_i) \Big\} \boldsymbol{m}_i^T(\tau_k)$$

with $d(\cdot)$ as indicator function. The score contribution $\boldsymbol{s}_i(\boldsymbol{\theta})$ is now approximated by

$$\boldsymbol{s}_i(\boldsymbol{\theta}) \quad = \quad \delta_i \boldsymbol{W}^T(Y_i) z_i - \sum_{k=0}^{K_i} \boldsymbol{W}^T(\tau_k) z_i \exp\{z_i^T \boldsymbol{W}(\tau_k)\boldsymbol{\theta} + o_{ik}\} \tag{7}$$

where o_{ik} are given offsets defined through $o_{i0} = \log\{\min(\tau_1, Y_i)\}$ and for $K_i > 1$ $o_{ik} = \log[1/2\{\min(\tau_{k+1}, Y_i) - \min(\tau_{k-1}, Y_i)\}]$ for $k = 1, \ldots, K_i$ with τ_{K+1} set to infinity. Approximation (7) shows the form of a Poisson model fitted to the independent

pseudo observations $\tilde{Y}_{ik} \sim Po(z_i \boldsymbol{W}(\tau_k)\boldsymbol{\theta} + o_{ik})$ taking values $\tilde{Y}_{ik} = 0$ for $k < K_i$ and $\tilde{Y}_{iK_i} = \delta_i$. Hence by trapezoid integration we get an approximate fit of model (1) by fitting a penalized Poisson regression model with given offset and pseudo data $\tilde{Y}_{ik}$, $k = 1, \ldots, K_i$, $i = 0, \ldots, n$

2.3 Practical Adjustments

Inserting approximation (7) in (3) yields the approximate score equation to be solved. The trapezoid integration is applied in the same way to approximate the second order derivative (4), so that solving the score equation can be carried out with a standard Newton procedure. In practice, however, there are a number of adjustments necessary, like how to choose q, the dimension of the basis, and how to specify K, the number of integration grid points. Finally, the ultimate question how to choose the right amount of penalization is postponed to the next section.

For the location of integration points τ_k we suggest to use the observed failure times. A coarser grid omits information in the data, a finer grid leads to identifiability problems. Moreover the choice of K and q should fulfill the restriction $K \geq q$ to achieve identifiability. We used the rule of thumb $q = \min\{n/4, 25, K + 1\}$ which is in line with Wand (2003) and showed satisfactory results in our examples. Practical experience in standard models has also shown that the actual choice of q has little influence on the fit (see also Ruppert, 2002). Finally, a starting value for the penalized fit can be obtained by fitting a model with an unpenalized baseline hazard but with covariate effects being constant, that is we set $\lambda_0 \to 0$ while $\lambda_l \to \infty$ for $l = 1, \ldots, p$. This mirrors a proportional hazard model and the fit is numerically stable.

2.4 Asymptotic Considerations

Let $\boldsymbol{\theta}^0$ be the "true" coefficient, that is $\boldsymbol{W}(t)\boldsymbol{\theta}^0$ is the best approximation to the smooth curve $\beta(t)$ based on the Kullback Leibler discrepancy

$$K\{\boldsymbol{W}(t)\boldsymbol{\theta}, \beta(t)\} \;=\; E\Big[l\{\boldsymbol{W}(t)\boldsymbol{\theta}^0\} - l\{\beta(t)\}\Big], \tag{8}$$

where the expectation $E(\cdot)$ is carried out with respect to model (1) and the true underlying censoring process. Differentiating (8) defines θ^0 implicitly through

$$0 \;=\; E\left\{\sum_{i=1}^{N} s_i(\boldsymbol{\theta}^0)\right\}. \tag{9}$$

Let $\widehat{\boldsymbol{\theta}}$ be the estimated coefficient resulting from (3). It is shown in the Appendix that the penalized estimate is consistent in the sense

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 \;=\; -\left\{\sum_{i=1}^{N} \nabla s_i(\boldsymbol{\theta})\right\}^{-1} \sum_{i=1}^{N} s_i(\boldsymbol{\theta})\,\{1 + o_p(1)\} \tag{10}$$

assuming that $\lambda_l = o(N^{1/2})$, $l = 0, \ldots, p$. The latter assumption is rather weak as it allows the smoothing parameter to increase with growing sample size. A data driven choice derived below is in fact suggests bounded, if the underlying function is not constant.

