Varying-coefficient models

Trevor HASTIE
Statistics Research Department
AT&T Bell Laboratories
and
Robert TIBSHIRANI
Department of Preventive Medicine and Biostatistics
and
Department of Statistics
University of Toronto

SUMMARY

We explore a class of regression and generalized regression models in which the coefficients are allowed to vary as smooth functions of other variables. General algorithms are presented for estimating the models in a flexible manner, and a number of examples are given. This class of models ties together generalized additive models and dynamic generalized linear models into one common framework. When applied to the proportional hazards model for survival data, this approach provides a new way of modelling departures from the proportional hazards assumption.

Keywords: regression, splines, smoothing, generalized linear models, dynamic generalized linear models, generalized additive models, survival analysis.

1. Introduction

In recent years, some progress has been made towards increasing the flexibility of linear regression models. One focus has been to replace some or all of the linear and parametric functions of regressors by smooth nonparametric functions—so called *generalized additive models*; Hastie and Tibshirani (1990) give a survey of some of this work.

Here we consider apparently different generalizations—models that are linear in the regressors, but their coefficients are allowed to change smoothly with the value of other variables, which we might call "effect modifiers". Suppose we have a random variable Y whose distribution depends on a parameter η, and we also have predictors $X_1, X_2, \ldots X_p$ and $R_1, R_2, \ldots R_p$. A varying-coefficients model has the form

$$\eta = \beta_0 + X_1\beta_1(R_1) + \ldots X_p\beta_p(R_p). \tag{1}$$

Model (1) says that $R_1, \ldots R_p$ change the coefficients of the $X_1, X_2, \ldots X_p$ through the (unspecified) functions $\beta_1(\cdot), \ldots \beta_p(\cdot)$. The dependence of $\beta_j(\cdot)$ on R_j implies a special kind of interaction between each R_j and X_j.

A common setting for the application of these ideas is the class of generalized linear models (Nelder and Wedderburn 1974, McCullagh and Nelder 1989). In that case, η is called the *linear predictor* and is related to the mean $\mu = EY$ via the link function $\eta = g(\mu)$. In the simplest case of the Gaussian model, $g(\mu) = \mu$ and Y is normally distributed with mean η, and model (1) has the form

$$Y = X_1\beta_1(r_1) + \dots X_p\beta_p(r_p) + \epsilon \tag{2}$$

where $E(\varepsilon) = 0, \mathrm{var}(\varepsilon) = \sigma^2$. Other common models are loglinear models for which $\eta = \log\mu$ and Y has a Poisson distribution, and the linear logistic model having $g(\mu) = \log\mu/(1-\mu)$ and Y a binomial variate. Generalized additive models extend generalized linear models by replacing the linear predictor by an additive sum of smooth functions. We will see that the generalized additive model is a special case of the varying coefficient model, as is the *dynamic generalized linear model* (West, Harrison, and Migon 1985). However the approach to inference is quite different in the latter model.

1.1 An Illustration

In order to motivate the presentation, we begin with an example. Cleveland *et al.* (1991) examine 88 observations on the exhaust from an engine fueled by ethanol. The response variable, denoted by NO_x is the concentration of nitric oxide and nitrogen dioxide, normalized by the work load of the engine. The two predictors are equivalence ratio E, a measure of the fuel/air mixture, and the compression ratio C of the engine. The data are presented in Figure 1a. The top left and right panels of Figure 1a show NO_x plotted against E and C.

There is a strong quadratic-like effect of E and seemingly little effect of C, suggesting the simple model $NO_x \sim E^2$. The bottom left panel shows NO_x versus E, with the levels of C coded as low, medium and high, and suggests that in fact C might be interacting with E. Figure 1b reveals the form of this interaction. The broken lines show the fitted linear regressions of NO_x on C in 4 nonoverlapping ranges of E. Within each range of E, a linear model in C seems to fit well. But as E varies, both the intercept and slope of the line vary.

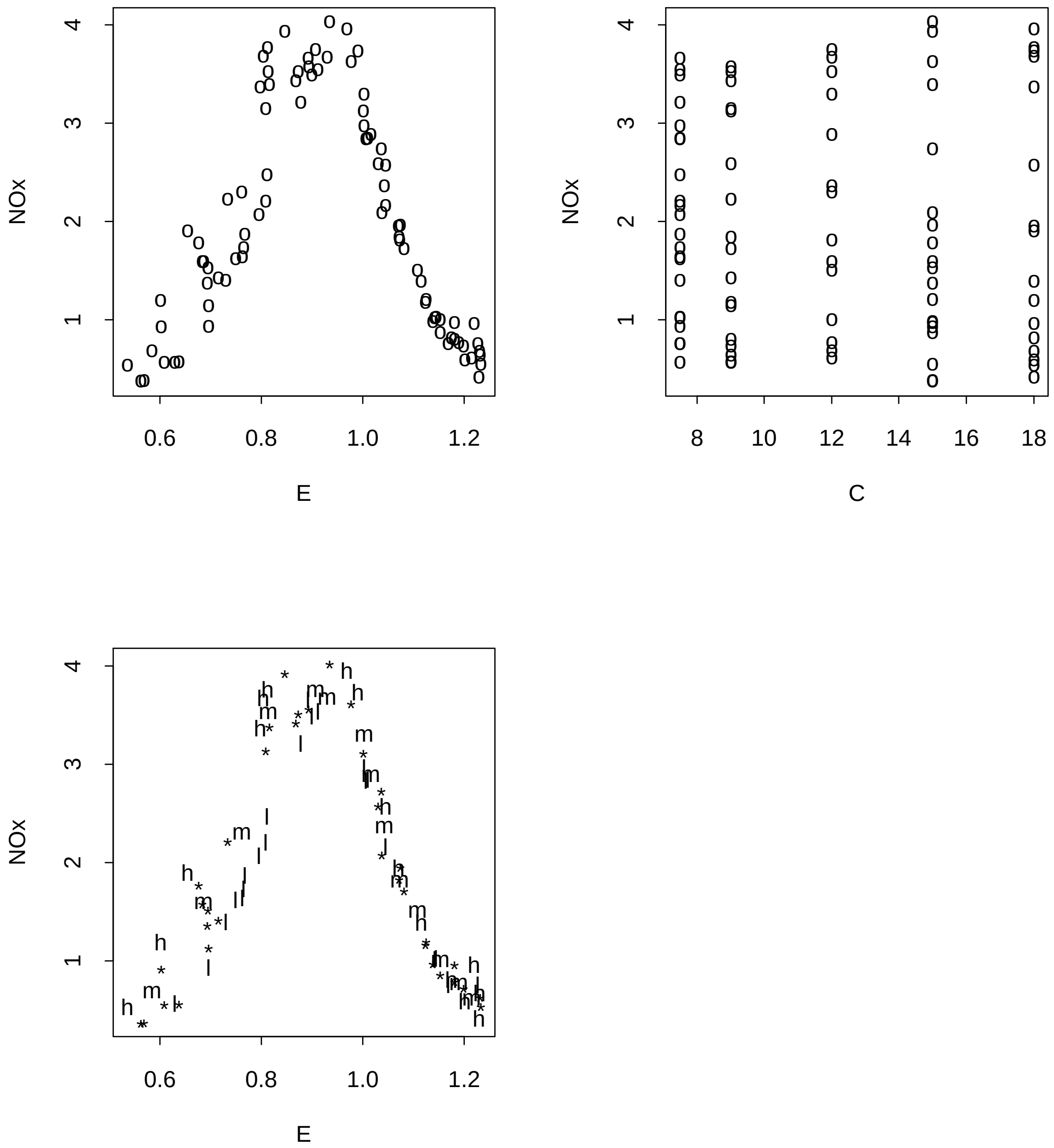

Fig. 1a. *Top two panels show plot of NO_x versus E and C. The bottom panel shows NO_X versus E with some values of C coded as low, medium or high. Intermediate values are coded with a $\star$.*

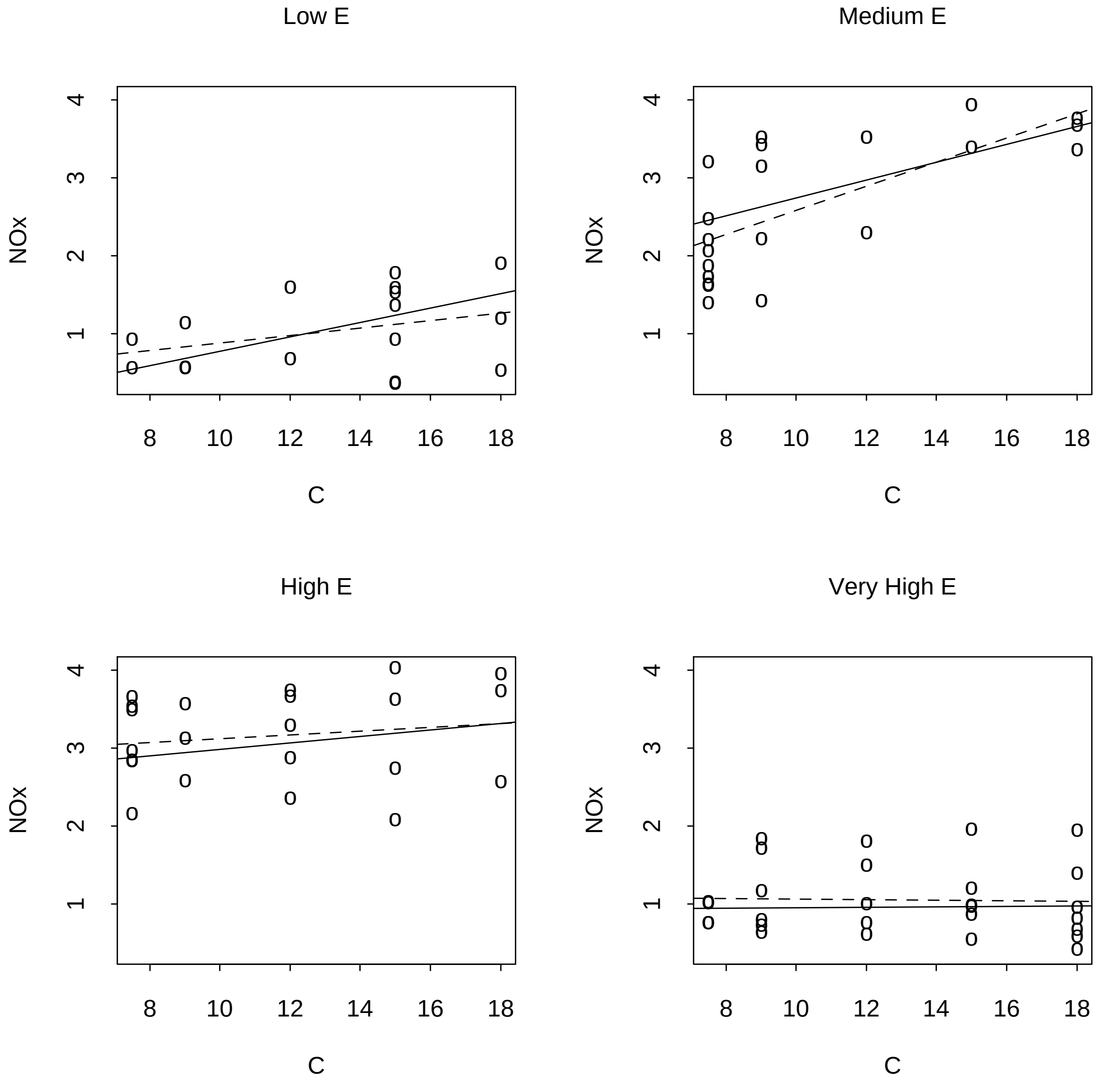

Fig. 1b. *Plots of NO_x versus C for low, medium, high and very high values of E. Included in each panel are the fitted lines (solid lines) from the varying coefficient model, taken at the median value of E for the data in the panel*

This leads us to consider a model of the form

$$NO_x = \beta_0(E) + \beta_1(E)C + \varepsilon, \tag{3}$$

Whereas the plots suggest $\beta_0(E) \sim E^2$, we will leave both $\beta_0(E)$ and $\beta_1(E)$ unspecified and fit them in a flexible manner. This is the model considered by Cleveland et al. and is an example of a varying-coefficient model.

