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.

© 1993 Royal Statistical Society 0035-9246/93/55757

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

$\eta = g(\mu)$. 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

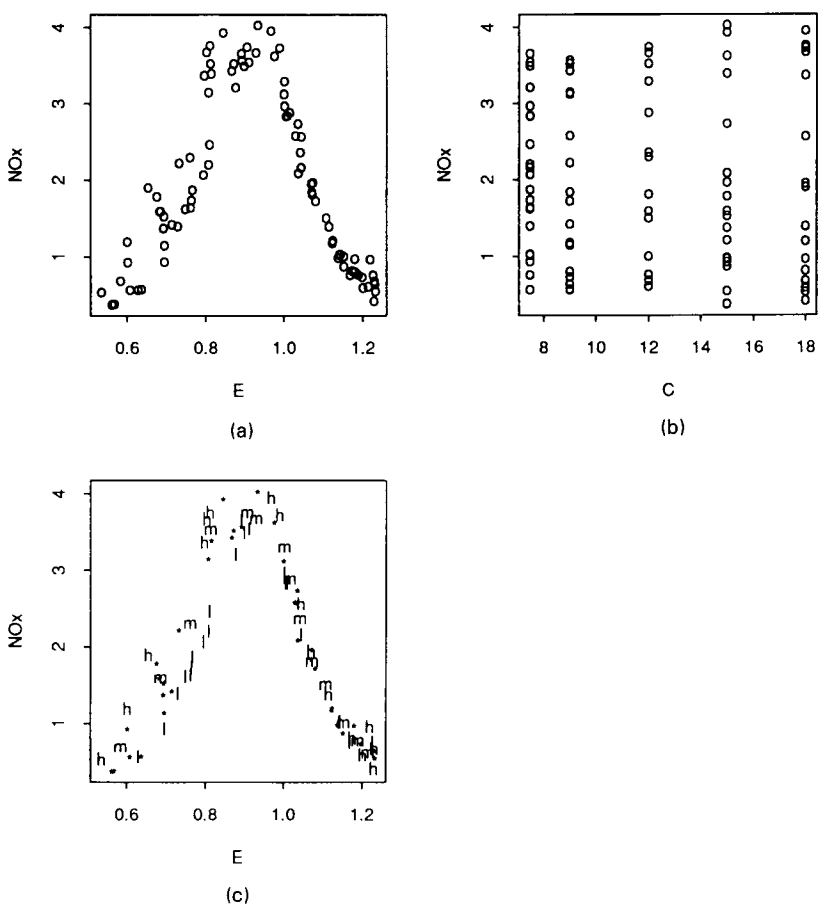


Fig. 1. (a) NO_x *versus* E; (b) NO_x *versus* C; (c) NO_x *versus* E with some values of C coded as low (l), medium (m) or high (h) (intermediate values are coded with *)

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

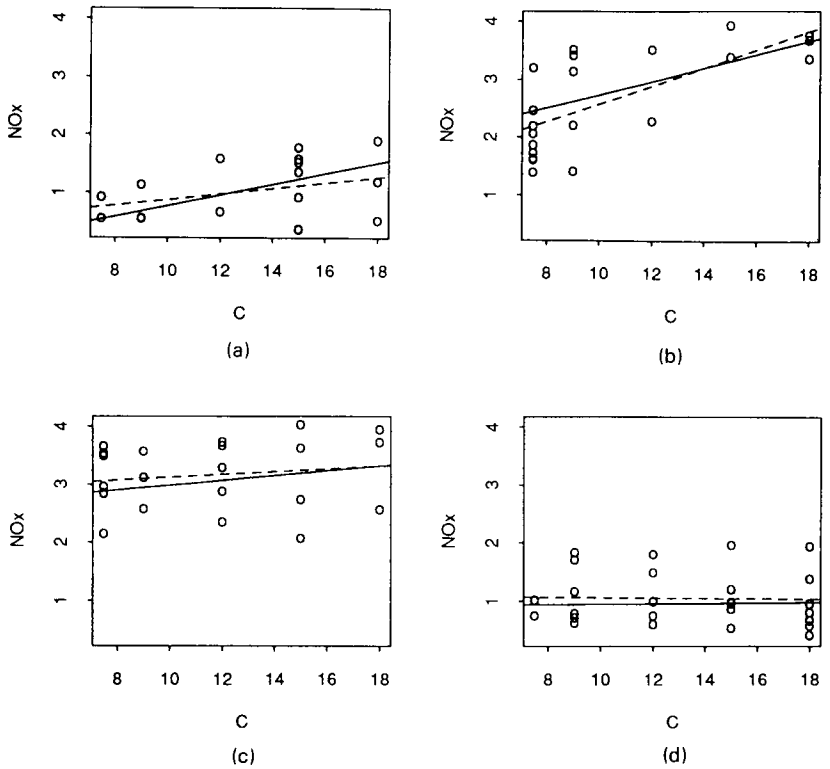


Fig. 2. NO_x *versus* C for (a) low, (b) medium, (c) high and (d) very high values of E: ———, fitted lines from the varying-coefficient model, taken at the median value of E for the data in the panel; – – –, fitted linear regression

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.

