The Annals of Statistics
1989, Vol. 17, No. 2, 453–555

LINEAR SMOOTHERS AND ADDITIVE MODELS

BY ANDREAS BUJA,[1] TREVOR HASTIE AND ROBERT TIBSHIRANI[2]

Bellcore, AT&T Bell Laboratories and University of Toronto

We study linear smoothers and their use in building nonparametric regression models. In the first part of this paper we examine certain aspects of linear smoothers for scatterplots; examples of these are the running-mean and running-line, kernel and cubic spline smoothers. The eigenvalue and singular value decompositions of the corresponding smoother matrix are used to describe qualitatively a smoother, and several other topics such as the number of degrees of freedom of a smoother are discussed. In the second part of the paper we describe how linear smoothers can be used to estimate the *additive model*, a powerful nonparametric regression model, using the "backfitting algorithm." We show that backfitting is the Gauss–Seidel iterative method for solving a set of normal equations associated with the additive model. We provide conditions for consistency and nondegeneracy and prove convergence for the backfitting and related algorithms for a class of smoothers that includes cubic spline smoothers.

1. Introduction. Consider a standard regression problem where we have n observations of a random variable Y, say $y_1, y_2, \dots, y_n$, at design points $x_1, x_2, \dots, x_n$. A regression procedure produces a decomposition of the form $y_i = \hat{y}_i + \text{residual}_i$, where the fit $\hat{y}_i$ is thought to estimate a systematic dependence of y_i on x_i. If the design points x_i are univariate real values, one usually makes the assumption that the dependence is smooth, and correspondingly, nonparametric regression methods are often called *scatterplot smoothers* in this case. Typically the fits are useful both for scatterplot enhancement as well as for estimating the regression model

$$E(Y|x) = f(x). \tag{1}$$

A *linear smoother* is special in that $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)^t$ can be written in the form $\hat{\mathbf{y}} = S\mathbf{y}$, where $\mathbf{y} = (y_1, y_2, \dots, y_n)^t$ and the $n \times n$ matrix S, called a *smoother matrix*, does not depend on $\mathbf{y}$. Examples of linear smoothers are running means, locally weighted running lines, kernel smoothers, smoothing splines, bin smoothers and even the least-squares line. The running median is an example of a nonlinear smoother for which a smoother matrix cannot be constructed. Most of the linear smoothers mentioned above depend on a *smoothing parameter*; if a data-driven technique such as cross-validation is used to select this parameter, they become nonlinear smoothers. Other examples of nonlinear smoothers are

Received May 1987; revised November 1988.

[1] Supported by NSF Grant DMS-85-04359 and DOE Grant DE-FG06-85-ER25006 while a member of the faculty of the Department of Statistics of the University of Washington.

[2] Supported by the Natural Sciences and Engineering Research Council of Canada.

AMS 1980 *subject classifications.* Primary 62G05; secondary 65D10.

Key words and phrases. Smoother, additive model, nonparametric, semiparametric, regression, Gauss–Seidel algorithm.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

robust smoothers ["Lowess," Cleveland (1979)] and cross-validated variable span smoothers ["Supersmoother," Friedman and Stuetzle (1981)]. Mallows (1980) discusses linear and nonlinear smoothers and methods for obtaining smoother matrices for the "linear" part of a nonlinear smoother. Examples of some linear smoothers, applied to a set of meteorological data, are shown in Figure 2. Each of these smoothers and these data are discussed later in this paper.

Because their smoother matrices do not depend on the response $\mathbf{y}$, linear smoothers lend themselves to relatively simple analyses. In the first part of this paper (Section 2) we use the eigenvalue and singular value decompositions of S to describe the qualitative behavior of a linear smoother and discuss some other descriptive tools for smoothers. We also discuss some statistical properties such as the number of "degrees of freedom" used by a smoother and the variance of the fit. The literature on linear smoothers for scatterplots is very large and we will not attempt a complete bibliography. Some notable papers include Watson (1964), Rosenblatt (1971), Reinsch (1967), Priestley and Chao (1972), Stone (1977), Craven and Wahba (1979), Cleveland (1979), Friedman and Stuetzle (1981) and Silverman (1985). Many others are given in the reference lists of these papers. A great deal of work on smoothing appears in the time-series literature [see Cleveland (1983)], where Whittaker (1923) first introduced spline smoothing.

In the second part of this paper (Sections 3–5) we study the use of linear smoothers as building blocks for nonparametric multiple regression models. In particular, we study the "additive model" in which the response is modeled as a sum of smooth functions of the covariates, for example,

$$E(Y|u, v, w) = f_1(u) + f_2(v) + f_3(w) \tag{2}$$

for three covariates u, v and w. The functions $f_i(\cdot)$ are unspecified in form and are estimated using linear smoothers in an iterative algorithm known as "backfitting." The additive model is more flexible than the standard linear model and at the same time is more interpretable than a general (nonadditive) regression surface. As an example, Figures 7(a), (b) and (c) show the estimated functions for three variables from the meteorological data set mentioned above. Note the nonlinearities that might be missed by standard parametric methods.

The additive model was suggested by Friedman and Stuetzle (1981) [see also Friedman, Grosse and Stuetzle (1983)], and forms the core of the "ACE" algorithm [Breiman and Friedman (1985)]. More recently the additive model and other related models have received renewed attention; see Wahba (1986), Engle, Granger, Rice and Weiss (1986), Burman (1988) and Hastie and Tibshirani (1986a). Here we describe the backfitting algorithm for estimating an additive model and study its properties. The backfitting algorithm is the Gauss–Seidel iterative method for solving a set of normal equations. For a class of smoothers containing cubic spline smoothers, we prove that the normal equations are consistent, and that the backfitting algorithm always converges to a solution. We give conditions for the uniqueness of these solutions, and in the case of degeneracies, we characterize them. We also propose more efficient versions of the algorithm. We extend the discussion of the statistical properties to this setting and explore the relationship between the additive model and generalized least squares. At various points in the paper, we explore connections of this work with that of Denby (1984), Green, Jennison and Seheult (1985), Green and Yandell (1985), O'Sullivan, Yandell and Raynor (1986), Mallows (1986) and Eubank (1984). Finally, in Section 6 some open problems are discussed.

Notation. Let $\mathscr{R}(S)$ denote the range of the linear mapping S and $\mathscr{N}(S)$ its nullspace. We use lowercase bold roman such as $\mathbf{v}$ to represent vectors, and uppercase roman for matrices such as S. In later sections we distinguish compound matrices by using bold uppercase, such as $\mathbf{P}$. We use $\mathscr{M}_\lambda(S)$ for the eigenspace corresponding to eigenvalue λ and hence $\mathscr{N}(S) = \mathscr{M}_0(S)$. By $V^\perp$ we mean the orthogonal complement of the subspace V. The spectral radius of S (largest absolute eigenvalue) is denoted by $\rho(S)$. The ith largest singular value of S is $\sigma_i(S)$, and $\lambda_i(S)$ the ith largest eigenvalue of S if all are real. An arbitrary matrix norm of S is denoted by $\|S\|$ and the two-norm of the matrix S is denoted by $\|S\|_2 = \sup_{\mathbf{a}\neq 0}\|S\mathbf{a}\|/\|\mathbf{a}\|$. We use $\mathbf{f}_+$ to abbreviate $\sum_{j=1}^p \mathbf{f}_j$ for $\mathbf{f}^t = (\mathbf{f}_1^t, \mathbf{f}_2^t, \dots, \mathbf{f}_p^t)$, $\mathbf{f}_j \in \mathbb{R}^n$. The orthogonal projection onto a subspace V will be denoted by H_V (the "hat" matrix in regression).