From (10) we can also derive a variance formula for the estimate. In practice however, the following sandwich version performs better:

$$\mathrm{var}(\widehat{\boldsymbol{\theta}}) \;=\; -\left\{\sum_{i=1}^{N} \nabla s_i(\boldsymbol{\theta}) - \boldsymbol{\Lambda}\boldsymbol{D}\right\}^{-1} \left\{\sum_{i=1}^{N} \nabla s_i(\boldsymbol{\theta})\right\} \left\{\sum_{i=1}^{N} \nabla s_i(\boldsymbol{\theta}) - \boldsymbol{\Lambda}\boldsymbol{D}\right\}^{-1}$$

3 Relation to Generalized Linear Mixed Models

3.1 Penalized Quasi Likelihood Estimation

Penalized spline smoothing has strong affinities to penalized quasi likelihood estimation in Generalized Linear Mixed Models (GLMM) as discussed in Breslow &

Clayton (1993) (see also McCulloch & Searle, 2001). For normal response models this link is illuminated in depth in Wand (2003) (see also Cai, Hyndman & Wand, 2002). For non-normal response models we achieve the link in the following way. We consider coefficients α_l, $l = 0, \ldots, p$, as independent normally distributed variables with

$$\alpha_l \sim N(0, \lambda_l^{-1} D_l^-) \tag{11}$$

where D_l^- is the (generalized) inverse of D_l. The bandwidth parameters λ_l now occur in the *a priori* variance of α_l. Conditional on α_l, $l = 0, \ldots, p$ and based on the trapezoid integration we model

$$\tilde{Y}_{ik} | (\alpha_0, \ldots, \alpha_p) \sim Po(z_i^T \boldsymbol{W}(\tau_k)\boldsymbol{\theta} + o_{ik}) \tag{12}$$

with $\boldsymbol{\theta}$ as above composed from β_{0l} and α_l. Apparently, (11) and (12) provide the ingredients of a Generalized Linear Mixed Model. The likelihood for parameters β_{0l} and λ_l, $l = 0, \ldots, p$, is then obtained by integrating out the random coefficients, i.e.

$$\begin{aligned} l(\beta_{00}, \ldots, \beta_{0p}, \lambda_0, \ldots, \lambda_p) &= \int \prod_{i=1}^{N} \prod_{k=1}^{K_i} Po(\tilde{Y}_{ik}; z_i^T \boldsymbol{W}(\tau_k)\boldsymbol{\theta} + o_{ik}) \\ &\quad \times \prod_{l=0}^{p} \phi(\alpha_l, \lambda_l^{-1} D_l^-) \mathrm{d}\alpha_l \end{aligned} \tag{13}$$

with $\phi(\cdot)$ as a normal density function. Using a Laplace approximation for the integral leads to penalized quasi likelihood estimation (see Breslow & Clayton, 1993). It is not difficult to show that this in turn gives the estimating equations given by (3), with scores $s_i(\theta)$ as listed in (7).

3.2 Maximum Likelihood based Estimates

The connection between smoothing and GLMMs is not only of theoretical nature but can be exploited practically to choose appropriate smoothing parameters λ_l,

$l = 0, \ldots, p$. The idea is to estimate λ_l based on the likelihood function (13). To do so we introduce the following notation. Let $\boldsymbol{U}_{ik} = z_i\{I_{p+1} \otimes B(\tau_k)\}$ for $k = 0, \ldots K_i$ and $\boldsymbol{U}_i = (\boldsymbol{U}_{i0}^T, \ldots \boldsymbol{U}_{iK_i}^T)^T$. That is $\boldsymbol{U}_i$ is the observed design for the pseudo Poisson variables of the i-th individual. Approximating the integral by Laplace integration and inserting estimates for β_{0l} provides the Laplace approximation for the log profile likelihood

$$\begin{aligned} l^P(\lambda_0, \ldots, \lambda_p) &= \sum_{i=1}^{N}\sum_{k=0}^{K_i} \log Po(\tilde{Y}_{ik}; \cdot) - \frac{1}{2}\sum_{l=0}^{p}\Big(\lambda_l \widehat{\alpha}_l^T D_l \widehat{\alpha}_l + \log|\lambda_l D_l|\Big) \\ &\quad -\frac{1}{2}\log|\sum_{i=1}^{N} \boldsymbol{U}_i^T V_i \boldsymbol{U}_i + \mathrm{diag}(\lambda_l D_l)| \end{aligned} \tag{14}$$

