Parametrization and Penalties in Spline Models with an Application to Survival Analysis

M. J. Costa and J.E.H.Shaw

Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

m.j.costa@warwick.ac.uk j.e.h.shaw@warwick.ac.uk

Abstract

In this paper we show how a simple parametrization, built from the definition of cubic splines, can aid in the implementation and interpretation of penalized spline models, whatever configuration of knots we choose to use. We call this parametrization value-first derivative parametrization. We perform Bayesian inference by exploring the natural link between quadratic penalties and Gaussian priors. However, a full Bayesian analysis seems feasible only for some penalty functionals. Alternatives include empirical Bayes methods involving model selection type criteria. The proposed methodology is illustrated by an application to survival analysis where the usual Cox model is extended to allow for time-varying regression coefficients.

Keywords: Penalized splines; Parametrization; Bayesian inference; Empirical Bayes; Survival analysis; Cox model.

1 Introduction

Spline smoothing methods have become a popular modeling tool in many statistical contexts. They enable the handling of complex nonlinear relationships, otherwise difficult to estimate with conventional parametric models. The number of different approaches to spline smoothing is quite wide, ranging from smoothing spline techniques (Hastie & Tibshirani, 1990a; Wahba, 1990; Green & Silverman, 1994), where a knot is placed at each observation, to regression splines with adaptive knot selection (Friedman, 1991). More recently, Eilers & Marx (1996) proposed the use of penalized splines (or P-splines), a different approach which can be seen as a compromise between smoothing and regression splines. There, the number of knots defining the spline function is larger than that justified by the data, but smaller than the number of observations. The level of overfitting is

controlled by a roughness penalty over the curve. The most common choice is a penalty based on the integral of a squared derivative of the spline curve.

Several Bayesian inference procedures for spline smoothing have been developed in the literature. In the framework of regression splines, Denison et al. (1998a,b) use transdimensional Markov chain Monte Carlo (MCMC) simulation techniques. Hastie & Tibshirani (2000) investigate Bayesian smoothing splines by exploring the relationship between the backfitting algorithm and the Gibbs sampler. Bayesian penalized splines have been studied by Biller & Fahrmeir (1997) and by Fahrmeir & Lang (2001), in the context of generalized additive and generalized additive mixed models respectively, using random walk and Markov random field priors. Lang & Brezger (2004) provide extensive simulation studies for Bayesian penalized splines in the context of Gaussian regression. Building on an idea first introduced by Gamerman (1997), Brezger & Lang (2006) propose an efficient MCMC sampling algorithm based on Fisher scoring type proposals. They develop the algorithm for generalized additive models but it can be modified so as to be applicable in other settings.

This paper focuses on Bayesian penalized spline models by exploring the relationship between penalty functionals and priors over the spline curve. Attention is restricted to cubic spline functions since we found these to provide the necessary amount of smoothing in the applications we have considered. Moreover, an intuitive parametrization exists that makes implementation of cubic spline functions straightforward. We call the parametrization *value-first derivative parametrization*. It is built from the definition of cubic polynomials. In the numerical analysis literature, cubic spline functions represented in this way are known as cubic Hermite splines (de Boor, 1978). We feel that 'value-first derivative parametrization' is more appropriate in the penalized likelihood framework considered here because it makes explicit reference to the components of the spline function being used. This will become particularly useful when interpreting penalty functionals of the spline as we shall see below.

In their paper, Eilers & Marx (1996) advocate the use of difference penalties and show that a difference penalty of order 2 approximates the standard penalty functional defined by the integral of the square of the 2nd derivative. The value-first derivative parametrization allows explicit use of integral based penalties. Because the parameters are directly related to properties of the spline curve, the penalty functionals have simple, interpretable expressions. One immediate consequence is that other forms of penalization, like double penalties as proposed in Eilers & Marx (2003, 2004), Eilers & Goeman (2004) or Aldrin (2006), are as easy to implement as single penalty functionals under the value-first derivative parametrization.

We present an application of the value-first derivative parametrization in the Cox model (Cox, 1972) with time-varying regression coefficients. This is a complex setting that has attracted a great deal of attention in recent years. See, for example, Gray (1994), Abrahamowicz et al. (1996), Kauermann (2005) or Tian et al. (2005). A Bayesian approach to inference for time-varying re-

gression coefficients can be found in Lambert & Eilers (2005). Because of the specific shape of the partial likelihood function in the Cox model we use a modified version of the algorithm proposed by Brezger & Lang (2006) to sample from the posterior of interest.

In the real data example we considered, based on the well-known PBC data set as found in Fleming & Harrington (1991, App. D), the value-first derivative parametrization proved to be a flexible tool that is simple to use yet able to yield smooth, plausible estimates.

The rest of the paper is organized as follows. Section 2 provides a review of penalized likelihood methods. The value-first derivative parametrization is characterized in detail in Section 3. We illustrate the use of the value-first derivative parametrization with an application to the Cox model with time-varying regression coefficients in Section 4. There we describe the Bayesian inference procedure we use and present alternatives to the usual full Bayes approach for the particular case of double penalty models. The analysis of the PBC data set is also included. Finally, we summarize our findings in Section 5, where we also outline directions for future research.

2 A Review of Penalized Likelihood Methods

For simplicity we introduce penalized likelihood methods in the univariate case. Let $g(t)$ be the function we wish to estimate. Denote by D the available data and by $l(g\,;D)$ the log-likelihood of the model under study. We assume that g is well approximated by a cubic spline function with knots $\{k_m\}_{m=1}^{\mathcal{K}}$ over the domain of t. Eilers & Marx (1996) proposed the use of a fairly large number of knots to avoid computationally intensive selection methods. To prevent overfitting they introduced a penalty on the roughness of g. For cubic splines this is typically of the form $P_2(g\,;\lambda) = \frac{\lambda}{2}\int g''^2$, a penalty that has proven to yield sensible estimates with good theoretical properties. The estimate $\widehat{g}$ is found by maximizing the penalized log-likelihood criterion

$$\mathcal{J}(g) = \ell(g\,;D) - P_2(g\,;\lambda). \tag{1}$$

The positive parameter λ in $P_2(g\,;\lambda)$ controls the bias-variability trade-off implicit in (1) and is therefore referred to as a *smoothing parameter*. As the value of λ increases, so does the 'weight' of the penalty $P_2(g\,;\lambda)$ when maximizing (1). As a result, the estimate $\widehat{g}$ becomes increasingly smoother, i.e., $\widehat{g}'' \approx 0$.

The penalty functional $P_2(g\,;\lambda)$ has no effect over linear and constant functions of t since, in this case, $g''(t) \equiv 0$. In some applications, however, it may additionally be desirable to penalize non-zero constants or non-zero slopes. In the latter case the penalty functional becomes

$$P_d(g\,;\lambda_1,\lambda_2) = P_1(g\,;\lambda_1) + P_2(g\,;\lambda_2) = \frac{\lambda_1}{2}\int g'^2 + \frac{\lambda_2}{2}\int g''^2. \tag{2}$$

If $P_2(g\,;\lambda)$ in (1) is replaced by $P_d(g\,;\lambda_1,\lambda_2)$, then as λ_1 and λ_2 become large, the corresponding spline fit $\widehat{g}$ converges to a polynomial of degree 0, i.e., a constant function of t.

The optimization problem defined by the criterion $\mathcal{J}(g)$, described in equation (1) for the penalty $P_2(g\,;\lambda)$, can be cast in a Bayesian framework whatever the penalty functional we choose to use. The penalty term has a natural interpretation as minus the log-prior for g. If $P_2(g\,;\lambda)$ is used, and we write $p\,(g \mid \lambda) \propto \exp\,(-P_2(g\,;\lambda))$ as the prior for g, then the criterion in (1) becomes the log-posterior for g and $\widehat{g}$ the MAP estimate.

