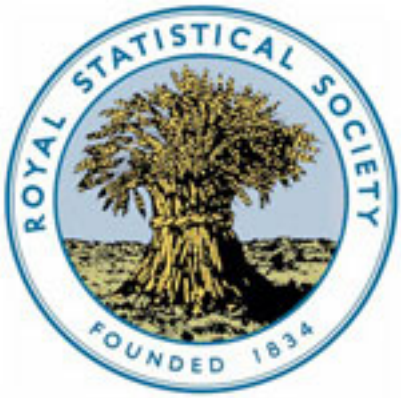

J. R. Statist. Soc. B (1993)
55, *No.* 4, *pp.* 757–796

Varying-coefficient Models

By TREVOR HASTIE and ROBERT TIBSHIRANI†

AT&T Bell Laboratories, Murray Hill, USA *University of Toronto, Canada*

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section on *Wednesday, February 10th, 1993,* Professor B. W. Silverman *in the Chair*]

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 flexibly and some 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: DYNAMIC GENERALIZED LINEAR MODELS; GENERALIZED ADDITIVE MODELS; GENERALIZED LINEAR MODELS; REGRESSION; SMOOTHING; SPLINES; 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 that 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 $\beta_1(\,), \ldots, \beta_p(\,)$. The dependence of $\beta_j(\,)$ on R_j implies a special kind of interaction between each R_j and X_j. In some cases, the variables R_j are indistinguishable from the variables X_j; in other cases R_j might be a special variable such as 'time'.

A common setting for the application of these ideas is the class of generalized linear models (Nelder and Wedderburn, 1972; 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

† *Address for correspondence*: Department of Preventive Medicine and Biostatistics, McMurrich Building, University of Toronto, Toronto, M5S 1A8, Canada.

 0035-9246/93/55757

$\eta = g(\mu)$. 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

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

where $E(\epsilon) = 0$, $\text{var}(\epsilon) = \sigma^2$. Other common models are log-linear 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 shall see that the generalized additive model is a special case of the varying-coefficient model, as is the *dynamic generalized linear model* (West *et al.*, 1985). However, the approach to inference is quite different in the latter model.

1.1. *Illustration*

To motivate the presentation, we begin with an example. Cleveland *et al.* (1991) examine 88 observations on the exhaust from an engine fuelled by ethanol. The

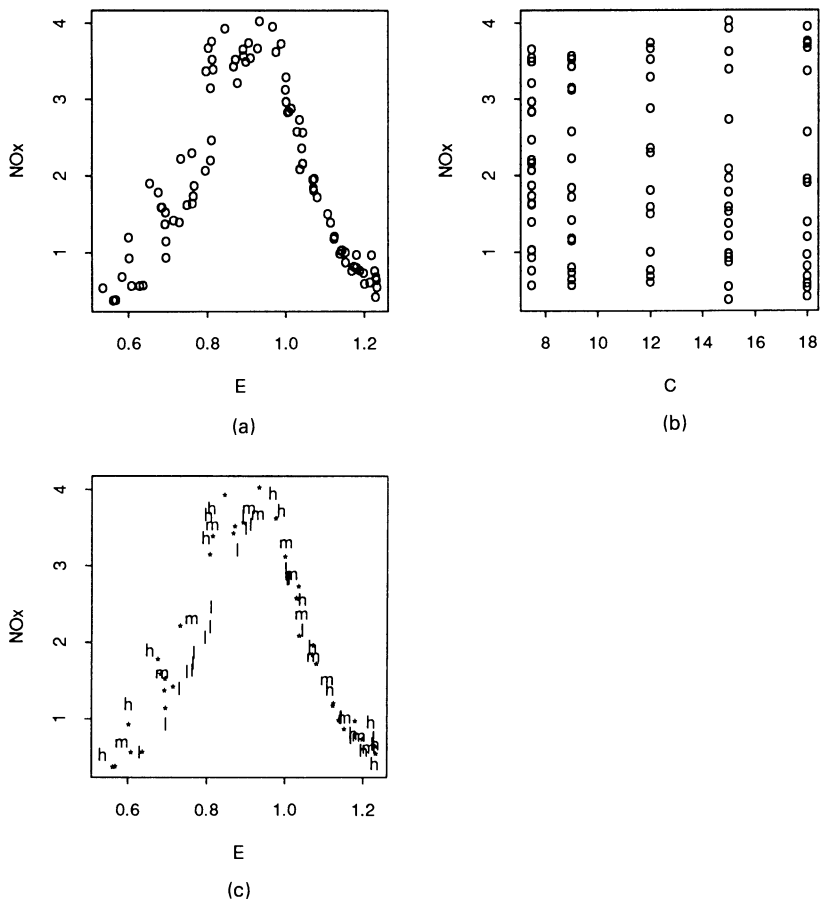

Fig. 1. (a) NO_x *versus* E; (b) NO_x *versus* C; (c) NO_x *versus* E with some values of C coded as low (l), medium (m) or high (h) (intermediate values are coded with *)

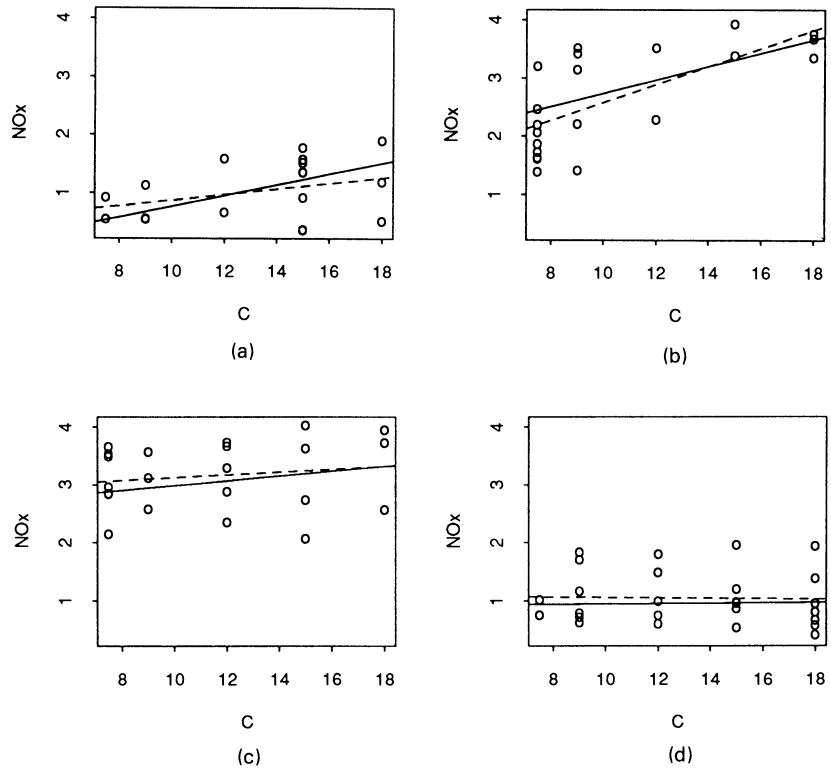

Fig. 2. NO_x *versus* C for (a) low, (b) medium, (c) high and (d) very high values of E: ———, fitted lines from the varying-coefficient model, taken at the median value of E for the data in the panel; - - -, fitted linear regression

response variable, denoted by NO_x is the concentration of nitric oxide and nitrogen dioxide, normalized by the workload of the engine. The two predictors are the equivalence ratio E, a measure of the fuel–air mixture, and the compression ratio C of the engine. The data are presented in Fig. 1. Figs 1(a) and 1(b) 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 $NO_x \sim E^2$. Fig. 1(c) shows NO_x *versus* E, with the levels of C coded as low, medium and high, and suggests that C might be interacting with E. Fig. 2 reveals the form of this interaction. The broken lines show the fitted linear regressions of NO_x on C in four non-overlapping 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 the slope of the line vary.

This leads us to consider a model of the form

$$NO_x = \beta_0(E) + \beta_1(E)C + \epsilon. \tag{3}$$

Whereas the plots suggest $\beta_0(E) \sim E^2$, we shall leave both $\beta_0(E)$ and $\beta_1(E)$ unspecified and fit them flexibly. This is the model considered by Cleveland *et al.* (1991) and is an example of a varying-coefficient model.

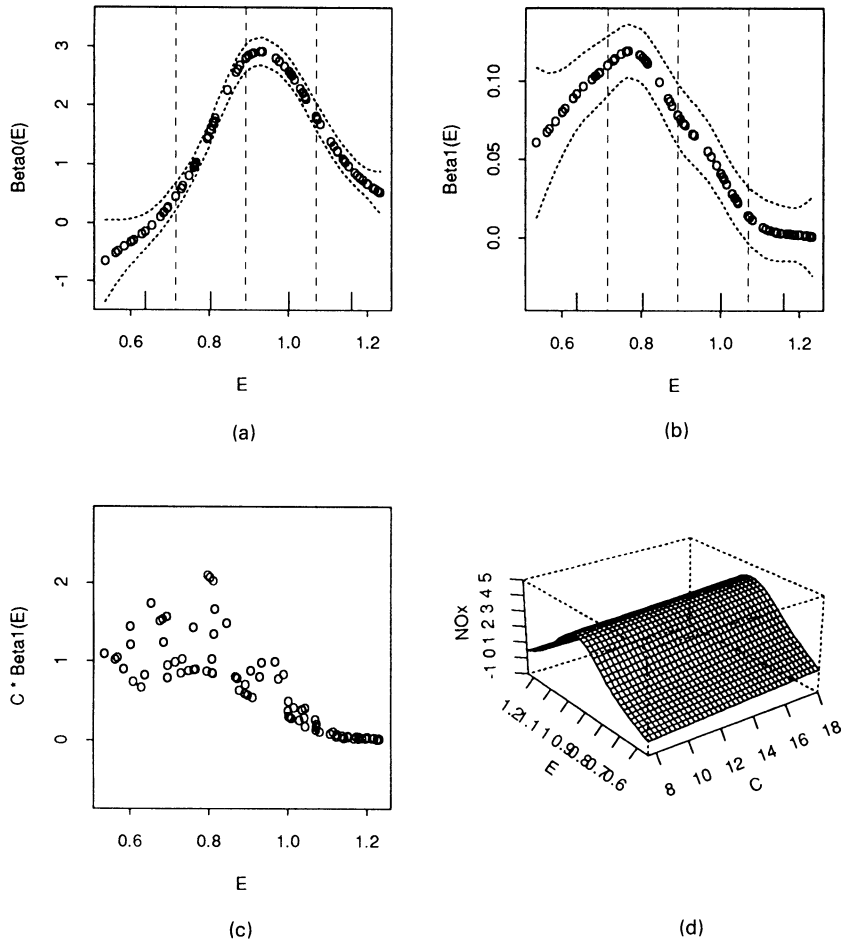

Fig. 3. (a) $\hat{\beta}_0(E)$ and (b) $\hat{\beta}_1(E)$ with pointwise 90% standard error bands defined by ± 1.645 times an estimate of the standard error (--------, regions of data used to construct the four conditioned plots of Fig. 2; the four spikes at the base of the plots show the four values of E used in representing the fit in Fig. 2); (c) fitted effects $C\hat{\beta}_1(E)$ on the same scale as the plot for $\hat{\beta}_0(E)$ (the latter is clearly more important); (d) perspective plot of the fitted (E, C)-surface

Using an estimation procedure described in this paper, we produced the curves $\hat{\beta}_0(E)$ and $\hat{\beta}_1(E)$ shown in Figs 3(a) and 3(b) (with 90% pointwise standard error bands); these correspond to the full lines in Fig. 2.

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 Fig. 3(d), and as shown by Cleveland *et al.* (1991) it fits the data as well as a general two-dimensional surface smoother, using fewer degrees of freedom.

Fig. 3(c) shows that the term $\hat{\beta}_1(E)C$ is not as important as $\hat{\beta}_0(E)$, and the perspective plot in Fig. 3(d) supports this claim.

TABLE 1
Analysis-of-variance table for a hierarchy of models for the NO_x data

	Model	*Residual sum of squares*	*Degrees of freedom*	*F-ratio*	*Pr(F)*
(i)	$NO_x = \beta_0(E) + \beta_1 C + \epsilon$	5.19	79	20.1	0.0
(ii)	$NO_x = \beta_0(E) + \beta_1 CE + \epsilon$	6.33	79	24.5	0.0
(iii)	$NO_x = \beta_0(E) + \beta_1 C + \beta_2 CE + \epsilon$	3.20	78	14.5	0.0
(iv)	$NO_x = \beta_0(E) + \beta_1(E) C + \epsilon$	2.65	72		

Some simpler models are suggested by the perspective plot and allow us to test for the interaction in model (3). Table 1 compares the various models with the conditionally parametric model (numbered (iv) in Table 1). 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 a non-linear form for the coefficient of C is necessary.

In this paper we discuss many aspects of the varying-coefficient model (1), showing connections to other published models and giving some general estimation procedures. The paper is organized as follows. In Section 2 we give examples of the varying-coefficient model in various 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 they 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 VARYING-COEFFICIENT MODEL

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

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

(b) 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 (a), then model (1) has the form of a *generalized additive model.*

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

(d) There is a different way of thinking about a model term $X\beta(R)$ when X is a binary (0–1) variable. Suppose that 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 to explore interactions with another quantitative variable in this way.