with $V_i = \mathrm{diag}(\mathrm{var}(\tilde{Y}_{i0}), \ldots, \mathrm{var}(\tilde{Y}_{iK_i}))$ resulting from the Poisson model and $\mathrm{diag}(\lambda_l D_l)$ as block diagonal matrix built from $\lambda_l D_l$, $l = 0, \ldots, p$. Ignoring the dependence of V_i on λ we get by differentiating (14)

$$0 = \widehat{\alpha}_l^T D_l \widehat{\alpha}_l - \frac{q}{\lambda} - \mathrm{tr}\left(\Big(\sum_{i=1}^{N} \boldsymbol{U}_i^T V_i \boldsymbol{U}_i + \mathrm{diag}(\lambda_l D_l)\Big)^l D_l\right) \tag{15}$$

where superscript l refers to the l-th block diagonal of matrix $(\sum_{i=1}^{N} \boldsymbol{U}_i^T V_i \boldsymbol{U}_i + \mathrm{diag}(\lambda_l D_l))^{-1}$. In asymptotic terms the latter component in (15) is of order $O(N^{-1})$ and could be neglected. Practical experience showed however that the term should not be omitted in finite samples and we make use of the approximate version

$$\frac{1}{\lambda}\{q + \mathrm{tr}\left((\sum_{i=1}^{N} \boldsymbol{U}_i^T V_i \boldsymbol{U}_i + \mathrm{diag}(\lambda_l D_l))^l D_l\right)\} = \frac{df_l}{\lambda_l} + O(N^{-1})$$

with

$$df_l := \mathrm{tr}\{(\sum_{i=1}^{N} z_{il}^2 B_i^T V_i B_i + \lambda_l D_l)^{-1} \sum_{i=1}^{N} z_{il}^2 B_i^T V_i B_i\}$$

as approximate for the degree of freedom of the l-th smooth component and $B_i = (B^T(\tau_0), \ldots, B^T(\tau_{K_i}))^T$. This finally yields the PQL estimate for λ_l via

$$\widehat{\lambda}_l = \frac{df_l}{\widehat{\alpha}_l^T D_l \alpha_l}, \tag{16}$$

The smoothing parameter estimate (16) depends on estimates $\widehat{\boldsymbol{\theta}}$ and vice versa. A convenient way to estimate both, λ and $\boldsymbol{\theta}$, is to cycle between estimation of $\boldsymbol{\theta}$ for given λ and estimation of λ for given $\boldsymbol{\theta}$. We denote with $\widehat{\lambda}^{(j)}$ and $\widehat{\boldsymbol{\theta}}^{(j)}$ the estimates in the j-th cycle of such algorithm.

3.3 Hybrid Smoothing Parameter Selection

Laplace approximation or PQL estimation, respectively, of the marginal likelihood can perform poorly, as pointed out in (Breslow & Lin, 1995) or Shun & McCullagh (1995). We observe unsatisfactory performance of the PQL estimates for low intensity Poisson data. To demonstrate this deficit we simulate 400 Poisson data $Y_i \sim Po\{\mu(t)\}$ with $\mu(t)$ as smooth but low intensity mean. Function $\mu(t)$ is fitted by P-spline smoothing using a truncated linear basis with 30 knots. In the left plot in Figure 1 we show the mean and pointwise 95% confidence intervals of 150 simulation with smoothing parameter λ_l estimated by (16). The true function is shown as dashed line. Apparently the PQL estimate over-smoothes and fails to detect the smooth structure. To overcome this deficit one can replace the Laplace approximation by a more accurate approach. We consider the Monte Carlo EM algorithm suggested by Booth & Hobert (1999). The result is shown in the middle plot in Figure 1. The improved behavior of the EM fit has however to be bought for the price of increased numerical effort. We therefore prefer to employ a hybrid strategy by taking advantages of the numerical simplicity of the PQL estimate, but to control the estimates with the Akaike criterion. This means at the j-th cycle of the PQL estimation we calculate the Akaike criterion $AIC(\widehat{\lambda}^{(j)})$ with

$$AIC(\lambda) \;=\; \sum_{i=1}^{N}\sum_{k=0}^{K_i} \log Po(\tilde{Y}_{ik}, \cdot) + 2df(\lambda)$$

where $df = \sum_{l=0}^{p} df_l$ is the degree of freedom of the model. We terminate the iterations if $AIC(\lambda^{j+1}) > AIC(\lambda^{(j)})$. The right hand plot in Figure 1 shows the behavior of the hybrid procedure. The performance appears promising, which also shows in further simulation in the next section. It is thereby important to point out that the hybrid estimate is numerically very handy which is advantageous in particular in multivariate smoothing parameter selection.