A comprehensive description of penalized likelihood methods is given by Green & Silverman (1994). See also Eilers & Marx (2003, 2004), Eilers & Goeman (2004) or Aldrin (2006) for applications of double penalty models. Wood (2003) describes the general case of multiple penalty functionals.

3 The Value-First Derivative Parametrization

Spline functions are typically represented as elements in the span of a chosen spline basis. Here we follow a different approach where the definition of cubic polynomials induces a local parametrization for cubic splines. The result is an intuitive parametrization whose elements are easy to interpret and relate naturally to the spline function - in particular, priors are easier to elicit. We call this the value-first derivative parametrization, denoted hereafter by *VFDP*.

3.1 *VFDP* Setup

Let $g(t)$ be the function we wish to estimate using a cubic spline with knots $\{k_m\}_{m=1}^{\mathcal{K}}$ which span the domain of t. By definition of a cubic spline, g agrees with a cubic polynomial within each knot interval $[k_m,\, k_{m+1})$. Such a polynomial can be uniquely defined by four conditions over its coefficients. For each knot k_m we define

$$a_m = g(k_m), \quad b_m = g'(k_m). \tag{3}$$

The parameters a_m, b_m, a_{m+1} and b_{m+1} define, according to (3), four equations over the coefficients of the polynomial that agrees with g within the knot interval $[k_m,\, k_{m+1})$. This means that, for $k_m \le t < k_{m+1}$, we can write $g(t)$ in terms of a_m, b_m, a_{m+1} and b_{m+1}. If we define the four cubic polynomials in t

$$\begin{aligned} \phi_{0m}(t) &= \frac{(u_m - \Delta_m)^2(2u_m + \Delta_m)}{\Delta_m^3}, \quad \phi_{1m}(t) = \frac{u_m^2(3\Delta_m - 2u_m)}{\Delta_m^3}, \\ \psi_{0m}(t) &= \frac{u_m(u_m - \Delta_m)^2}{\Delta_m^2}, \quad \psi_{1m}(t) = \frac{u_m^2(u_m - \Delta_m)}{\Delta_m^2}, \end{aligned} \tag{4}$$

where $u_m = t - k_m$ and $\Delta_m = k_{m+1} - k_m$, then it is straightforward to show that, for $m = 1, \ldots, \mathcal{K} - 1$,

$$g(t) = a_m\phi_{0m}(t) + b_m\psi_{0m}(t) + a_{m+1}\phi_{1m}(t) + b_{m+1}\psi_{1m}(t), \quad t \in [k_m,\, k_{m+1}). \tag{5}$$

Hence, the 4-dimensional vector of parameters $\boldsymbol{\alpha}_m = (a_m, b_m, a_{m+1}, b_{m+1})^T$ completely specifies the spline $g(t)$ within the knot interval $[k_m,\, k_{m+1})$, $m = 1, \ldots, \mathcal{K}-1$, and the $2\mathcal{K}$-dimensional vector $\boldsymbol{\alpha} = (a_1, b_1, \ldots, a_{\mathcal{K}}, b_{\mathcal{K}})^T$ defines $g(t)$ in $[k_1, k_{\mathcal{K}})$.

A general expression for g is readily available from (5). Let $\boldsymbol{\eta}_m(t)$ be the 4-dimensional row vector with components the four polynomials in (4), i.e.,

$$\boldsymbol{\eta}_m(t) = (\phi_{0m}(t), \psi_{0m}(t), \phi_{1m}(t), \psi_{1m}(t))\,.$$

The value of the spline function g evaluated at some point $t \in [k_1, k_{\mathcal{K}})$ is

$$g(t) = \sum_{m=1}^{\mathcal{K}-1} \mathbb{I}_m(t)\, \boldsymbol{\eta}_m(t)\, \boldsymbol{\alpha}_m, \tag{6}$$

where $\mathbb{I}_m(t)$ is an indicator function taking value 1 if $k_m \le t < k_{m+1}$, and zero otherwise.

One of the advantages of spline models when compared to single polynomial ones is local influence. The *VFDP* highlights this property since each parameter in $\boldsymbol{\alpha}$ affects the fitted curve in the span of two consecutive knot intervals only. Any form of correlation among the parameters is thus likely to be small.

Note that the definition of a_m and b_m in (3) automatically imposes g and g' to be continuous everywhere. However, g'' may be discontinuous across the knots, which brings additional flexibility to the fitting process. This is not the case in the parametrization used by Green & Silverman (1994), where the estimated curve is constrained to have continuous curvature throughout its domain.

3.2 Penalty Interpretation and Implementation

The penalty functional $P_2(g\,;\lambda)$ can be written as

$$P_2(g\,;\lambda) = \frac{\lambda}{2} \sum_{m=1}^{\mathcal{K}-1} \int_{k_m}^{k_{m+1}} g''^2 = \sum_{m=1}^{\mathcal{K}-1} P_{2m}(g\,;\lambda) \tag{7}$$

using the fact that $\{k_m\}_{m=1}^{\mathcal{K}}$ constitutes a partition of the domain of g and that within each knot interval g'' is squared integrable. The local penalties $P_{2m}(g\,;\lambda)$ have simple, interpretable expressions as we shall see below.

For $m = 1, \ldots, \mathcal{K}-1$ we define

$$\begin{aligned} d_{m,m+1} &= a_{m+1} - (a_m + \Delta_m b_m), \\ d_{m+1,m} &= a_m - (a_{m+1} - \Delta_m b_{m+1}). \end{aligned} \tag{8}$$

The value of $d_{m,m+1}$ is the difference between the value of g at k_{m+1} and the tangent to g at k_m evaluated at k_{m+1}. Similar reasoning applies to $d_{m+1,m}$. It turns out that there exists a relationship between the degree of g within $[k_m,\, k_{m+1})$ and the value of $d_{m,m+1}$ and $d_{m+1,m}$: the curve g is a quadratic polynomial between k_m and k_{m+1} if, and only if, $d_{m,m+1} = d_{m+1,m}$; g is linear within $[k_m,\, k_{m+1})$ if, and only if, $d_{m,m+1} = d_{m+1,m} = 0$.

Shaw (1987) showed that we can write each $P_{2m}(g\,;\lambda)$ in (7) in terms of $d_{m,m+1}$ and $d_{m+1,m}$,

$$P_{2m}(g\,;\lambda) = \frac{\lambda}{2}\,\frac{3\,(d_{m,m+1} - d_{m+1,m})^2 + (d_{m,m+1} + d_{m+1,m})^2}{\Delta_m^3}, \quad m = 1, \ldots, \mathcal{K}-1. \tag{9}$$

The impact of the local penalty $P_{2m}(g\,;\lambda)$ on the portion of the spline curve g between $[k_m,\, k_{m+1})$ is now clear from (9). It penalizes generalizations of linear relationships, the strength of the penalization increasing as these generalizations become more complex. Hence, linear polynomials yield a value of zero for $P_{2m}(g\,;\lambda)$. Parabolas are penalized only through the term $(d_{m,m+1} + d_{m+1,m})^2$. Cubic polynomials are fully penalized since both terms in the numerator of $P_{2m}(g\,;\lambda)$ are different from zero. How much these terms affect the estimated curve is governed by the value of the smoothing parameter λ as described in Section 2.