We will represent fitted functions at n points as n vectors $\mathbf{f}_j$. The unsubscripted $\mathbf{f}$ will denote the np vector consisting of the concatenation of p such fitted functions $\mathbf{f}_j$, and let $\mathbf{f}_+ = \sum_k \mathbf{f}_k$.

2. Linear smoothers.

2.1. *Definition and examples.* Suppose we have data of the form $(x_1, y_1), \dots, (x_n, y_n)$ and let $\mathbf{x} = (x_1, x_2, \dots, x_n)^t$, $\mathbf{y} = (y_1, y_2, \dots, y_n)^t$. A scatterplot smoother of $\mathbf{y}$ against $\mathbf{x}$ is a function $\mathscr{S}(x_0|\mathbf{x}, \mathbf{y})$ which at each x_0 estimates the dependence of $\mathbf{y}$ on $\mathbf{x}$. Often we are only interested in the fit at the observed x_i, in which case $\hat{y}_i = \mathscr{S}(x_i|\mathbf{x}, \mathbf{y})$. The smoother is *linear* if $\mathscr{S}(x_0|\mathbf{x}, \mathbf{y}_1 + a\mathbf{y}_2) = \mathscr{S}(x_0|\mathbf{x}, \mathbf{y}_1) + a\mathscr{S}(x_0|\mathbf{x}, \mathbf{y}_2)$. This in turn implies that $\mathscr{S}(x_0|\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n s(i, x_0, \mathbf{x}) y_i$ for some weights $s(i, x_0, \mathbf{x})$. Alternatively, if we are given a method for producing the fit $\hat{y}_i$, we can always extend it to a full scatterplot smoother using, for example, linear or quadratic interpolation or extrapolation.

For the remainder of this paper we will concentrate on the computation of the fit at the points x_i, in which case we can write a linear smoother as a linear map $S\colon \mathbb{R}^n \mapsto \mathbb{R}^n$ defined by $\hat{\mathbf{y}} = S\mathbf{y}$. The *smoother matrix* S depends on $x_1, x_2, \dots, x_n$ as well as the particular smoother, but not on $\mathbf{y}$.

Given a linear smoothing algorithm, we can produce the corresponding smoother matrix S by smoothing unit basis vectors: The result of smoothing the ith unit vector is the ith column of S [sometimes referred to as the "impulse response function"; see O'Sullivan (1986b)]. Note that this cannot be done for a nonlinear smoother since the estimates depend on $\mathbf{y}$ in a nonlinear way. The present paper is concerned with properties of these smoother matrices and their use in iterative procedures. Our focus does not imply that in practice we always prefer to use linear rather than nonlinear smoothers; we are simply dealing with the analytically more tractable situation.

The following examples contrast some of the properties of linear smoothers with which we will be concerned: (a) speed and simplicity of computation, (b) endpoint bias, (c) influence of individual points, (d) symmetry of the smoother matrix and (e) "shrinkage" properties.

Our "running example" will be a data set from meteorology, consisting of 330 observations and 9 covariates, analyzed by Breiman and Friedman (1985). The response is Ozone Concentration (ppm) and the goal is to investigate its relationship with a number of atmospheric measurements. For our purposes we will restrict attention to three covariates: Daggot Pressure Gradient (mm Hg), Inversion Base Height (ft) and Inversion Base Temperature (°F × 10).

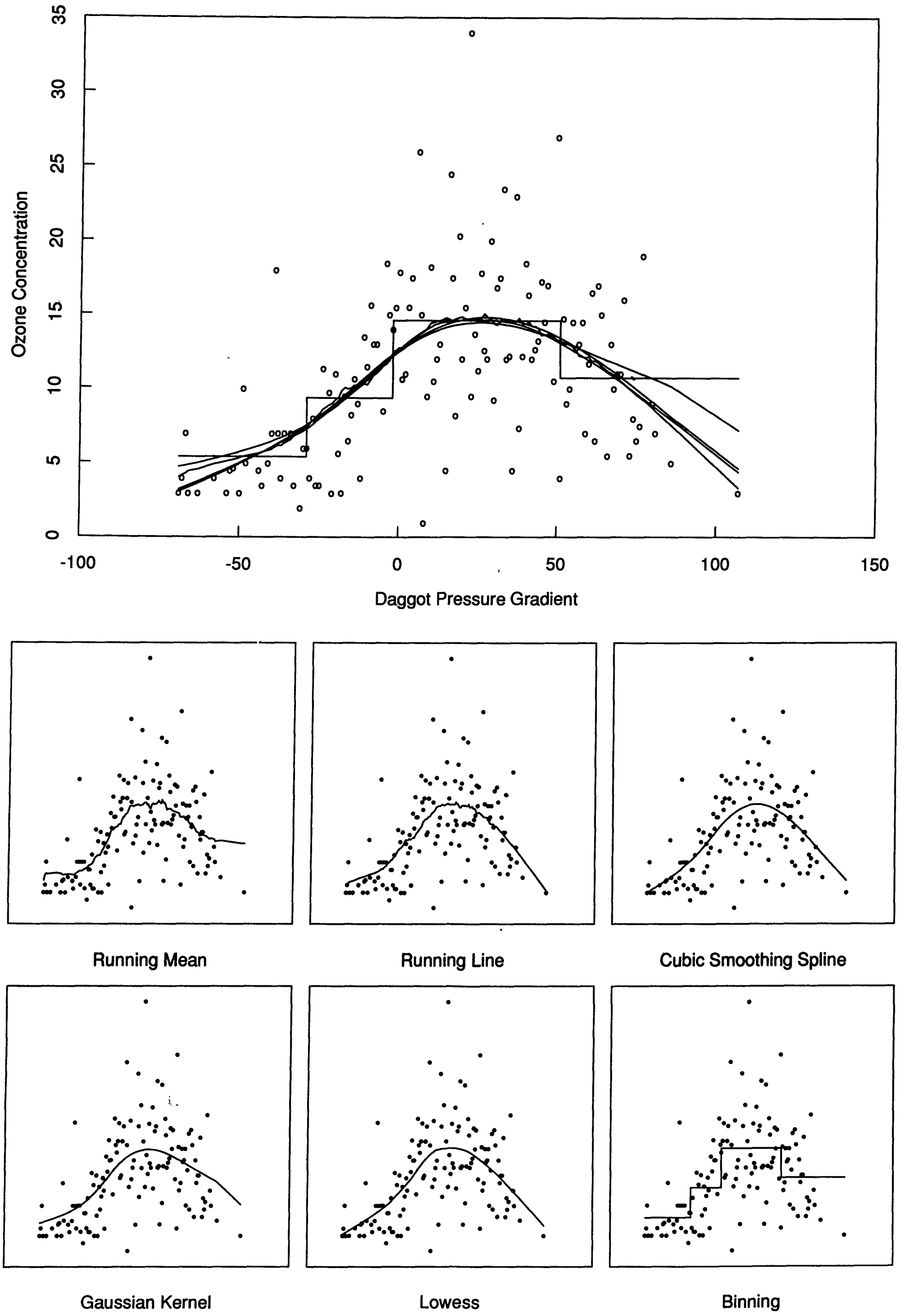
EXAMPLE 1. *Running-mean smoother.* A running-mean smoother produces a fit at x_i by averaging the data points in a *neighborhood* N_i around x_i. The neighborhoods that are commonly used are *symmetric nearest neighborhoods.* Assuming, for w between 0 and 1, that $[wn]$ is odd ($[\cdot]$ denoting integer part), these consist of $[wn]$ points, $([wn]-1)/2$ to the left and right of x_i plus x_i itself. The number w is called the *span* and controls the smoothness of the resultant estimate—larger spans tend to produce smoother functions. Assuming the data pairs are sorted by increasing x_i, a formal definition of the symmetric nearest neighborhood is