4 Application

4.1 Simulation

We simulate survival data for $N = 400$ individuals on a discrete time grid $t = 1, 2, 3, \ldots$ using a constant drop out probability of 3 % for each time interval t to $t+1$. The two binary covariates x_1 and x_2 are randomly chosen with $P(x_1 = 1) = 0.5$ and $P(x_2 = 1) = 0.3$. As dynamic effects we include $\beta_0(t) = -5$ as constant baseline hazard and $\beta_1(t) = -1+t/30$ and $\beta_2(t) = 1.5\sin(\pi t/60)$. In Figure 2 we show for one simulation the principle of the hybrid smoothing parameter selection. Smoothing parameter estimation is started with bandwidth $\widehat{\lambda}_l^{(0)} = \exp(-5)$ for $l = 0, 1, 2$ and updated with $\lambda_l^{(t)}$ as long as the Akaike criterion decreases. Figure 2 shows the Akaike function for λ_1 and λ_2 (with λ_0 set to its optimal value infinity due to the constant baseline). The steps of the algorithm are indicated with numbers, where $\widehat{\lambda/}^{(6)}$ is the final estimate based on the stopping rule. The crosses show the further steps of the PQL iteration which apparently steers towards oversmoothing.

In Figure 3 (two left plots) we show for 100 simulation the final estimates $\widehat{\lambda}^{(t)}$ based on the the hybrid approach (top row) and the PQL estimates (bottom row). The tendency of oversmoothing for PQL is obvious. The PQL estimates have a high probability of omitting the dynamic structure of the effects. In this respect

the hybrid approach performs better. This can also be seen from the estimated curves $\widehat{\beta}_l(t)$ shown in the middle plots for $\beta_1(t)$ and in the right hand side plots for $\beta_2(t)$. We show the simulation estimates corresponding to the 5, 15, 25, 50, 75, 85 and 95% quantiles of $\widehat{\lambda}_l^{(t)}$. The true curve is included as thick line. It appears that the hybrid routine performs well by detecting non-proportional hazards, while the PQL estimate is less sensitive. This impression remains unchanged in a slightly modified simulation. We set $\beta_2(t)$ to zero yielding the results shown in Figure 4. The hybrid approach now detects the proportional (zero) hazards for $\beta_2(t)$. Overall the interpretation does not change and the hybrid smoothing parameter selection performs promising.

4.2 Example

We demonstrate the modeling approach with data from the so called Munich founder study. In this study a sample of size $N = 1123$ is drawn from firms which have been founded during the years 1985 and 1986 in the state of Bavaria. The firms were followed up until 1990 and the measurement of interest is the time the companies stays in the market without going bankrupt. Details on the study can be found in Brüderl, Preisendörfer & Ziegler (1992), data are available from the Central Archive for Empirical Social Research, University Cologne, Germany (`http://www.gesis.org/ZA/`). We consider a subsample of 369 firms with their founders aged 30 years and younger. About 50 % (185) of the firms went bankrupt within the first 5 years of follow up. We model the survival time T of the enterprise to depend on the following indicator variables:

• *start capital:* =1 if the company started with capital, =0 otherwise,

• *plan:* = 1 if the planing process for the venture took longer than 6 months, =0 otherwise,

• *branch knowledge:* =1 indicates that the founder had previous knowledge and expertise in the branch of the firm, =0 otherwise,

• *innovation:* =1 if the product produced or sold by the company is an innovation, =0 if the product is on the market already,

• *purpose:* =1 if the business was started and is run as main source of income for the founder, =0 otherwise,

• *degree:* =1 indicates whether the founder is holding a degree (university or craftsmen degree), =0 otherwise,

• *gender:* =1 for male.

The resulting fits with smoothing parameters selected by the hybrid approach are shown in Figure 5. The baseline uncovers a dereasing risk of failure with the company being on the market. If the business started with positive capital and was planed in advance it has a reduced risk of failure. These effects however fade away after about 2-3 years. Branch experience has a constant risk decreasing effect. Innovative products induce an increased risk with hardly any time variation. Companies which have been founded to provide the main source of income for the founder have better survival chances. This effect gets strengthened over time. Finally, the degree of the founder has a weak effect only and gender does not appear to be significant.