Using an estimation procedure described in this paper, we produced the curves $\hat{\beta}_0(E)$ and $\hat{\beta}_1(E)$ shown in the first two panels of Figure 1c (with 90% pointwise standard error bands); these correspond to the solid lines in Figure 1b.

Table 1 *Analysis of variance table for a hierarchy of models for the NOx data*

	Model	*RSS*	*df*	*F-ratio*	*Pr(F)*
(i)	$NO_x = \beta_0(E) + \beta_1 C + \varepsilon$	5.19	79	20.1	0.0
(ii)	$NO_x = \beta_0(E) + \beta_1 CE + \varepsilon$	6.33	79	24.5	0.0
(iii)	$NO_x = \beta_0(E) + \beta_1 C + \beta_2 CE + \varepsilon$	3.20	78	14.5	0.0
(iv)	$NO_x = \beta_0(E) + \beta_1(E)C + \varepsilon$	2.65	72		

The $\hat{\beta}_0(E)$ and $\hat{\beta}_1(E)$ curves depend in a roughly parabolic fashion on E; the fitted model has a residual sum of squares of 2.65 and explains about 97% of the variation in the data. Roughly 8 degrees of freedom were used in estimating each of the curves; section 6 contains the details of how this number is obtained.

The corresponding surface is depicted in the bottom right panel of Figure 1c, and as shown by Cleveland *et al.*, it fits the data as well as a general two dimensional surface smoother, using fewer degrees of freedom.

The third plot in Figure 1c shows that the term $\hat{\beta}_1(E)C$ is not as important as $\hat{\beta}_0(E)$, and the perspective plot in Figure 1c supports this claim.

A number of simpler models are suggested by the perspective plot, and allow us to test for the interaction in (3). Table 1 compares these different models to the conditionally parametric model (numbered iv in the table). We use this model as the basis for an F-test in the table. Our conclusion there is that an interaction term is needed between E and C, and that it should be nonlinear in E.

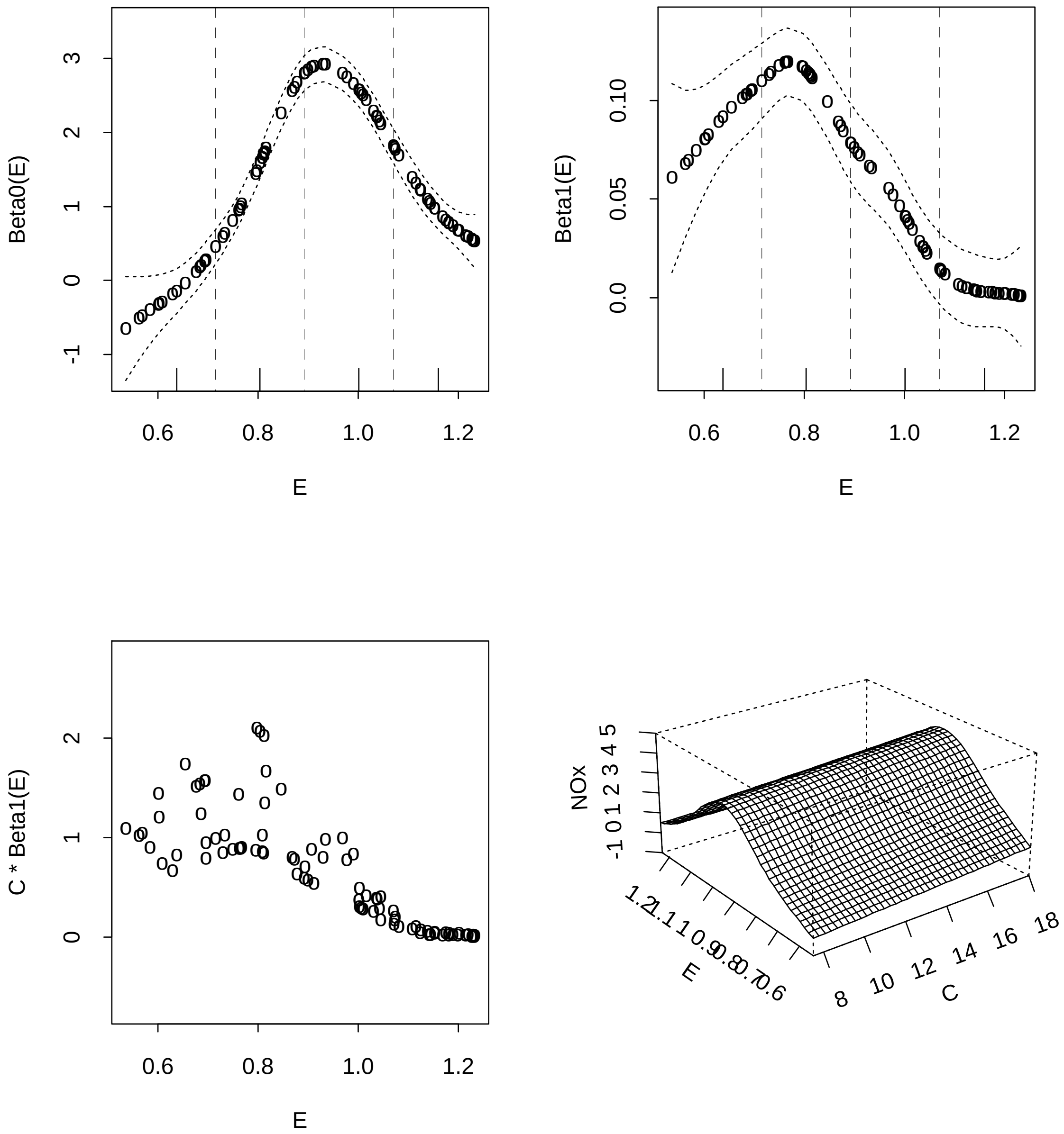

Fig. 1c. *The first two plots show* $\hat{\beta}_0(E)$ *and* $\hat{\beta}_1(E)$ *with pointwise* 90% *standard error bands defined by* ± 1.645 *times an estimate of the standard error. The broken lines delineate regions of data that are used to construct the four conditioned plots in Figure 1b. The four spikes at the base of the plots show the four values of* E *used in representing the fit in Figure 1b. The third panel shows the fitted effects* $C\hat{\beta}_1(E)$ *on the same scale as the plot for* $\hat{\beta}_0(E)$*, and the latter is clearly more important. The fourth panel is a perspective plot of the fitted* (E, C) *surface.*

In this paper we discuss many aspects of the varying-coefficient model (1), showing connections to other models in the literature and giving some general estimation procedures. The paper is organized as follows. In section 2 we give examples of the varying-coefficient model in different settings. In section 3 we describe some fitting procedures. Models with a single effect-modifying variable are somewhat simpler, as is their interpretation. Section 3.4 is devoted to them, and include the dynamic generalized linear model. Section 4 illustrates the model in a generalized regression setting and shows how the approach can be used to specify separate curves for one variable for different levels of another. Section 5 discusses the proportional hazards model in which "time" is the natural choice for the modifying variables R_j; this provides a new way of checking the proportional hazards assumption. Methods for inference are briefly discussed in section 6, and section 7 contains discussion. The paper ends with a technical appendix.

2. Examples of the varying-coefficient model

Although the varying-coefficients model looks quite specific, it is in fact rather general; many of the instances of (1) listed below will be familiar, and have appeared before in the literature. One purpose of this paper is show the common structure of these different models.

(i) If $\beta_j(R_j) = \beta_j$ (the constant function), then that term is linear in X_j. If all of the terms are linear, then (1) is the usual linear model or generalized linear model.

(ii) If $X_j = c$, (say $c = 1$), then the jth term is simply $\beta_j(R_j)$, an unspecified function in R_j. If all the terms have this form or are linear as in (i), then (1) has the form of a *generalized additive model.*

(iii) A linear function $\beta_j(R_j) = \beta_j R_j$ leads to a product interaction of the form $\beta_j X_j R_j$.

(iv) There is a different way of thinking about a model term $X\beta(R)$ when X is a binary (0/1) variable. Suppose there is also a term $\beta_0(R)$ in the model. This amounts to having a separate curve corresponding to each of the two values of X. More generally such terms are generated by a factor variable F, and each X_j represents a coding for the levels of F. Binary coding (dummy variable) is not essential; other contrasts can also be used. Symbolically a collection of such terms might be denoted by $F * \beta(R)$, and represents an interaction between the factor

F and the function $\beta(R)$. In fact, it is often useful to "cut" a quantitative variable into a number of groups in order to explore interactions with another quantitative variable in this way.

(v) Often the R_js will be the same variable, a factor such as "Time" or "Age", that we suspect could modify the effects of $X_1, \ldots X_p$. Suppose, for example, that the data consists of repeated measurements of the variables $\{Y, X_1, \ldots, X_p\}$ over n time points $t \in (t_1, \cdots t_n)$. Then we might model this as

$$\eta_t = \beta_0(t) + X_1(t)\beta_2(t) + \cdots X_p(t)\beta_p(t) \tag{4}$$

over time. West, Harrison, and Migon (1985) call this a "dynamic generalized linear model"; a comprehensive treatment is given in West and Harrison (1989). Cleveland (1991) calls such models "conditionally parametric", and allows conditioning variables (possibly vector valued) other than time. As we will see, Cleveland treats the $\beta_j(t)$s as nonparametric functions; given t, the model is a parametric and hence the name 'conditionally parametric".

(vi) Suppose the modifying variable R_j is taken to be X_j, and for simplicity suppose that the model is a normal linear model with only one term. Thus we have

$$Y = X\beta(X) + \epsilon \tag{5}$$

This is a common model for smoothing or nonparametric regression of Y versus X, and is discussed in various forms by Stone(1977), O'Hagan (1978), and Cleveland (1979).

(vii) Each R_j can be scalar or vector-valued. For most of the paper we will assume the R_j are scalar; extensions to the vector-valued case are mentioned in sections 3.4.1 and 7.