$$(3)\qquad \begin{aligned} N_i = \Biggl\{\max\left(i - \frac{[wn]-1}{2}, 1\right), \dots, i-1, i, i+1, \\ \dots, \min\left(i + \frac{[wn]-1}{2}, n\right)\Biggr\}. \end{aligned}$$

Notice that the neighborhoods are truncated near the endpoints if $([wn]-1)/2$ points are not available. Figure 1 shows the smoother matrix for a running mean of span 0.5, with $n = 10$.

$$\begin{pmatrix}
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3}
\end{pmatrix}$$

FIG. 1. *Smoother matrix for a running-mean smoother,* $n = 10$, *span* $= 0.5$. *Notice the truncated neighborhoods near the boundaries.*



Running Mean

Running Line

Cubic Smoothing Spline

Gaussian Kernel

Lowess

Binning

FIG. 2. *Plot of Ozone Concentration vs. Daggot Pressure Gradient along with various scatterplot smooths. The upper panel superimposes all the fits, whereas the lower panel identifies the individual curves.*

Figure 2(a) shows a plot of 128 values of Ozone Concentration and Daggot Pressure Gradient, from the meteorological data set (the original 330 data points were collapsed onto 128 points with unique values of Daggot Pressure Gradient for the demonstrations in this section). Also shown are a series of scatterplot smooths, one of which is a running mean [identified in Figure 2(b)] with a span of 0.27. Each smooth shown has a smoothing parameter, chosen in this plot so that they all do about the same amount of smoothing (see Section 2.7; in these examples the "degrees of freedom" was set at 4). Running-mean smoothers produce somewhat wiggly functions and are biased at the endpoints. This is the price to be paid for simplicity, speed and the local nature of the fit.

EXAMPLE 2. *Running-line smoother.* A running-line smoother fits a line by least squares to the data points in a symmetric nearest neighborhood N_i around each x_i. The estimated smooth at x_i is the value of the fitted line at x_i. This is done for each x_i. Figure 2 shows a running-line smooth of span 0.45. The running-line smoother is considered to be an improvement over the running mean because it reduces bias near the endpoints. Through the use of updating formulas, a running-line smoother can be computed with only $O(n)$ calculations (once the data are sorted). The running-line smoother matrix is also zero outside the banded diagonal, and the nonzero elements in the ith row are given by

$$s_{ij} = \frac{1}{n_i} + \frac{(x_i - \bar{x}_i)(x_j - \bar{x}_i)}{\sum_{k \in N_i}(x_k - \bar{x}_i)^2},$$

where n_i denotes the number of observations in the neighborhood of the ith point, j subscripts the points in this neighborhood and $\bar{x}_i$ denotes their mean.

The running-line smoother often produces quite jagged output. When used in an iterative procedure, it is often desirable to resmooth the final function. Alternatively, it can be modified to produce smoother output, at the cost of increased computations (see locally weighted running lines below).

EXAMPLE 3. *Bin smoother.* A bin smoother is similar to a running-mean smoother, the difference being that the average is computed in nonoverlapping neighborhoods. The data are partitioned into contiguous regions, each containing $[wn]$ data points (the rightmost region might contain fewer than $[wn]$ points). The fit for a given point is the mean of the points in its neighborhood. A bin smoother is not a very practical tool but is an example of an *orthogonal projection*. Figure 2 (step function) shows the result of a bin smoother, each partition consisting of one-fifth of the data.

EXAMPLE 4. *Simple and polynomial regression.* Denote by X the design matrix $(\mathbf{1}, \mathbf{x})$, and by H the matrix that projects onto the space spanned by the columns of X. The simple least-squares line is a linear smoother, with $S = H$. This S is an orthogonal projection matrix. Least-squares polynomial regression results in linear smoothing as well, with a smoother matrix which is the hat matrix of $X = (\mathbf{1}, \mathbf{x}, \mathbf{x}^2, \dots, \mathbf{x}^q)$; in this case, the order of the polynomial is the smoothing parameter. The use of monomials for polynomial regression has poor numerical properties however, and orthogonal polynomials are recommended instead.

EXAMPLE 5. *Cubic smoothing spline.* Consider the following minimization problem: Find g to minimize

$$\sum_1^n (y_i - g(x_i))^2 + \lambda \int_{-\infty}^{+\infty} [g''(z)]^2 \, dz \tag{4}$$

over the Sobolev space W_2 of functions with g' absolutely continuous and $g'' \in L_2$, where λ is a fixed tuning constant.

The solution $\hat{g}(x)$ is a cubic spline with knots at each distinct x_i [Reinsch (1967) and de Boor (1978)]. The constant λ plays the role of the smoothing parameter, trading off the smoothness of the curve with its closeness to the y values. When $\lambda = 0$, the solution is any interpolating function, while if $\lambda = +\infty$, the solution is the least-squares line. One can use other orders of derivatives in the penalty term, and these in turn generate different degree piecewise polynomials. Since cubic splines are by far the most popular, we will use the term smoothing spline to refer to cubic smoothing splines. One can also show that the smoothing spline is a linear smoother, and hence we can write down a smoother matrix. The following is taken from Green and Yandell (1985). Let $h_i = x_{i+1} - x_i$, $i = 1, 2, \dots, n-1$, Δ be a tridiagonal $(n-2) \times n$ matrix with $\Delta_{ii} = 1/h_i$, $\Delta_{i,i+1} = -(1/h_i + 1/h_{i+1})$, $\Delta_{i,i+2} = 1/h_{i+1}$ and let C be a symmetric tridiagonal matrix of order $n-2$ with $C_{i-1,i} = C_{i,i-1} = h_i/6$, $C_{ii} = (h_i + h_{i+1})/3$. Then if $\hat{y}_i = \hat{g}(x_i)$, it can be shown that solving (4) is equivalent to minimizing

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \lambda \hat{\mathbf{y}}^t K \hat{\mathbf{y}}, \tag{5}$$

where $K = \Delta^t C^{-1} \Delta$, with solution $\hat{\mathbf{y}} = S\mathbf{y}$, where $S = (I + \lambda K)^{-1}$. Figure 2 also shows a cubic spline smooth with λ chosen so that roughly the same amount of fitting is performed (degrees of freedom = 4).