5 Discussion

We demonstrated the use of P-splines for fitting non-proportional hazard models. Numerical integration was pursued which led to Poisson data. Multivariate smoothing parameter selection was carried out by a hybrid procedure, utilizing the link between P-spline smoothing and Generalised Linear Mixed Models. In particular complicated grid searching was avoided and the routine is numerically simple. A

data example demonstrated the new insight which could be gained by allowing hazard functions to be dynamic in time.

A Technical Details

The penalized estimate is defined through $0 = \sum_{i=1}^{N} s_i(\widehat{\boldsymbol{\theta}}) - \Lambda \boldsymbol{D}\boldsymbol{\theta}$. Expansion provides

$$
\begin{aligned}
0 &= \sum_{i=1}^{N} \boldsymbol{s}_i(\boldsymbol{\theta}^0) + \sum_{i=1}^{N} \nabla \boldsymbol{s}_i(\boldsymbol{\theta}^0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \Lambda \boldsymbol{D}\widehat{\boldsymbol{\theta}} && (17)\\
&\quad + \Big[\sum_{i=1}^{N} \nabla^2 \boldsymbol{s}_i(\boldsymbol{\theta}^0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2 \Big] + \ldots && (18)
\end{aligned}
$$

where brackets $[\cdot]$ here and in the following embrace terms which are written in a symbolic manner since $\nabla^2 \boldsymbol{s}_i(\cdot) = \partial \nabla \boldsymbol{s}_i(\cdots)/\partial \boldsymbol{\theta}$ is a three dimensional array. Exact notation is possible by employing the Einstein summation convention (see e.g. McCullagh, 1987), for simplicity of notation however we here prefer the obvious symbolic notation using brackets. Moreover, we will subsequently drop the parameter argument if components are calculated at the "true" parameter value, e.g. we write $\boldsymbol{s}_i$ shortly for $\boldsymbol{s}_i(\boldsymbol{\theta}^0)$. Inversion of (17) then provides

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 &= \Big(- \sum_{i=1}^{N} \nabla \boldsymbol{s}_i + \Lambda \boldsymbol{D} \Big)^{-1} \Big(\sum_{i=1}^{N} \boldsymbol{s}_i + \lambda \boldsymbol{D}\boldsymbol{\theta} \Big) && (19)\\
&\quad - \frac{1}{2} \Big[\Big(- \sum_{i=1}^{N} \nabla \boldsymbol{s}_i + \Lambda \boldsymbol{D} \Big)^{-3} \Big(\sum_{i=1}^{N} \boldsymbol{s}_i + \Lambda \boldsymbol{D}\boldsymbol{\theta} \Big) \Big(- \sum_{i=1}^{N} \nabla^2 \boldsymbol{s}_i \Big) \Big] + \ldots
\end{aligned}
$$

We decompose $\nabla \boldsymbol{s}_i$ in its mean and stochastic part via $-\nabla \boldsymbol{s}_i = \boldsymbol{F}_i + \boldsymbol{\epsilon_i}$ with $\boldsymbol{F}_i = E(-\nabla \boldsymbol{s}_i)$. From (19) we get

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 &= \Big(\sum_{i=1}^{N} \boldsymbol{F}_i + \boldsymbol{\Lambda D} \Big)^{-1} \Big(\sum_{i=1}^{N} \boldsymbol{s}_i + \boldsymbol{\Lambda D}\boldsymbol{\theta} \Big) && (20)\\
&\quad - \Big[\Big(\sum_{i=1}^{N} \boldsymbol{F}_i \Big)^{-2} \Big(\sum_{i=1}^{N} \boldsymbol{\epsilon}_i \Big) \Big(\sum_{i=1}^{N} \boldsymbol{s}_i + \boldsymbol{\Lambda D}\boldsymbol{\theta} \Big) \Big] && (21)\\
&\quad - \frac{1}{2} \Big[\Big(\sum_{i=1}^{N} \boldsymbol{F}_i \Big)^{-3} \Big(\sum_{i=1}^{N} \boldsymbol{s}_i + \boldsymbol{\Lambda D}\boldsymbol{\theta} \Big) \Big(- \sum_{i=1}^{N} \nabla^2 \boldsymbol{s}_i \Big) \Big] + \ldots && (22)
\end{aligned}
$$