(viii) In all the above cases, there are many ways to model the so-far unspecified functions $\beta_j(R_j)$. For example one could use flexible parametric representations such as polynomials, fourier series or piecewise polynomials, or otherwise and more generally nonparametric functions. The latter can be approached in a variety of ways, for example using kernel methods, penalization, or stochastic Bayesian formulations.

As can be seen, there are a bewildering array of interesting special cases of the varying coefficients model, including generalized linear models, dynamic generalized linear models and generalized additive models. Many of these have been thoroughly studied and specific estimation procedures for them have been developed. Our second objective is to present an estimation procedure applicable to the fully general model (1). This procedure, along with specialized procedures for particular models, are described in the next section.

3. Estimation

Model (2) as it stands is too general for most applications, in that no restrictions are imposed on the coefficient functions. Unrestricted nonparametric estimation of these functions would not likely be possible except for special designs, and the problems encountered seem no different than those encountered in the nonparametric estimation of a single regression function. For observational data we are likely to see a different value for each R_j in each sample, and thus without any further assumptions it is not clear how to average.

For this reason we impose restrictions of one form or another on the coefficient functions; e.g., piecewise constant, smooth with known parametric form, or else smooth but nonparametric. One approach would be through parametric bases such as polynomial or trignometric functions. Typically these do not provide enough flexibility and local adaptiveness, and a set of regression spline bases with a fixed arrrangement of knots is likely to be preferable. One then proceeds exactly as for the linear model, only with (many) more variables defined by the products of each X_j and the bases for $\beta_j(R_j)$. Using this approach, all of the standard inferential tools can be used to evaluate sets of coefficients, influential points, etc. In this regard they are more convenient than the nonparametric procedures described in this section, for which these tools are not nearly as developed. Unfortunately the characteristics of the fitted curves can be quite different with minor changes in the positions of the knots, especially if only a few can be afforded.

In this section we present a general nonparametric procedure for the varying-coefficients model (1) based on a penalized least squares criterion, that does not suffer from the problems mentioned above. For simplicity of exposition, we give details of estimation of the regression form (2) rather than the generalized regression form (1). Extensions to the latter case usually involve imbedding a procedure for model (1) into a Newton-Raphson -type algorithm and is illustrated in section 4.

Following that, we describe two other approaches for the nonparametric estimation of the β_js applicable to the special case of a single effect-modifying variable as in (4). One approach is local plane-fitting, suggested by Cleveland *et.al.* (1991); the other is the dynamic linear model (West, Harrison, and Migon 1985, West and Harrison 1989). We point out some interesting relationships between these approaches and the general method described next.

3.1 Estimation in L_2

Here we present a population version of an algorithm that forms the basis of our nonparametric algorithms that follow. Suppose we decide to estimate $\beta_1(\cdot)\ldots\beta_p(\cdot)$ in (1) by minimizing

$$E(Y-\sum_1^p X_j\beta_j(R_j))^2$$

Conditioning on each R_j, a sufficient condition for the solutions is

$$E\{X_j(Y-\sum_1^p X_j\beta_j(R_j)\mid R_j\}=0,\ \ j=1,2,\ldots,p$$

To find $\beta_j(\cdot)$ say, we can rearrange the above equation and solve:

$$\begin{aligned}\beta_j(R_j)=&\frac{E\{X_j(Y-\sum_{k\neq j}X_k\beta_k(R_k))\mid R_j\}}{E(X_j^2\mid R_j)}\\ =&\frac{E\{X_j^2(Y-\sum_{k\neq j}X_k\beta_k(R_k))/X_j\mid R_j\}}{E(X_j^2\mid R_j)}\end{aligned}\tag{6}$$

Equation (6) is a ratio of two conditional expectations, and can be viewed as a conditional weighted mean, where the conditional weights are supplied by the term X_j^2; the term in the denominator ensures that the weights integrate to 1. There is a similar equation for each function $\beta_j,\ j=1,\cdots,p$, and the set of p equations must be solved simultaneously for the β_j. Since a scatterplot smoother can be viewed as a flexible estimate of a conditional expectation, this suggests that each function β_j can be estimated in an iterative "one-at-a-time" manner by smoothing $\{Y-\sum_{k\neq j}X_k\beta_k(R_k)\}/X_j$ on R_j, with weights X_j^2. This emerges as the central idea from the formal framework discussed next.

3.2 Penalized Least Squares

Suppose we have observations $y_1, \ldots y_n$ from the varying-coefficient model (1) and denote by x_{ij} and r_{ij} the observed values of X_j and R_j for the ith case. Then our model has the form

$$y_i = x_{i1}\beta_1(r_{i1}) + x_{i2}\beta_2(r_{i2}) \ldots x_{ip}\beta_p(r_{ip}) + \varepsilon_i$$

For estimation of $\beta_1, \ldots \beta_p$ we propose to minimize the penalized least squares criterion

$$J(\beta_1, \ldots \beta_p) = \sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j(r_{ij})\Big)^2 + \sum_{j=1}^{p}\lambda_j \int \{\beta_j''(r_j)\}^2 dr_j \tag{7}$$

The first term measures the goodness of fit and the second term penalizes the roughness of each β_j with a fixed parameter λ_j (to be chosen later).

Equation (7) is an example of an inverse problem, as discussed by O'Sullivan (1986) or Wahba (1990). We observe neither the β_js nor their sum, but the linear functionals

$$(L_j\beta_j)(r_{ij}) = x_{ij}\beta_j(r_{ij}), \quad j = 1, 2, \ldots p.$$

Alternatively, terms of this type can be viewed as a particular element in an expansion of an interaction term using tensor-product splines; again see Wahba (1990, Chapter 10). In Theorem 3.1 of Wahha (1990) a general result is given concerning bounded linear functionals in reproducing kernel Hilbert spaces. As a consequence of this result, the minimizers $\beta_1, \ldots \beta_p$ of (7) are cubic splines with each β_j having knots at the unique values of $r_{1j}, \ldots r_{nj}$. If the criterion has a unique solution with each β_j restricted to be linear, then a unique solution exists in the unrestricted problem. Practically speaking, this means that the model matrix

$$\begin{pmatrix} x_{11} & x_{11}r_{11} & x_{12} & x_{12}r_{12} & \cdots & x_{1p} & x_{1p}r_{1p} \\ x_{21} & x_{21}r_{21} & x_{22} & x_{22}r_{22} & \cdots & x_{2p} & x_{2p}r_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n1}r_{n1} & x_{n2} & x_{n2}r_{n2} & \cdots & x_{np} & x_{np}r_{np} \end{pmatrix} \tag{8}$$

must be of full rank. This will usually be the case; if not, the solution will not be unique.

3.3 Computational details

For convenience we parametrize the problem in terms of the natural cubic spline basis. To characterize the solution, we need some additional notation. Denote by $\mathbf{y}$ the n observed values of Y, and let $\mathbf{D}_j$ be a diagonal matrix with the n observed values of X_j on the diagonal. Denote the number of unique values of R_j by n_j, let $N_j^1(r_j), \ldots N_j^{n_j}(r_j)$ be the natural cubic spline basis functions for the jth variable, and let the basis matrix $\mathbf{N}_j$ have ikth element $N_j^k(r_{ij})$. We can express each β_j in terms of its basis functions $\beta_j(r_{ij}) = \sum_{\ell=1}^{n_j} \gamma_{\ell j} N_j^{\ell}(r_{ij})$. If we let $\boldsymbol{\beta}_j$ denote the function $\beta_j(r_{ij})$ evaluated at the n observed values of R_j, we can write $\boldsymbol{\beta}_j = \mathbf{N}_j \boldsymbol{\gamma}_j$. The penalized least squares equation (7) can be written as in terms of the finite dimensional parameters $\boldsymbol{\gamma}_j$

$$J(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_p) = \left\| \mathbf{y} - \sum_1^p \mathbf{D}_j \mathbf{N}_j \boldsymbol{\gamma}_j \right\|^2 + \sum_1^p \lambda_j \left\| \boldsymbol{\gamma}_j \right\|_{\Omega_j}^2 \tag{9}$$

where the penalty seminorm $\left\| \boldsymbol{\gamma}_j \right\|_{\Omega_j}^2 = \boldsymbol{\gamma}_j^T \boldsymbol{\Omega}_j \boldsymbol{\gamma}_j$, and $\boldsymbol{\Omega}_j$ has ikth element $\int N_j^{i''}(r) N_j^{k''}(r)\, dr$.

The score equations are

$$\frac{\partial J}{\partial \boldsymbol{\gamma}_j} = (\mathbf{N}_j^T \mathbf{D}_j^2 \mathbf{N}_j + \lambda_j \boldsymbol{\Omega}_j) \boldsymbol{\gamma}_j = \mathbf{N}_j^T \mathbf{D}_j (\mathbf{y} - \sum_{k \neq j} \mathbf{D}_k \mathbf{N}_k \boldsymbol{\gamma}_k),\ j = 1, \cdots, p \tag{10}$$

Direct solution of (10) involves solution of an $\sum n_j \times \sum n_j$ linear system, a task requiring $O((\sum n_j)^3)$ computations. A more efficient procedure can be derived by writing

$$\begin{aligned} \hat{\boldsymbol{\beta}}_j = \mathbf{N}_j \hat{\boldsymbol{\gamma}}_j &= \mathbf{N}_j (\mathbf{N}^T \mathbf{D}_j^2 \mathbf{N}_j + \lambda_j \boldsymbol{\Omega}_j)^{-1} \mathbf{N}_j^T \mathbf{D}_j (\mathbf{y} - \sum_{k \neq j} \mathbf{D}_k \mathbf{N}_k \boldsymbol{\gamma}_k) \\ &= \mathbf{S}_j(\lambda_j) \mathbf{D}_j^- (\mathbf{y} - \sum_{k \neq j} \mathbf{D}_k \mathbf{N}_k \boldsymbol{\gamma}_k), \end{aligned} \tag{11}$$

where the matrix operator

$$\mathbf{S}_j(\lambda_j) = \mathbf{N}_j (\mathbf{N}_j^T \mathbf{D}_j^2 \mathbf{N}_j + \lambda_j \boldsymbol{\Omega}_j)^{-1} \mathbf{N}_j^T \mathbf{D}_j^2$$

computes a weighted cubic smoothing spline with weights given by the diagonal elements of $\mathbf{D}_j^2$ (see Silverman 1985 or Wahba 1990 for a details of cubic smoothing splines). Hence the minimizers $\hat{\boldsymbol{\beta}}_1 \ldots \hat{\boldsymbol{\beta}}_p$ can be found one at a time, by successive smoothing operations applied to $\mathbf{D}_j^- (\mathbf{y} - \sum_{k \neq j} \mathbf{D}_k \mathbf{N}_k \boldsymbol{\gamma}_k)$ with smoothers $\mathbf{S}_j(\lambda_j)$. Algorithms of this kind are known as 'backfitting" procedures (See Green, Jennison and Seheult 1985, Breiman and Friedman 1985, Buja, Hastie and

Tibshirani 1989). Since a weighted cubic smoothing spline can be computed in $O(n_j)$ operations, this reduces the order of computations to solve the estimation equations from $O((\sum n_j)^3)$ to $O(\sum n_j)$

Appendix B gives a general result concerning the convergence of this procedure and other related procedures. The results are a generalization of those given in Buja, Hastie and Tibshirani (1989).