The reasoning applied above for the penalty $P_2(g\,;\lambda)$ can be extended to the penalty $P_1(g\,;\lambda)$. Again we make use of the local structure of the *VFDP* and write

$$P_1(g\,;\lambda) = \frac{\lambda}{2}\sum_{m=1}^{\mathcal{K}-1}\int_{k_m}^{k_{m+1}} g'^2 = \sum_{m=1}^{\mathcal{K}-1} P_{1m}(g\,;\lambda). \tag{10}$$

The local penalty $P_{1m}(g\,;\lambda)$ can be written in terms of $d_{m,m+1}$ and $d_{m+1,m}$ as follows:

$$P_{1m}(g\,;\lambda) = \frac{\lambda}{2}\,\frac{(a_{m+1} - a_m)^2 + \frac{1}{20}\,(d_{m,m+1} - d_{m+1,m})^2 + \frac{1}{12}\,(d_{m,m+1} + d_{m+1,m})^2}{\Delta_m}. \tag{11}$$

The term $(a_{m+1} - a_m)^2$ in (11) comes as no surprise in the light of the discussion involving $P_{2m}(g\,;\lambda)$. It penalizes linear functions of t. Thus, $P_{1m}(g\,;\lambda)$ is increasingly penalizing curves that grow in complexity compared to a constant function of t, $t \in [k_m,\, k_{m+1})$, for which $P_{1m}(g\,;\lambda) \equiv 0$.

Given $\boldsymbol{\alpha}$, evaluation of the $\mathcal{K}-1$ penalties $P_{2m}(g\,;\lambda)$ in (7) and $P_{1m}(g\,;\lambda)$ in (10) is straightforward using the expressions in (9) and (11) respectively, which are valid for any configuration of knots. The double penalty $P_d(g\,;\lambda_1, \lambda_2)$ in (2) is simply the sum of $P_1(g\,;\lambda_1)$ and $P_2(g\,;\lambda_2)$ above.

3.3 Computational Details

Suppose we have observations $t_1, \ldots, t_n$ of the random variable T which defines the domain of g. Our aim is to find the parameter vector $\boldsymbol{\alpha}$ associated with the spline function $g(t)$ that maximizes the penalized log-likelihood criterion in (1). In order to characterize the solution we need some additional notation. We start by building the design matrix $\boldsymbol{T}$ associated with $t_1, \ldots, t_n$. Denote by $\boldsymbol{I}$ the $n \times (\mathcal{K}-1)$ incidence matrix whose ith row has zeros everywhere except for the column corresponding to the knot interval containing observation t_i, where it takes value 1. For each t_i we define the matrix

$$\boldsymbol{\Omega}^i = \begin{pmatrix} \phi_{01}(t_i) & \psi_{01}(t_i) & \phi_{11}(t_i) & \psi_{11}(t_i) & 0 & 0 & 0 & \ldots \\ 0 & 0 & \phi_{02}(t_i) & \psi_{02}(t_i) & \phi_{12}(t_i) & \psi_{12}(t_i) & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The ith row of the $n\times(2\mathcal{K})$ design matrix $\boldsymbol{T}$ is given by $\boldsymbol{I}_{i\bullet}\,\boldsymbol{\Omega}^i$, where the superscript $i\bullet$ represents the ith row of the matrix. The vector of function evaluations $\boldsymbol{g} = (g(t_1),\ldots,g(t_n))^T$ can be expressed as $\boldsymbol{T}\boldsymbol{\alpha}$, i.e., $g(t_i) = \boldsymbol{T}_{i\bullet}\,\boldsymbol{\alpha}$.

The penalty $P_2(g\,;\lambda)$ in (7) defines a quadratic form in $\boldsymbol{\alpha}$ through the expressions in (9)

$$P_2(g\,;\lambda) = \frac{\lambda}{2}\,\boldsymbol{\alpha}^T\boldsymbol{P}_2\,\boldsymbol{\alpha}, \tag{12}$$

where $\boldsymbol{P}_2$ is the $(2\mathcal{K})\times(2\mathcal{K})$ penalty matrix

$$\begin{pmatrix}
\frac{12}{\Delta_1^3} & \frac{6}{\Delta_1^2} & -\frac{12}{\Delta_1^3} & \frac{6}{\Delta_1^2} & 0 & 0 & \ldots\\
\frac{6}{\Delta_1^2} & \frac{4}{\Delta_1} & -\frac{6}{\Delta_1^2} & \frac{2}{\Delta_1} & 0 & 0 & \ldots\\
-\frac{12}{\Delta_1^3} & -\frac{6}{\Delta_1^2} & \frac{12}{\Delta_1^3}+\frac{12}{\Delta_2^3} & -\frac{6}{\Delta_1^2}+\frac{6}{\Delta_2^2} & -\frac{12}{\Delta_2^3} & \frac{6}{\Delta_2^2} & \ldots\\
\frac{6}{\Delta_1^2} & \frac{2}{\Delta_1} & -\frac{6}{\Delta_1^2}+\frac{6}{\Delta_2^2} & \frac{4}{\Delta_1}+\frac{4}{\Delta_2} & -\frac{6}{\Delta_2^2} & \frac{2}{\Delta_2} & \ldots\\
0 & 0 & -\frac{12}{\Delta_2^3} & -\frac{6}{\Delta_2^2} & \frac{12}{\Delta_2^3}+\frac{12}{\Delta_3^3} & -\frac{6}{\Delta_2^2}+\frac{6}{\Delta_3^2} & \ldots\\
0 & 0 & \frac{6}{\Delta_2^2} & \frac{2}{\Delta_2} & -\frac{6}{\Delta_2^2}+\frac{6}{\Delta_3^2} & \frac{4}{\Delta_2}+\frac{4}{\Delta_3} & \ldots\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix}. \tag{13}$$

The matrix $\boldsymbol{P}_2$ is non-negative definite and has rank $2\mathcal{K}-2$. The two zero eigenvalues correspond to linear and constant functions of t.

The matrix $\boldsymbol{P}_1$ associated with the penalty functional $P_1(g\,;\lambda)$ can also be easily derived using the $\mathcal{K}-1$ expressions in (11). It will be a non-negative definite matrix with rank $\mathcal{K}-1$. The unique 0 eigenvalue corresponds to constant functions of t. The penalty matrix for $P_d(g\,;\lambda_1,\lambda_2)$ in (2) follows directly from $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$.

4 Application to the Cox Model with Time-varying Regression Coefficients

In this section we illustrate the use of the *VFDP* with an application to the Cox model (Cox, 1972) with time-varying regression coefficients.

The Cox proportional hazards model assumes that the hazard ratio, which describes the effect of a predictor on survival, is constant over the follow-up period. In some survival studies, however, the proportional hazards assumption is not valid as the effect of some or all of the covariates on survival varies with time. This happens, for example, if an initially significant predictor gradually becomes non-significant after a certain period of follow-up.

In what follows we assume that the covariate effects are smooth functions of time that can be well approximated by a cubic spline.

4.1 Model Specification

The data available is of the form $D = (t_i, \delta_i, \boldsymbol{x}_i)_{i=1}^N$, the survival time t_i being complete if $\delta_i = 1$ and (right) censored if $\delta_i = 0$. The vector $\boldsymbol{x}_i$ contains the measurements on p selected time-constant covariates for individual i. The observed failure times are $\tilde{t}_1 < \cdots < \tilde{t}_n$. Throughout the rest of the paper we assume for simplicity that there are no ties present in the data. The Cox model with time-varying regression coefficients assumes the hazard relation

$$h\left(t \mid \boldsymbol{x}\right) = h_0(t) \exp\left(\sum_{j=1}^{p} g_j(t)\, x_j\right), \tag{14}$$

where $h_0(t)$ is an unspecified baseline hazard function, and $g_j(t)$ is a smooth function of t which defines the logarithm of the hazard ratio at time t corresponding to a unit increase in x_j. We are interested in estimating the coefficient functions $g_j(t)$, $j = 1, \ldots, p$.