(e) 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 consist of repeated measurements of the variables $\{Y, X_1, \ldots, X_p\}$ over n time points $t \in (t_1, \ldots, t_n)$. Then we might model this as

$$\eta_t = \beta_0(t) + X_1(t)\beta_1(t) + \ldots + X_p(t)\beta_p(t) \tag{4}$$

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

(f) Suppose that 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).

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

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

As can be seen, there is 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 that is 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 probably not be possible except for special designs, and the

problems encountered seem no different from 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 trigonometric functions. Typically these do not provide enough flexibility and local adaptiveness, and a set of regression spline bases with a fixed arrangement of knots is likely to be preferable. We then proceed exactly as for the linear model, only with (many) more variables defined by the products of each X_j and the bases for $\beta_j(R_j)$. With this approach, all 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 well 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, we give details of estimation of the regression form (2) rather than the generalized regression form (1). Extensions to the latter case usually involve embedding 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 that are applicable to the special case of a single effect modifying variable as in model (4). One approach is local plane fitting, suggested by Cleveland *et al.* (1991); the other is the dynamic linear model (West *et al.*, 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 that we decide to estimate $\beta_1(\,) \ldots \beta_p(\,)$ in model (1) by minimizing

$$E\left\{Y-\sum_1^p X_j\beta_j(R_j)\right\}^2.$$

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

$$E\left[X_j\left\{Y-\sum_1^p X_j\beta_j(R_j)\right\}\Big|R_j\right]=0, \qquad j=1, 2,\ldots, p.$$

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

$$\beta_j(R_j)=\frac{E\left[X_j\left\{Y-\sum_{k\neq j} X_k\beta_k(R_k)\right\}\Big|R_j\right]}{E(X_j^2|R_j)}$$

$$= \frac{E\left[X_j^2\left\{Y - \sum_{k \neq j} X_k \beta_k(R_k)\right\} \Big/ X_j \mid R_j\right]}{E(X_j^2 \mid R_j)}. \tag{6}$$

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 β_j, $j = 1, \ldots, p$, 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 $\{Y - \Sigma_{k \neq j} X_k \beta_k(R_k)\}/X_j$ 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 that 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

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

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

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

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

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

Alternatively, terms of this type can be viewed as a particular element in an expansion of an interaction term by using tensor product splines; again see Wahba (1990), chapter 10. In theorem 3.1 of Wahba (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 criterion (7) are natural 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 when each β_j is restricted to be linear, then a unique solution exists in the unrestricted problem. In practice, this means that the model matrix

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

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 parameterize 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

$$\beta_j(r_{ij}) = \sum_{l=1}^{n_j} \gamma_{lj} N_j^l(r_{ij}).$$

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 $\boldsymbol{\gamma}_j$

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

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

The score equations are

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

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

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

where the matrix operator

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

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 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 $\mathbf{D}_j^{-} (\mathbf{y} - \Sigma_{k \neq j} \mathbf{D}_k \mathbf{N}_k \boldsymbol{\gamma}_k)$ with smoothers $\mathbf{S}_j(\lambda_j)$. Algorithms of this kind are known as 'backfitting' procedures (see Green *et al.* (1985), Breiman and Friedman (1985) and Buja *et al.* (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 $O\{(\Sigma n_j)^3\}$ to $O(\Sigma n_j)$.

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 *et al.* (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 0. 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 that these observations are given 0 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 equation (9), and to side-step 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, whereas the alternative 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 to 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 the use of other smoothers in place of the weighted cubic smoothing spline in the iterative procedure above. For example, we 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 Single Effect Modifying Variable*

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

$$Y = X_1 \beta_1(R) + \ldots + X_p \beta_p(R) + \epsilon. \tag{12}$$

This model is a special case of model (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 us look back at the first example and the model used there:

$$\mathrm{NO}_x = \beta_0(E) + \beta_1(E) C + \epsilon.$$

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 neighbourhood in the (E, C)-plane that is an infinite strip in the C-direction, and just sufficiently wide 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 that

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

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

For the moment let us focus on estimation in L_2, and therefore assume that $\mathbf{Z}$, R and Y are integrable random variables, and we wish to minimize

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

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

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

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 equation (13). Note that $\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$ is a 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 neighbourhoods 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.

(a) In addition to the restriction to the r-neighbourhoods, observations are assigned weights from a *tricube* weight function based on their distance from r (the closer to r the bigger the weight). In particular, the observation furthest from r receives almost 0 weight. This ensures that as we move r smoothly over the range of R the estimates change smoothly as well.

(b) R itself is used in the local hyperplane fit, and possibly bilinear terms are included between R and the other predictors $\mathbf{Z}$, i.e. the local model that is fitted might be

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

The motivation for the last item is concerned 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 or median of values of R in its neighbourhood. 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-hand boundary. For this reason, Cleveland (1979) and others proposed the fitting of 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 neighbourhoods. The required hyperplane is simply obtained by fixing the value of $R=r$ in expression (14).

How does the procedure described by equation (13) relate to that of Section 3.3? When Z is scalar, we have

$$\hat{\beta}(r)=\frac{E(ZY|r)}{E(Z^2|r)}$$

$$=\frac{E\{(Y/Z)Z^2|r\}}{E(Z^2|r)}.$$

This is estimated by smoothing Y/Z on R with weights Z^2, in agreement with the method proposed in Section 3.3.

3.4.2. *Bayesian perspective: 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), p. 108) is defined by the equations

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

where t refers to time. The first expression in equations (15) is called the *observation equation* whereas the second is 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 sequentially. Simple updating formulae 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 $G_t=1$ 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, $0\leqslant j\leqslant t-1$. 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 that of Cleveland *et al.* (1991) described in the previous section. Other related approaches have been published. 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 an 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. However, the Markov assumption in equations (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 straightforwardly 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 autoregressive moving average model for Y_t can depend on the previous state of the system Y_{t-1}. This would correspond to a varying-coefficient model in which Y played the roles of response, regressor *and* effect modifier. It is analogous to a locally parametric model for smoothing (see for example Cleveland (1979)).

4. SECOND 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 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 behaviour; we shall 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 limits the interpretations dramatically, and the interactions shown here are in fact a demonstration of the pitfalls of a naïve investigation of such data. A larger set of risk factors is described in Hastie and Tibshirani (1987).

Figs 4(a) and 4(b) show the fit of the generalized additive model

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

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.

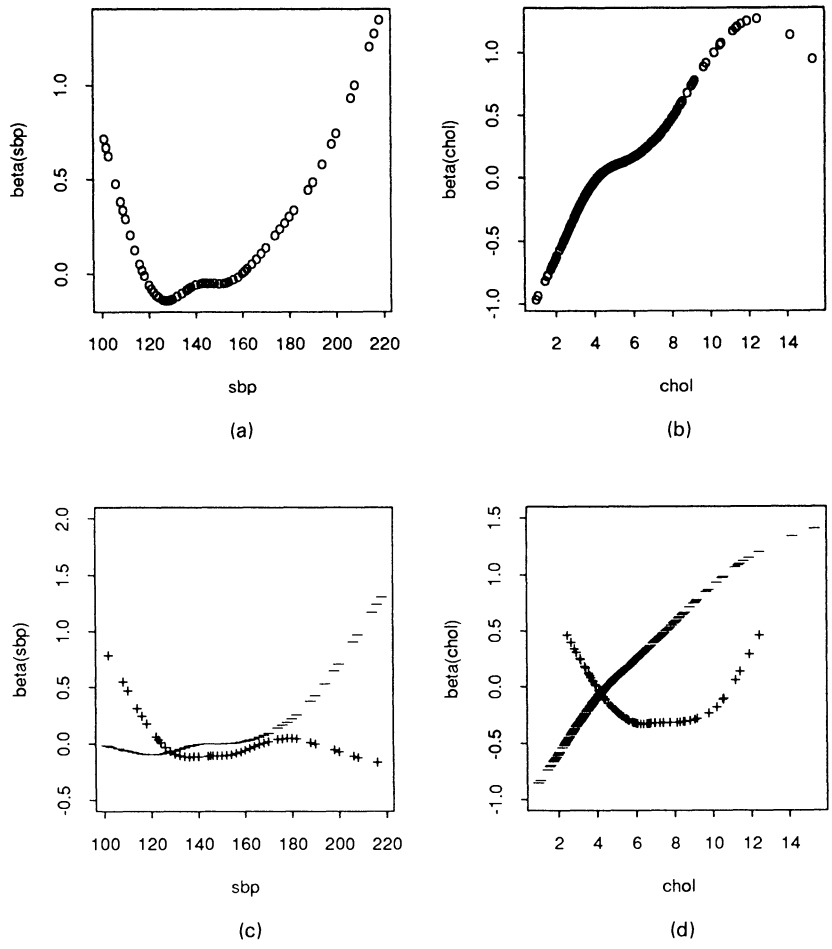

Fig. 4. (a) Fitted curve for S and (b) fitted curve for C in a generalized additive model, ignoring the treatment variable T; (c) pairs of functions $\hat{\beta}_{S0}(S)$ and $\hat{\beta}_{S0}(S)+\hat{\beta}_{S1}(S)$ for S ($-$, no treatment ($T=0$); $+$, treatment); (d) plot equivalent to (c) for C

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

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

The functions $\beta_{S1}(S)$ and $\beta_{C1}(C)$ represent the effect of the treatment on the curves for S and C. Figs 4(c) and 4(d) show the estimated functions $\hat{\beta}_{S0}(S)$ and $\hat{\beta}_{S0}(S)+\hat{\beta}_{S1}(S)$, and $\hat{\beta}_{C0}(C)$ and $\hat{\beta}_{C0}(C)+\hat{\beta}_{C1}(C)$ respectively.

As mentioned, the interpretation of these interactions is tricky, since the data were collected retrospectively and the treatment was given after the heart attack. 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 procedure described in the present paper and used in 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 $t_{(1)} < \ldots < t_{(d)}$, 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 base-line hazard function.

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

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

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

Unless each $\beta_j(t)$ is a constant, this model represents non-proportional hazards. Hence it 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 ($X = 0$ or $X = 1$), model (19) reduces to

$$\lambda(t \mid 0) = \lambda_0(t); \qquad \lambda(t \mid 1) = \lambda_0(t) \exp \beta(t).$$

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 base-line 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 $\beta(t)$ in contrast is modelled smoothly and measures the difference between the hazards for the two groups.

Goodness-of-fit and diagnostic tests for proportional hazards model (18) are difficult to develop because of the arbitrary base-line 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 smoothly.

The partial likelihood for model (19) is given by

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

In this expression, $\mathscr{R}_k$ 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 keeping with the approach of this paper, we focus on a penalized partial likelihood method.

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

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

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 equation (21) in which the squared mth derivative of β_j is penalized. Their results require $m \geqslant 3$ for consistency of the maximizer of the penalized log-partial-likelihood and $m \geqslant 4$ 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. The form of the score equations is interesting and has the appearance of a ridge regression style of modification to the usual partial likelihood score equations.

5.1. *Third Example: 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 2) indicates that performance status and cell type are highly related to survival, whereas treatment is marginally significant. The maximized value of the likelihood is -475.2.

We tried to add a term of the form $x\beta(t)$ for performance status to this model, choosing the target degrees of freedom to be 5. The resulting estimate $\hat{\beta}(t)$ is shown in Fig. 5(a) and indicates a decreasing then increasing advantage with time. The decrease

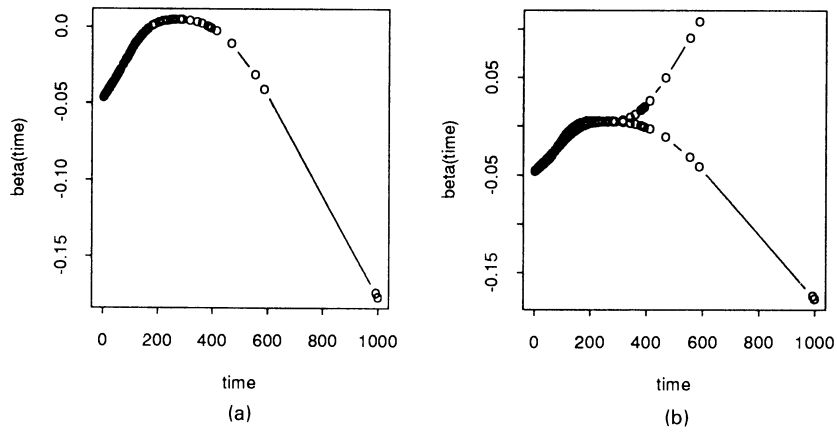

Fig. 5. Plots of $\hat{\beta}(t)$ for the third example: (a) with all the data the plot is markedly quadratic; (b) effect of removing the two values of time on the extreme right-hand side and refitting the model—the tail of the function increases rather than decreases

TABLE 2
Proportional hazards fit for the third example

Variable	*Coefficient*	*Standard error*
Treatment	0.289	0.207
Cell type—squamous	0.400	0.282
Cell type—small cell	0.457	0.266
Cell type—adeno	0.788	0.303
Months from diagnosis	-9.2×10^{-5}	0.009
Age	−0.009	0.009
Prior therapy	0.007	0.023
Performance status	−0.032	0.006

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 the log-likelihood ratio of 15.0 on about 4 degrees of freedom.