In expression (11) we have used the notation $\mathbf{D}_j^-$ to denote the generalized inverse of $\mathbf{D}_j$; this handles the possibility that some of the x_{ij}s are zero. In fact, such points drop out of the penalized likelihood and the solution for β_j has knots at only those values of r_{ij} for which $x_{ij} \neq 0$. By using the weight matrix $\mathbf{D}_j^2$ which ensures these observations are given zero weight, we can use (for convenience) the full basis and obtain an identical fit.

It may seem more natural to use the basis matrices $\mathbf{N}_j^* = \mathbf{D}_j\mathbf{N}_j$ in (9), and sidestep the need for weights entirely. Although this would be equivalent, it is not convenient. Standard efficient software for fitting weighted cubic smoothing splines can be used in our approach, while the former would require specialized software which would be unlikely to achieve the same efficiency.

The smoothing parameters $\lambda_1 \ldots \lambda_p$ used above are considered fixed. These parameters control the amount of smoothing, and in practice we must choose or estimate them in some fashion. One approach would be to use a criterion such as cross-validation or generalized cross-validation for their estimation. In the examples in this paper, we prefer to prespecify a target "degrees of freedom" for a function and then find the value of λ_j that achieves this target. This strategy is used for generalized additive models by Hastie and Tibshirani (1990). Some additional details on degrees of freedom are given in section 6.

One might consider using other smoothers in place of the weighted cubic smoothing spline in the iterative procedure above. For example, one might use a kernel or locally weighted running line smoother, or for time varying coefficients an exponentially weighted moving average (section 3.4.2). Convergence of the procedure can only be guaranteed for certain smoothers however, and we give details in Appendix B. Although our solution fits the general problem (1), it is not difficult to see that the algorithm reduces to those already developed for the special cases. In particular, when the X_j are all identically 1, we obtain the usual backfitting algorithm using unweighted smoothing splines.

3.4 Models with a single effect-modifying variable

Consider a varying coefficient model with a single effect-modifying variable

$$Y = X_1\beta_1(R) + \dots X_p\beta_p(R) + \varepsilon \tag{12}$$

This model is a special case of (1), and the general fitting method can be used for its estimation. However, the presence of only one effect-modifying variable suggests that more specialized fitting procedures might be used. In this section we outline two specialized approaches. The first involves plane-fitting, and this bears close relation to the work of Cleveland *et al.* (1991). The second is based on the dynamic linear model and the Kalman filter.

3.4.1 Conditionally parametric smoothing

For motivation let's look back at Example 1, and the model used there:

$$NO_x = \beta_0(E) + \beta_1(E)C + \varepsilon.$$

For each value of E, this model specifies a linear regression of NO_x on C, with the slope function varying (smoothly) with E. To fit the model, for each value $E = E_0$ we might fit a line (in C) to the data points having almost the same E value. To achieve this, we construct a neighborhood in the (E, C) plane that is an infinite strip in the C direction, and just wide enough in the E direction to capture a fraction or *span* of the E values around E_0. We then fit a simple linear regression of NO_x on C for these data, giving us point estimates $\hat{\beta}_0(E_0)$ and $\hat{\beta}_1(E_0)$, and this can be repeated for different values E_0.

More generally, suppose

$$Y = \mathbf{Z}^T\boldsymbol{\beta}(R) + \varepsilon$$

where $Z = (X_1, \dots X_p)$ and $\boldsymbol{\beta}(R) = (\beta_1(R), \dots \beta_p(R))$

For the moment let us focus on estimation in L^2, and therefore assume $\mathbf{Z}$, R and Y are integrable random variables, and we wish to minimize

$$E(Y - \mathbf{Z}^T\boldsymbol{\beta}(R))^2.$$

A sufficient requirement for the solution is that it minimize $E\{(Y - \mathbf{Z}^T\boldsymbol{\beta}(R))^2 \,|\, R = r\}$ for every r. This latter problem has solution

$$\hat{\boldsymbol{\beta}}(r) = E(\mathbf{Z}\mathbf{Z}^T \,|\, r)^{-1} E(\mathbf{Z}^T Y \,|\, r) \tag{13}$$

which is a linear regression of Y on $\mathbf{Z}$ for each value r. With observed data, we might estimate $\hat{\boldsymbol{\beta}}(r)$ by using a smoother to estimate each of the conditional expectations in (13). Note that $\mathbf{Z}\mathbf{Z}^T$ is matrix and we would require the conditional expectation of each of its elements.

Equivalently, expression (13) suggests estimation of $\boldsymbol{\beta}$ by fitting a hyperplane to Y as a function of $\mathbf{Z}$ in neighborhoods of each r value. This is an extension of the local linear fits of the previous section.

The approach of Cleveland *et al.* (1991) makes at least two important enhancements to this procedure:

- in addition to the restriction to the r-neighborhoods, observations are assigned weights from a *tri-cube* weight function based on their distance from r (the closer to r the bigger the weight). In particular, the observation furthest from r gets almost zero weight. This ensures that as we move r smoothly over the range of R, the estimates change smoothly as well.

- R itself is used in the local hyperplane fit, and possibly bilinear terms are included between R and the other predictors $\mathbf{Z}$. In other words the local model that is fit might be

$$\gamma_1 \mathbf{Z} + \gamma_2 R + \gamma_{12} \mathbf{Z} R. \tag{14}$$

The motivation for the last item has to do with bias reduction. To understand this, consider first the simple running mean of Y on R, ignoring $\mathbf{Z}$. The local neighbourhoods can be quite asymmetric at times, depending on the distribution of R, and thus the target point r might be far from the mean/median of values of R in its neighborhood. This is specially true at the boundaries, where in the extreme case, *all* the values of R are to the right of r at say the left boundary. For this reason, Cleveland (1979) and others proposed fitting a local linear regression in R rather than a constant, and then take the estimated conditional expectation of Y given $R = r$ to be the value from this regression line at $R = r$. The procedure here is an extension of this idea. Our goal is the local hyperplane regression in $\mathbf{Z}$, given R. Instead, we fit a hyperplane in $\mathbf{Z}$ and R (with bilinear cross-terms) and thereby reduce the bias due to asymmetry in the neighborhoods. The required hyperplane is simply obtained by fixing the value of $R = r$ in (14).

How does the procedure described by (13) relate to that of section 3.3? Note that when Z is scalar, we have

$$\begin{aligned}\hat{\beta}(r) =& \frac{E(ZY \mid r)}{E(Z^2 \mid r)} \\ =& \frac{E(\frac{Y}{Z}Z^2 \mid r)}{E(Z^2 \mid r)}\end{aligned}$$

This is estimated by a smooth of Y/Z on R with weights Z^2, in agreement with the method proposed in section 3.3.

3.4.2 A Bayesian perspective: the dynamic linear model

Consider the simple varying coefficient model $Y = X\beta(r) + \epsilon$ where ϵ is normally distributed with mean 0 and variance σ^2. One way to impose some structure on the process $\beta(r)$ is to assume an appropriate prior. The dynamic linear model has such a formulation, where the modifying variable is typically time. One version (West and Harrison 1989, page 108) is defined by the equations

$$\begin{aligned} Y_t =& X_t\beta_t + \nu_t, \quad \nu_t \sim N(0, V_t) \\ \beta_t =& G_t\beta_{t-1} + t\omega_t, \quad \omega_t \sim N(0, W_t) \end{aligned} \tag{15}$$

where t refers to time. The first expression in (15) is called the *observation equation* while the second is the called the *evolution equation.* The evolution equation specifies a Markov process for the regression parameter, which can be viewed as a function of time t. Formally, this differs from the approach above only in the choice of prior for $\beta_1, \ldots \beta_t$. The Markov assumption makes it convenient to carry out inference in a sequential manner. Simple updating formula are available for the mean and variance of β_t and also for the mean and variance of the predictive distribution of Y_t, based on the Kalman filter.

In equation (15) X_t and β_t can be vectors or scalars; for simplicity we focus on the scalar case. When G_t is one and the variance V_t is constant, West and Harrison (1989) show that the mean of β_t is approximately an exponentially weighted moving average of Y_j/X_j, $0 \le j \le t-1$. The weight assigned to Y_{t-j}/X_{t-j} is of the form $\delta^j X_j^2$, where $0 < \delta < 1$ is called a *discount factor*. Notice that this has the same form as the nonparametric regression method of section 3.3. The smoother used is an exponentially weighted average, with an asymmetric window that uses only past observations. When X_t and β_t are vectors, a similar correspondence exists between the Kalman filter approach and Cleveland *et. al.*'s approach described in the previous section. Other related approaches appear in the literature. Brown *et al.* (1975) propose tests for constancy of regression effects over time. Another approach is the "flexible least squares" method of Kalaba and Tesfatsion (1989). They minimize the residual sum of squares plus the squared first difference of the regression coefficient by sequential updating.

One advantage of the dynamic linear model is its rich Bayesian structure. This facilitates not only estimation of the mean vector but also forecasting of the process. This is especially useful when the data arrive sequentially as in the monitoring of clinical measurements of patients. On the other hand, the Markov assumption in (15) may not always be reasonable, and the inferences might also be heavily dependent on the normality assumptions in the model.

Smoothing splines also have a Bayesian interpretation (Kimeldorf and Wahba, 1970; Wahba, 1990). The prior process is assumed to be twice-integrated Brownian motion plus a deterministic linear trend, and the sampling distribution is assumed to be Gaussian. The smoothing spline estimate emerges as the mean of the posterior distribution for β. This framework can be extended in a straightforward manner to the varying-coefficients models, with the penalized least-squares estimate $\hat{\beta}_j$ of section 3.2 representing the mean of the posterior distribution for β_j.

In related work, Priestley (1980) discusses state dependent models for time series, in which the coefficients of an ARMA model for Y_t can depend on the previous state of the system Y_{t-1}, . For the time series setting, the models described here are less general as they would allow the coefficients of an ARMA model to depend only on $t-1$.

4. Example: Heart Disease Data

In this example we illustrate the use of a varying-coefficient model in a generalized regression setting, specifically a logistic model for binary data. The example also illustrates the use of a varying-coefficient model to specify separate curves in one variable for different levels of another. The estimation procedure is Newton-Raphson in the iteratively reweighted form used in generalized linear models (McCullagh and Nelder 1989, chapter 2) and generalized additive models (Hastie and Tibshirani 1990, chapter 6). The difference here is that the inner loop consists of the generalized backfitting algorithm of section 3.3.

In this particular example we investigate the interaction of a continuous variable with a binary treatment variable. The data come from a study of myocardial infarction (*MI*) and are described in some detail in Hastie and Tibshirani (1987). The risk factors under study here are systolic blood pressure S and cholesterol ratio C and treatment for high blood pressure T, ($T = 0$ denotes treatment absent, $T = 1$ denotes treatment present). In addition, there are confounding variables family history, age, and type A behavior; we will denote these collectively by X. This is a retrospective study; all the measurements were taken at a particular time period which was *after* the *MIs* had occurred for the cases. This of course limits the interpretations dramatically, and the interactions shown here are in fact a demonstration of the pitfalls of a naive investigation of such data. A larger set of risk factors is described in Hastie and Tibshirani (1987).