The partial-likelihood for model (14) is given by

$$L_p(g_1, \ldots, g_p\,; D) = \prod_{f=1}^{n} \frac{\exp\left(\sum_j g_j(\tilde{t}_f)\, \tilde{x}_{fj}\right)}{\sum\limits_{l \in R_f} \exp\left(\sum_j g_j(\tilde{t}_f)\, x_{lj}\right)}, \tag{15}$$

where R_f is the set of individuals at risk just before time $\tilde{t}_f$ and $\tilde{x}_{fj}$ is the value of the jth covariate for the individual who fails at $\tilde{t}_f$. We assume that the censoring mechanism is noninformative with respect to the coefficient functions $g_j(t)$ so that an analysis based on the partial likelihood can be justified.

We parameterize each $g_j(t)$ using the *VFDP*. Because only information at observed failures contributes to the partial-likelihood, we place the knots $\{k_m\}_{m=1}^{\mathcal{K}}$ in the following way: we fix $k_1 = 0$, the time origin, and $k_{\mathcal{K}} = \max\{t_i\}$. For the remaining $\mathcal{K} - 2$ interior knots we take k_m, $m = 2, \ldots, \mathcal{K} - 1$, to be the $m/(\mathcal{K} - 1)$ quantile of the observed failures. This ensures roughly the same amount of information between knots and thus stable estimates. Moreover, because we do not expect $g_j(t)$ to have many local maxima or minima, the number $\mathcal{K}$ of knots considered here is moderate, say 10 or so. The partial-likelihood only depends on $g_j(t)$ through $g_j(\tilde{t}_f)$, $f = 1, \ldots, n$, $j = 1, \ldots, p$. Therefore we build the design matrix $\boldsymbol{T}$ defined in Section 3 using the knot configuration $\{k_m\}_{m=1}^{\mathcal{K}}$ described above and the set of observed failures $\{\tilde{t}_f\}_{f=1}^{n}$. Hence, $g_j(\tilde{t}_f) = \boldsymbol{T}_{f\boldsymbol{\cdot}}\, \boldsymbol{\alpha}_j$. The partial-likelihood in (15) becomes

$$L_p(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p\,; D) = \prod_{f=1}^{n} \frac{\exp\left(\sum_j \boldsymbol{T}_{f\boldsymbol{\cdot}}\, \boldsymbol{\alpha}_j\, \tilde{x}_{fj}\right)}{\sum\limits_{l \in R_f} \exp\left(\sum_j \boldsymbol{T}_{f\boldsymbol{\cdot}}\, \boldsymbol{\alpha}_j\, x_{lj}\right)}. \tag{16}$$

4.2 Bayesian Inference via MCMC

Bayesian inference relies on the posterior distribution of the model. It is a convenient framework that allows joint estimation of all the parameters in the model. Here we use MCMC simulation

techniques to sample from the posterior of interest. Proposals are based on the full conditionals of blocks of parameters given the rest of the data (Brezger & Lang, 2006). Sections 4.2.1 and 4.2.2 lay down the tools for Bayesian inference in the Cox model with time-varying regression coefficients when the single penalty functional $P_2(g\,;\lambda)$ is used. Inference for double penalty models will be addressed separately at a later stage in Section 4.2.3.

4.2.1 Prior Definitions

As we have pointed out in Section 2, there exists a relationship between penalty functionals and roughness priors for the spline parameters. Given that all the penalties discussed so far can be written as quadratic forms in $\boldsymbol{\alpha}_j$, an intuitive choice is a Gaussian type prior for $\boldsymbol{\alpha}_j$. For the single penalty model based on $P_2(g\,;\lambda)$ the prior is

$$p\left(\boldsymbol{\alpha}_j \mid \lambda_j\right) \propto \lambda_j^{\mathrm{rk}(\boldsymbol{P}_2)/2} \exp\left(-\frac{\lambda_j}{2}\,\boldsymbol{\alpha}_j^T \boldsymbol{P}_2\,\boldsymbol{\alpha}_j\right). \tag{17}$$

Note that the prior in (17) is improper as $\boldsymbol{P}_2$ is rank deficient.

For the hyperparameter λ_j a gamma prior (the conjugate prior of (17)) is assumed, i.e., $\lambda_j \sim G(a_j, b_j)$. The constants a_j and b_j are usually chosen so that the prior is vague, and therefore expresses our ignorance, while yielding a proper posterior distribution. Typical values include $a_j = 1$ and $b_j = 0.0001$, $b_j = 0.00001$ or $b_j = 0.000001$.

The joint posterior distribution of the single penalty model is given by

$$p\left(\boldsymbol{\theta} \mid D\right) \propto L_p(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p\,; D) \prod_j \lambda_j^{\mathrm{rk}(\boldsymbol{P}_2)/2} \exp\left(-\frac{\lambda_j}{2}\,\boldsymbol{\alpha}_j^T \boldsymbol{P}_2\,\boldsymbol{\alpha}_j\right) \prod_j \lambda_j^{a_j - 1} \exp\left(-b_j \lambda_j\right), \tag{18}$$

where $\boldsymbol{\theta}$ is the vector of all the parameters in the model.

4.2.2 Estimating the Spline Parameters

We explore the high-dimensional posterior distribution in (18) using a hybrid Metropolis-Hastings sampler. The main idea is to approximate the full conditional of $\boldsymbol{\alpha}_j$ by a Gaussian distribution obtained by accomplishing one Fisher scoring step as proposed in Brezger & Lang (2006). This process is repeated in every iteration of the sampler. Because of the specific shape of the partial-likelihood function, direct implementation of the algorithm in Brezger & Lang (2006) is not possible. The parameters of the proposal density need to be modified to account for the presence of the risk sets (Hastie & Tibshirani, 1993). Finally, the hyperparameter λ_j is updated using a Gibbs step.

Denote by $\boldsymbol{\alpha}_j^c$ and λ_j^c the current value of the parameters defining $g_j(t)$. A new value $\boldsymbol{\alpha}_j^p$ is proposed by drawing from the multivariate Gaussian proposal distribution $q\left(\boldsymbol{\alpha}_j^c, \boldsymbol{\alpha}_j^p\right)$ with variance matrix and mean

$$\boldsymbol{\Sigma}_j^c = \left[\boldsymbol{T}^T \boldsymbol{W}_{jj}^c\, \boldsymbol{T} + \lambda_j^c\, \boldsymbol{P}_2\right]^{-1}, \qquad \boldsymbol{m}_j^c = \boldsymbol{\Sigma}_j\, \boldsymbol{T}^T \boldsymbol{W}_{jj}^c \left(\boldsymbol{z}_j^c - \left[\boldsymbol{W}_{jj}^c\right]^* \sum_{\substack{s=1 \\ s \neq j}}^{p} \boldsymbol{W}_{js}^c\, \boldsymbol{T} \boldsymbol{\alpha}_s^c\right).$$

The matrix $\boldsymbol{W}_{js}^c = \text{diag}\,(w_{js1}^c, \ldots, w_{jsn}^c)$ has elements the weighted covariances of (x_j, x_s) in the risk set R_f,

$$w_{jsf}^c = \sum_{l \in R_f} \zeta_{fl}^c \, x_{lj} x_{ls} - \left(\sum_{l \in R_f} \zeta_{fl}^c \, x_{lj} \right) \left(\sum_{l \in R_f} \zeta_{fl}^c \, x_{ls} \right).$$