The 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.

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 obtain an idea of which features are estimated with precision.

For a broad class of smoothers, including kernels and nearest neighbours, as well as splines, we 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 $\nu = \mathrm{tr}(2\mathbf{S} - \mathbf{S}\mathbf{S}^{\mathrm{T}})$ 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 $\mathbf{D}\mathbf{N}(\mathbf{N}^{\mathrm{T}}\mathbf{D}^2\mathbf{N} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{N}\mathbf{D}$ (in the notation of Section 3.3). It can be seen that

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

where $\mathbf{S}' = \mathbf{N}(\mathbf{N}^{\mathrm{T}}\mathbf{D}^2\mathbf{N} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{N}\mathbf{D}^2$ 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.

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

Finally, predictions from the model outside the range of the observed R_1, R_2, . . ., R_p will sometimes be of interest. The straightforward approach would be to extrapolate each estimate $\hat{\beta}_j(R_j)$ linearly (see for example Hastie and Tibshirani (1990), chapter 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 probabilistic structure is likely to be more useful.

7. DISCUSSION

The varying-coefficient model described in this paper is 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 in which 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.* (1991) automatically allow for this case when all the terms are modelled conditionally on the same $\mathbf{R}$.

Extensions to non-linear regression models could also be explored. Another extension (suggested by Mike Leblanc) would be to allow terms of the form $X\beta(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{r})$ 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 large changes in the coefficients.

ACKNOWLEDGEMENTS

We thank Scott Emerson for providing us with 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 referees greatly improved the exposition. The second author gratefully acknowledges the support of the Natural Science and Engineering Research Council of Canada.

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 $\mathbf{n}_k=\mathbf{n}(t_k)$, $\Omega_{ij}=\int n_i''(t)\, n_j''(t)\, dt$. 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 $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$ are

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

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

$$p_{ki}=\frac{\exp\left(\sum_j x_{ij}\mathbf{n}_k^T\boldsymbol{\gamma}_j\right)}{\sum_{i\in\mathscr{R}_k}\exp\left(\sum_j x_{kj}\mathbf{n}_k^T\boldsymbol{\gamma}_j\right)}$$

and $\mu_{kj}=d_k\sum_{i\in\mathscr{R}_k}p_{ki}x_{ij}$. 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 $\mathscr{R}_k$, with respect to the model probabilities p_{ki}. Then the score equations can be expressed as

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

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 $\mathbf{n}_k^T=1, 2, \ldots, d$. Now if $\mathbf{N}$ consisted only of a column of 1s and the penalty $\lambda_j\boldsymbol{\Omega}\boldsymbol{\gamma}_j$ were absent, equation (22) would simply require that the sum over the risk sets of the 'residuals' $(\mathbf{s}_j-\boldsymbol{\mu}_j)$ be 0, the usual stationary conditions 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.

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 equation (22) we obtain the $pd\times pd$ system

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

where Σ_{jl} is a diagonal matrix with diagonal elements the covariances of (x_{ij}, x_{il}) in the risk set $\mathscr{R}_k$: $\Sigma_{jl}=\mathrm{diag}(\sigma_{jl1}, \ldots, \sigma_{jld})$ where

$$\sigma_{jlk}=d_k^2\left\{\sum_{i\in\mathscr{R}_k}p_{ki}x_{ij}x_{il}-\left(\sum_{i\in\mathscr{R}_k}p_{ki}x_{ij}\right)\left(\sum_{i\in\mathscr{R}_k}p_{ki}x_{il}\right)\right\}.$$

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:

(a) an inner backfitting loop that successively smooths $\mathbf{z}_j$ on the time variable, to obtain new values of $\hat{\boldsymbol{\beta}}_j$;
(b) an outer loop that updates $\mathbf{z}_j$, $j = 1, \ldots, p$, by using the current values of $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p$.

Step (a) solves the linear system (23) in $O(kn)$ computations, where k is the number of iterations (typically fewer than 10). Step (a) 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 *et al.* (1989) can still be applied to establish convergence. Details are given in Appendix B.

APPENDIX B

In this appendix we study the system of equations

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

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 $\int N_i''(r)N_k''(r)\,dr$, Σ_{ij} are all submatrices of the $\Sigma n_j \times \Sigma n_j$ symmetric non-negative definite matrix Σ and $\Delta_j = \boldsymbol{\gamma}_j^{\text{new}} - \boldsymbol{\gamma}_j^{\text{old}}$. 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:

(a) if $\mathbf{u}_j = \mathbf{y}$, $n_j = n$ for all j, $\Sigma_{ij} = \mathbf{I}$ (the identity matrix) for all i, j, and $\boldsymbol{\gamma}_j^{\text{old}} = 0$, then equations (24) are the score equations for an additive model fitted by cubic smoothing splines (this is described in Buja *et al.* (1989));
(b) if $\mathbf{u}_j = \mathbf{u}$ is the derivative of the log-likelihood and $\Sigma_{ij} = \mathbf{A}$, the matrix with the Fisher information components on its diagonal, then system (24) represents the Newton–Raphson update from $\boldsymbol{\gamma}_j^{\text{old}}$ to $\boldsymbol{\gamma}_j^{\text{new}}$ for a generalized additive model (this is the 'local scoring' procedure of Hastie and Tibshirani (1990), chapter 6; see also Green (1987));
(c) as in (a) but with $\Sigma_{ij} = \mathbf{D}_i\mathbf{D}_j$ ($\mathbf{D}_i$ is defined in Section 3.3), and $\mathbf{u}_i = \mathbf{D}_i\mathbf{y}$—then system (24) represents the score equations for the varying-coefficient model described in Section 3.2;
(d) as in (b), but with $\Sigma_{ij} = \mathbf{D}_i\mathbf{A}\mathbf{D}_j$—then system (24) represents the Newton–Raphson update from $\boldsymbol{\gamma}_j^{\text{old}}$ to $\boldsymbol{\gamma}_j^{\text{new}}$ for the varying-coefficient model in the generalized regression setting;

(e) if $\mathbf{N}_j = \mathbf{N}$, $j = 1, \ldots, p$, the evaluated natural spline bases for the unique failure times, and Σ_{ij} are the risk set covariances defined in Section 5, then system (24) represents the Newton–Raphson update from $\boldsymbol{\gamma}_j^{\text{old}}$ to $\boldsymbol{\gamma}_j^{\text{new}}$ for the survival model of Section 5;

(f) 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 Σ_{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 non-diagonal matrices Σ_{ij}.

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

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

where $\boldsymbol{\beta}_j = \mathbf{N}_j\boldsymbol{\gamma}_j$,

$$\mathbf{z}_j = \Sigma_{jj}^{-}\mathbf{y}_j + \Sigma_{jj}^{-}\left(\sum_{i=1}^{p}\Sigma_{ji}\boldsymbol{\beta}_j^{\text{old}}\right)$$

and $\mathbf{S}_j = \mathbf{N}_j(\mathbf{N}_j^{\mathrm{T}}\Sigma_{jj}\mathbf{N}_j + \lambda_j\boldsymbol{\Omega}_j)^{-1}\mathbf{N}_j^{\mathrm{T}}\Sigma_{jj}$, a weighted smoothing spline 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:

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

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

Theorem. If Σ is symmetric and non-negative definite, then the system of equations (24) (or system (25)) is consistent and the Gauss–Seidel iteration (26) converges to a solution for any starting values.

Proof: consistency. Let $z^* = \Sigma^{-}\mathbf{z}$ where $\mathbf{z} = (z_1, \ldots, z_p)^{\mathrm{T}}$. Then the solutions to equations (25) can be expressed as the minimizers of the quadratic function

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

where $\boldsymbol{\Omega} = \operatorname{diag}(\Omega_1, \ldots, \Omega_p)$ and $\mathbf{N} = (N_1, \ldots, N_p)$. Since both Σ and $\boldsymbol{\Omega}$ are non-negative definite, expression (27) is non-negative. A real-valued multivariate function bounded below must have a minimum and therefore system (25) is consistent.

Proof: convergence of Gauss–Seidel iteration. (Typical convergence results for the Gauss–Seidel procedure require that the system matrix be symmetric and positive definite. The left-hand matrix in system (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) who proved the convergence of the Gauss–Seidel procedure through a 'seminorm descent principle'. We make use of this principle here to establish convergence.)

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

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

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

Lemma: seminorm descent principle (Buja *et al.*, 1989). If $|\mathbf{f}|$ is a complex seminorm and $\mathbf{T}$ a linear mapping on $\mathcal{E}^{np}$ 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 system (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 $Q(\ )$ along the jth co-ordinate, 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 the theorem we require that Σ be symmetric and non-negative definite. This is immediate in all the cases listed at the beginning of the appendix, except the survival model (case (e)). For that model, let $\mathcal{P}$ be the permutation matrix that sorts the columns of Σ by failure time and regressor within failure time. Then the matrix $\mathcal{P}^T \Sigma \mathcal{P}$ is block diagonal, with ith block the covariance matrix of $x, \ldots, x_p$ at time $t_{(i)}$. Hence $\mathcal{P}^T \Sigma \mathcal{P}$ is non-negative definite and it follows that Σ is also non-negative definite.

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

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

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

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

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

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 (a) and (c) listed above, and implies that $\Sigma \boldsymbol{\beta}_j = 0$. So in this case the concurvity space corresponds to exact collinearities among the regressors. For cases (b) and (d), a similar situation occurs: the β_js are linear and $\Sigma \mathbf{D}_j \boldsymbol{\beta}_j = 0$, 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 irrespectively the β_j are linear and so the interpretation is the same as for the corresponding (bi)linear model.

REFERENCES

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

Breiman, L. and Friedman, J. (1985) Estimating optimal transformations for regression. *J. Am. Statist. Ass.*, **80**, 580–619.

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

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

Cleveland, W. S. (1979) Robust locally-weighted regression and smoothing scatterplots. *J. Am. Statist. Ass.*, **74**, 829–836.
Cleveland, W. S. and Devlin, S. (1988) Locally-weighted regression: an approach to regression analysis by local fitting. *J. Am. Statist. Ass.*, **83**, 597–610.
Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991) Local regression models. In *Statistical Models in S* (eds J. M. Chambers and T. Hastie). Pacific Grove: Wadsworth and Brooks/Cole.
Cox, D. R. and Oakes, D. (1984) *Analysis of Survival Data*. London: Chapman and Hall.
Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Am. Statist. Ass.*, **76**, 817–823.
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**, 245–260.
Green, P., Jennison, C. and Seheult, A. (1985) Analysis of field experiments by least squares smoothing. *J. R. Statist. Soc.* B, **47**, 299–315.
Hastie, T. and Tibshirani, R. (1987) Generalized additive models: some applications. *J. Am. Statist. Ass.*, **82**, 371–386.
——(1990) *Generalized Additive Models*. London: Chapman and Hall.
Kalaba, R. and Tesfatsion, L. (1989) Time-varying linear regression via flexible least squares. *Comput. Math. Applic.*, **17**, 125–145.
Kalbfleisch, J. D. and Prentice, R. L. (1980) *The Statistical Analysis of Failure Time Data*. New York: Wiley.
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*, 2nd edn. London: Chapman and Hall.
Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.
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 Ser. Anal.*, **1**, 47–71.
Silverman, B. W. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. R. Statist. Soc.* B, **47**, 1–52.
Stone, C. J. (1977) Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–645.
Wahba, G. (1990) *Spline Models for Observing Data*. Philadelphia: Society for Industrial and Applied Mathematics.
West, M. and Harrison, P. J. (1989) *Bayesian Forecasting and Dynamic Models*. New York: Springer.
West, M., Harrison, P. J. and Migon, H. S. (1985) Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Am. Statist. Ass.*, **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.

DISCUSSION OF THE PAPER BY HASTIE AND TIBSHIRANI

P. J. Green (University of Bristol): There is nothing very hard conceptually in inventing methodology for nonparametric multiple regression, but to do this usefully means insisting on some structure in the modelling. That structure can build on substantive theory and the elicitation of prior belief, it automatically controls the expenditure of degrees of freedom, and its exploitation facilitates computation, presentation and interpretation. In a long series of papers running forwards from 1984 and in their book, the authors have made many contributions to good structured nonparametric regression methodology, much of it in the framework of generalized additive models. It is a great pleasure to welcome the authors' latest work on this theme to the Society's proceedings.

Nonparametric regression does not always mean smoothing, but when it does there is a rich variety of basic methods to choose from. These fall into two groups: those specified operationally, by the algorithms by which they are implemented—running means, kernel methods, etc.—and those defined

by the objective function that they optimize. Methods in the former group are perhaps more easily understood, and their performance in simple situations more readily assessed, but the second group of methods has other attractions. Such methods, based on penalized least squares or likelihood, are unambiguously characterized, without reference to algorithmic details, and are straightforwardly adapted to novel situations to build more flexible methodology. The resulting smoothing operators, being symmetric and non-negative definite, have convenient theoretical properties, that account for the large part they play in the present paper.