If we isolate a particular region (in x) of the curve, there is some additional intuition in the objective function (4). If the data density is high in that region, the first term dominates (4) there, and the curve reflects its local behavior; if the data are sparse, the penalty term dominates and the function will be nearly linear in the region [Silverman (1984) and O'Sullivan (1986a)]. By exploiting the banded nature of the matrices involved, cubic spline smoothing takes $O(n)$ operations to compute; in fact approximately $35n$ [Silverman (1985)] vs. $7n$ for the running-line smoother.

EXAMPLE 6. *Regression spline.* Regression spline smoothing [de Boor (1978), Stone and Koo (1985) and Ramsay (1988)] is a projection method for fitting splines. In this case S is a projection onto k basis or B-splines placed at judiciously chosen *knots* in the range of x. The number k and positions of the knots are all parameters of the procedure. We find fixed knot cubic splines less appealing than their immediate competitors, smoothing splines. Although k, the number of knots, is usually considered to be the smoothing parameter, one has also to determine the placement of the knots. The nature of these "parameters" makes it hard to vary the smoothness of the resultant estimate in a somewhat continuous fashion [Hastie and Tibshirani (1988)].

EXAMPLE 7. *Kernel smoother.* The kernel smoother matrix has elements $s_{ij} = c_i d_\lambda(x_i, x_j)$, where d is an inverse distance measure, λ is the *window size* and c_i is chosen so that the rows sum to unity. An example is the Gaussian kernel with

$$d_\lambda(x_i, x_j) = \exp\left(-\left(\frac{x_i - x_j}{\lambda}\right)^2\right).$$

Kernel smoothers are expensive to compute [$O(n^2)$ for the whole sequence], but are visually smooth if the kernel is smooth. A Gaussian kernel smooth with $\lambda = 0.13$ is shown in Figure 2, and we notice that it also has bias problems at the ends. We could easily correct this by using running lines, weighted by a Gaussian kernel. Some key references for kernel smoothers are Rosenblatt (1971), Priestley and Chao (1972) and Härdle (1987) who also provides an $O(n \log n)$ approximation for computing a kernel smooth.

EXAMPLE 8. *Locally weighted running-line smoother.* This smoother combines the strict local nature of running lines, and the smooth weights of kernel smoothers, in a locally weighted running-line smoother. Cleveland's (1979) implementation ("LOWESS") uses the tricube weight function in each neighborhood. Specifically, if h_i is the distance to the wnth nearest neighbor, then the points x_j in the neighborhood get weights $w_{ij} = (1 - |(x_j - x_i)/h_i|^3)^3$. The fit at point i is then computed by a weighted least-squares straight line, using these weights on the points in the neighborhood. Since the weights have to be recomputed for each neighborhood, locally weighted running-line smoothers require $O(n^2)$ computations; however, the current implementation of LOWESS in the S language [Becker and Chambers (1984)] computes the running lines at a default (50) number of "knots," and evaluates the fits at other points by interpolation. This makes it $O(n)$ to compute, like splines and running lines.

LOWESS has a robustness option which can be used to downweight outlying responses, which, when used, causes it to be a nonlinear smoother. On a more subtle note, LOWESS uses nearest neighbors, whereas the running means and lines described earlier use *symmetric* nearest neighbors. Without smoothness weights, running-line fits tend to be jagged; the symmetric neighborhoods tend to alleviate this (Werner Stuetzle, personal communication).

2.2. *Choice of smoothing parameters.* Many of the aforementioned smoothers require a choice of a smoothing parameter. The running-mean and running-line smoothers rely on a span size w, the cubic spline smoother has a penalty factor λ and the kernel smoother has an inverse penalty factor, also λ. In practice these parameters are chosen either a priori, through visual inspection of the curve, or by an automatic method such as cross-validation. Some details may be found in Silverman (1985) and Craven and Wahba (1979). If the smoothing parameter is chosen a priori, then the resultant smoothers are linear for the above examples. If the smoothing parameter is selected by using the y-values as in cross-validation, the smoothers are (strictly speaking) nonlinear, and the results developed here for linear smoothers do not apply. This paper avoids the issue entirely, and we assume the smoothing parameters are known and fixed. In Section 2.7 we suggest a linear method for fixing the degree of smoothing.

2.3. *Smoother matrix plots*. One way to compare the various linear smoothers is to plot the rows of their smoother matrices against $\mathbf{x}$ [e.g., see Silverman (1984), O'Sullivan (1986b) and Rice (1986)]. These plots, also known as the equivalent kernels, show explicitly the form of neighborhoods used and the weighting function. Figure 3 gives such a display for some of the smoothers used in Figure 2, and plots rows 1, 20 and 64 of the 128 × 128 smoother matrices.

The rectangular windows of the running-mean and running-line smoothers account for the discontinuous appearance of their output. Locally weighted running lines appear to be a hybrid, mixing kernels and nearest neighbors. Spline fits have global support, and splines, running lines and locally weighted running lines can have negative weights. The slope of the "roof" for the running-line smoother can change frequently, depending on which side of the neighborhood mean $\bar{x}_i$ the target point x_i falls.

Spline smoother matrices are symmetric, a property we will find increasingly useful in later sections. One consequence of this symmetry together with the fact that splines reproduce lines is that we get the global least-squares line whether we fit a straight line to the original data or to the spline smoothed data: $HS\mathbf{y} = (S^tH^t)^t\mathbf{y} = (SH)^t\mathbf{y} = H^t\mathbf{y} = H\mathbf{y}$, where H is the least-squares "hat" matrix for straight-line regression. This is not true of nonsymmetric smoothers such as the running-line and locally weighted running-line smoothers (although one can correct them to ensure this property if desired).

Figure 4 contains "self-influence" plots for various smoothers. These show the diagonal elements of the smoother matrix as a function of $\mathbf{x}$, and are especially useful for understanding the endpoint behavior. We see that self-influence is very small except near the endpoints of the data, where the running lines have the most self-influence. It is clear that as the endpoint becomes more separated from the rest of the data, the self-influence of the endpoint for running lines and locally weighted running lines can approach 1. The self-influence for running means, however, does not depend on the data ($1/n_i$ for the ith neighborhood).