The weights ζ_{fl}^c depend on $\boldsymbol{\alpha}_j^c$ and represent the model probabilities

$$\zeta_{fl}^c = \frac{\exp\left(\sum_j \boldsymbol{T}_{f\cdot} \, \boldsymbol{\alpha}_j^c \, x_{lj} \right)}{\sum\limits_{r \in R_f} \exp\left(\sum_j \boldsymbol{T}_{f\cdot} \, \boldsymbol{\alpha}_j^c \, x_{rj} \right)}.$$

The vector $\boldsymbol{z}_j^c$ also depends on $\boldsymbol{\alpha}_j^c$ and plays the role of the usual working response vector in the Fisher scoring algorithm

$$\boldsymbol{z}_j^c = \left[\boldsymbol{W}_{jj}^c \right]^* \left(\tilde{\boldsymbol{x}}_j - \bar{\boldsymbol{x}}_j^c \right) + \left[\boldsymbol{W}_{jj}^c \right]^* \sum_{s=1}^{p} \boldsymbol{W}_{js}^c \, \boldsymbol{T} \boldsymbol{\alpha}_s^c.$$

The fth component of the n-dimensional vector $\bar{\boldsymbol{x}}_j^c$ is the weighted mean of x_j in the risk set R_f, i.e., $\bar{x}_{fj}^c = \sum_{l \in R_f} \zeta_{fl}^c \, x_{lj}$. The matrix $\left[\boldsymbol{W}_{jj}^c \right]^*$ denotes the generalized inverse of $\boldsymbol{W}_{jj}^c$.

We use Brezger & Lang (2006) modified algorithm which replaces $\boldsymbol{\alpha}_j^c$ by the current posterior mode approximation $\boldsymbol{m}_j^c$. This makes the proposal scheme independent of the current value of $\boldsymbol{\alpha}_j$, i.e., $q\,(\boldsymbol{\alpha}_j^c, \boldsymbol{\alpha}_j^p) \equiv q\,(\boldsymbol{\alpha}_j^p)$. Convergence to the stationary distribution is fast even with poor starting values for $\boldsymbol{\alpha}_j$.

Once $\boldsymbol{\alpha}_j$ has been updated, a new value for the hyperparameter λ_j is obtained through a Gibbs step by sampling from its full conditional, a gamma distribution with parameters

$$\check{a}_j = a_j + \frac{\text{rk}(\boldsymbol{P}_2)}{2}, \quad \check{b}_j = b_j + \frac{1}{2} \, \boldsymbol{\alpha}_j^T \, \boldsymbol{P}_2 \, \boldsymbol{\alpha}_j.$$

4.2.3 Double Penalty Models and Empirical Bayes Methods

The coefficient functions $g_j(t)$ in (14) generalize constant functions of t corresponding to the proportional hazards model. It therefore seems reasonable to consider penalty functionals that shrink towards a constant. This can be achieved using the double penalty $P_d(g_j\,; \lambda_j^1, \lambda_j^2)$ in (2). The improper prior for $\boldsymbol{\alpha}_j$ becomes

$$p\,(\boldsymbol{\alpha}_j \mid \lambda_j^1, \lambda_j^2) \propto \exp\left(-\frac{1}{2} \, \boldsymbol{\alpha}_j^T \left[\lambda_j^1 \boldsymbol{P}_1 + \lambda_j^2 \boldsymbol{P}_2 \right] \boldsymbol{\alpha}_j \right). \tag{19}$$

The hyperparameters λ_j^1 and λ_j^2 deserve special attention here. Since they are associated with the same spline function, their values are likely to be correlated, making independence a priori an implausible assumption. However, it is not clear how one should elicit a joint prior for λ_j^1 and λ_j^2. We therefore resort to empirical Bayes methods as proposed in Ruppert & Carroll (2000). The main idea is to estimate hyperparameters in a prior using the data at hand and then to plug-in those estimates in the prior as though they were known.

Let $\boldsymbol{\lambda}_j = \left\{\lambda_j^1, \lambda_j^2\right\}$, $j = 1, \ldots, p$. We use Akaikes information criterion (AIC) to select the smoothing parameters $\boldsymbol{\lambda}_j$,

$$\text{AIC}\,(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p) = \ell_p(\boldsymbol{\alpha}_1(\boldsymbol{\lambda}_1), \ldots, \boldsymbol{\alpha}_p(\boldsymbol{\lambda}_p)\,; D) - \text{tr}(\boldsymbol{R}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p))/n, \tag{20}$$

where ℓ_p is the logarithm of the partial-likelihood in (16) and $\boldsymbol{R}$ is the smoother (or hat) matrix of the additive fit in (14) whose trace provides an estimate of the total number of degrees of freedom in the model (Hastie & Tibshirani, 1990b). Since computation of $\boldsymbol{R}$ is too cumbersome we use the approximation in Hastie & Tibshirani (1990b) and write

$$\text{tr}\,(\boldsymbol{R}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p)) \approx 1 + \sum_{j=1}^{p} \left[\text{tr}(\boldsymbol{S}_j(\boldsymbol{\lambda}_j)) - 1\right], \tag{21}$$

where $\boldsymbol{S}_j(\boldsymbol{\lambda}_j) = \boldsymbol{T}\left[\boldsymbol{T}^T\boldsymbol{W}_{jj}\,\boldsymbol{T} + \left(\lambda_j^1\boldsymbol{P}_1 + \lambda_j^2\boldsymbol{P}_2\right)\right]^{-1}\boldsymbol{T}^T\boldsymbol{W}_{jj}$ is the smoother matrix associated with the fitted curve $\boldsymbol{g}_j(\boldsymbol{\lambda}_j) = \boldsymbol{T}\boldsymbol{\alpha}_j(\boldsymbol{\lambda}_j)$. We estimate the parameters $(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p)$ by

$$(\hat{\boldsymbol{\lambda}}_1, \ldots, \hat{\boldsymbol{\lambda}}_p) = \arg\max \text{AIC}\,(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p). \tag{22}$$

Simultaneous maximization of $\text{AIC}\,(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p)$ is very intensive computationally. We overcome this by using an adaptive algorithm that estimates each $\boldsymbol{\lambda}_j$ individually by maximizing the global criterion $\text{AIC}\,(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p)$. In practice, the smoothing parameters in $\boldsymbol{\lambda}_j$ are varied over a pre-specified grid. For each pair $\boldsymbol{\lambda}_j$ we estimate $\boldsymbol{\alpha}_j(\boldsymbol{\lambda}_j)$ using the Fisher scoring algorithm with weights and working response $\boldsymbol{W}_{js}$ and $\boldsymbol{z}_j$ described in the previous section. We then select $\hat{\boldsymbol{\lambda}}_j$ that maximizes $\text{AIC}\,(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p)$. Each iteration of the algorithm updates all p smoothing parameter pairs $\boldsymbol{\lambda}_j$, $j = 1, \ldots, p$. The algorithm stops when $\text{AIC}\,(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p)$ converges.