Varying-coefficient models, in the form of equations (1) or (2), are examples of structured non-parametric regression and thus offer all the benefits that I mention above. The examples in the paper are quite convincing examples of the application of such models. However, they use only very special cases of models (1) and (2), and my main criticism of this class as a whole is that it emphasizes coefficients rather than fitted values. One consequence is a lack of invariance to transformations of the variables. Invariance to rotations (of the space of Xs, for example) is not always important, but it seems to me that effective invariance to translations is usually needed. The model

$$Y=\beta_1(R_1)+\beta_2(R_2)X+\epsilon,$$

in which the intercept and slope of regression on X have different effect modifiers R_1 and R_2, can only make sense when the origin for X is physically meaningful. Otherwise, I would want to allow a shift in this origin, leading to

$$\begin{aligned} Y&=\beta_1(R_1)+\beta_2(R_2)(X+\gamma)+\epsilon \\ &=\{\beta_1(R_1)+\gamma\beta_2(R_2)\}+\beta_2(R_2)X+\epsilon, \end{aligned}$$

a model of different structure.

Having done so, procedures familiar to the authors can be used to fit the model. Using backfitting with a three-part cycle, the estimates of the curves $\beta_1(R_1)$ and $\beta_2(R_2)$ are updated by weighted cubic spline smoothing, while γ is updated using a calculation by regression through the origin. This is no longer an additive model and perhaps in consequence convergence is slow. As applied to an artificial data set (see Fig. 6) contrived to have structure similar to that of the example in Section 1.1, the resulting fitted curves β_1 and β_2 are displayed in Figs 7(a) and 7(b), superimposed on the dotted curves representing the true values used in the simulation. For comparison, Figs 7(c) and 7(d) are the corresponding results with γ fixed at 0: the form of β_2 is completely lost.

The moral that I draw from this is that some considerable discipline is needed in writing down varying-coefficient models. I would welcome further work which will produce exploratory and diagnostic methods for selecting such models.

Turning now to purely computational matters, I hope that Sections 3.2 and 3.3 will not be read as implying that backfitting is a universal solution. As Appendix B shows, full rank of the matrix in equation (8) is sufficient to ensure mathematical convergence of backfitting, but unfortunately this does not guarantee convergence in practical terms. Pathological counter-examples can always be constructed, but problems can occur with genuine data sets too. A reanalysis of the price–volume study described in Daniel and Wood (1980) using, very naturally, a partially linear model in which price, price differential and day of week appear linearly, and date nonparametrically, is a case in point. The spectral radius of the convergence matrix for the iteration turns out to be 0.9992 for this example, using an appropriate degree of smoothing for the cubic spline, and a backfitting alternation between date and the linear variables. This is impossibly slow for any practical purpose. Consideration of equation (8), or even of the corresponding iteration in which date is fitted linearly, provides no warning: the spectral radius here is only 0.8838. With quadratic dependence on date, this rises to 0.9815, and alarms begin to ring.

General purpose alternatives to backfitting, or perhaps modifications that increase orthogonality, would be desirable. A modest step in this direction is the observation that the non-iterative method in Green *et al.* (1985) generalizes somewhat to certain other varying-coefficient models. Consider model (2), in which only one of the βs is not constant: it is convenient to rewrite this as

$$Y=x^{\mathrm{T}}\beta+z\,g(r)+\epsilon.$$

Then the appropriate penalized sum of squares

$$\|\mathbf{Y}-\mathbf{X}\beta-\mathbf{Z}g\|^2+\lambda\int g''(r)^2\,\mathrm{d}r,$$

(where $\mathbf{Z}$ is the diagonal matrix of zs), is minimized by

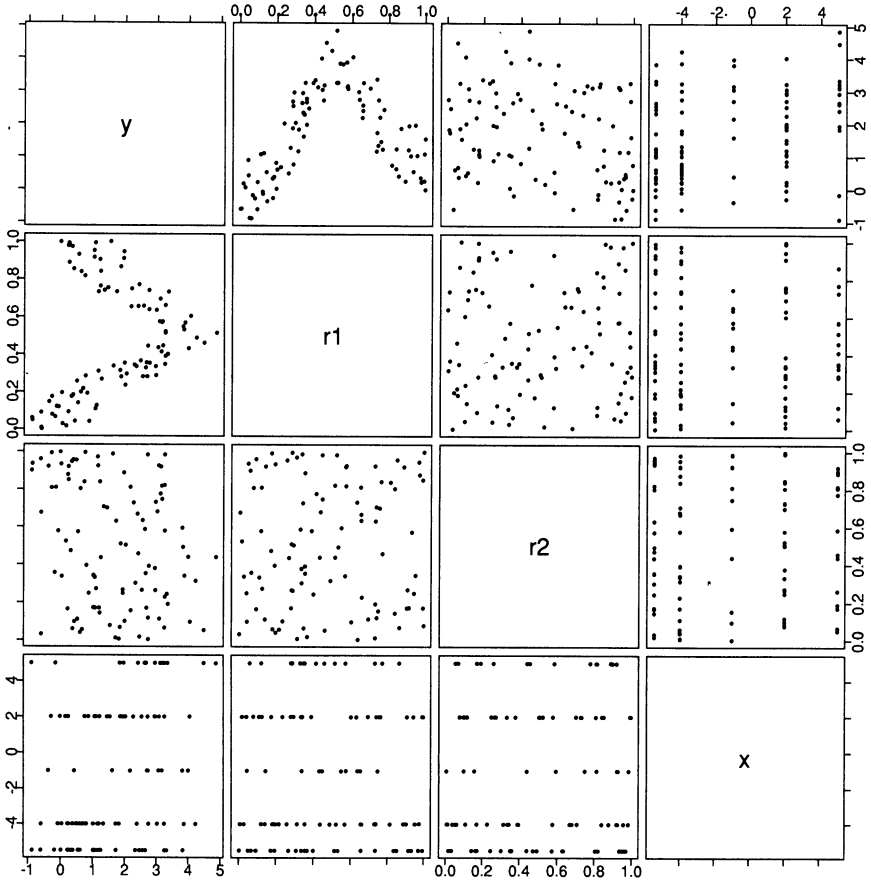

Fig. 6. Scatterplot matrix for an artificial data set

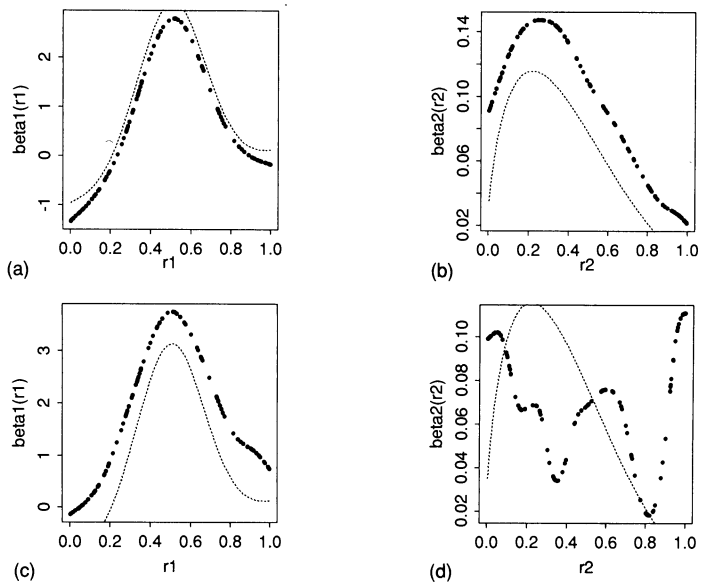

Fig. 7. (a), (b) Estimated regression curves $\beta_1(R_1)$ and $\beta_2(R_2)$ respectively, when γ is also estimated, for the data in Fig. 6; (c), (d) as for (a) and (b) but with γ fixed at 0, using the authors' analysis

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}(\mathbf{I}-\mathbf{S}^*)\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{I}-\mathbf{S}^*)\mathbf{Y},$$
$$\hat{g} = \mathbf{S}\mathbf{Z}^{-1}(\mathbf{Y}-\mathbf{X}\hat{\beta})$$

where

$$\mathbf{S}^* = \mathbf{Z}\mathbf{S}\mathbf{Z}^{-1}$$

and $\mathbf{S}$ is the usual weighted cubic spline smoother. This solution can be computed, without iteration, in $O(n)$ steps for fixed p.

I hope that it is evident that I have been greatly interested by this paper, and I have much pleasure in proposing a warm vote of thanks to the authors.

J. Cuzick (Imperial Cancer Research Fund, London): What we have been offered here is a family of models for studying interactions. Additive models offer an important extension of multivariate linear regression for studying main effects, and the present approach offers a structured but flexible approach for interactions which avoids the abyss of general nonparametric multivariate models. However, the choice of what structure to impose is much wider here. The most natural generalization of additive models are models of the form

$$E(y \mid x_1, \ldots, x_k) = \alpha + \sum_{i=1}^{k} f_i(x_i) + \sum_{1 \leqslant i < j \leqslant k} g_{ij}(x_i, x_j)$$

where $Ef_i(X_i) = E\{g_{ij}(X_i, X_j) \mid X_j\} = 0$ for all $1 \leqslant i < j \leqslant k$. This is already problematic because of the difficulties in fitting two-dimensional functions nonparametrically. An alternative is to replace $g_{ij}(x_i, x_j)$ by $g_{ij}(x_i)\, h_{ij}(x_j)$ leading to the need to fit k^2 one-dimensional functions, compared with the k required in additive regression. The authors stop short of this and propose models in which only one factor of a quadratic form is nonparametric. Theoretical analysis of their model is not particularly enlightening and its utility can only be judged by its value in specific applications.

One area where these sorts of model have application is survival analysis. Although the proportional hazards model has become standard for these problems, in very many cases it is found that differences between hazards attributed to covariates tend to converge as follow-up time elapses. This can be accounted for by postulating the existence of unobserved covariates or frailties with prognostic value. Thus the survival model can be written as

$$\lambda(t \mid z, \xi) = \lambda_0(t) \exp(\beta z)\, \xi$$

where z are the observed covariates with coefficients β and ξ is an unobserved frailty or random effect. If ξ is assumed independent of z and is normalized so that $E(\xi) = 1$, then the observed marginal hazard takes the form

$$\lambda(t \mid z) = \lambda_0(t) \exp(\beta z)\, \gamma(t, z)$$

where

$$\gamma(t, z) \equiv \gamma\{\theta\, \Lambda_0(t)\} = E(\xi \mid z, T \geqslant t)$$
$$= \frac{E[\, \xi \exp\{-\Lambda_0(t)\, \theta \xi\}\,]}{E[\, \exp\{-\Lambda_0(t)\, \theta \xi\}\,]}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)\, \mathrm{d}s$, $\theta = \exp(\beta z)$ and T is the survival time. Taking another derivative of $\gamma(t, z)$ with respect to t yields

$$\frac{\mathrm{d}\gamma}{\mathrm{d}t} = -\lambda_0(t)\, \theta\, \mathrm{var}(\xi \mid z, T \geqslant t) \leqslant 0,$$

so that an independent frailty always leads to converging hazards. This leads to a specialization related to the authors' form (12) where only one interaction function is introduced, but it now depends on $\theta\, \Lambda_0(t)$. When the range of variation of θ is small compared with $\Lambda_0(t)$, this might usefully be approximated by a function of t alone.

If a measurement error model is considered,

$$z_i = z_i^{\mathrm{true}} + \epsilon_i, \qquad \epsilon_i \text{ independent}, \ i = 1, \ldots, k,$$

with $\lambda(t \mid z^{\text{true}}) = \lambda_0(t) \exp(\beta z^{\text{true}})$ then

$$\lambda(t \mid z) = \lambda_0(t) \exp\{\beta_1(t, z_1) z_1 + \ldots + \beta_k(t, z_k) z_k\}$$

where

$$\beta_i(t, z_i) = z_i^{-1} \log E\{\exp(\beta_i z_i^{\text{true}}) \mid z_i, T \geqslant t\}$$

which is more like model (4) or (12).

When frailty interacts with treatment, models of the form

$$\lambda(t \mid z, \xi) = \lambda_0(t) \exp(\beta z \xi)$$

are more appropriate and lead to crossing or diverging hazards (Cuzick and Trejo, 1992). When covariates exert their effects at different times of the disease process, then additive hazard models (Aalen, 1980; Huffer and McKeague, 1991; McKeague and Sasieni, 1993), i.e.

$$\lambda(t \mid z) = \lambda_0(t) + \sum_{i=1}^{k} \lambda_i(t) \exp(\beta_i z_i)$$

are probably more relevant, since factors affecting different mechanisms are more likely to add and not to multiply on the hazard scale.

A perplexing example where flexible models of interaction are needed is the modelling of serum β_2-microglobulin, which is the major prognostic factor for myelomatosis. Previous analyses (Cuzick *et al.*, 1990) have shown that its prognostic value diminishes rapidly with follow-up time and have also suggested that the relative risk function is not log-linear. Today's paper suggests that we try flexible interaction models and an interesting possibility is

$$\lambda(t \mid z) = \lambda_0(t) \exp\{g(z)\, \beta(t)\}, \qquad \beta(0) = 1,\ g(0) = 0$$

where $g(z)$ and $\beta(t)$ are modelled nonparametrically. Preliminary estimates for g and β using regression splines on 1072 patients in the Medical Research Council's fourth and fifth trials are shown in Fig. 8. A steady decline in prognostic value is seen for the first 4 years of follow-up, at which time all prognostic value has disappeared. After that time the curve became unstable, which was also true in the author's Fig. 5. The log(relative hazard) for serum β_2-microglobulin appears to have a flat spot in the middle of its range after a preliminary logarithmic transformation. This can be explained by its dependence on two separate mechanisms—myeloma cell tumour and renal failure. However, the detailed shape of the curve is dependent on the location of the knots for both z and t, and use of smoothing splines as suggested by the authors may help to stabilize things.