It will become obvious that components (21) and (22) are of negligible order compared to the leading term. We therefore concentrate on (20) only, which by expansion yields

$$\begin{aligned}\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 &= \Big(\sum_{i=1}^N \boldsymbol{F}_i\Big)^{-1}\Big(\sum_{i=1}^N \boldsymbol{s}_i + \boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\theta}\Big) - \boldsymbol{\Lambda}\Big(\sum_{i=1}^N \boldsymbol{F}_i\Big)^{-1}\boldsymbol{D}\Big(\sum_{i=1}^N \boldsymbol{F}_i\Big)^{-1}\Big(\sum_{i=1}^N \boldsymbol{s}_i + \boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\theta}\Big) \\ &\quad + \Big[\boldsymbol{\Lambda}^2\Big(\sum_{i=1}^N \boldsymbol{F}_i\Big)^{-3}\boldsymbol{D}^2\Big(\sum_{i=1}^N \boldsymbol{s}_i + \boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\theta}\Big)\Big].\end{aligned}$$

We assume that $\lambda_l = o(N^{1/2}), l = 0, \ldots, p$, that is the penalty λ_l may tend to infinity but at a smaller rate than $N^{1/2}$. Note that this is a very weak condition and in fact the ML estimate of λ_l is of order $O(1)$ (as long as $\lambda = O(1)$. Using this assumption, the variance of the first term is of order $O(N^{-1})$ and dominates the variance of the second term which has order $O(N^{-2}\Lambda) = o(N^{-3/2})$. Moreover, with (9) we find the bias of $\widehat{\boldsymbol{\theta}}$ to be given by $(\sum_{i=1}^N \boldsymbol{F}_i)^{-1}\boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\theta}$ which is of order $O(N^{-1}\Lambda)$. As mean squared error of $\widehat{\boldsymbol{\theta}}$ we therefore get

$$\mathrm{var}(\widehat{\boldsymbol{\theta}}) + \mathrm{bias}(\widehat{\boldsymbol{\theta}})^2 = O(N^{-1}) + O(\Lambda^2 N^{-2}) \tag{23}$$

which is dominated by the variance as long as $\Lambda = o(N^{1/2})$.

Finally, reflecting that $\boldsymbol{s}_i$, $i = 1, \ldots N$, are independent it is easily seen with arguments similar to those above that (21) is of order $O_p(N^{-1}) + O_p(N^{-1}\Lambda)$. Analogously, (22) is found to have negligible order.

References

Abrahamowicz, M., MacKenzie, T., and Esdaile, J. M. (1996). Time-dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association* **91**, 1432–1439.

de Boor, C. (1978). *A Practical Guide to Splines*. Berlin: Springer.

Booth, J. and Hobert, J. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo EM algorithm. *Journal of the Royal Statistical Society, Series B* **B 62**, 265–285.

Breslow, N. E. (1972). Comment on "regression and life tables" by D. R. Cox. *Journal of the Royal Statistical Society, Series B* **34**, 216–217.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association.* **88**, 9–25.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

Brüderl, J., Preisendörfer, P., and Ziegler, R. (1992). Survival chances of newly founded business organizations. *American Sociological Review* **57**, 227–242.

Cai, T., Hyndman, R., and Wand, M. (2002). Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics* **11**, 784 – 798.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics* **30**, 93 – 111.

Cox, D. (1975). Partial likelihood. *Biometrika* **62**, 187–220.

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

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Stat. Science 11*(2), 89–121.

Fan, J., Gijbels, I., and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *Annals of Statist.* **25**, 1661–1690.

Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics* **50**, 640–652.

Green, D. J. and Silverman, B. W. (1994). *Nonparametric Regression and generalized linear models*. Chapman & Hall.

Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazard model. *Biometrics* **46**, 1005–1016.

Hastie, T. and Tibshirani, R. (1993). Varying–coefficient models. *Journal of the Royal Statistical Society, Series B* **55**, 757–796.

Hess, K. R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine* **13**, 1045–1062.

Kooperberg, C., Stone, C., and Troung, Y. (1995). Hazard regression. *Journal of the American Statistical Association.* **90**, 78–94.

Lin, D. Y. and Wei, L. J. (1991). Goodness-of-fit tests for the general Cox regression model. *Statistica Sinica* **1**, 1–17.

McCullagh, P. (1987). *Tensor Methods in Statistics*. London: Chapman & Hall.