Given $\hat{\boldsymbol{\lambda}}_1, \ldots, \hat{\boldsymbol{\lambda}}_p$, estimation of $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$ proceeds as in the single penalty case. The joint posterior distribution of the model is now

$$p\,(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p \mid D, \hat{\boldsymbol{\lambda}}_1, \ldots, \hat{\boldsymbol{\lambda}}_p) \propto L_p(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p\,; D) \prod_j \exp\left(-\frac{1}{2}\,\boldsymbol{\alpha}_j^T\left[\hat{\lambda}_j^1\boldsymbol{P}_1 + \hat{\lambda}_j^2\boldsymbol{P}_2\right]\,\boldsymbol{\alpha}_j\right), \tag{23}$$

and the variance matrix of the proposal density in Section 4.2.2 becomes

$$\boldsymbol{\Sigma}_j = \left[\boldsymbol{T}^T\boldsymbol{W}_{jj}\,\boldsymbol{T} + \hat{\lambda}_j^1\,\boldsymbol{P}_1 + \hat{\lambda}_j^2\,\boldsymbol{P}_2\right]^{-1}.$$

4.3 Predicting Individual Survival

In survival analysis it is usually of interest to predict the survival experience of an individual with given covariate vector $\boldsymbol{v}$. More specifically, we would like to know this individual's survival probability beyond some time point t. This information is summarized in the survival function $S(t \mid \boldsymbol{v}) = \Pr\,(T > t \mid \boldsymbol{v}) = \exp\left\{-\int_0^t h\,(u \mid \boldsymbol{v})\,du\right\}$. For the Cox model with time-varying regression coefficients in (14) $S(t \mid \boldsymbol{v})$ becomes

$$S(t \mid \boldsymbol{v}) = \exp\left\{-\int_0^t h_0(u) \exp\left(\sum_{j=1}^{p} g_j(u)\,v_j\right) du\right\}. \tag{24}$$

An estimate of $S(t \mid \boldsymbol{v})$ can be obtained from estimates of $h_0(t)$ and $g_j(t)$, $j = 1, \ldots, p$. The parameter vector $\boldsymbol{\alpha}_j$ defining the coefficient function $g_j(t)$ is estimated using the Bayesian framework based on the partial-likelihood described in Section 4.2, where the baseline hazard $h_0(t)$ was treated as a nuisance parameter. The joint posterior distribution of $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$ using the partial-likelihood function can be seen as the marginal posterior distribution of $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$ when a gamma process is assigned to the cumulative baseline hazard (Sinha et al., 2000). Here we take a much simpler approach by using maximum likelihood techniques.

The full likelihood of the model in (14) can be written in terms of the hazard and survival functions as follows

$$L(h_0(t), g_1, \ldots, g_p \,; D) = \prod_{f=0}^{n} \left\{ \prod_{r \in F_f} h(\tilde{t}_r \mid \tilde{\boldsymbol{x}}_r)\, S(\tilde{t}_f \mid \tilde{\boldsymbol{x}}_f) \prod_{c \in C_f} S(t_c \mid \boldsymbol{x}_c) \right\}, \tag{25}$$

where F_f is the set of labels associated with individuals who fail at $\tilde{t}_f$, and C_f is the set of labels corresponding to individuals censored in $(\tilde{t}_f, \tilde{t}_{f+1})$, $f = 0, \ldots, n$, with $\tilde{t}_0 = 0$ (or k_1) and $\tilde{t}_{n+1} = \max\{t_i\}$ (or $k_{\mathcal{K}}$). Following Breslow (1972), we adopt the convention that all censored observations were censored at the preceding observed failure time. The set F_0 is empty and F_f, $f = 1, \ldots, n$, has only one element, $\tilde{t}_f$, since we assume that no tied observations exist in the data.

In his discussion of Cox's paper, Breslow (1972) suggested to take $h_0(t)$ to be a left continuous step function with jumps possibly only at the points in time where failures occurred, i.e.

$$h_0(t) = h_{0f}, \qquad \tilde{t}_{f-1} < t \leq \tilde{t}_f, \quad f = 1, \ldots, n. \tag{26}$$

Let the last observed failure before time t be $\tilde{t}_L$. The integral in (24) can be written as

$$\int_0^t h_0(u) \exp\left(\textstyle\sum_j g_j(u)\, v_j\right) du = \sum_{f:\, \tilde{t}_f \leq \tilde{t}_L} \int_{\tilde{t}_{f-1}}^{\tilde{t}_f} h_{0f} \exp\left(\textstyle\sum_j g_j(u)\, v_j\right) du. \tag{27}$$

The integrals in (27) are not analytically tractable and therefore have to be estimated through numerical integration. We follow Kauermann (2005) and apply the trapezium rule

$$\int_{\tilde{t}_{f-1}}^{\tilde{t}_f} \exp\left(\textstyle\sum_j g_j(u)\, v_j\right) du \approx (\tilde{t}_f - \tilde{t}_{f-1}) \, \frac{\exp\left(\sum_j g_j(\tilde{t}_f)\, v_j\right) + \exp\left(\sum_j g_j(\tilde{t}_{f-1})\, v_j\right)}{2}. \tag{28}$$

If $\hat{\boldsymbol{\alpha}}_j$ is the Monte Carlo estimate obtained from the MCMC algorithm of Section 4.2, we replace $g_j(\tilde{t}_{f-1})$ and $g_j(\tilde{t}_f)$ in (28) by their estimates, $\boldsymbol{T}_{f-1,\cdot}\hat{\boldsymbol{\alpha}}_j$ and $\boldsymbol{T}_{f\cdot}\hat{\boldsymbol{\alpha}}_j$ respectively. Conditional on $\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_p$, differentiation of (25) with respect to h_{0f}, $f = 1, \ldots, n$, yields the profile maximum likelihood estimate

$$\widehat{h}_{0f} = \left[(\tilde{t}_f - \tilde{t}_{f-1}) \sum_{l \in R_f} \frac{\exp\left(\sum_j \boldsymbol{T}_{f\cdot}\hat{\boldsymbol{\alpha}}_j\, x_{lj}\right) + \exp\left(\sum_j \boldsymbol{T}_{f-1,\cdot}\hat{\boldsymbol{\alpha}}_j\, x_{lj}\right)}{2} \right]^{-1}, \tag{29}$$

resulting in the following estimate for $S(t \mid \boldsymbol{v})$

$$\widehat{S}(t \mid \boldsymbol{v}) = \prod_{f:\, \tilde{t}_f \leq \tilde{t}_L} \exp\left\{ - \frac{\exp\left(\sum_j \boldsymbol{T}_{f\boldsymbol{.}}\, \hat{\boldsymbol{\alpha}}_j\, v_j\right) + \exp\left(\sum_j \boldsymbol{T}_{f-1,\boldsymbol{.}}\, \hat{\boldsymbol{\alpha}}_j\, v_j\right)}{\sum_{l \in R_f} \left[\exp\left(\sum_j \boldsymbol{T}_{f\boldsymbol{.}}\, \hat{\boldsymbol{\alpha}}_j\, x_{lj}\right) + \exp\left(\sum_j \boldsymbol{T}_{f-1,\boldsymbol{.}}\, \hat{\boldsymbol{\alpha}}_j\, x_{lj}\right)\right]} \right\}. \tag{30}$$

4.4 Analysis of Clinical Data

We now apply the proposed methodology to the PBC data set described in Fleming & Harrington (1991, App. D). It results from a Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. The data set contains measurements on 418 individuals. Besides the patient's survival time and censoring indicator, 17 potential prognostic factors were recorded. These include clinical, biochemical, serologic, and histologic measurements made at the time of randomization to one of the two treatments: placebo or D-penicillamine. See Fleming & Harrington (1991) for a detailed description of the study. We focus here on the five covariates found to be important by Fleming & Harrington (1991): age (`age` - in years), edema (`edema` - 0 if no edema, 1 if edema is present), log (albumin) (`albumin` - in gm/dl), log (bilirubin) (`bilirubin` - in mg/dl), and log (prothrombin time) (`protime` - in seconds). The hazard model we consider is thus

$$h\,(t \mid \boldsymbol{x}) = h_0(t) \exp\,(g_1(t)x_1 + g_2(t)x_2 + g_3(t)x_3 + g_4(t)x_4 + g_5(t)x_5), \tag{31}$$

where $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5)^T = (\texttt{age}, \texttt{edema}, \texttt{albumin}, \texttt{bilirubin}, \texttt{protime})^T$.