The choice of models for interaction is vast and knowledge of the subject-matter at hand can help to focus on the appropriate class for particular application. The models discussed in this paper are best considered as illustrative examples of what might be attempted. Nevertheless, the need for structured but flexible models of interaction is great and the authors do us a great service by bringing the issues to our attention and by providing a general method for fitting them. I am very pleased to second the vote of thanks.

The vote of thanks was passed by acclamation.

R. A. Rigby and D. M. Stasinopoulos (University of North London): We wish to draw attention to an important subclass of varying-coefficient models: 'break point' models where a sudden change occurs in the dependence of η on one of the predictor variables. Specifically in model (1), a $\beta_j(R)$ can be modelled by

$$\beta_j(R) = \beta_{j1}(R)\, \text{IF}(R < \gamma) + \beta_{j2}(R)\, \text{IF}(R \geqslant \gamma) \tag{28}$$

where β_{j1} and β_{j2} may be parametric or nonparametric functions in R, and the value γ of R at which the break point occurs may be unknown and need to be estimated. Models with break points are, generally, not smooth and consequently may provide *better* fits than smooth functions.

Consider two specific cases.

(a) If $X = 1$ then equation (28) models a break in the dependence of η on R, using piecewise parametric (e.g. piecewise polynomial; Stasinopoulos and Rigby (1992)) or piecewise nonparametric models in R.

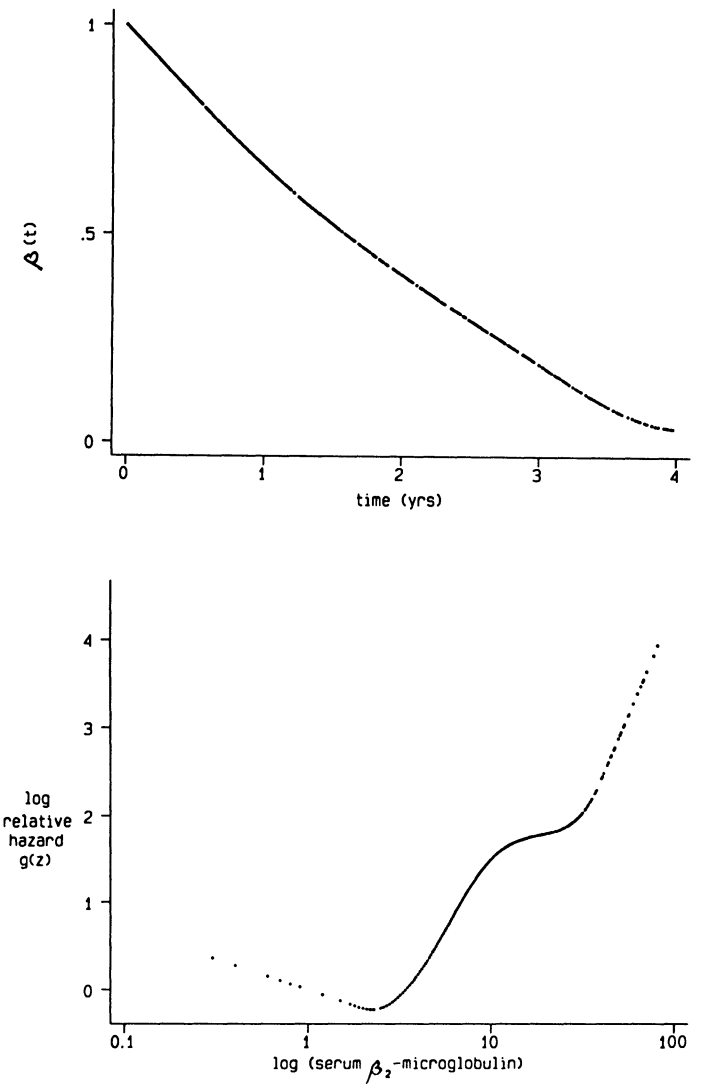

Fig. 8. Plots of $\hat{\beta}(t)$ and $\hat{g}(z)$ estimated from the myelomatosis data using the model $\lambda(t \mid z) = \lambda_0(t) \exp\{g(z)\,\beta(t)\}$

(b) If both X and R are variables, then equation (28) models a break, at $R = \gamma$, in the dependence of η on the interaction between X and R.

In the authors' example in Section 1.1, we were interested in whether there was a break point in the dependence of coefficient $\beta_1(E)$ on E. We fitted

$$\mathrm{NO}_x = \beta_0(E) + \beta_1 \,\mathrm{IF}(E < \gamma)\,(C - \bar{C}) + \epsilon, \tag{29}$$

$$\mathrm{NO}_x = \beta_0(E) + \{\beta_{10} + \beta_{11} E + \beta_{12}(E - \gamma)\,\mathrm{IF}(E > \gamma)\}(C - \bar{C}) + \epsilon \tag{30}$$

where $\beta_0(E)$ is a smoothing function with 8 degrees of freedom and $\bar{C}$ is the mean of C.

Hence model (29) uses a piecewise constant model for $\beta_1(E)$, and implies that C has no effect on NO_x if E is greater than the break point γ, and was suggested by the exploratory plots such as Fig. 1(c). Model (30) uses a piecewise linear model for $\beta_1(E)$, with continuity at the break point, and

was suggested by the smooth $\beta_1(E)$ function in Fig. 3(b). The fitted break points in models (29) and (30) were obtained from their profile deviance (e.g. Fig. 9 for model (29)). The fitted models for $\beta_1(E)$ corresponding to models (29), (30) and the linear model (iii) from Table 1 (with their residual sum of squares RSS and degrees of freedom DF) were

$$\beta_1(E) = 0.10 \text{ IF}(E<1), \quad \text{RSS} = 2.84, \text{ DF} = 78,$$

$$\beta_1(E) = -0.059 + 0.234E - 0.528(E-0.765) \text{ IF}(E>0.765), \quad \text{RSS} = 2.74, \text{ DF} = 76,$$

$$\beta_1(E) = 0.253 - 0.208E, \quad \text{RSS} = 3.20, \text{ DF} = 78.$$

The fitted break point models for $\beta_1(E)$ were superimposed on Fig. 3(b) giving Fig. 10. The break point models for $\beta_1(E)$ fit 'significantly better' than the linear model (iii) and provide an adequate fit to the data when compared with the smooth model (iv).

How well can smoothing functions detect break points? Our experience with generalized linear models (Rigby and Stasinopoulos, 1992; Stasinopoulos and Rigby, 1992) suggests that break points, including discontinuities, may be successfully detected by using smoothing functions in an exploratory fashion, provided that the smoothing parameter is sufficiently relaxed (i.e. using high degrees of freedom).

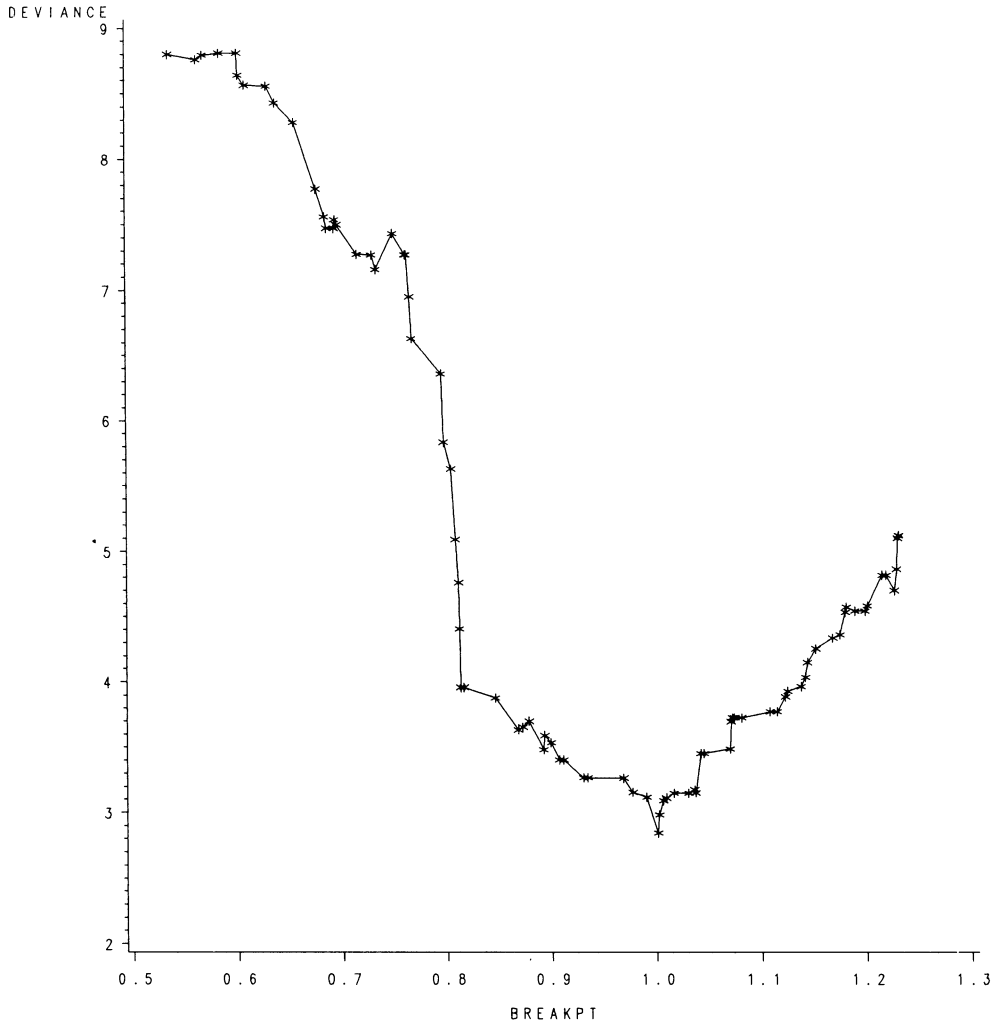

Fig. 9. Profile deviance plot for the break point γ in model (29)

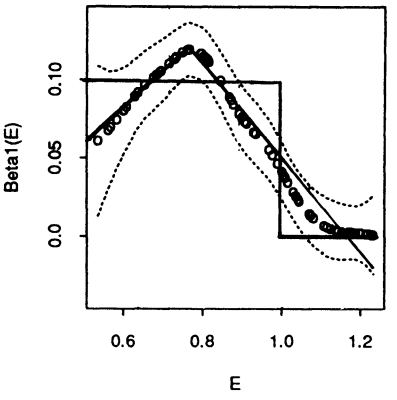

Fig. 10. Fitted break point models for $\beta_1(E)$ from models (29) and (30)

Stasinopoulos and Francis (1993) have adapted programs of the authors so that generalized additive models can be fitted very easily in GLIM4 by using a GLIM library macro. A macro to fit varying-coefficient models may be obtained from us.

Stuart Young and Adrian Bowman (University of Glasgow): The authors are to be congratulated on proposing a class of models which is an attractive extension of additive and other models, with a potentially wide area of application. We see a need to develop the important subject of inference, which is mentioned briefly in Section 6.

With the exhaust data from Section 1.1, the eventual model fitted is

$$NO_x = \beta_0(E) + \beta_1 C + \beta_2 CE + \epsilon.$$

This is chosen through a series of approximate F-tests in an analysis-of-variance table. Where comparison is to be made with a fully parametric model, an alternative model of inference is provided by Azzalini and Bowman's (1993) residual smoothing approach. In this, the residuals from a parametric regression (the null model) are taken as a response variable and smoothed against an explanatory variable of interest. Evidence of a relationship suggests the null model to be inadequate. A test statistic based on residual sums of squares can be formulated and, with an assumption of normally distributed errors, a highly accurate moment-based approximation to the null distribution can be used. With the exhaust data, Fig. 1(a) suggested a quadratic effect of E. If we wish to test this, we can obtain the residuals from the regression,

$$NO_x = \alpha + \beta_1 C + \beta_2 CE + \beta_3 E + \beta_4 E^2 + \epsilon,$$

and proceed as above. In this example, the simple plot of the residuals against E gives a clear indication that the quadratic model is unsatisfactory; nevertheless, it serves as an illustration of the method. The residual smoothing approach returns p-values of less than 0.0001 for all plausible choices of smoothing parameter, suggesting, as we expected, that a simple quadratic term in E is insufficient, and agreeing with the authors' choice of the more general form.

Point (d) in Section 2 mentions analysis of covariance. Consider the simple case where there is one continuous and one discrete covariate, and we wish to test whether a single (nonparametric) curve is sufficient. We can obtain a test statistic based on individual estimates of each curve contrasted with a common curve. Judicious choice of smoothing method can, at least asymptotically, eliminate bias, and we can again use a moment-based approach to approximating the distribution under the null hypothesis. This method can be extended to test parallelism, where the shift parameter is estimated parametrically. We are currently investigating how this approach might extend to more general models.