The top two plots in Figure 3 show the fit of the generalized additive model

$$logit\{pr(MI)\} = X\beta + \beta_S(S) + \beta_C(C) \tag{16}$$

to these data, ignoring the treatment T. The "U" shape blood pressure curve is suggestive, and this shape persists even when an additive treatment effect for the dummy variable T is present.

To investigate how the effects of S and C might differ whether or not the person was treated for high blood pressure, we fit the model

$$logit\{pr(MI)\} = X\beta + \beta_{S0}(S) + T \cdot \beta_{S1}(S) + \beta_{C0}(C) + T \cdot \beta_{C1}(C) \tag{17}$$

The functions $\beta_{S1}(S)$ and $\beta_{C1}(C)$ represent the effect of the treatment on the curves for S and C. The lower two plots in Figure 3 shows the estimated functions $\hat{\beta}_{S0}(S)$ and $\hat{\beta}_{S0}(S) + \hat{\beta}_{S1}(S)$ on the left, and $\hat{\beta}_{C0}(C)$ and $\hat{\beta}_{C0}(C) + \hat{\beta}_{C1}(C)$ on the right.

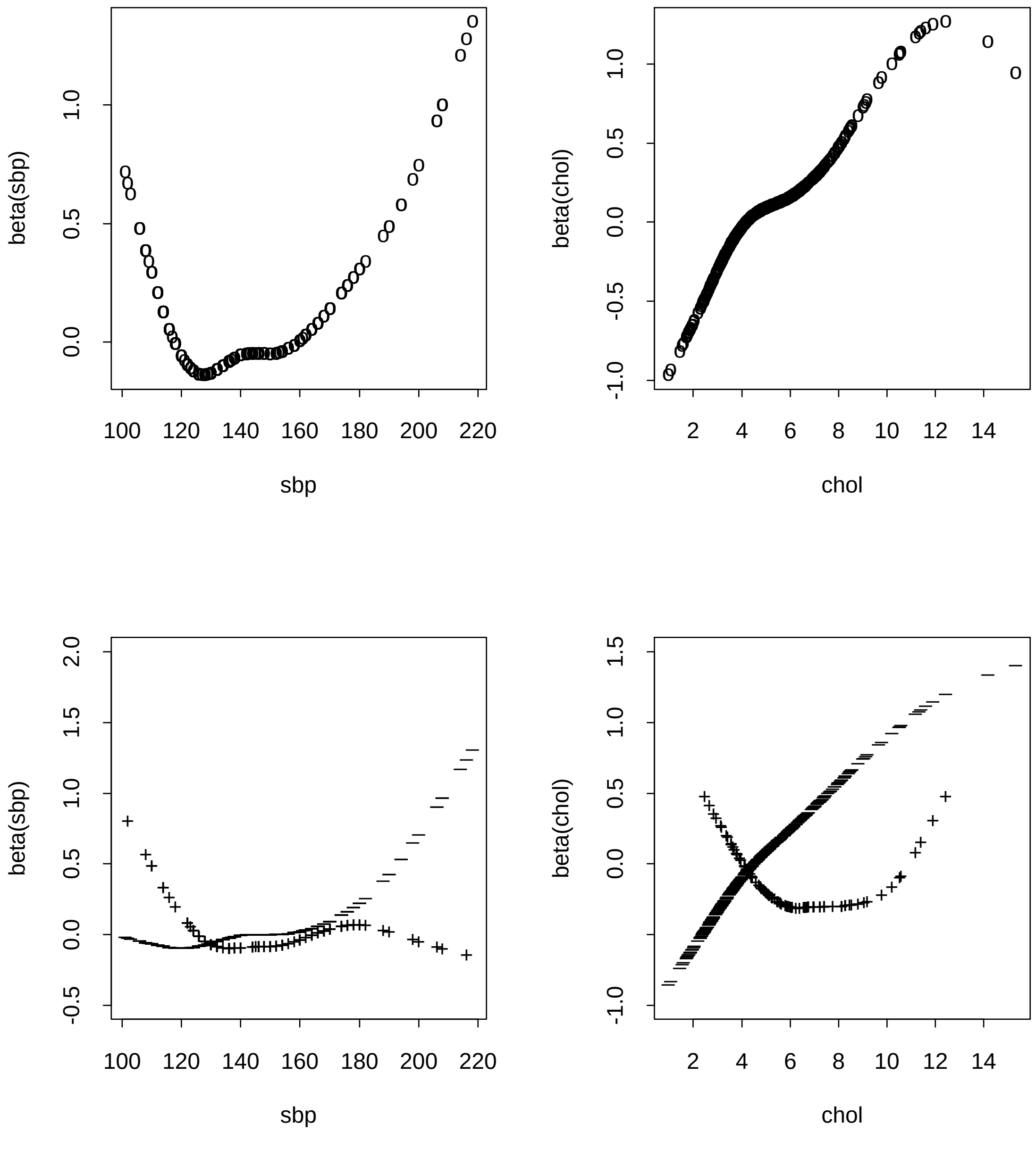

Fig. 3. *Example 3: The top two plots show the fitted curves for S and C in a generalized additive model, ignoring the treatment variable T. The lower left plot shows the pairs of functions $\hat\beta_{S0}(S)$ and $\hat\beta_{S0}(S)+\hat\beta_{S1}(S)$ for S with "$-$" denoting no treatment ($T=0$), "+" denoting treatment. The lower right panel is the equivalent plot for C.*

As mentioned, the interpretation of these interactions is tricky, since the data were collected retrospectively and the treatment was given after the heart attack occurred. They probably suggest that people who had heart attacks were more likely to receive treatment for high systolic blood pressure and cholesterol ratio, and the treatment successfully reduced both of those measures.

Hastie and Tibshirani (1990 section 9.5.2) describe a procedure for fitting a separate curve to each level of a group variable, in a generalized additive model. The method inserts a separate term for the risk factor (say S) for each group, and when the curve for group j is estimated in the backfitting procedure, the smoother uses only the data for that group. If the groups are coded by a set of dummy variables X, that procedure can easily be shown to be equivalent to the one described in the present paper and used in the this example.

5. Models for survival data

In problems involving a possibly censored lifetime, the available data are of the form $(y_1, x_{11}, \ldots x_{1p}, \delta_1,), \ldots, (y_n,, x_{n1}, \ldots x_{np}, \delta_n)$. The survival time y_i is complete if $\delta_i = 1$ and censored if $\delta_i = 0$, and $(x_{i1}, \ldots x_{ip})$ denotes the usual vector of predictors for the ith individual. The distinct failure times are denoted by $t_{(1)} < \cdots < t_{(d)}$, with d_k individuals failing at $t_{(k)}, k = 1, \ldots d$.

The proportional-hazards (or "Cox" model) assumes that

$$\lambda(t \mid X_1 \ldots X_p) = \lambda_0(t) \exp\left(\sum_j X_j \beta_j\right) \tag{18}$$

where $\lambda(t \mid X_1 \ldots X_p)$ is the hazard at time t given predictor values $X_1, \ldots, X_p$, and $\lambda_0(t)$ is an arbitrary baseline hazard function.

The methods of the previous sections could be applied here in order to make the model more flexible. In particular, we could allow the coefficients β_j to depend on a variable R_j, for example a disease severity score.

However "time" is a more compelling choice for r_j. Consider a model of the form

$$\lambda(t \mid X_1 \ldots X_p) = \lambda_0(t) \exp\Big(\sum_j X_j \beta_j(t)\Big) \tag{19}$$

Unless each $\beta_j(t)$ is a constant, this model represents non-proportional hazards. Hence its provides a way of assessing the proportional hazards assumption and describing any departures that are present.

In the special case of a single group variable ($X = 0/1$), the model (19) reduces

$$\lambda(t \mid 0) = \lambda_0(t); \quad \lambda(t \mid 1) = \lambda_0(t)e^{\beta(t)}$$

Thus $\beta(t)$ measures the difference in log relative risk between the two groups. In this case it might seem equivalent and simpler to model the hazard separately in each group. There is a difference, however, in the approach outlined here. The baseline hazard $\lambda_0(t)$ is estimated with the usual minimal restrictions and is thus piecewise constant and monotone, with jumps at the observed death times. The curve $\beta(t)$ on the other hand is modelled smoothly, and measures the difference between the hazards for the two groups.

Goodness of fit and diagnostic tests for the proportional hazards model (18) are difficult to develop because of the presence of the arbitrary baseline function; Cox and Oakes (1984) contains a fairly recent survey of some proposed approaches. Gore *et.al.* (1984) present a simple exploratory method that uses a discretization of the time axis.

Gamerman (1991) describes the dynamic linear model approach to estimation of model (19). He assumes that $\lambda_0(t)$ and the $\beta_j(t)$s are piecewise constant functions, constant between the distinct failure times. He uses the full likelihood for estimation, updating the terms sequentially in time. Here we instead use the partial likelihood for estimation, leaving $\lambda_0(t)$ unspecified but modelling the $\beta_j(t)$s in a smooth way.

The partial likelihood for model (19) is given by

$$L(\beta_1 \ldots \beta_p; \mathbf{y}) = \prod_{k=1}^{d} \frac{\exp\left(\sum_{j=1}^{p} s_{kj}\beta_j(t_{(k)})\right)}{\left\{\sum_{i\in\mathcal{R}_k} \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j(t_{(k)})\right)\right\}^{d_k}} \tag{20}$$

In the above expression, $\mathcal{R}_k$ is the set of indices of the individuals at risk at time $t_{(k)} - 0$, and s_{kj} is sum of the values of the jth covariate for the individuals failing at $t_{(k)}$.

One approach to estimation would be to use a parametric basis set for each of the $\beta_j(t)$s; this leads to a time-dependent covariate model, for which numerical algorithms are available. In our experience this approach is numerically unstable, and the algorithm often fails to converge . Here we focus on the penalized partial likelihood method which seems to have better numerical behaviour.

Denoting the logarithm of L by l, we estimate the $\beta_1, \ldots \beta_p$ by maximization of

$$J(\beta_1 \ldots \beta_p) = l(\beta_1 \ldots \beta_p; \mathbf{y}) - \tfrac{1}{2}\sum_{j=1}^{p} \lambda_j \int \{\beta_j''(r)\}^2 dr \tag{21}$$

Zucker and Karr (1990) study the mathematical properties of this model. They show that the solution is a cubic spline with knots at the unique failure times. They also study the general case of (21) in which the squared mth derivative of β_j is penalized. Their results require $m \geq 3$ for consistency of the maximizer of the penalized log-partial likelihood and $m \geq 4$ for asymptotic normality. However their conditions are sufficient (not necessary) and it is plausible that consistency and asymptotic normality also hold for the case $m = 2$ (given the success of squared derivative penalties in other settings).

Since the modifying variable (time) changes for each individual as the risk set changes, we cannot apply the method of the previous sections directly. A specialized algorithm for this problem is developed in Appendix A.