2.4. *Eigenvalue and singular value decompositions of a smoother matrix.* For a symmetric smoother, the eigendecomposition of S can be used to describe the smoother based on S [Demmler and Reinsch (1975); see also O'Sullivan (1986b) and Utreras (1979)]. This is much like the use of a *transfer function* or *spectrum* to describe a linear filter for time series. Let $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ be an orthonormal basis of eigenvectors of S with eigenvalues $\theta_1 \geq \theta_2 \cdots \geq \theta_n$,

$$S\mathbf{u}_i = \theta_i\mathbf{u}_i, \qquad i = 1, 2, \dots, n, \tag{6}$$

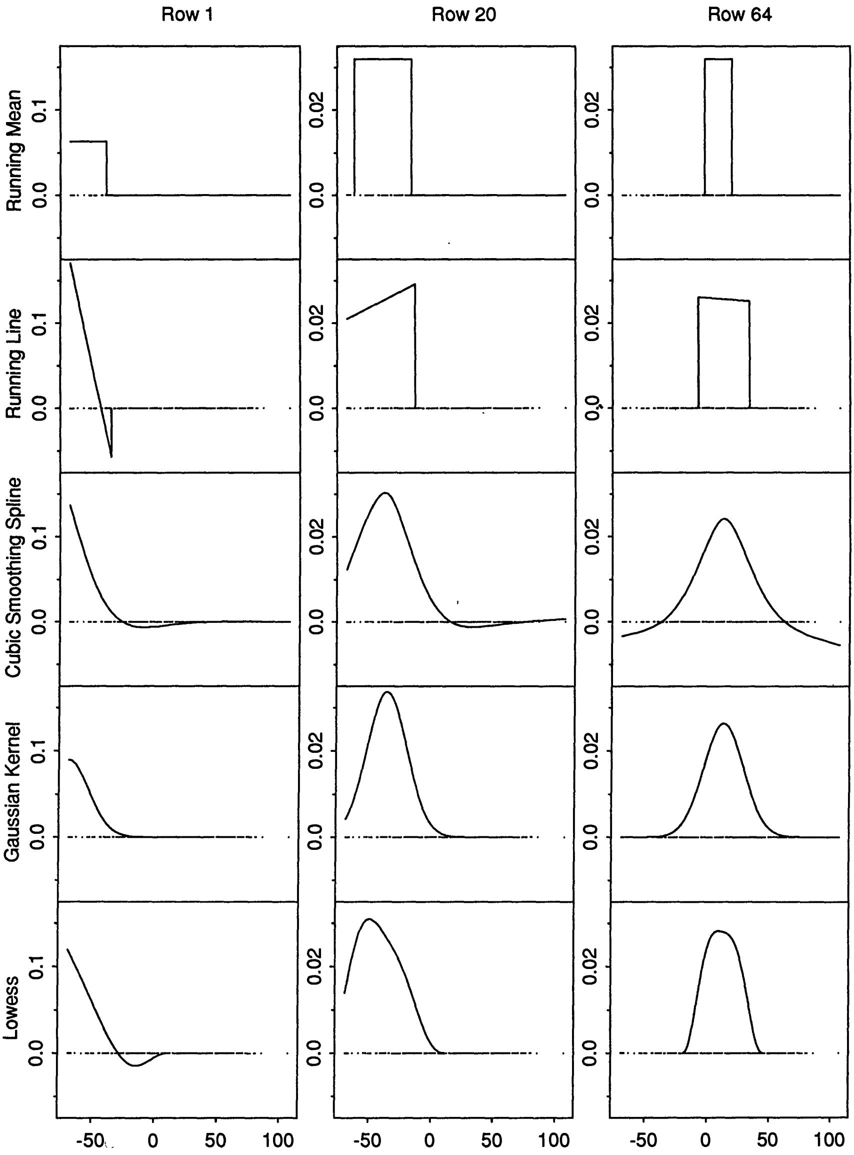


FIG. 3. *Selected rows of the smoother matrix for a variety of linear smoothers. These are the matrices used to smooth the air pollution data in Figure 2. For the ith row, we graph $s(j, x_i, \mathbf{x})$ against x_j. Each column of figures is plotted on the same scale, and all the smoothers are calibrated so that the overall amount of smoothing is approximately the same.*

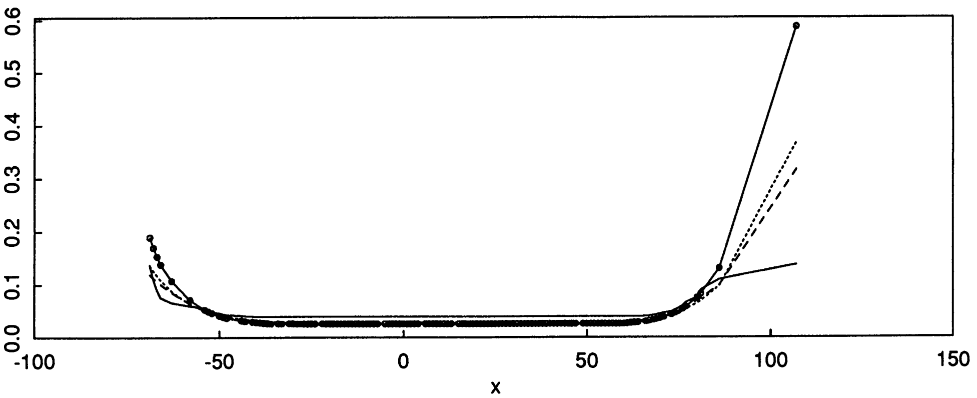


FIG. 4. *Self-influence plots for the running lines (solid), splines (dashes), locally weighted running lines (dots) and running means (solid—lowest).*

or

$$S = UD_\theta U^t. \tag{7}$$

The cubic spline smoother is an important example of a symmetric smoother, and its eigenvectors look approximately like polynomials of increasing degree [see also Eubank (1984) and Mallows (1986)]. In particular, it is easy to show that the first two eigenvalues are 1, with eigenvectors which correspond to linear functions of x. Figure 5(a) shows the eigenvalues for the cubic spline smoother used in Figure 2. Figure 5(b) shows the third to sixth eigenvectors. Demmler and Reinsch (1975) give some theoretical support for these empirical findings. They show that for $k \geq 3$, the number of sign changes in the kth eigenvector of a cubic spline smoother is $k - 1$. They also derive asymptotic approximations for the eigenvalues which show that they decrease fairly rapidly with increasing order.

The bin smoother, least-squares line, polynomial regression and regression splines are other symmetric smoothers which we have discussed. They are all in fact orthogonal projections arising from different spaces of fits. Thus their eigenvalues are 0 or 1 only, with corresponding eigenspaces consisting of the spaces of residuals and fits, respectively. The smoother or projection matrices are the familiar hat matrices of one-way analysis of variance and simple and multiple regression.