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

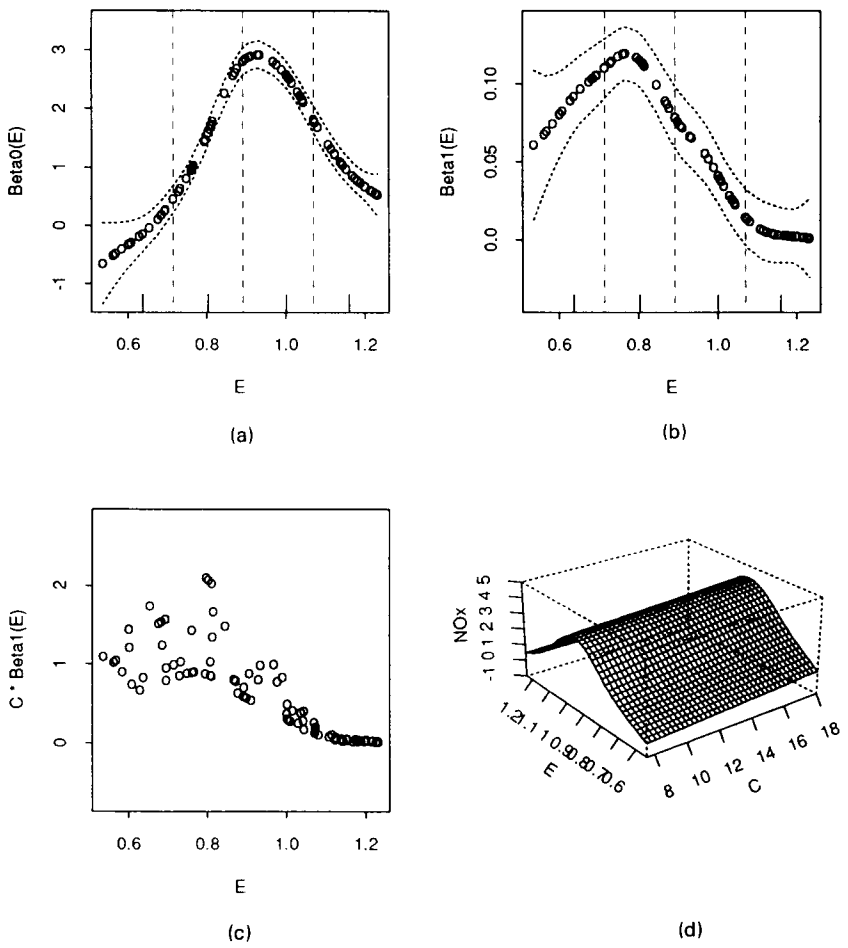


Fig. 3. (a) $\hat{\beta}_0(E)$ and (b) $\hat{\beta}_1(E)$ with pointwise 90% standard error bands defined by ± 1.645 times an estimate of the standard error (--------, regions of data used to construct the four conditioned plots of Fig. 2; the four spikes at the base of the plots show the four values 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.

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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

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

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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

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

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

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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).

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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 $E = E_0$

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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 $\mathbf{Z}$,

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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.

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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

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

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.