5.1 Example 4: Lung Cancer Data

Kalbfleisch and Prentice (1980) examined some data from a Veteran's Administration lung cancer trial. The outcome is survival time; the regressors are performance status, disease duration, age, prior therapy (yes or no), cell type (squamous, small, adeno, large), and treatment (yes or no). A fit of the proportional hazards model (Table 1) indicates that performance status and cell type are highly related to survival, while treatment is marginally significant. The maximized value of the likelihood is -475.2.

Table 1. *Proportional Hazards fit for Example 4*

Variable	*Coefficient*	Standard Error
Treatment	.289	.207
Cell type-squamous	.400	.282
Cell type- small cell	.457	.266
Cell type- adeno	.788	.303
Months from diagnosis	-9.2e-5	.009
Age	-.009	.009
Prior therapy	.007	.023
Perf Status	-.032	.006

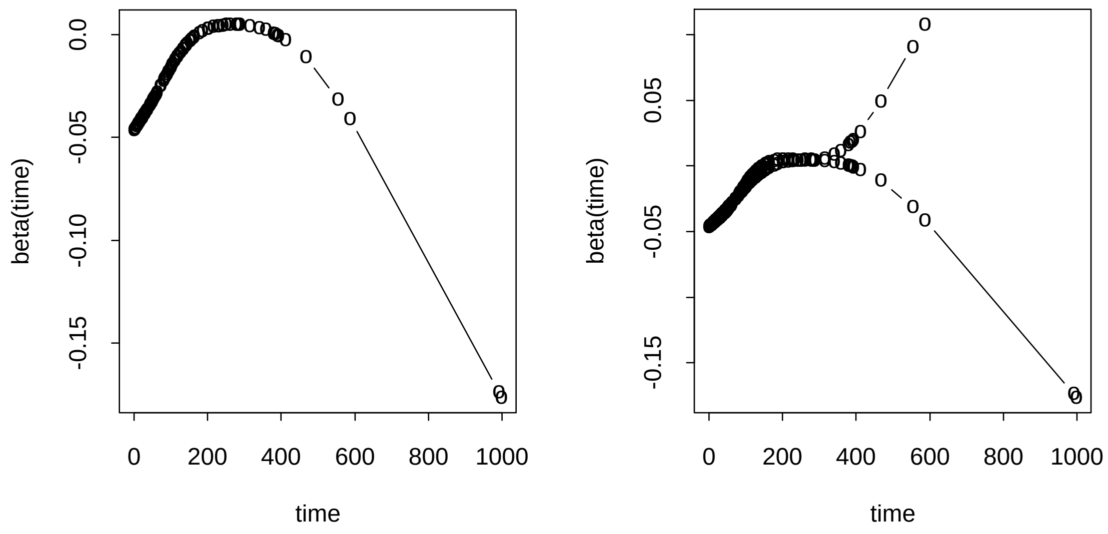

Fig. 4. *Plots of $\hat{\beta}(t)$ for example 4. The plot on the left uses all the data, and has a marked quadratic shape. The plot on the right shows the effect of removing the two values of time on the extreme right, and refitting the model; the tail of the function increases rather than decreases.*

We tried adding to this model a term of the form $x\beta(t)$ for performance status, choosing the target degrees of freedom to be 5. The resulting estimate $\hat{\beta}(t)$ is shown in Figure 4 (left plot) and indicates a decreasing then increasing advantage with time. The decrease after about 400 days is quite surprising but is based on relatively few observations. The log-likelihood increased to -467.7, corresponding to a change in twice log-likelihood ratio of 15.0 on about 4 degrees of freedom.

Inclusion of terms $x\beta(t)$ for other covariates did not significantly improve the fit; hence we conclude that these effects are well described by proportional hazards.

Figure 4 suggests the inclusion of time dependent covariates xt and xt^2 in the standard proportional hazards model. This model gives a value of -467.2 for the log-likelihood, and a very similar estimate to that in Figure 4. The point is that one can think of each $\beta_j(t)$ in model (19) as either a time-varying coefficient or rather view $x\beta_j(t)$ as a time dependent covariate.

The steep decline in the fitted function is rather suspicious, especially in light of the sparsity of data in that region. We refit the model, excluding the two points on the extreme right of the time axis. The resulting $\hat{\beta}(t)$ is shown in the right plot of Figure 4 (with the original curve superimposed). The right tail increases now, while the initial increase and levelling off of the function remains much the same. The log-likelihood for this deleted fit is -458.9; with the varying coefficient term removed it decreases to -468.7, a change of 9.8 for 4 degrees of freedom.

6. Inference for the model

Having estimated a curve $\beta(r)$, we would like to be able to test whether $\beta(r)$ is well approximated by a constant or linear function of r. The test of constant $\beta(r)$ is of particular interest because it indicates whether the simple linear model is adequate. Furthermore it is often of interest to include standard-error curves about a fitted curve, to get an idea of which features are estimated with precision.

For a broad class of smoothers, including kernels, nearest neighbors, as well as splines, one can use the approximate methods described in Hastie and Tibshirani (1990). These rely on the unifying concept of the "approximate degrees of freedom" of a smooth term, which depends in turn monotonically on the smoothing parameter. For a simple smooth fit $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, Hastie and Tibshirani (1990) argue that the appropriate degrees of freedom for the fit is $\nu = tr(2\mathbf{S} - \mathbf{S}\mathbf{S}^T)$ and that the distribution of the test statistic is roughly $F_{\nu,n-\nu}$ (more refined approximations can be found in Cleveland and Devlin, 1988). In the present case, the operator $\mathbf{S}$ that produces the fitted term $x\hat{\beta}(R)$ is $\mathbf{D}\mathbf{N}(\mathbf{N}^T\mathbf{D}^2\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}\mathbf{D}$ (in the notation of section 3.3). It can be seen that

$$tr(2\mathbf{S} - \mathbf{S}\mathbf{S}^T) = tr(2\mathbf{S}' - \mathbf{S}'\mathbf{S}'^T)$$

where $\mathbf{S}' = \mathbf{N}(\mathbf{N}^T\mathbf{D}^2\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}\mathbf{D}^2$ is the weighted cubic spline smoother that is used in the algorithm described in section 3.3. Conveniently, then, the degrees of freedom of a term $x\hat{\beta}(R)$ is simply that of the smoother used in computing it.

While $tr(\mathbf{S}')$ is easy to compute, $tr(\mathbf{S}'\mathbf{S}'^T)$ is not, and for this reason Hastie and Tibshirani (1990, appendix B) develop the approximation $tr(2\mathbf{S} - \mathbf{S}\mathbf{S}^T) \approx 1.25tr(\mathbf{S}) - 0.5$. We have made use of this approximation in the present work as well.

Finally, predictions from the model outside of the range of the observed $R_1, R_2, \ldots R_p$ will sometimes be of interest. The straightforward approach would be to extrapolate each estimate $\hat{\beta}_j(R_j)$ linearly (see e.g. Hastie and Tibshirani 1990, chap.2). The resulting estimates may be quite unstable, and a reasonable estimate of its accuracy would be difficult to obtain. If extrapolation is the main objective, the dynamic linear model with its richer probablistic structure is likely to be more useful.

7. Discussion

The varying-coefficient model described in this paper appears to be a potentially useful extension of regression and generalized regression models. It allows the coefficients that describe the effect of a regressor to vary as a function of another factor.

There are some directions that this work could be extended. The effect modifier R might be vector-valued, in which case a multidimensional smoother would be used in the estimation procedure for the function $\beta(\mathbf{R})$. The conditionally parametric models of Cleveland *et al.* automatically allow for this case when all the terms are modelled conditionally on the same $\mathbf{R}$.

Extensions to nonlinear regression models could also be explored. Another extension (suggested by Mike Leblanc) would be to allow terms of the from $X\beta(\gamma \boldsymbol{r})$ where $\boldsymbol{\gamma}$ is an unknown unit vector; this would be a generalization of projection pursuit regression (Friedman and Stuetzle, 1981). Hence one would look for directions in the effect-modifier space that result in larges changes in the coefficients.

Acknowledgements

We thank Scott Emerson for providing us a copy of his S function (Becker *et al.*, 1988) "coxrgrss" for fitting proportional hazards models and William Cleveland for sharing his work on conditionally parametric fitting before its publication. Comments by the editors and referees greatly improved the exposition. The second author gratefully acknowledges the support of the Natural Science and Engineering Research Council of Canada.

REFERENCES

Becker, R., Chambers, J. and Wilks. A. (1988) *The S language.* Wadsworth, Belmont CA.

Brown, R.L., Durbin, J., Evans, M.J. (1975) Techniques for testing the constancy of regression effects over time (with discussion). *J. R. Statist. Soc. B* **37**, 149–192.

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

Cleveland, W.S. (1979) Robust locally-weighted regression and smoothing scatterplots. *J. Am. Statist. Assoc.* **74**, 829–36.

Cleveland, W.S. and Devlin, S. (1988) Locally-weighted regression: an approach to regression analysis by local fitting. *J. Am. Statist. Assoc.* **83**, 597–610.

Cleveland, W.S., Grosse, E., Shyu, W. and Terpenning, I. (1991) Local Regression Models, In *Statistical models in S.* edited by Chambers, J. and Hastie, T., Wadsworth and Brooks/Cole, Pacific Grove, California.

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

Friedman, J.H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Am. Statist. Assoc.* **76**, 817–23.

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

Gore, S.M., Pocock, S.J. and Kerr, G.R¿ (1984) Regression models and non-proportional hazards in the analysis of breast cancer survival. *Appl. Statist.* **33**, 176–195.

Green, P.J. (1987) Penalized likelihood for general semi-parametric regression models. *Int. Statist. Rev.* **55**, 245–60.

Green, P.J., Jennison, C. and Seheult, A. (1985) Analysis of field experiments by least squares smoothing. *J. R. Statist. Soc. B* **47**, 299–315.

Härdle, W. (1989) *Applied nonparametric regression.* Cambridge University Press.

Hastie, T. and Tibshirani, R. (1987) Generalized additive models: some applications, *J. Am. Statist. Assoc.* **82**, 371–86.

Hastie, R. and Tibshirani, R. (1990) *Generalized additive models.* Chapman and Hall, London.

Kalaba, R. and Tesfatsion, L. (1989) Time-varying linear regression via flexible least squares. *Computer Math. Applic.* **17**, 125–1245.

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

Kimeldorf, G.S. and Wahba, G. (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.* **2**, 495–502.

McCullagh, P. and Nelder, J. A. (1989) *Generalized linear models.* Chapman and Hall, London.

Nelder. J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *J. R. Statist. Soc. A* **153**, 370–94.

O'Hagan, A. (1978) Curve fitting and optimal design for prediction (with discussion), *J. R. Statist. Soc. B* **40**, 1–42.

O'Sullivan, F. (1986) A statistical perspective on ill-posed inverse problems, *Statist. Sci.* **1**, 502–527.

Priestley, M.B. (1980) State dependent models: a general approach to non-linear time series. *J. Time Series Anal.* **1**, 47–71.

Silverman, B.W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. R. Statist. Soc. B* **36**, 111–47.

Stone, C.J. and Koo, C.Y. (1985) Additive splines in statistics. Proceedings of the Stat. Comp. Sec., ASA 45-8.

Wahba, G. (1990) *Spline models for observational data.* SIAM, Philadelphia.