What if S is not symmetric? Then the eigendecomposition is no longer useful because the eigenvalues and vectors may be complex, and algebraic and geometric multiplicities may differ. One can, however, turn to the singular value decomposition of S, which is always real. Figure 5(a) also includes the series of singular values for both the running-line and locally weighted running-line smoothers. The smoothing parameters for these smoothers were chosen such that they all do approximately the same amount of smoothing (in terms of total

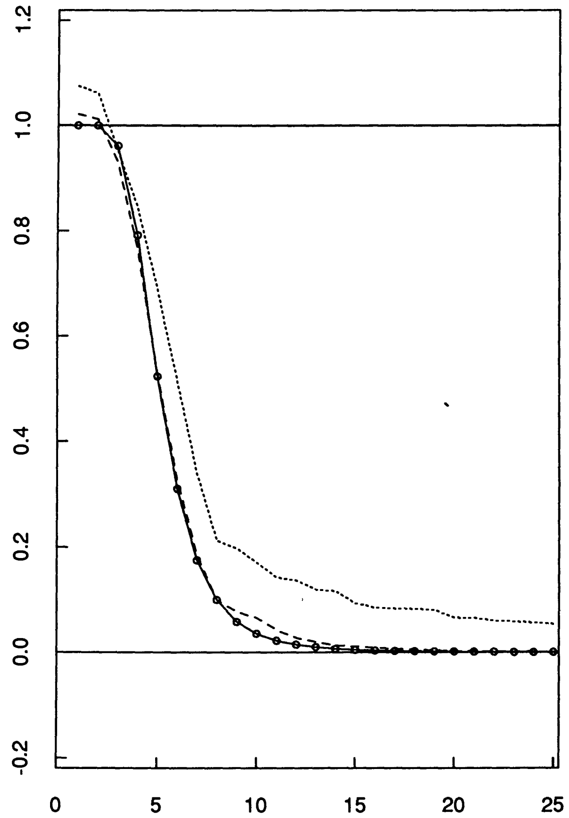


FIG. 5(a). *The first* 25 *ordered eigenvalues for the* (*symmetric*) *cubic spline smoother used in Figure* 2 (*solid curve with dots*). *Also shown are the first* 25 *singular values for the running-line* (*dotted*) *and locally weighted running-line* (*dashed*) *smoothers.*

variance of the fit). We will see later that this implies that tr(SS^t) is the same for all the smoothers (approximately 4 for all the smoothers in Figure 2), and thus the sum of squares of the eigen/singular values is 4. The most noteworthy feature of this singular value decomposition is the largest singular value. For the running-line smoother used in Figure 2, this is 1.065. Since this value is larger than one, the singular value decomposition tells us that running-line smoothers (including locally weighted running lines) are not members of the class of *shrinking* smoothers, which we discuss in the next section. Since the running-line output is typically rough, we are not surprised that the singular values approach 0 more slowly than for splines. In fact, if one smooths a genuinely smooth curve, such as a cubic polynomial, the running-line smoother can put wiggles in the output! (This is due to the discrete nature of the neighborhoods.) For the running-line smoother, the first two (left and right) singular vectors are approxi-

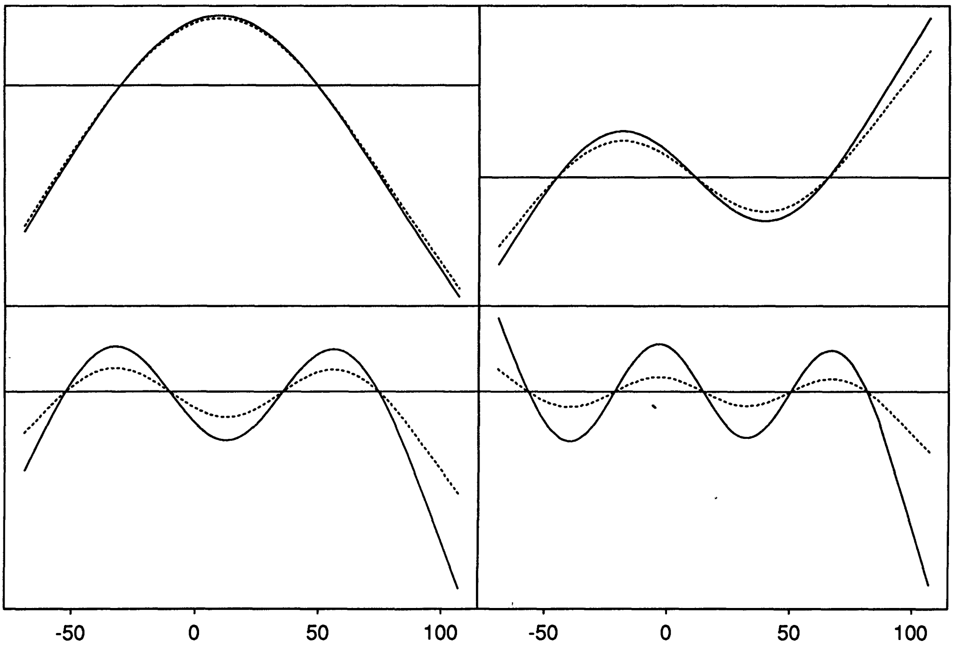


FIG. 5(b). *The eigenvectors corresponding to the third to sixth eigenvalues of the cubic spline smoother used in Figure 2 (since constants and linear are not shrunk, they are omitted). The dotted curves are the smoothed functions, and demonstrate the amount of shrinking.*

mately linear, while the remaining ones are approximately orthogonal polynomials of increasing degree (not shown).

2.5. *Shrinking and strictly shrinking smoothers.* We call the smoother based on S "shrinking" if $||S\mathbf{y}|| \leq ||\mathbf{y}||$ and "strictly shrinking" if $||S\mathbf{y}|| < ||\mathbf{y}||$ for all $\mathbf{y}$. This will be the case if all of its singular values are ≤ 1 and < 1, respectively. We use the Euclidean norm, but other vector norms could be used. In the second part of this paper we discuss procedures that use smoothers iteratively, and show in some cases that their convergence is guaranteed if shrinking or strictly shrinking smoothers are used.

In some special cases we can give conditions to guarantee that S is shrinking. For example, if S is doubly stochastic (all elements nonnegative, rows and columns sum to 1), then it is easy to show by Jensen's inequality that S is shrinking. Therefore symmetric smoothers with nonnegative elements are shrinking if each row adds to 1. We saw earlier, however, that not many smoothers are likely to have smoother matrices with exclusively nonnegative elements, so this condition is limited. A case in point is the cubic spline smoother; its smoother matrix typically has negative elements. We can show directly, however, that a cubic spline smoother matrix has real positive eigenvalues less than or equal to 1 (and hence is shrinking) and furthermore, $\|S\mathbf{y}\| < \|\mathbf{y}\|$ unless $\mathbf{y}$ is a linear function of $\mathbf{x}$. We can verify this through the representation $S = (I + \lambda\Delta^t C^{-1}\Delta)^{-1}$, where C, λ and Δ are defined in Example 5. First note that C is positive definite since it is diagonally dominant [Golub and van Loan (1983), page 7]. Thus C^{-1} exists, $\Delta^t C^{-1}\Delta$ is nonnegative definite and hence $(I + \lambda\Delta^t C^{-1}\Delta)^{-1}$ has eigenvalues ≤ 1. Now $\|\mathbf{y}\| = \|S\mathbf{y}\|$ can only hold if $S\mathbf{y} = \mathbf{y}$ since S has nonnegative eigenvalues. Suppose then that $(I + \lambda\Delta^t C^{-1}\Delta)^{-1}\mathbf{y} = \mathbf{y}$. Then $\Delta^t C^{-1}\Delta\mathbf{y} = 0$, $\mathbf{y}^t\Delta^t C^{-1}\Delta\mathbf{y} = 0$ and thus $\Delta\mathbf{y} = 0$ (since C and hence C^{-1} are positive definite). Now Δ takes second differences and hence $\mathbf{y}$ must be a linear function of $\mathbf{x}$. This result is also contained in Craven and Wahba (1979), Lemma 4.3.