Ian McKeague (Florida State University, Tallahassee) **and Peter Sasieni** (Imperial Cancer Research Fund, London): The authors propose a flexible family of models for regression with interactions between covariates. In survival analysis the interactions may be with time, which is neither an 'independent' nor the 'dependent' variable. The general algorithm provided enables one to use structured models that are far from parametric.

Attention should be paid to the interpretation and to the prospects of inference in such models. The rest of these comments cover time-varying-coefficient models for conditional hazard functions.

In the time-dependent proportional hazards model (19), the 'parameters' can be interpreted in terms of instantaneous relative risks. The following points arise naturally.

(a) The model does not fit the structure of model (1). An equivalent form is given by 'the transformation

$$\int_0^T \exp\{X'\beta(t)\}\lambda_0(t)\,dt$$

is exponentially distributed'.

(b) Can a particular component $\beta_j(\ )$ be approximated by a time-independent term? See Murphy (1993).

(c) Even with prognostic index $X'\beta(t)+Z'\gamma$, it seems necessary to smooth β to estimate γ. This makes inference about γ difficult. One would not smooth the base-line hazard to make inference about the regression parameters in the Cox model.

The additive risk model (Aalen, 1980), or more generally $\lambda(t\,|\,X,\ Z)=\lambda_0(t)+X'\beta(t)+Z'\gamma$, can be viewed as an alternative to model (19). The additive structure is natural for competing risks with unknown failure type, or whenever failure arises from the cumulative damage to several components. Interestingly, estimation and inference in this model can be carried out without smoothing: the estimators are even unbiased. Efficient estimation is possible with weights obtained from a consistent preliminary estimator (McKeague and Sasieni, 1993). Thus the degree of smoothing is of little importance.

More generally consider

$$\lambda(t\,|\,X)=\alpha(t)'\,r(X,\ \beta)=\alpha_1(t)\,r_1(X_1'\beta_1)+\ldots+\alpha_p(t)\,r_p(X_p'\beta_p),$$

where the functions r_j are known, but $\alpha(\ )$ and β are unknown. Special cases include

(a) proportional hazards ($p=1$),

(b) excess (additive) hazards ($r_j\equiv$identity, $\dim(X_j)=1$, $j=1,\ \ldots,\ p$) and

(c) additive hazards with blocked covariates ($r_j\equiv$identity), as might arise when using dummy variables for categorical factors.

Such models provide a rich family for testing the fit of proportional hazards and can be used to investigate departures. Recent study of the efficient estimating equations for this model suggests that β can be estimated non-iteratively at a $\sqrt{n}$-rate without smoothing or explicitly estimating α. Efficient estimation may then follow by using weights based on consistent preliminary estimates of β and α.

Another approach is to estimate the β_j iteratively 'one at a time' (cf. Section 3.1) using the partial likelihood corresponding to r_j in a competing risks model, estimating the jth cause-specific counting process by the residual

$$\tilde{N}_j(t)=N(t)-\sum_{k\neq j}\int_0^t r_k(X_k'\beta_k)\,Y(s)\,dA_k(s)$$

where $N(t)$ counts uncensored failures, $Y(t)$ is the at-risk indicator and

$$A_k(t)=\int_0^t \alpha_k(s)\,ds.$$

The cumulative functions A_k can be estimated, without smoothing, by Aalen's (1980) least squares estimator in the full model with the unknown β_k replaced by their current estimates.

A. O'Hagan (University of Nottingham): I join with other speakers in congratulating the authors on presenting a stimulating paper of impressive scope. The variety of applications that they have outlined shows how general and practically useful the varying-coefficient models are. The authors kindly mention O'Hagan (1978). They will, I hope, forgive me for pointing out that their model (2) is a special case of mine. I wrote

$$y=h_1(z)\,\beta_1(z)+\ldots+h_p(z)\,\beta_p(z)+e,$$

where the $h_j(z)$ are known functions and the $\beta_j(z)$ are unknown. The underlying variable z can be anything at all and could certainly be $z=(x_1,\ \ldots,\ x_p,\ r_1,\ \ldots,\ r_p)$. Then making the $h_j(z)$ depend only on $x_1,\ \ldots,\ x_p$ and the $\beta_j(z)$ depend only on $r_1,\ \ldots,\ r_p$ gives the authors' model (2). Of course, they

have gone further in two respects. First they embed such a relationship within generalized linear models, and second they give some enlightening examples of the model's versatility.

I am struck, as I always am when I see sensible and elegant statistics being done from a non-Bayesian standpoint, by the authors' ingenuity, and also by the arbitrariness of their approach. Maximizing a penalized likelihood is, of course, analogous to finding the mode of a posterior distribution when the prior is proportional to the penalty function. But the Bayesian approach leads one to think about the prior, and to try to represent realistic prior information about the β_js. Apparently, a classical statistician can pick an arbitrary penalty function. Bayesians are not allowed to be *ingenious* in this way; they have to think about what is *right* for the problem at hand.

The techniques presented here have some restrictions. Hypothesis tests and other inferences are asymptotic and approximate. The authors say in Section 6 that one would like to have standard errors about the curve, but they do not say how this could be done or give any examples. And there remain restrictions on how the model might be generalized.

Bayesian methods, with modern computational techniques like Gibbs sampling, can overcome these difficulties with relative ease. Posterior inferences can be computed to any desired accuracy, rather than relying on approximations. Almost any kind of inference could be computed, such as bounds on the curve, or the probability that it is positive over some range. As an example of a generalization that is easy within a Bayesian framework, but probably much more difficult for a non-Bayesian because it departs from the exponential family, we could allow Cauchy-distributed errors. This would have the result (see, for example, O'Hagan (1988)) of effectively rejecting outliers and could produce interesting answers to the questions posed by Fig. 5.

M. J. R. Healy (London University Institute of Education): Singularity of the coefficient matrix does not preclude convergence of the Gauss–Seidel iteration, at least in the linear case. In the days before computers, Gauss–Seidel iteration was the method of choice for solving the large sets of simultaneous equations arising from the least-squares analysis of non-orthogonal many way tables. Since the nature of the singularity was known, it was possible to utilize it to good effect in the iteration (Yates, 1934; Healy and Dyke, 1952), a fact known to Gauss himself (Forsythe, 1951). If exact singularity can be harmless or even useful, it may be advantageous to replace 'small' eigenvalues by 0, after the manner of principal component regression.

R. J. Verrall (City University, London): I would like to make some comments on the connection between generalized additive models and dynamic generalized linear models.

Consider first equation (2), with one parameter which depends on a single predictor, i.e.

$$Y = X\beta(r) + \epsilon.$$

Assuming that the predictor r is equally spaced, we may index this by i and write the model for the ith observation as

$$y_i = x_i \beta(i) + \epsilon_i$$

corresponding to the first equation in Section 3.2. The penalized least squares criterion (equation (7)) now becomes

$$J(\beta) = \sum_{i=1}^{n} \{y_i - x_i \beta(i)\}^2 + \lambda \sum_{i=3}^{n} \{\Delta^2 \beta(i)\}^2$$

where $\Delta\beta(i) = \beta(i) - \beta(i-1)$.

This is equivalent to the following Bayesian model:

$$Y_i \,|\, \beta(i) \sim N\{x_i \beta(i),\ \sigma^2\},$$
$$\beta(i) - \beta(i-1) = \gamma(i),$$
$$\gamma(i) \,|\, \gamma(i-1) \sim N\{\gamma(i-1),\ \sigma^2/\lambda\}.$$

This can be seen since the prior distribution implies that

$$\Delta^2 \beta(i) = \Delta\gamma(i) \sim N(0,\ \sigma^2/\lambda)$$

which gives the second term in $J(\beta)$.

As a dynamic generalized linear model this would be written as

$$y_i = F\boldsymbol{\theta}_i + \epsilon_i,$$
$$\boldsymbol{\theta}_i = G\boldsymbol{\theta}_{i-1} + \mathbf{w}_i$$

where $F = (1, 0)$,

$$\boldsymbol{\theta}_i = \begin{pmatrix} \beta(i) \\ \gamma(i) \end{pmatrix},$$
$$G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$
$$\epsilon_i \sim N(0, \sigma^2),$$
$$\mathbf{w}_i \sim N\left\{ \mathbf{0}, \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2/\lambda \end{pmatrix} \right\}.$$

This is the familiar straight line model for the parameter with a stochastic gradient.

Important and very useful generalizations include the use of variance terms for $\mathbf{w}_i$ which depend on i allowing changes in the form of the model to be incorporated. I do not see how it would be possible to generalize the λ in equation (7) in this way.

Also, the state equation can be generalized to the perhaps more familiar form

$$\mathbf{w}_i \sim N(\mathbf{0}, W)$$

where W contains variances for both $\gamma(i)$ and $\beta(i)$. This is equivalent to including the first differential in the penalized likelihood. Also, higher differentials can be included in a similar way.

A. C. Davison (University of Oxford): I congratulate the authors on yet another important and stimulating contribution to nonparametric modelling.

My question relates to the combination of information from different sets of data that bear on the same substantive phenomenon, a procedure sometimes rather grandly called meta-analysis. In medical contexts, for example, the presence of important treatment effects may only be settled by the very large effective sample size obtained by combining suitably weighted estimates from many smaller data sets.

In the golden future when methods such as those described are available on the meanest laptop computer, and are widely used by practitioners, one might envisage the need to combine many nonparametric curves. Do the authors think that this might be a worthwhile activity, and, if so, how might it be done? Some sort of analysis of variance for curves, such as mentioned in the verbal presentation or along the lines of Rice and Silverman (1991), would be one way to proceed, but what do the authors suggest, if anything? This is of course tied to the question of how to produce standard errors and other measures of accuracy for these curves.

More fundamentally, and perhaps more provocatively, are these methods anything more than a form of elegant and computer-intensive—and undoubtedly very useful—exploratory data analysis?

The following contributions were received in writing after the meeting.

William S. Cleveland (AT&T Bell Laboratories, Murray Hill): A nonparametric regression surface is conditionally parametric with parametric family $\mathcal{F}$ if we can divide the factors into two subsets A and B with the following property: given any value of the factors in B, the surface is a member of $\mathcal{F}$ as a function of the factors in A (Cleveland *et al.*, 1991). We say that the surface is conditionally parametric in A.

Luckily, conditionally parametric surfaces are easy to fit; they require only small modifications to existing nonparametric methods. For example, for local fitting methods, one way to produce a conditionally parametric surface is simply to ignore A in computing distance in the space of the factors. Furthermore, co-plots, which are a particularly useful graphical method for regression studies, provide a diagnostic for determining when a conditionally parametric surface is likely to provide a good fit to the data. In such a case, using conditionally parametric modelling in place of fully nonparametric modelling provides a more parsimonious fit that saves degrees of freedom.

Conditionally parametric fits can be written as varying-coefficient models, and varying-coefficient models are conditionally parametric. One of the results of the authors' energetic and important paper is that there are many cases where intuition is better served by addressing the varying-coefficient formulation.

There are also many cases where intuition is better served by the conditionally parametric formulation. The reason involves one of the common ways in which conditionally parametric surfaces arise in practice (Cleveland, 1993). Suppose that a subset A of the factors has a limited effect on the response, and the other factors B have a substantial effect. Taylor series arguments suggest that it is reasonable to expect that the surface will be well approximated by terms involving only low powers, say 1 and 2, of the factors in A, and thus be amenable to modelling as conditionally parametric in A. After all, if the effect of the factors in A were so limited as to be non-existent, the approximation would involve terms with factors of A to 0 powers. The next step beyond this extreme case is powers of 1 and 2.

Dani Gamerman (Universidade Federal do Rio de Janeiro): I would like to make three brief comments.

Hastie and Tibshirani comment in Section 5 on the work of Gamerman (1991) where use of the full likelihood coupled with a dynamic model for regression coefficients *and* base-line hazard led to smooth estimates of both. That approach is in line with the methodology proposed by them but instead they only allowed regression coefficients to be smooth. No reason was put forward for not having a smooth base-line hazard.

The data set of Section 5.1 was analysed with the models of Gamerman (1991). Information from previous analyses suggests the removal of disease duration, age and prior therapy indicator as covariates. The model was specified with a grid determined by every other survival time (61 time intervals), a vague prior, fixed regression coefficients (proportional hazards (PH)) and a discounted base-line hazard. The posterior estimates are given in Table 3.

Apart from the negative sign for the squamous-type cell, the results are similar to Table 2 of Section 5.1. I then proceeded as the authors did to check for time variation of the effect of performance status. The dynamic model with this variation has Bayes factor of 1.36, only slightly preferable to the PH model.