In total there were 160 deaths (approximately 60% censoring). We exclude the three individuals for whom information on one or more of the above selected covariates is missing. We center all continuous covariates around their means and randomly break the ties at each iteration of the MCMC sampler.

All five covariate effects are modeled as time-varying regression coefficients using the *VFDP* with $\mathcal{K} = 8$ knots. We set $a_j = 1$ and $b_j = 0.00000001$, $j = 1, \ldots, 5$, as parameters in the prior for λ_j in the single penalty case. The reasoning behind the choice of b_j is as follows: any change in the effect of a prognostic factor on the patient's survival is likely to be very mild, yielding $\boldsymbol{\alpha}_j^T \boldsymbol{P}_2\, \boldsymbol{\alpha}_j \approx 0$; the parameter $\bar{b}_j$ in the full conditional for λ_j is therefore influenced by its value a priori, b_j. Thus, in order to obtain clinically plausible covariate effect estimates, we choose a smaller b_j than what is usually advocated. The posterior distribution of the model is still proper (Hennerfeind et al., 2006).

For double penalty models, we have to pre-specify a grid of values for λ_j^1 and λ_j^2. The grid $10^{-5}, 10^{-4}, \ldots, 10^4, 10^5$ provided satisfactory results. We denote by $\widehat{g}_j^{\,\mathrm{SP}}(t)$ and $\widehat{g}_j^{\,\mathrm{DP}}(t)$ the estimated time-varying regression coefficient functions obtained using the single and double penalty models based on the penalty functionals $P_2(g_j\,; \lambda_j)$ and $P_d(g_j\,; \lambda_j^1, \lambda_j^2)$ respectively.

The estimated coefficient functions in Figure 1 were obtained using the output of a chain of length 20,000 for the spline parameters $\boldsymbol{\alpha}_j$, $j = 1, \ldots, 5$ (after an initial burn-in period of length

2,000). Convergence of the chain was determined by examining the plot of its path. The plots

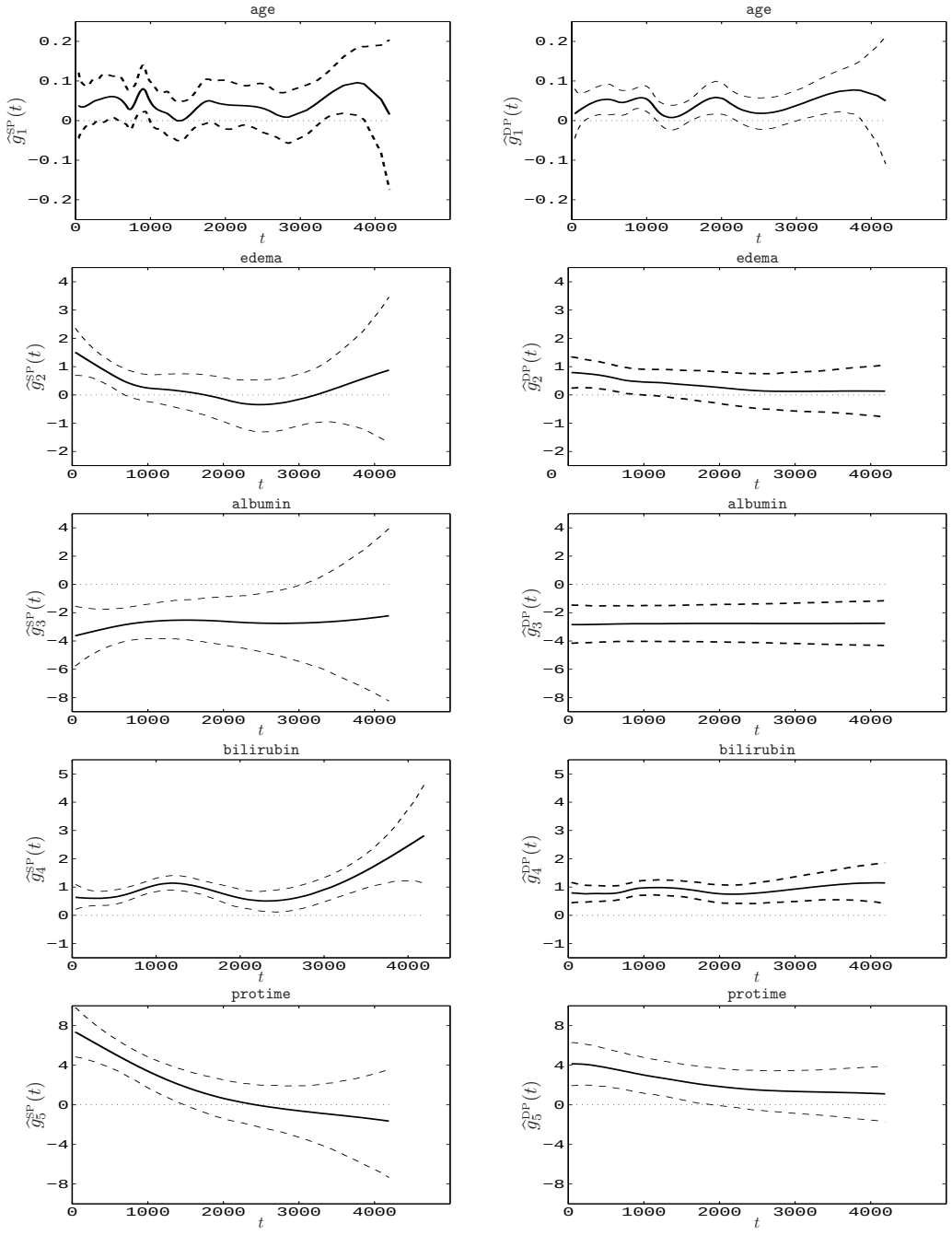

Figure 1: Estimated time-varying regression coefficients as a function of time t (in days) for PBC data set (solid line) together with 95% pointwise credibility intervals (dashed line) using both the single and double penalty models.

in Figure 1 suggest that the effects of `age`, `albumin` and `bilirubin` on survival are essentially constant throughout the study period. Older patients, with lower values of `albumin` and higher values of `bilirubin`, have worst survival prognosis. The regression coefficients associated with

covariates `edema` and `protime` seem to vary with time. This variation can be characterized as a loss of prognosis ability as the follow-up time increases. Initially, the presence of edema and larger values of prothrombin time have a negative effect on survival, but this eventually vanishes as time progresses. This is in line with the findings in previous analysis of the PBC data set (Fleming & Harrington, 1991; Martinussen & Scheike, 2006).

In general, the 95% pointwise credible intervals are narrower at the tails for double penalty estimates when compared to the ones obtained with a single penalty functional. This is not surprising giving that the uncertainty associated with the estimation of the smoothing parameters λ_j^1 and λ_j^2 is not taken into account if a double penalty is used.

Note that the estimate $\widehat{g}_3^{\mathrm{DP}}(t)$ clearly supports the proportional hazards assumption throughout, whereas $\widehat{g}_3^{\mathrm{SP}}(t)$ starts off constant but becomes non-significant at the end of the follow-up. This may be related to the fact that proportional hazards models arise as the smoothing limit of the double penalty model in (2). The estimate $\widehat{g}_4^{\mathrm{SP}}(t)$ suggests an increased risk for patients with high values of `bilirubin` around 1,000 days. This is probably the result from the large number of deaths between 500 and 1,500 days of follow-up (69, almost 44% of the total number of deaths), which correspond to patients that tend to have higher values of `bilirubin` than average. This is illustrated in Figure 2 (left plot), where the values of `bilirubin` corresponding to failures are plotted against time. A 'lowess' smooth is also shown. The increase in the values of `bilirubin` around day 1,000 is clear from this plot. This apparent increase in risk fades away in $\widehat{g}_4^{\mathrm{DP}}(t)$.