2.6. *Smoothers and penalized least squares.* The minimization problem (4) leading to cubic spline smoothing can be reexpressed in terms of a quadratic penalty function,

$$\|\mathbf{y} - \mathbf{f}\|^2 + \lambda\mathbf{f}^t K\mathbf{f}, \tag{8}$$

where $K = \Delta^t C^{-1}\Delta$ as defined in Example 5. The solution to (8) is $\hat{f}_i = \hat{g}(x_i)$, where $\hat{g}(x)$ is the minimizer of (4). This leads us to ask: Is there a larger class of smoothers which can be characterized as solutions to penalized least-squares problems? The penalization term $\lambda\mathbf{f}^t K\mathbf{f}$ depends only on the symmetric part of K, since it is a quadratic form. Hence only symmetric penalization matrices K should be considered. Assuming that inverses exist, the stationarity condition for (8) implies $\mathbf{f} = (I + \lambda K)^{-1}\mathbf{y}$, that is, $S = (I + \lambda K)^{-1}$. We see that only symmetric smoothers S can be obtained by penalized least squares. Conversely, given a symmetric invertible smoother matrix S, we can characterize $\mathbf{f} = S\mathbf{y}$ as a stationary solution of

$$\|\mathbf{y} - \mathbf{f}\|^2 + \mathbf{f}^t(S^{-1} - I)\mathbf{f}. \tag{9}$$

In order to cover noninvertible smoothers as well, we have to resort to a linear constraint. Let S be an arbitrary symmetric matrix with range $\mathscr{R}(S)$ and nullspace $\mathscr{N}(S)$, and let S^- be some generalized inverse ($SS^-S = S$). In this case we can obtain $\mathbf{f} = S\mathbf{y}$ as a stationary solution of

$$Q(\mathbf{f}) = \|\mathbf{y} - \mathbf{f}\|^2 + \mathbf{f}^t(S^- - I)\mathbf{f} \tag{10}$$

under the constraint $\mathbf{f} \in \mathscr{R}(S)$.

It is illuminating to work with the eigendecomposition of S, although a coordinate-free approach using directional derivatives confined to $\mathscr{R}(S)$ is equally straightforward. Let $S = UDU^t$, and partition $U = (U_1\colon U_2)$ where U_2 corresponds to the zero eigenvalues of S. Thus $S = U_1 D_\theta U_1^t$, where D_θ is diagonal of size $\dim(\mathscr{R}(S))$. The matrix U_1 spans $\mathscr{R}(S)$, and we can represent any $\mathbf{f} \in \mathscr{R}(S)$ as $\mathbf{f} = U_1\boldsymbol{\beta}$ for some vector $\boldsymbol{\beta}$ of length $\dim(\mathscr{R}(S))$. Any generalized inverse can be written as $U_1 D_\theta^{-1} U_1^t + L$, where L operates in $\mathscr{N}(S)$. Then (10) reduces to

$$Q = \|\mathbf{y} - U_1\boldsymbol{\beta}\|^2 + \boldsymbol{\beta}^t U_1^t\big(U_1 D_\theta^{-1} U_1^t - I\big)U_1\boldsymbol{\beta}$$

with stationarity condition $dQ/d\boldsymbol{\beta} = \mathbf{0} \Rightarrow \boldsymbol{\beta} = D_\theta U_1^t\mathbf{y}$, or $\mathbf{f} = U_1 D_\theta U_1^t\mathbf{y} = S\mathbf{y}$.

We should add a warning: The characterization of smooths has been expressed in terms of *stationarity* w.r.t. the penalized least-squares criteria. Such smooths are *minimizers* only under additional conditions on the smoother matrices. Writing

$$Q(\mathbf{f}) = \|\mathbf{y}\|^2 - 2\mathbf{y}^t\mathbf{f} + \mathbf{f}^t S^- \mathbf{f},$$

we see that the quadratic behavior of the criterion is solely determined by $\mathbf{f}^t S^- \mathbf{f}$. The stationary solution $\mathbf{f} = S\mathbf{y}$ becomes a minimizer iff S is nonnegative definite, since S^- as a linear map is positive definite on $\mathscr{R}(S)$ in this case. If S is indefinite, we have the curious situation that $\mathbf{f} = S\mathbf{y}$ minimizes the penalized least-squares criterion in some directions, but maximizes it in others, that is, the stationary solution is a true saddle point.

The eigenvalues of $S^- - I$ in $\mathscr{R}(S)$ are $1/\theta_i - 1$, where $\theta_i \neq 0$ with eigenvectors the same as those of S. Hence the term $\mathbf{f}^t(S^- - I)\mathbf{f}$ does not penalize in directions contained in $\mathscr{M}_1(S)$. If, for example, the eigenvectors of S are orthogonal polynomials of increasing degree with first two eigenvalues 1, and the rest decreasing from 1 to 0, the criterion does not penalize the constant and linear components, but puts an increasing penalty on the higher-order polynomial components. From this we also see that if $K = S^- - I$ is positive semidefinite, then the resulting smoother will be shrinking.

We can go a step further and make explicit a smoothing parameter in (10):

$$Q_\lambda(\mathbf{f}) = \|\mathbf{y} - \mathbf{f}\|^2 + \lambda\mathbf{f}^t(S^- - I)\mathbf{f}, \tag{11}$$

with the stationary solution $\mathbf{f} = S^{(\lambda)}\mathbf{y}$, where $S^{(\lambda)}$ has the same eigenvectors as S but eigenvalues

$$\mu_i^{(\lambda)} = \frac{1}{1 + \lambda\left(\dfrac{1}{\theta_i} - 1\right)}$$

for eigenvalues $\theta_i > 0$ of S. In addition $S^{(\lambda)}$ has the properties that

$$S^{(\lambda)} \underset{\infty}{\overset{\lambda}{\rightarrow}} H_{\mathscr{M}_1(S)},$$

$$S^{(\lambda)} \underset{0}{\overset{\lambda}{\rightarrow}} H_{\mathscr{R}(S)} = S^{(0)},$$

where H_M denotes the projection onto the subspace M.

In a trivial sense ordinary least squares is just a special case of penalized constrained least squares. If S is an orthogonal projection (hat matrix), then $\mathscr{R}(S)$ is just the space of fits and the penalization term vanishes: Since S is its own generalized inverse, we have $\mathbf{f}^t(S^- - I)\mathbf{f} = \mathbf{0}$ for $\mathbf{f} \in \mathscr{R}(S)$, and thus the problem reduces to ordinary least squares,

$$\min_{f \in \mathscr{R}(S)} \|\mathbf{y} - \mathbf{f}\|^2.$$

A final remark concerns the generality of the penalized constrained least-squares approach. Although it helps our intuition to think of S as a smoother matrix, all that matters is that S represents a symmetric linear mapping, and many other applications unrelated to smoothing are conceivable.