TABLE 3
Summary of estimation

Variable	*Posterior mean*	*Posterior standard deviation*
Treatment	0.236	0.192
Cell type—squamous	−0.451	0.284
Cell type—small cell	0.332	0.277
Cell type—adeno	0.442	0.303
Performance status	−0.023	0.006

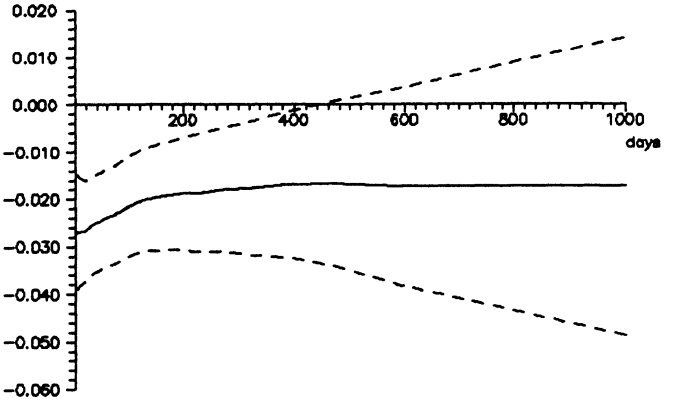

Fig. 11. Estimated trajectory of the performance status coefficient: ———, smoothed mean; - - - - -, 1.645 standard deviations limits

The estimated mean trajectory of the coefficient (shown in Fig. 11) varies smoothly from -0.028 in early stages to -0.017 at about 1 year. After that, it stabilizes at that value but its uncertainty increases with time. Similar results are obtained after removing the two extremely large survival times confirming lack of information in later periods. On comparison with Figs 5(a) and 5(b), there may be an indication of a larger reaction of Hastie and Tibshirani's estimates in the presence of little information (confidence limits are not provided there).

Another modelling approach related to varying-coefficient models is hierarchical modelling (Lindley and Smith, 1972). The richness of the Bayesian structure, mentioned in Section 3.4.2, allows here the explanation of regression coefficients by additional covariates through stochastic relations. Dynamic hierarchical models (Gamerman and Migon, 1993) allow, in addition, coefficient variation with time. However, parametric relationships are an integral part of hierarchical (or dynamic) models and have a strong effect on the results even when the prior for the higher stage (or initial) parameter is vague. Smoothness in the appropriate direction is a consequence of the model.

Wolfgang Härdle and Marlene Müller (Humboldt University, Berlin): We would like to congratulate the authors for an excellent and interesting paper which gives a framework for a wide range of flexible regression models. The varying-coefficient model as presented in this paper is very powerful indeed. Its application in the examples in Sections 4 and 5 speak for the method proposed.

Our comment will address some aspects of inference for the estimation method described. Once the varying-coefficient regression model has been estimated it is natural to compare it with competing fits. Since the coefficients of the model are functions $\beta_j(\ )$ the comparison could be based on confidence bands for the coefficient functions. Another proposal would be a squared distance between competing coefficient functions. Suppose that the nonparametric $\hat{\beta}$ has to be tested against a parametric fit $\hat{g}$. Härdle and Mammen (1993) have derived the distribution of

$$nh^{1/2} \int (\hat{\beta} - \mathscr{K}\hat{g})^2$$

where h denotes the kernel bandwidth and $\mathscr{K}\hat{g}$ denotes the smoothed parametric model. Simulations suggest that this test (based on the quantile of the asymptotic normal distribution) is not very powerful. The correct bootstrap (the so-called 'wild bootstrap') yields much better results. Have the authors similar experiences for their test based on the 'approximate degrees of freedom'? The same comment applies to uniform confidence bands.

M. C. Jones (The Open University, Milton Keynes): My remarks concern only a rather technical point which may be of little practical consequence. Consider, for simplicity, model (5) with univariate X. Write $V = Y/X$ so that $V = \beta(X) + \epsilon/X$. One might, appropriately, fit a parametric β to (X, V) by weighted least squares using weights proportional to X^2. This *global* experience does not, it seems to me, necessarily carry over immediately to *local* nonparametric regression. In Jones (1993), I show how weighting affects Nadaraya–Watson estimators, in particular, and the answer (asymptotically) is only in terms of bias and not at all in terms of variance. Moreover, there is no argument for choosing weights inversely proportional to error variance. In fact, swift calculations involving (preferable) local linear fitting suggest no effect of weights whatsoever (asymptotically), and that the bias effect is one of Nadaraya–Watson's peculiarities. It seems, however, that there may be some sense in inverse variance weighting for splines (essentially as used by the authors), but only because of splines' effective local bandwidth choice. This appears to involve weight$(x) \times f(x)$ (Silverman, 1984); since variance of smoothers depends inversely on $\sigma^2(x)/f(x)$, inverse variance weighting is suggested.

All that I am trying to say is that the authors' weighting, which is applied to general versions of their methodology, is not quite that obviously appropriate, and that it is an issue that might repay further investigation; for example, perhaps it can be done without, although I would not expect great differences to result.

I do not mean to detract in the slightest from a most interesting and worthwhile further contribution to an important area of the subject, one to which the current authors continue to contribute enormously.

Charles Kooperberg (University of Washington, Seattle) **and Charles J. Stone** (University of California, Berkeley): It is implicit in the discussion in Section 5 of the application of varying-coefficient models to survival data that the penalized partial likelihood estimate for β_j is a natural cubic spline and hence linear in the right-hand tail. When there are scant data in this tail, and especially when there is a substantial

amount of censoring, the methodology could be improved by incorporating the extra constraint that β_j be constant in the tail. In particular, this should prevent anomalies such as that displayed in Fig. 5. Alternatively, as in Gray (1992), $m=1$ could be used in equation (21) instead of $m=2$.

The authors mention regression spline bases with a fixed arrangement of knots. Recently, we have been working on hazards regression (HARE), which involves a MARS-like methodology (Friedman, 1991) for coming up with a model for the conditional log-hazard function having the form $\log \lambda(t \mid X_1, \ldots, X_p) = \Sigma_j \beta_j g_j(t, X_1, \ldots, X_p)$; each g_j involves at most two of the variables $t, X_1, \ldots, X_p$ and has the form of a linear spline or tensor product of two linear splines with the linear splines in t being constant in the right-hand tail and the knots selected by stepwise addition–deletion and the B information criterion (Kooperberg *et al.*, 1993). This methodology is combined with hazard estimation with flexible tails (HEFT), which similarly uses adaptive cubic splines that are constant in the right-hand tail possibly together with one or two log-terms to model the logarithm of the unconditional hazard function (Kooperberg and Stone, 1993). Specifically, HEFT is used to transform time so that the transformed unconditional hazard function is approximately equal to 1. In this manner we obtain a fitted model having the form

$$\log \hat{\lambda}(t \mid X_1, \ldots, X_p) = \hat{h}_0(t) + \sum_j \hat{\beta}_j \hat{g}_j(\hat{q}_0(t), X_1, \ldots, X_p)$$

for the conditional log-hazard function, where $\hat{h}_0$ is the HEFT fit to the logarithm of the unconditional hazard function and $\hat{q}_0 = -\log(1-\hat{F}_0)$ with $\hat{F}_0$ being the distribution function corresponding to $\hat{h}_0$.

We applied our methodology to the lung cancer data discussed in Section 5.1, choosing the options to obtain a model similar to that of the authors. In our fit, $\hat{h}_0(t) \approx -1.667 - 0.579 \log(t+144)$ (144 is the upper quartile of the survival times), whereas the basis functions $\hat{g}_j$ and their coefficients β_j are shown in Table 4 with $r_+ = \max(r, 0)$.

In Fig. 12(a) we show the fitted coefficient $-0.003 - 0.043(1.229 - \hat{q}_0(t))_+$ of performance status

TABLE 4

HARE fit to the transformed lung cancer data

Basis function	*Coefficient*	*Standard error*
1	−0.106	0.662
Cell type—small cell	0.756	0.224
Cell type—adeno	0.993	0.259
Performance status	−0.003	0.009
$(1.229 - \hat{q}_0(t))_+$	1.884	0.672
(Performance status)$(1.229 - \hat{q}_0(t))_+$	−0.043	0.011

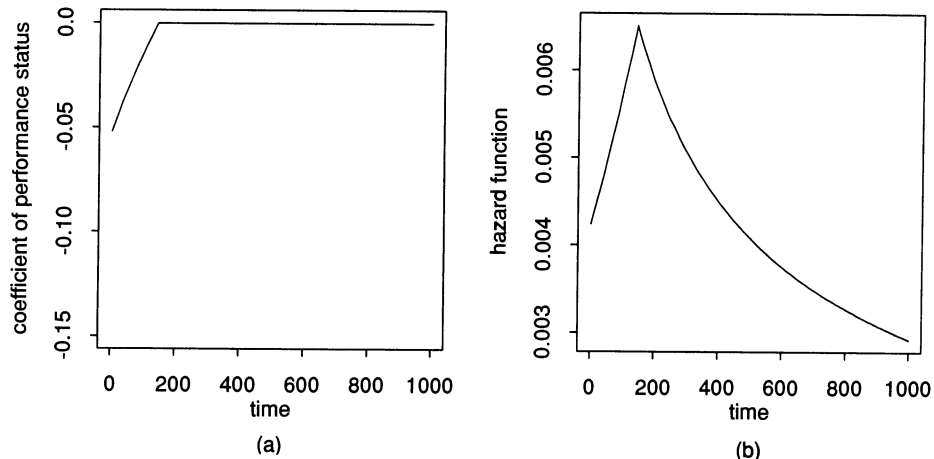

Fig. 12. (a) Fitted coefficient of performance status as a function of time; (b) fitted hazard function for a person with cell type squamous and performance status 60

as a function of time, which should be compared with the authors' Fig. 5, and in Fig. 12(b) we show the fitted hazard function for a person with specified values of the relevant variables.

Susan A. Murphy (Pennsylvania State University, University Park): The following comments concern Section 6, 'Inference for the model'. It is important to realize the disadvantages of the statistical tests in this section. These tests are usually used to detect a large variety of alternatives. However, higher priority should be placed on how to order the alternatives of interest carefully. Because the alternative space ($\beta(r)$ non-constant) is very large and because the true parameter may not be of a form that is easily approximated by the chosen class of smoothers, the test statistic for constant $\beta(r)$ can have very low power. Even the asymptotic analysis suffers from the large dimensionality of the alternative space. When the alternative space is finite dimensional, we can expect that a likelihood ratio test will be consistent against local alternatives of order 1 over the square root of the sample size. However, in the regression setting, Eubank and Spiegelman (1990) and in the proportional hazards setting with time varying coefficient Murphy (1993) find that only local alternatives of lower order can be detected. These results point to a need to prioritize the alternatives of interest and to choose the class of smoothers to contain the higher priority alternatives. It would be of interest to have the authors address this issue.

Grace Wahba (University of Wisconsin, Madison): We thank the authors for telling us interesting things about the class of 'varying-coefficient models'. Since we now have such a wealth of possible semiparametric models in several variables it becomes more important to have subclasses of such models that have an easily interpretable structure, so that possible models can be sensibly matched to the application at hand. These models are surely such a subclass. I found it interesting that the penalized least squares model (7) can be viewed both as an example of an inverse problem and as a particular element in a tensor product spline expansion. Gu and Wahba (1993a, b) and Wahba *et al.* (1993) generalize the tensor product spline expansions to smoothing spline analysis of variance in function spaces, which among other things allow the r_j to be spatial variables, extend the Bayesian 'confidence intervals' to the components of these models and consider binomial responses in the context of penalized generalized linear models. The results in each case would specialize to this very interesting class of varying-coefficient models. The interesting computational procedures provided by the authors should generalize to the spatial data case.

The issue of testing whether $\beta(r)$ is well approximated by a constant or linear function of r raises some philosophical questions in this context. The authors note that the tests they propose involve the degrees of freedom for signal and that this depends monotonically on the smoothing parameter(s). Should we think of these tests differently according to whether the smoothing parameter(s) are chosen

(a) by eyeball,
(b) by a rule of thumb based on experience or
(c) by some data-based method such as generalized cross-validation or C_p in the case σ^2 is known?

To continue in a philosophical vein, to what extent would we use different tests if our goal is to test the null hypothesis of the parametric model (β linear), or, to build the best model we can for prediction, under the assumption that no parametric model is exactly true?

The **authors** replied later, in writing, as follows.

An objective of this work was to extend the class of generalized additive models, and in doing so to reveal some interesting relationships that exist between various classes of models that have been proposed. We had hoped that this would generate discussion that would help to clarify these relationships. The excellent contributions of the discussants suggest that we have been successful. We thank them and the Editors for their interest and efforts.

In the limited space provided, we would like to reply briefly to some of the issues that have been raised.

Professor Green shows that the varying-coefficient model, with more than one effect modifying variable R_j, is not invariant to translations of the variables X. This is an important practical point, and Professor Green proposes an alternative model that is invariant.