West, M. and Harrison, P. J. (1989) *Bayesian forecasting and dynamic models.* Springer-Verlag, New York.

West, M., Harrison, P. J. and Migon, H. S. (1985) Dynamic generalized linear models and Bayesian forecasting (with discussion), *J. Am. Statist. Assoc.* , **80**,, 73—97.

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

APPENDIX A

In this Appendix we derive an algorithm for solving the penalized partial likelihood problem for the varying-coefficient Cox model. The notation is described in section 3.

Denote the vector of natural spline basis functions with knots at the unique failure times by $\mathbf{n}(t) = (n_1(t), \ldots n_d(t))$ and let $\mathbf{n}_k = \mathbf{n}(t_{(k)})$, $\mathbf{\Omega} = \int n_i(t) n_j(t) dt$. Then we can express $\beta_j(t_{(k)})$, the jth function evaluated at $t_{(k)}$, as $\mathbf{n}_k^T \boldsymbol{\gamma}_j$. The score equations for the unknown parameters $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_p$ are

$$\frac{\partial J}{\partial \boldsymbol{\gamma}_j} = \sum_{k=1}^{d} \Big(s_{kj} \mathbf{n}_k - d_k \frac{\sum_{i \in \mathcal{R}_k} x_{ij} \mathbf{n}_k^T \exp(\sum_j x_{ij} \mathbf{n}_k^T \boldsymbol{\gamma}_j)}{\sum_{i \in \mathcal{R}_k} \mathbf{n}_k^T \exp(\sum_j x_{ij} \mathbf{n}_k^T \boldsymbol{\gamma}_j)} \Big) - \lambda_j \mathbf{\Omega} \boldsymbol{\gamma}_j = 0; \;\; j = 1, 2, \ldots p$$

These score equations can be written in a more appealing form.

Let $p_{ki} = \exp(\sum_j x_{ij} \mathbf{n}_k^T \gamma_j) / \sum_{i \in \mathcal{R}_k} \mathbf{n}_k^T \exp(\sum_j x_{kj} \mathbf{n}_k^T \gamma_j)$ and $\mu_{kj} = d_k \sum_{i \in \mathcal{R}_k} p_{ki} x_{ij}$. Ignoring the d_k which account for ties, we can think of μ_{kj} as the mean of the jth regressor in the risk set $\mathcal{R}_k$, with respect to the model probabilities p_{ki}. Then the score equations can be expressed as

$$\frac{\partial J}{\partial \boldsymbol{\gamma}_j} = \mathbf{N}^T (\mathbf{s}_j - \boldsymbol{\mu}_j) - \lambda_j \mathbf{\Omega} \boldsymbol{\gamma}_j = 0; \;\; j = 1, 2, \ldots, p \tag{22}$$

where $\mathbf{s}_j = (s_{1j}, \ldots s_{dj})^T$, $\boldsymbol{\mu}_j = (\mu_{1j} \ldots . \mu_{dj})$, and the matrix $\mathbf{N}$ has rows $\mathbf{n}_k^T, k = 1, 2, \ldots, d$. Now if $\mathbf{N}$ consisted only of a column of ones and the penalty $\lambda_j \mathbf{\Omega} \boldsymbol{\gamma}_j$ was absent, (22) would simply require that the sum over the risk sets of the "residuals" $(\mathbf{s}_j - \boldsymbol{\mu}_j)$ be zero, the usual score equations for the proportional hazards model. The generalized score (22) has the form of a ridge regression: it measures the inner product of the residuals with the column space of $\mathbf{N}$, with a penalty for smoothness subtracted off.

If we compute the second derivative matrix $\partial^2 J / \partial \boldsymbol{\gamma}_j \partial \boldsymbol{\gamma}_l^T$, and work out the Newton-Raphson correction $(\Delta_1, \ldots \Delta_p)^T$ for solving (22) we obtain the $pd \times pd$ system

$$\begin{pmatrix} \mathbf{N}^T \mathbf{\Sigma}_{11} \mathbf{N} + \lambda_1 \mathbf{\Omega} & \mathbf{N}^T \mathbf{\Sigma}_{12} \mathbf{N} & \cdots & \mathbf{N}^T \mathbf{\Sigma}_{1p} \mathbf{N} \\ \mathbf{N}^T \mathbf{\Sigma}_{21} \mathbf{N} & \mathbf{N}^T \mathbf{\Sigma}_{22} \mathbf{N} + \lambda_2 \mathbf{\Omega} & \cdots & \mathbf{N}^T \mathbf{\Sigma}_{2,p} \mathbf{N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{N}^T \mathbf{\Sigma}_{p1} \mathbf{N} & \mathbf{N}^T \mathbf{\Sigma}_{p2} \mathbf{N} & \cdots & \mathbf{N}^T \mathbf{\Sigma}_{pp} \mathbf{N} + \lambda_p \mathbf{\Omega} \end{pmatrix} \begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_p \end{pmatrix} = \begin{pmatrix} \mathbf{N}^T (\mathrm{s}_1 - \boldsymbol{\mu}_1) - \lambda_1 \mathbf{\Omega} \boldsymbol{\gamma}_1^0 \\ \mathbf{N}^T (\mathrm{s}_2 - \boldsymbol{\mu}_2) - \lambda_2 \mathbf{\Omega} \boldsymbol{\gamma}_2^0 \\ \vdots \\ \mathbf{N}^T (\mathrm{s}_p - \boldsymbol{\mu}_p) - \lambda_p \mathbf{\Omega} \boldsymbol{\gamma}_p^0 \end{pmatrix} \tag{23}$$

where $\mathbf{\Sigma}_{jl}$ is a diagonal matrix with diagonal elements the covariances of (x_{ij}, x_{il}) in the risk set $\mathcal{R}_k$: $\mathbf{\Sigma}_{jl} = diag(\sigma_{jl1}, \ldots \sigma_{jld})$ where $\sigma_{jlk} = d_k^2 \{ \sum_{i \in \mathcal{R}_k} p_{ki} x_{ij} x_{il} - (\sum_{i \in \mathcal{R}_k} p_{ki} x_{ij})(\sum_{i \in \mathcal{R}_k} p_{ki} x_{il}) \}$. Direct solution of this system requires $O((pd)^3)$ computations , a formidable number in large samples since typically d is a significant fraction of n.

Here is an efficient computational strategy for solution of the system. Let $\mathbf{z}_j = \Sigma_{jj}^{-}(\mathbf{s}_j - \boldsymbol{\mu}_j) + \Sigma_{jj}^{-}\left\{\sum_{i=1}^{p}\Sigma_{ji}\boldsymbol{\beta}_j^0\right\}$. Then we can write each line of the system (23) as

$$\boldsymbol{\beta}_j = \mathbf{S}_j(\mathbf{z}_j - \Sigma_{jj}^{-}\sum_{i\neq j}\Sigma_{ji}\boldsymbol{\beta}_i)$$

where as before $\boldsymbol{\beta}_j = \mathbf{N}\boldsymbol{\gamma}_j$, and $\mathbf{S}_j = \mathbf{N}(\mathbf{N}^T\Sigma_{jj}\mathbf{N} + \lambda_j\boldsymbol{\Omega})^{-1}\mathbf{N}^T\Sigma_{jj}$ computes a weighted cubic smoothing spline.

This suggests the following iterative strategy similar to that for generalized regression models:

1. An inner backfitting loop that successively smooths $\mathbf{z}_j$ on the time variable, to get new values of $\hat{\boldsymbol{\beta}}_j$.
2. An outer loop that updates $\mathbf{z}_j, j = 1,\dots p$ using the current values of $\hat{\boldsymbol{\beta}}_1,\dots\hat{\boldsymbol{\beta}}_p$.

Step 1 solves the linear system (23) in $O(kn)$ computations, were k is the number of iterations (typically < 10). Note that step 1 differs from the usual backfitting-type procedure in that the "dependent" variable $\mathbf{z}_j$ and the terms subtracted from the dependent variable are different for each j. However, the results of Buja, Hastie and Tibshirani (1989) can be still be applied to establish convergence. Details are given in Appendix B.

APPENDIX B

In this appendix we study the system of equations

$$\begin{pmatrix} \mathbf{N}_1^T\boldsymbol{\Sigma}_{11}\mathbf{N}_1 + \lambda_1\boldsymbol{\Omega}_1 & \mathbf{N}_1^T\boldsymbol{\Sigma}_{12}\mathbf{N}_2 & \cdots & \mathbf{N}_1^T\Sigma_{1p}\mathbf{N}_p \\ \mathbf{N}_2^T\Sigma_{21}\mathbf{N}_1 & \mathbf{N}_2^T\Sigma_{22}\mathbf{N}_2 + \lambda_2\boldsymbol{\Omega}_2 & \cdots & \mathbf{N}_2^T\Sigma_{2,p}\mathbf{N}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{N}_p^T\Sigma_{p1}\mathbf{N}_1 & \mathbf{N}_p^T\Sigma_{p2}\mathbf{N}_2 & \cdots & \mathbf{N}_p^T\Sigma_{pp}\mathbf{N}_p + \lambda_p\boldsymbol{\Omega}_p \end{pmatrix} \begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_p \end{pmatrix} = \begin{pmatrix} \mathbf{N}_1^T\mathbf{u}_1 - \lambda_1\boldsymbol{\Omega}_1\boldsymbol{\gamma}_1^{old} \\ \mathbf{N}_2^T\mathbf{u}_2 - \lambda_2\boldsymbol{\Omega}_2\boldsymbol{\gamma}_2^{old} \\ \vdots \\ \mathbf{N}_p^T\mathbf{u}_p - \lambda_p\boldsymbol{\Omega}_p\boldsymbol{\gamma}_p^{old} \end{pmatrix} \tag{24}$$

where each $\mathbf{N}_j$ has columns that are evaluated natural spline basis functions for a set of n_j unique real values, $\boldsymbol{\Omega}_j$ has $ikth$ element $\int N_i''(r)N_k''(r)\,dr$, $\boldsymbol{\Sigma}_{ij}$ are all submatrices of the $\sum n_j \times \sum n_j$ symmetric non-negative definite matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\Delta}_j = \boldsymbol{\gamma}_j^{new} - \boldsymbol{\gamma}_j^{old}$. This system is quite general and special cases of it arise in the estimation of the models of this paper, and also in additive and generalized additive models. Specifically:

1. If $\mathbf{u}_j = \mathbf{y}, n_j = n$ for all j, $\boldsymbol{\Sigma}_{ij} = \mathbf{I}$ (the identity matrix) for all i,j, and $\boldsymbol{\gamma}_j^{old} = 0$, then (24) are the score equations for an additive model fitted by cubic smoothing splines. This is described in Buja, Hastie and Tibshirani (1989).