McCulloch, C. and Searle, S. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

O'Quigley, J. and Pessione, F. (1989). Score tests for homogeneity of regression effect in the proportional hazards model. *Biometrics* **45**, 135–144.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (c/r: P519-527). *Statistical Science* **1**, 502–518.

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

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

Ruppert, R., Wand, M., and Carroll, R. (2003). *Semiparametric Modelling*. Cambridge University Press.

Sasieni, P. (1999). Cox regression model. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics*, Volume 1, pp. 1006–1020. New York: Wiley.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B* **57**, 749–760.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Series B* **40**, 364–372.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics* **18**, 223–249.

Wong, W. (1986). Theory of partial likelihood. *Annals of Statist.* **14**, 88–123.

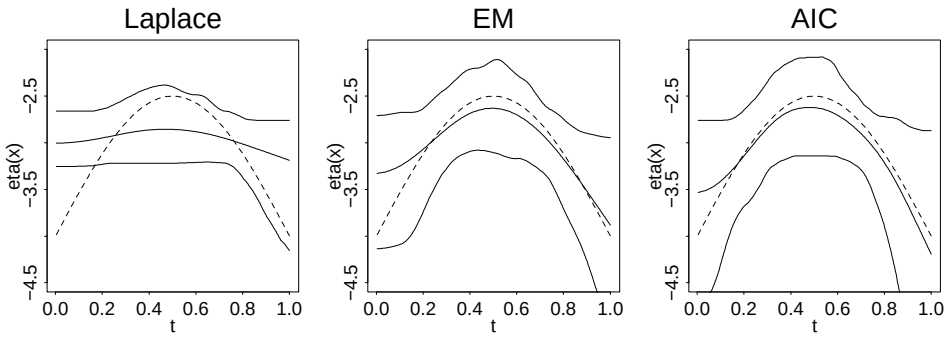

Figure 1: Mean and pointwise empirical 90 % confidence intervals based on 100 simulated estimates of low intensity Poisson data. The dashed curve gives the true function

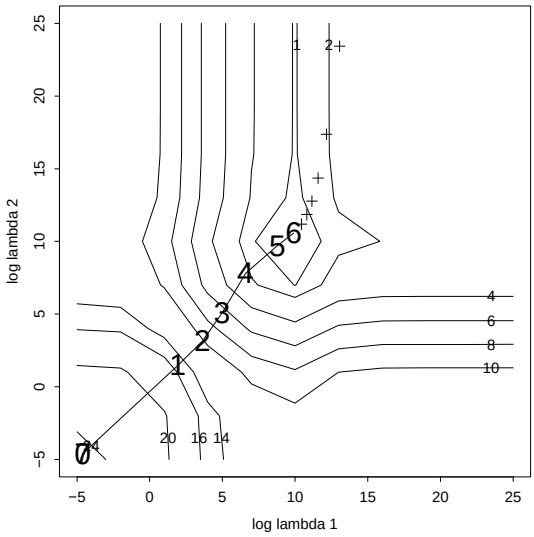

Figure 2: Akaike function $AIC(\lambda)$ for a single simulation. Shown is the difference to the minimum. The line with the thick numbers indicate the steps of the hybrid estimate $\widehat{\lambda}^{(j)}$ with highest number as final estimate. The crosses show the further divergence of Laplace estimates, i.e. if the stopping rule is ignored.

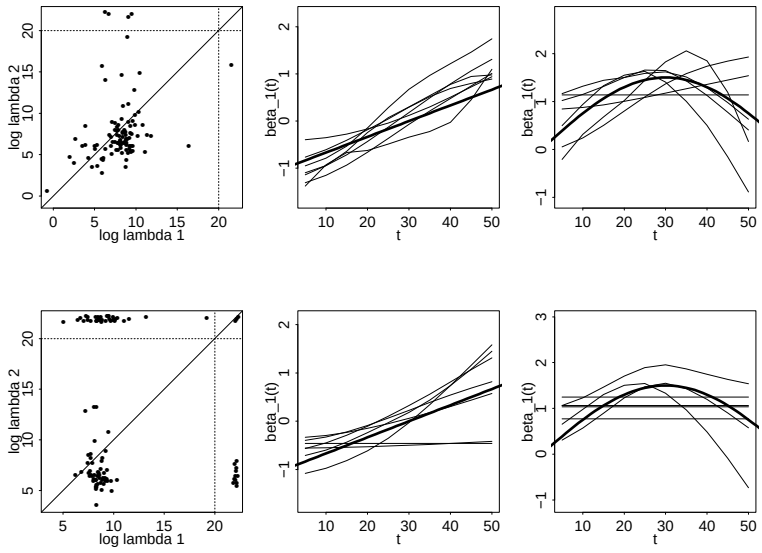

Figure 3: Estimated smoothing parameter (left plots) for hybrid approach (top row) and PQL (bottom row) with corresponding fits $\beta_1(t)$ (middle plots) and $\beta_2(t)$ (right plots).