Another solution is to *require* an intercept term $\beta_0(R_j)$ for every term $X\beta_j(R_j)$ that appears in the model, thus making the model equivariant under location changes to X. Unfortunately, if we estimate this model by penalized least squares, the equivariance is lost. Consider the simpler example $\beta_1(R)+X\beta_2(R)$, and suppose that we have penalized least squares estimates $\hat{\beta}_1(R)$ and $\hat{\beta}_2(R)$. If X is

replaced by $X^* = X + c$, we would expect our estimates to be modified accordingly: $\hat{\beta}_2^* = \hat{\beta}_2(R)$ and $\hat{\beta}_1^* = \hat{\beta}_1(R) - c\,\hat{\beta}_2(R)$. But, if both $\hat{\beta}_1(R)$ and $\hat{\beta}_2(R)$ are 'smooth', then $\hat{\beta}_1^*$ as prescribed above could be arbitrarily 'rough' depending on the size of c. Another way of saying this is that the penalty functionals in criterion (7) are not linear in their arguments. For similar reasons, estimates using other 'shrinking' smoothers (Buja *et al.*, 1989) would also lack equivariance. These problems disappear as the smoothers approach consistency (conditional expectations), and indeed the L_2-estimates in Section 3.2 are equivariant. Interestingly, for a single effect modifier, estimates produced by the conditionally parametric approach in Section 3.4.1 are equivariant. As a practical solution to the problem, we propose that, in addition to including an intercept coefficient as above, the X-variables are centred to have overall mean 0; this helps with convergence of the backfitting algorithm as well. In the more general model with different modifying variables, there is a further lack of equivariance and it may make sense to orthogonalize the variables as well as to centre them.

Professor Green also shows how the model containing only a single non-constant $B(r)$ can be estimated non-iteratively, a fact he originally discovered for the additive model. In both settings, this provides a useful simplification.

Dr Cuzick wonders about our particular choice for representing interactions. We do not find terms such as $g_{ij}(x_i, x_j) = h_{ij}(x_i)\,k_{ij}(x_j)$ easy to interpret. The varying-coefficient formulation allows at least two different and interpretable versions:

(a) $g_{ij}(x_i, x_j) = h_{ij}(x_i)x_j$ or $g_{ij}(x_i, x_j) = k_{ij}(x_j)x_i$, whichever is more appropriate; here we can see how the slopes change as a function of x_i or x_j respectively, as in Fig. 2 in the text;

(b) $g_{ij}(x_i, x_j) = \Sigma_{m=1}^{M} I(x_i \in I_m)\,k_{ijm}(x_j)$ which assumes that the surface is piecewise constant over M non-overlapping intervals in x_i, as in Fig. 4 in the text; in this case x_i is replaced by several 'dummy' versions $I(x_i \in I_m)$ that define the intervals.

In the survival analysis setting, Dr Cuzick, and Dr McKeague and Dr Sasieni explore the underlying processes that might lead to the varying-coefficient model of Section 5. This is an important area to understand. We took an empirical view of this model, thinking simply that non-constant parameter functions would be a useful alternative to proportional hazards. It would be interesting to examine whether those models could be estimated in a manner similar to the approach in our paper. A general question for Dr Cuzick, and Dr McKeague and Dr Sasieni is how far can partial likelihood estimation be pushed? The model of Section 5 is quite far away from Cox's original proportional hazards model, yet partial likelihood can still be useful for estimation. There might be some point at which the loss of information in partial likelihood relative to full likelihood becomes unacceptably large.

Dr Gamerman asks why we do not model the base-line hazards smoothly. We see no reason for doing this (other than the potential information loss mentioned above), if interest lies in the regression effects. The use of partial rather than full likelihood alleviates the need for smoothness assumptions about the base-line hazard and also makes possible the inclusion of time-dependent covariates. This is not something that we have tried, but in principle it should work.

Dr Kooperberg and Dr Stone outline their interesting new work on MARS-type models for log-hazard functions. An appealing aspect of their approach is the joint modelling of time and the covariates of interest. When there are no time × covariate interactions (but main effects in both), their model becomes a proportional hazards model. In this case the main effect for time allows one to model the base-line hazard with splines, thus giving a compromise between the fully nonparametric Cox model and the parametric exponential model. When these interactions are present, their model has varying coefficients in the sense of Section 5, with a more parametric approach to estimation (via the full likelihood).

Dr Rigby and Dr Stasinopoulos suggest break point models as an alternative to those in the paper. Smoothing splines are not well suited to detecting break points: better choices would be the split linear smoother of McDonald and Owen (1986) and wavelet smoothing (Donoho and Johnstone, 1992). Because our algorithm is modular, we could just 'plug in' either one of these smoothers for any variable for which a break point model is desired.

Mr Young and Dr Bowman having been making progress in the important area of inference, in particular the comparison of parametric submodels with a larger nonparametric model. Inference tools are vital if these methods are to gain practical acceptance. The advantage of residual smoothing is not clear to us: why not fit models with and without a nonparametric term for E and compare the quality of the fits?

We thank Professor O'Hagan for pointing out the full generality of the models suggested in O'Hagan (1978). With regard to his other points, we can compute standard error curves, as shown in Fig. 3 of the paper. There is not space here to delve into the multifaceted issue of Bayes *versus* non-Bayes inference

in this setting: some of our recent views on this issue appear in the reply to the discussants in Buja *et al.* (1989). However, we would like to point out that whatever Gibbs sampling can do for the Bayesian the bootstrap can do for the non-Bayesian!

We thank Dr Verrall for further clarifying the connection between the varying-coefficient model and the dynamic generalized linear model. As stated in our paper, the dynamic generalized linear model has more rich probabilistic structure, and hence in some ways it offers more flexibility. However, our paper shows that time varying or more general varying-coefficient models can be estimated in a static, non-sequential and modular way. This offers both conceptual and practical advantages. On a technical note, although we cannot vary λ in equation (7), we can vary σ^2 and often do via observation weights.

Dr Davison raises the interesting question about meta-analysis and the more specific question of analysis of variance applied to curves. We have explored the latter in informal ways, and Professor Wahba reports on a formal approach via tensor product splines. On the bigger issue of exploratory data analysis *versus* inference: we have used generalized additive models, for both purposes (see Hastie and Tibshirani (1990)) and expect that the same will be true for the varying-coefficient model. Some care is clearly needed when flexible adaptive methods are used for inference, and further work along the lines of that reported by Young and Bowman is necessary.

Dr Cleveland reminds us of his definition of conditionally parametric models but then goes on to say that varying-coefficient models are also conditionally parametric. For example $X_1\beta_1(R_1)+X_2\beta_2(R_2)$ is linear, conditioned on (R_1, R_2); we can stretch this even further and argue that any regression model is conditionally constant given the entire predictor set! Although strictly correct, we do not feel that this interpretation captures the essence of models such as those above, and hence we prefer to call these models *varying coefficient*. For us the name *conditionally parametric* suggests models of the form (4) in the paper, and connects strongly with Dr Cleveland's local regression approach to fitting such models.

The issue of smoothing weights raised by Dr Jones is interesting, and we look forward to seeing his 1993 paper. It seems that weights serve two purposes in models such as these (and generalized additive models):

(a) efficiency weights—if we are smoothing against R and we know that the response variance is a function $\sigma^2(R)$, the Gauss–Markov theorem tells us to use weights $\sigma^{-2}(R)$; it is also intuitively clear that the benefit is asymptotically negligible;
(b) model weights—even in the L_2-case in equation (6), where we are computing conditional expectations, the weights are needed (and indeed we can view the ratio as a weighted conditional expectation); we doubt whether these weights can be ignored.

Professor Härdle and Professor Müller suggest a new way of testing parametric *versus* nonparametric fits: their suggestion looks especially useful in non-nested comparisons. Our approximate degrees of freedom test, based on asymptotic χ^2- or F-distributions, seems to perform reasonably well. Use of the bootstrap might provide better accuracy but we have not investigated this.

Professor Wahba enquires about the choice of smoothing parameters when testing $\beta(r)$, or when prediction is the goal. Although we do not put a great emphasis on formal testing in this setting, we currently favour the use of fixed degrees of freedom tests to reduce the adaptivity factor. When our goal is prediction, we currently use an adaptive backfitting algorithm, BRUTO, driven by a global generalized cross-validation criterion (Hastie and Tibshirani, 1990). For each term BRUTO selects whether it should be included, linear or non-linear, and if non-linear how smooth.

Professor Murphy warns that omnibus tests will have low power against specific alternatives. We would like to read her paper on this subject and to learn the specifics of her argument. At present, it is not clear to us that the dimensionality of the alternative space is that high when say 5 degrees of freedom are used to estimate $\beta(r)$.

Professor Healy points us to his 1952 paper on Gauss–Seidel iteration. We obtained the paper from the library and *déjà vu*—the computational problems in fitting unbalanced main effects models in the 1950s scale up to the problems in fitting nonparametric additive models in the 1980s. We recommend this interesting paper on the early use of computers and wish that we had been aware of it when writing Buja *et al.* (1989).

REFERENCES IN THE DISCUSSION

Aalen, O. O. (1980) A model for nonparametric regression analysis of counting processes. *Lect. Notes Statist.*, **2**, 1–25.
Azzalini, A. and Bowman, A. (1993) On the use of nonparametric regression for checking linear relationships. *J. R. Statist. Soc.* B, **55**, 549–557.

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

Cleveland, W. S. (1993) Coplots, nonparametric regression, and conditionally parametric fits. In *Multivariate Analysis and Its Applications* (eds T. W. Anderson, K. T. Fang and I. Olkin). Hayward: Institute of Mathematical Statistics.

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991) Local regression models. In *Statistical Models in S* (eds J. M. Chambers and T. Hastie), pp. 309–376. Pacific Grove: Wadsworth and Brooks/Cole.

Cuzick, J., De Stavola, B. L., Cooper, E. H., Chapman, C. and MacLennan, I. C. M. (1990) Long-term prognostic value of β_2 serum microglobulin in myelomatosis. *Br. J. Haem.*, **75**, 506–510.

Cuzick, J. and Trejo, B. (1992) Analysis of trials with treatment—individual interactions. In *Survival Analysis: State of Art* (eds J. P. Klein and P. K. Goel), pp. 65–76. Dordrecht: Kluwer.

Daniel, C. and Wood, F. S. (1980) *Fitting Equations to Data*, 2nd edn, pp. 142–145. New York: Wiley.

Donoho, D. and Johnstone, I. (1992) Ideal spatial adaptation via wavelet shrinkage. *Technical Report*. Department of Statistics, Stanford University, Stanford.

Eubank, R. L. and Spiegelman, C. H. (1990) Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Am. Statist. Ass.*, **85**, 387–392.

Forsythe, G. E. (1951) Gauss to Gerling on relaxation. *Math. Tab. Wash.*, **5**, 255–258.

Friedman, J. H. (1991) Multivariate regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.

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

Gamerman, D. and Migon, H. S. (1993) Dynamic hierarchical models. *J. R. Statist. Soc.* B, **55**, 629–642.

Gray, R. J. (1992) Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Am. Statist. Ass.*, **87**, 942–951.

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

Gu, C. and Wahba, G. (1993a) Semiparametric analysis of variance with tensor product thin plate splines. *J. R. Statist. Soc.* B, **55**, 353–368.

——— (1993b) Smoothing spline ANOVA with component-wise Bayesian confidence intervals. *J. Comput. Graph. Statist.*, to be published.

Härdle, W. and Mammen, E. (1993) Comparing nonparametric *versus* parametric regression fits. *Ann. Statist.*, to be published.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*, p. 262. London: Chapman and Hall.

Healy, M. J. R. and Dyke, G. V. (1952) A Hollerith technique for the solution of normal equations. *J. Am. Statist. Ass.*, **48**, 809–815.

Huffer, F. W. and McKeague, I. W. (1991) Weighted least squares estimation for Aalen's additive risk model. *J. Am. Statist. Ass.*, **86**, 38–53.

Jones, M. C. (1993) Do not weight for heteroscedasticity in nonparametric regression. *Aust. J. Statist.*, to be published.

Kooperberg, C. and Stone, C. J. (1993) Hazard estimation with flexible tails. *Technical Report 388*. Department of Statistics, University of California, Berkeley.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1993) Hazard regression. *Technical Report 389*. Department of Statistics, University of California, Berkeley.

Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, **34**, 1–41.

McDonald, J. and Owen, A. (1986) Smoothing with split linear fits. *Technometrics*, **28**, 195–208.

McKeague, I. W. and Sasieni, P. D. (1993) A partly parametric additive risk model. Submitted to *Biometrika*.

Murphy, S. A. (1993) Testing for a time-dependent coefficient in Cox's regression model. *Scand. J. Statist.*, to be published.

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

——— (1988) Modelling with heavy tails. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 345–359. Oxford: Clarendon.

Rice, J. A. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B, **53**, 233–243.

Rigby, R. A. and Stasinopoulos, D. M. (1992) Detecting break points in the hazard function in survival analysis. In *Statistical Modelling* (eds B. Francis, G. U. H. Seeber, P. G. M. van der Heijden and W. Jansen), pp. 303–312. North-Holland: Elsevier.

Silverman, B. W. (1984) Spline smoothing: the equivalent variable kernel method. *Ann. Statist.*, **12**, 898–916.

Stasinopoulos, D. M. and Francis, B. (1993) Generalised additive models in GLIM4. *GLIM Newslett.*, to be published.

Stasinopoulos, D. M. and Rigby, R. A. (1992) Detecting break points in generalised linear models. *Comput. Statist. Data Anal.*, **13**, 461–471.

Wahba, G., Gu, C., Wang, Y. and Chappell, R. (1993) 'Soft' classification, a.k.a. penalized log likelihood risk estimation with smoothing spline analysis of variance. In *Proc. Conf. Supervised Machine Learning*. Santa Fe Institute.

Yates, F. (1934) The analysis of multiple classifications with unequal numbers in the different classes. *J. Am. Statist. Ass.*, **29**, 51–66.