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

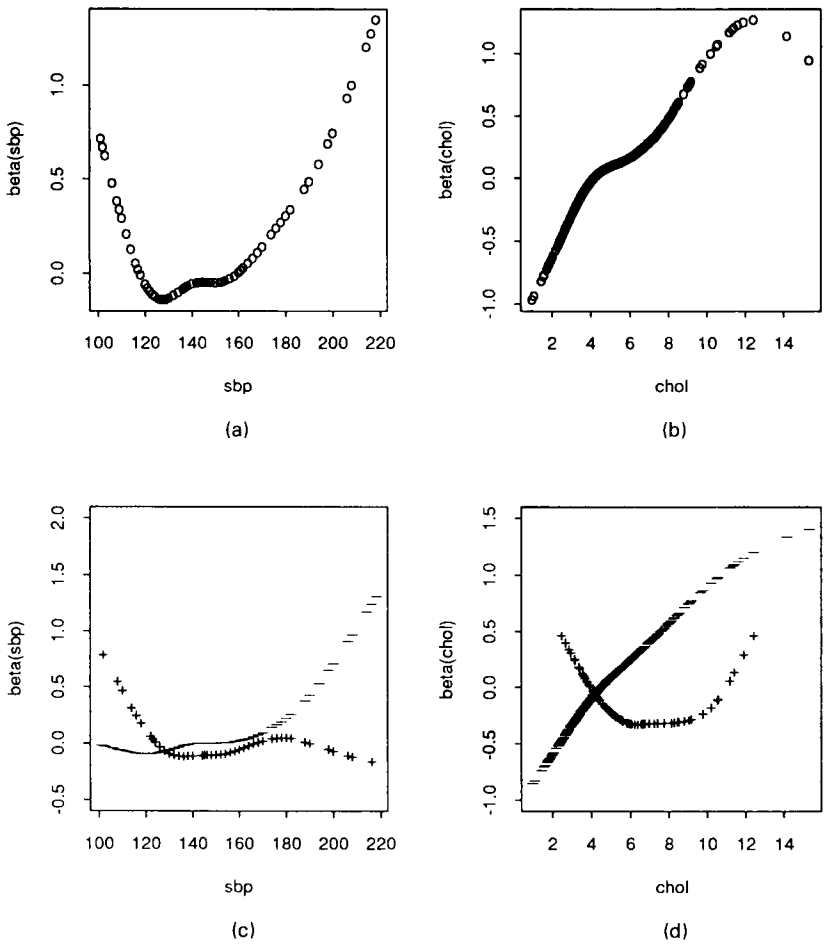


Fig. 4. (a) Fitted curve for S and (b) fitted curve for C in a generalized additive model, ignoring the treatment variable T; (c) pairs of functions $\hat{\beta}_{S0}(S)$ and $\hat{\beta}_{S0}(S)+\hat{\beta}_{S1}(S)$ for S ($-$, no treatment ($T=0$); $+$, treatment); (d) plot equivalent to (c) for C

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.

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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

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

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,

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

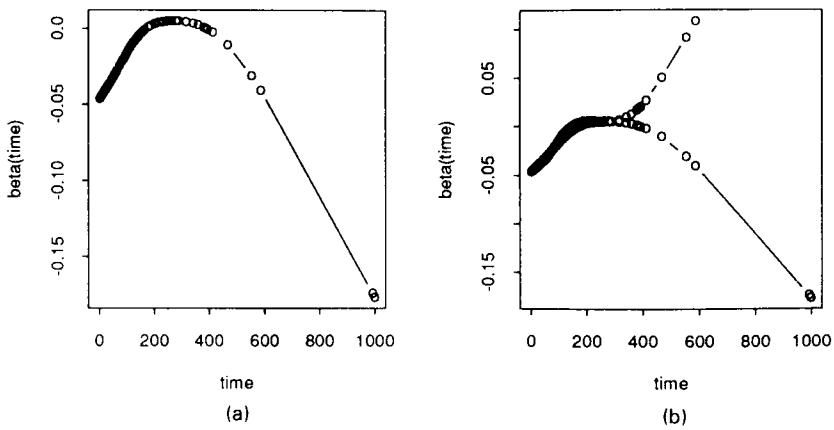


Fig. 5. Plots of $\hat{\beta}(t)$ for the third example: (a) with all the data the plot is markedly quadratic; (b) effect of removing the two values of time on the extreme right-hand side and refitting the model—the tail of the function increases rather than decreases

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'

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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 $R_1, R_2, \ldots, R_p$ 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.

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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.

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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

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

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 $\Delta_j = \boldsymbol{\gamma}_j^{\mathrm{new}} - \boldsymbol{\gamma}_j^{\mathrm{old}}$. 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 $\boldsymbol{\gamma}_j^{\mathrm{old}} = 0$, 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 $\boldsymbol{\gamma}_j^{\mathrm{old}}$ to $\boldsymbol{\gamma}_j^{\mathrm{new}}$ 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 $\boldsymbol{\gamma}_j^{\mathrm{old}}$ to $\boldsymbol{\gamma}_j^{\mathrm{new}}$ for the varying-coefficient model in the generalized regression setting;

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

(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 $\mathbf{z}^* = 0$):

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

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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 $\mathscr{E}^{np}$ 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 $\mathbf{TT}^\infty = \mathbf{T}^\infty \mathbf{T} = \mathbf{T}^\infty$.

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 $\mathscr{P}$ be the permutation matrix that sorts the columns of Σ by failure time and regressor within failure time. Then the matrix $\mathscr{P}^T \Sigma \mathscr{P}$ is block diagonal, with ith block the covariance matrix of $x, \ldots, x_p$ at time $t_{(i)}$. Hence $\mathscr{P}^T \Sigma \mathscr{P}$ 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($\mathbf{U}$) the *concurvity space* (expressed in terms of $\boldsymbol{\beta} = \mathbf{N}\boldsymbol{\gamma}$). In general, elements of null($\mathbf{U}$) must satisfy

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

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.

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026

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

Downloaded from academic.oup.com/jrsssb/article/55/4/757/7028270 by Stanford University user on 28 October 2026