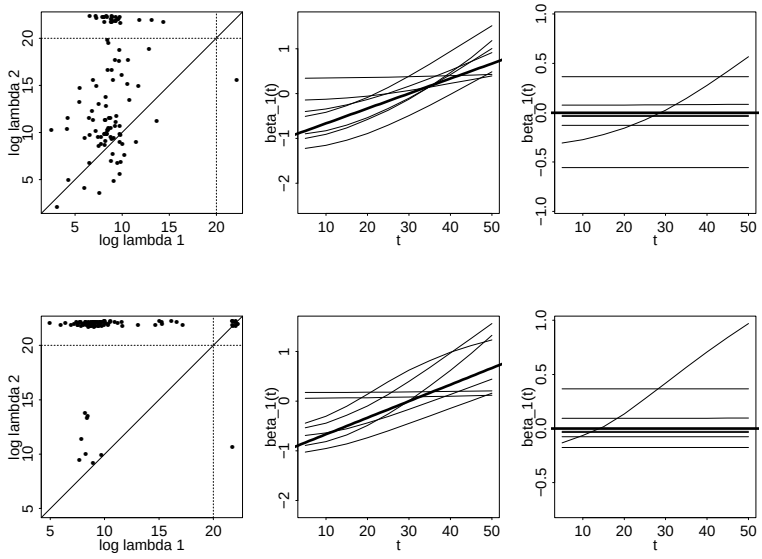

Figure 4: Estimated smoothing parameter (left plots) for hybrid approach (top row) and PQL (bottom row) with corresponding fits $\beta_1(t)$ (middle plots) and $\beta_2(t)$ (right plots).

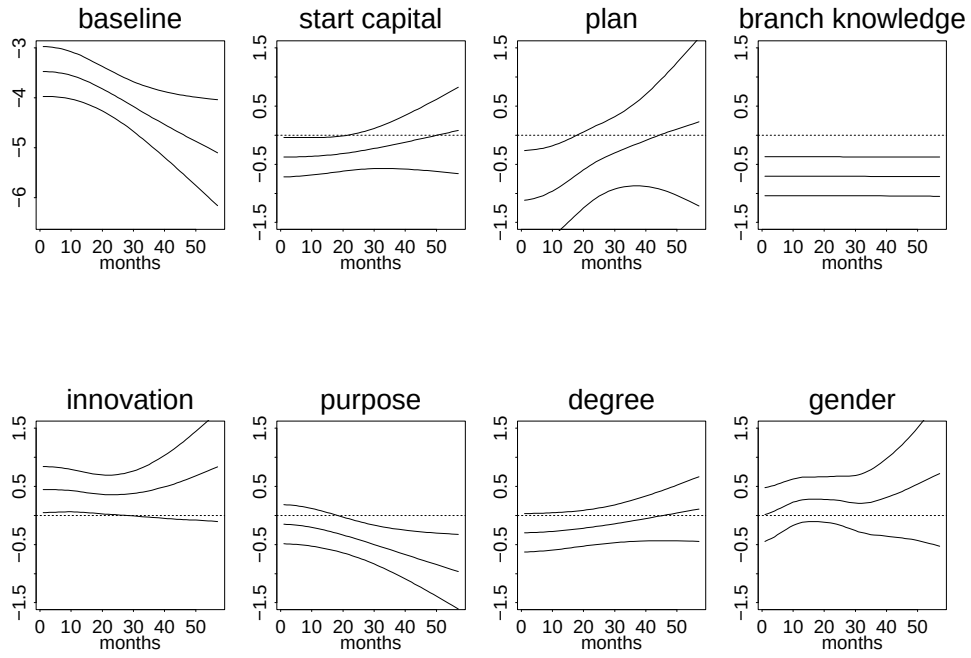

Figure 5: Baseline and dynamic effects for Munich founder study. As reference the zero line is indicated as dotted line. Shown are penalised estimates and pointwise 95 % confidence intervals.