The estimates for the effect of `age` show too much variation, even when a double penalty functional is used. The fact that the values for covariate `age` are evenly spread around their mean might explain the excessive uncertainty when estimating $g_1(t)$. This can be seen in Figure 2 (right plot), which displays the values of `age` associated with observed failures versus the survival time together with a 'lowess' smooth which is essentially flat. Nevertheless one can still plausibly fit a constant function of time.

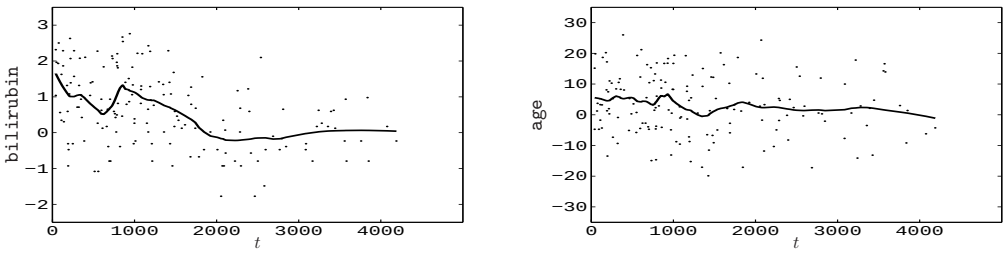

Figure 2: `bilirubin` (left plot) and `age` (right plot) corresponding to observed failures *vs* time t (in days) for the PBC data set. The solid lines in both plots correspond to 'lowess' smooths.

The added penalty functional in the double penalty model leads to covariate effect estimates that are, in general, smaller in magnitude. This is particularly true for `edema` and `protime`, the

covariates whose effects vary with time. This is reflected on the survival curve estimates. In Figure 3 we have represented the survival curve estimates for an average patient with and without edema for the two penalized spline models we studied. The loss of prognosis ability of the covariate `edema` is clear, particularly in the single penalty model estimate. Here, the two curves start off apart, with the presence of edema leading to lower survival probability, but eventually collide later in the follow up, when `edema` is no longer a significant predictor. This effect is less visible for the double penalty estimates as the magnitude of `edema`'s effect is smaller.

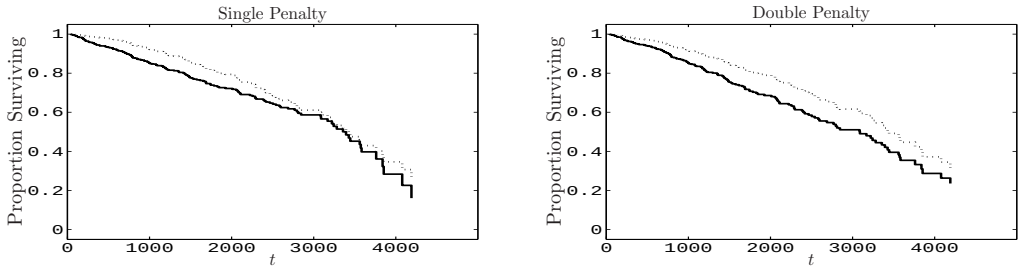

Figure 3: Survival function for a 51-year-old patient with 3.5 gm/dl albumin, 1.7 mg/dl bilirubin, 10.6 seconds of prothrombin time with edema (solid line), and no edema (dotted line), using the single penalty (left plot) and double penalty (right plot) models for the PBC data set.

5 Discussion

This paper explores a parametrization for cubic spline functions that is intuitive and allows easy interpretation and implementation of standard quadratic penalty functionals. This parametrization, which we call value-first derivative parametrization, is defined locally and its parameters are directly related to the spline function. It has been used previously in the context of numerical analysis as an interpolation tool under the name of cubic Hermite interpolation.

We focused here on penalized likelihood methods. We consider not only the standard case, where a single penalty functional is subtracted to the model's log-likelihood, but also an extension where double penalization is applied to the spline function. The latter becomes useful in situations where single penalty models do not attain the desired limit of smoothness. This is the case of the application we considered, where the Cox proportional hazards model is extended to allow time-varying regression coefficients. The model is applied to the well known PBC data set in Fleming & Harrington (1991). The estimates obtained with double penalization tend to be lower in magnitude than those obtained with a single penalty. In the cases where proportional hazards seems to be a plausible assumption, the double penalty estimates tend to behave better than the single penalty ones. We only considered the model where all effects are modeled as time-varying but a more flexible approach, including both static and dynamic effects is also possible with no

additional effort. The *VFDP* proved to be a flexible parametrization that is easy to implement and yields satisfactory estimates.

Bayesian inference is particularly useful within penalized likelihood methods because it allows the simultaneous estimation of the spline and smoothing parameters. However, if a double penalty is used, setting a prior for pairs of smoothing parameters is not straightforward. For this reason we turned to empirical Bayes methods. The spline parameters are updated using a Metropolis-Hastings algorithm with Fisher scoring proposals.

We are currently investigating adaptive and bivariate smoothing with the *VFDP*, two natural extensions for any smoothing method. Regarding adaptive smoothing, and in the single penalty case, this can be accomplished by, for example, taking the smoothing parameter to be, a priori, a correlated process through time.

Acknowledgments

M. J. Costa gratefully acknowledges financial support from Fundação para a Ciência e a Tecnologia through the grant SFRH/BD/16955/2004

References

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

Aldrin, M. (2006). Improved predictions penalizing both slope and curvature in additive models. *Computational Statistics and Data Analysis* 50 267–284.

Biller, C. & Fahrmeir, L. (1997). Bayesian spline-type smoothing in generalized regression models. *Computational Statistics* 12 135–151.

Breslow, N. E. (1972). Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B* 34 216–217.

Brezger, A. & Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis* 50 957–991.

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

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

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998a). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society, Series B* 60 333–350.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998b). Bayesian MARS. *Statistics and Computing* 8 337–346.

Eilers, P. H. C. & Goeman, J. J. (2004). Enhancing scatterplots with smoothed densities. *Bioinformatics* 20 623–628.

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

Eilers, P. H. C. & Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems* 66 159–174.

Eilers, P. H. C. & Marx, B. D. (2004). Splines, knots, and penalties. URL `http://www.stat.lsu.edu/faculty/marx/splines_knots_penalties.pdf`.

Fahrmeir, L. & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Applied Statistics* 50 201–220.

Fleming, T. R. & Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley-Interscience, New York.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* 19 1–67.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* 7 57–68.

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

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall.

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

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

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

Hastie, T. & Tibshirani, R. (2000). Bayesian backfitting. *Statistical Science* 15 196–213.

Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association* 101 1065–1075.

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics and Data Analysis* 49 169–186.

Lambert, P. & Eilers, P. H. C. (2005). Bayesian proportional hazards model with time-varying regression coefficients: A penalized Poisson regression approach. *Statistics in Medicine* 24 3977–3989.

Lang, S. & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* 13 183–212.

Martinussen, T. & Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data*. Springer.

Ruppert, D. & Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* 42 205–223.

Shaw, J. E. H. (1987). Numerical Bayesian analysis of some flexible regression models. *The Statistician* 36 147–153.

Sinha, D., Ibrahim, J. G. & Chen, M.-Hui. (2000). A Bayesian justification of Cox's partial likelihood. *Biometrika* 90 629–641.

Tian, L., Zucker, D. & Wei, L. J. (2005). On the Cox model with time-varying regression coefficients. *Journal of the American Statistical Association* 100 172–183.

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

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B* 65 95–114.