2. If $\mathbf{u}_j = \mathbf{u}$ is the derivative of the log-likelihood and $\boldsymbol{\Sigma}_{ij} = \mathbf{A}$, the matrix with the Fisher information components on its diagonal, then (24) represents the Newton-Raphson update from $\boldsymbol{\gamma}_j^{old}$ to $\boldsymbol{\gamma}_j^{new}$ for a generalized additive model. This is the "local scoring" procedure of Hastie and Tibshirani (1990, chapter 6); see also Green (1987).
3. As in (1) but with $\boldsymbol{\Sigma}_{ij}$ equal to $\mathbf{D}_i\mathbf{D}_j$, ($\mathbf{D}_i$ is defined in section 3.3), and $\mathbf{u}_i = \mathbf{D}_i\mathbf{y}$. Then (24) represents the score equations for the varying-coefficient model described in section 3.2.
4. As in (2), but with $\boldsymbol{\Sigma}_{ij}$ equal to the matrix $\mathbf{D}_i\mathbf{A}\mathbf{D}_j$. Then (24) represents the Newton-Raphson update from $\boldsymbol{\gamma}_j^{old}$ to $\boldsymbol{\gamma}_j^{new}$ for the varying-coefficient model in the generalized regression setting.
5. If $\mathbf{N}_j = \mathbf{N}, j = 1, \ldots p$, the evaluated natural spline bases for the unique failure times, and $\boldsymbol{\Sigma}_{ij}$ are the risk set covariances defined in section 5, then (24) represents the Newton-Raphson update from $\boldsymbol{\gamma}_j^{old}$ to $\boldsymbol{\gamma}_j^{new}$ for the survival model of section 5.
6. A more general version of the Cox model arises if we have different effect modifiers, other than time t, or simply other additive smooth terms in the model. Then the matrices $\boldsymbol{\Sigma}_{ij}$ are typically full and unstructured, but the results described below still carry through. Other models involving dependent contributions to the likelihood would also result in nondiagonal matrices $\boldsymbol{\Sigma}_{ij}$.

The following system of equations is equivalent to (24):

$$\begin{pmatrix} \mathbf{I} & \mathbf{S}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} & \mathbf{S}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13} & \cdots & \mathbf{S}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{1p} \\ \mathbf{S}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{I} & \mathbf{S}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{23} & \cdots & \mathbf{S}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p\boldsymbol{\Sigma}_{pp}^{-1}\boldsymbol{\Sigma}_{p1} & \mathbf{S}_p\boldsymbol{\Sigma}_{pp}^{-1}\boldsymbol{\Sigma}_{p2} & \mathbf{S}_p\boldsymbol{\Sigma}_{pp}^{-1}\boldsymbol{\Sigma}_{p3} & \cdots & \mathbf{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1^{new} \\ \boldsymbol{\beta}_2^{new} \\ \vdots \\ \boldsymbol{\beta}_p^{new} \end{pmatrix} = \begin{pmatrix} \mathbf{S}_1\mathbf{z}_1 \\ \mathbf{S}_2\mathbf{z}_2 \\ \vdots \\ \mathbf{S}_p\mathbf{z}_p \end{pmatrix} \tag{25}$$

where $\boldsymbol{\beta}_j = \mathbf{N}_j\boldsymbol{\gamma}_j$, $\mathbf{z}_j = \Sigma_{jj}^{-}\mathbf{y}_j + \Sigma_{jj}^{-}\left\{\sum_{i=1}^{p}\Sigma_{ji}\boldsymbol{\beta}_j^{old}\right\}$ and $\mathbf{S}_j = \mathbf{N}_j(\mathbf{N}_j^T\boldsymbol{\Sigma}_{jj}\mathbf{N}_j + \lambda_j\boldsymbol{\Omega}_j)^{-1}\mathbf{N}_j^T\boldsymbol{\Sigma}_{jj}$, a weighted smoothing spline smoother operator. In this form, we can see how other smoothers different from smoothing splines might be substituted for the $\mathbf{S}_j$. It is also easy to write down the form of the Gauss-Seidel iterations for solving the system:

$$\boldsymbol{\beta}_j = \mathbf{S}_j(\mathbf{z}_j - \boldsymbol{\Sigma}_{jj}^{-1}\sum_{i\neq j}\boldsymbol{\Sigma}_{ji}\boldsymbol{\beta}_i) \tag{26}$$

We extend the results of Buja *et. al* (1989) concerning the existence of solutions to (24) or (25), and convergence of the Gauss Seidel algorithm:

Theorem.

If $\mathbf{\Sigma}$ is symmetric and nonnegative definite, then the system of equations (24) (or (25)) is consistent and the Gauss-Seidel iteration (26) converges to a solution for any starting values.

Proof.

Consistency: Let $\mathbf{z}^* = \mathbf{\Sigma}^{-}\mathbf{z}$ where $\mathbf{z} = (\mathbf{z}_1, \ldots \mathbf{z}_p)^T$. Then the solutions to (25) can be expressed as the minimizers of the quadratic function

$$(\mathbf{z}^* - \mathbf{N}\boldsymbol{\gamma})^T \mathbf{\Sigma}(\mathbf{z}^* - \mathbf{N}\boldsymbol{\gamma}) + \boldsymbol{\gamma}^T \mathbf{\Omega}\boldsymbol{\gamma} \tag{27}$$

where $\mathbf{\Omega} = \mathrm{diag}(\mathbf{\Omega}_1, \cdots, \mathbf{\Omega}_p)$ and $\mathbf{N} = (\mathbf{N}_1 : \cdots \mathbf{N}_p)$. Since both $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ are non-negative definite, (27) is nonnegative. A real-valued multivariate function bounded below must have a minimum and therefore the system (25) is consistent.

Convergence of Gauss-Seidel iteration. [Aside: typical convergence results for the Gauss-Seidel procedure require that the system matrix be symmetric and positive definite. The left hand matrix in (24) is symmetric but not necessarily positive definite; the matrix may possess a non-trivial null space, a phenomenon called “concurvity” by Buja *et. al* (1989). These authors proved the convergence of the Gauss-Seidel procedure through a “semi-norm descent principle”. We make use of this principle here to establish convergence.]

Without loss of generality we may consider the homogeneous version of the function (27) (that is, take $\mathbf{z}^* = 0$):

$$Q(\boldsymbol{\gamma}) = \boldsymbol{\gamma}^T(\mathbf{N}^T\mathbf{\Sigma}\mathbf{N} + \mathbf{\Omega})\boldsymbol{\gamma}.$$

The function $Q(\boldsymbol{\gamma})$ is a non-negative quadratic form and $|\,\boldsymbol{\gamma}\,| = \sqrt{Q(\boldsymbol{\gamma})}$ defines a complex semi-norm. To prove consistency, we use the following:

Lemma: *Seminorm descent principle* (Buja *et al.*, 1989)

If $|\mathbf{f}|$ is a complex seminorm and $\mathbf{T}$ a linear mapping on $\mathcal{C}^{np}$ satisfying $|\mathbf{Tf}| < |\mathbf{f}|$ unless $|\mathbf{f}| = 0$, and $\mathbf{Tf} = \mathbf{f}$ for $|\mathbf{f}| = 0$, then $\mathbf{T}^m$ converges to a limit $\mathbf{T}^\infty$ with the properties $|\mathbf{T}^\infty\mathbf{f}| = 0$ for all $\mathbf{f}$, $(\mathbf{T}^\infty)^2 = \mathbf{T}^\infty$, and $\mathbf{T}\mathbf{T}^\infty = \mathbf{T}^\infty\mathbf{T} = \mathbf{T}^\infty$.

In the present problem, the Gauss Seidel iterations are defined by $\mathbf{T} = \mathbf{T}_p\mathbf{T}_{p-1}\ldots\mathbf{T}_1$ where $\mathbf{T}_j$ leaves $\boldsymbol{\gamma}_i$ unchanged for $i \neq j$, and maps $\boldsymbol{\gamma}_j$ according to (26). To apply the Lemma, we need to show that $|\,\mathbf{T}\boldsymbol{\gamma}\,| < |\,\boldsymbol{\gamma}\,|$ unless $|\,\boldsymbol{\gamma}\,| = 0$, and $\mathbf{T}\boldsymbol{\gamma} = \boldsymbol{\gamma}$ for $|\,\boldsymbol{\gamma}\,| = 0$. The first fact follows since the iteration (26) minimizes $Q(\cdot)$ along the jth coordinate, and at the global minimum $Q(\boldsymbol{\gamma}) = 0$. The second fact is verified by checking directly that if $|\,\boldsymbol{\gamma}\,| = 0$ then $\mathbf{T}_j\boldsymbol{\gamma} = \boldsymbol{\gamma}$ for all j. Hence convergence is established. ♣

To apply this theorem we require that $\boldsymbol{\Sigma}$ be symmetric and non-negative definite. This is immediate in all of the cases listed at the beginning of the appendix, except the survival model (case 5). For that model, let $\mathcal{P}$ be the permutation matrix that sorts the columns of $\boldsymbol{\Sigma}$ by failure time and regressor within failure time. Then the matrix $\mathcal{P}^T\boldsymbol{\Sigma}\mathcal{P}$ is block diagonal, with ith block the covariance matrix of $x_1, \ldots x_p$ at time $t_{(i)}$. Hence $\mathcal{P}^T\boldsymbol{\Sigma}\mathcal{P}$ is nonnegative definite and it follows that $\boldsymbol{\Sigma}$ is also nonnegative definite.

These results have the following practical implications. If the matrix $(\mathbf{N}^T\boldsymbol{\Sigma}\mathbf{N} + \boldsymbol{\Omega})$ is of full rank, the solution to (24) is unique and the Gauss-Seidel method converges to that solution for any set of starting values. If $(\mathcal{N}^T\boldsymbol{\Sigma}\mathcal{N} + \boldsymbol{\Omega})$ is not of full rank, the final iterate depends on the starting values.

It is interesting to examine the null space of matrix $\mathbf{U} = \mathcal{N}^T\boldsymbol{\Sigma}\mathcal{N} + \boldsymbol{\Omega}$, that is,

$$\text{Null}(\mathbf{U}) = \{\boldsymbol{\gamma} : (\mathcal{N}^T\boldsymbol{\Sigma}\mathcal{N} + \boldsymbol{\Omega})\boldsymbol{\gamma} = 0\}$$

For the additive model (case 1 above), Buja *et al.*(1989) call Null($\mathbf{U}$) the *concurvity space* (expressed in terms of $\boldsymbol{\beta} = \mathbf{N}\boldsymbol{\gamma}$). In general, elements of Null($\mathbf{U}$) must satisfy

$$\boldsymbol{\Omega}\boldsymbol{\gamma} = 0, \quad \boldsymbol{\Sigma}\mathbf{N}\boldsymbol{\gamma} = 0$$

The first equation, which by construction is the same as $\boldsymbol{\Omega}_j\boldsymbol{\gamma}_j = 0\ \forall j$, defines the null space of the penalty functionals, and for the second derivative penalties these correspond to functions $\beta_1, \ldots, \beta_p$ that are linear in their arguments. The interpretation of the second equation is simplest for the cases 1 and 3 listed above, and implies $\sum \boldsymbol{\beta}_j = 0$. So in this case the concurvity space corresponds to exact collinearities among the regressors. For the cases 2 and 4, a similar situation occurs: the β_js are linear and $\sum \mathbf{D}_j\boldsymbol{\beta}_j = 0$, which also means that the model matrix (8) in section 3.2 will not have full rank. In the general case the concurvity space is more difficult to interpret, but irrespective the β_j are linear and so the interpretation is the same as for the corresponding (bi-)linear model.