2.7. *Remarks on inference for smoothers.* In this section we discuss consistency, the variance and the number of parameters or "degrees of freedom" of the fitted smooth. We have already made use of the latter quantity in order to render different smoothers comparable with respect to the amount of "fitting" they do; both of these notions are also useful when a smoother is used in a data analysis. For the remainder of this section we assume that $y_i = f_i + \varepsilon_i$, where f_i is the true function and the errors ε_i are uncorrelated with zero expectation and common variance σ^2.

2.7.1. *Bias and consistency.* An interesting question arises in smoothing situations: What is being estimated? All the smoothers we consider are biased for arbitrary f, and we can represent the bias at the n sample $x_i s$ by $\mathbf{b} = \mathbf{f} - S\mathbf{f}$. They will be unbiased for a restricted class of functions, for example, cubic smoothing splines and running lines are exactly unbiased for linear functions. Linear least-squares estimates, by comparison, are unbiased for the elements of their spaces of fits.

One approach is to define the estimand as the expected value of the estimator. With this definition, different smoothers might be estimating different quantities in a given problem. Another approach is to consider what happens asymptotically. If the smoothing parameters are held fixed, asymptotic bias will generally result. If however the amount of smoothing is decreased at an appropriate rate as we approach the limit, then under regularity conditions, the estimates should be consistent for the underlying functions.

We do not go into details here. See Stone (1977) for consistency results for nearest-neighbor-type smoothers. Results for cubic splines are given by many authors, for example Cox (1983) and Rice and Rosenblatt (1983), while results for kernel smoothers are derived, for example, by Gasser and Müller (1979).

2.7.2. *Variance.* The covariance matrix of the fits $\hat{\mathbf{y}} = S\mathbf{y}$ is simply

$$\operatorname{cov}\hat{\mathbf{y}} = SS^t\sigma^2. \tag{12}$$

Under normality assumptions, this can be used to form pointwise *standard error* bands for the estimated smooth (see Figure 7, section 3). These standard error bands should not be confused with confidence bands, since they apparently contain no information on the bias of $\hat{\mathbf{y}}$; they are confidence bands for what the smoother is estimating, that is, $E(S\mathbf{y})$. We do have to estimate σ^2, however, and if it is based on the residual sum of squares, it will be biased (see below). So, perversely, the standard error bands are spread out due to this bias, although in an average sense. Wahba (1983) discusses a Bayesian approach to confidence bands for smoothing splines in some detail, and gives a frequentist interpretation.

2.7.3. *Degrees of freedom.* Given an estimate $\mathbf{y}$, it would be useful to know how many "degrees of freedom" we have fitted to the data, a notion borrowed from parametric linear regression. It is not immediately clear how to quantify this notion. There are at least three possible definitions of degrees of freedom depending on the context in which it is to be used. All three are derived by analogy to the linear regression model.

DEFINITION 1. Degrees of freedom $= \operatorname{tr}(SS^t)$. For the linear model, $\Sigma \operatorname{var}(\hat{y}_i) = p\sigma^2$, where degrees of freedom $= p$ is the number of parameters. The analogous definition for smoothers is degrees of freedom $= \operatorname{tr}(SS^t)$. This is the definition we have used to calibrate the various smoothers in the previous sections with regard to the choice of smoothing parameter, where by trial and error we achieved $\operatorname{tr}(SS^t) \approx 4$. The more "parameters" we fit, the rougher will be the function and the higher its variance.

DEFINITION 2. Degrees of freedom $= \operatorname{tr}(2S - S^tS)$. The residual sum of squares RSS $= (\mathbf{y} - \hat{\mathbf{y}})^t(\mathbf{y} - \hat{\mathbf{y}})$ has expectation

$$E(\text{RSS}) = [n - \operatorname{tr}(2S - S^tS)]\sigma^2 + \mathbf{f}^t(I - S)^t(I - S)\mathbf{f}, \tag{13}$$

where the last term measures bias. Here we are motivated to define degrees of freedom $= \operatorname{tr}(2S - S^tS)$ since once again in the linear regression case this is p. If we are smoothing noise ($\mathbf{f} = \mathbf{0}$), then degrees of freedom corresponds to the expected drop in the RSS due to overfit.

This definition is useful when comparing two smooths. Suppose that we have two fitted responses $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$, say $\hat{\mathbf{y}}_1 = S_1\mathbf{y} = H\mathbf{y}$, the linear fit, and $\hat{\mathbf{y}}_2 = S_2\mathbf{y}$, a running-line smooth at some span. Then we ask the following: Given that the fit based on smoother 1 is adequate, what is the expected decrease in RSS due to fitting the second smooth? Our "null hypothesis" says the "model" $E(\mathbf{y}) = a\mathbf{x} + b$ is correct.

Letting RSS_1 and RSS_2 be the residual sum of squares for the two fitted responses, it is clear from (13) that

$$E[\text{RSS}_1 - \text{RSS}_2] = [\operatorname{tr}(2S_2 - S_2^tS_2) - \operatorname{tr}(2S_1 - S_1^tS_1)]\sigma^2. \tag{14}$$

Thus in comparing two smooths, we can compare the decrease in RSS due to fitting a more complex smooth with the increase in the degrees of freedom $\operatorname{tr}(2S_i - S_i^tS_i)$, in units of σ^2.

This definition involves only the expectation of the residual sum of squares. It turns out that the distribution of the residual sum of squares is not χ^2 as in the linear case, but fairly close to it. For results on χ^2 approximations to the RSS see Cleveland (1979), Cleveland and Devlin (1988) and Tibshirani and Hastie (1987). Devlin (1986) explores these issues in considerably more detail. She considers using RSS terms from different smoother-based models in approximate F-tests, and explores two moment corrections for the relevant distributions.

Definition 3. Degrees of freedom = tr(S). One interpretation of the C_p statistic [Mallows (1973)] is that it corrects RSS to make it unbiased for the true MSE for prediction by adding a quantity $2p\hat{\sigma}^2$, where $\hat{\sigma}^2$ is an unbiased estimate of σ^2 and p is once again the number of parameters. The appropriate number for smoothers in this context is degrees of freedom = tr(S). This is the popular definition in the spline smoothing literature [Green and Yandell (1985), O'Sullivan, Yandell and Raynor (1986) and Silverman (1985)], where $S\sigma^2$ emerges as the posterior covariance of $\hat{\mathbf{y}}$, after appropriate Bayesian assumptions are made.

For symmetric smoothers with eigenvalues θ_i, the following relationships are immediate:

$$\operatorname{tr}(S) = \sum_{i=1}^{n} \theta_i,$$

$$\operatorname{tr}(SS^t) = \sum_{i=1}^{n} \theta_i^2,$$

$$\operatorname{tr}(2S - S^tS) = \sum_{i=1}^{n} \left(2\theta_i - \theta_i^2\right).$$

Consequently, for symmetric shrinking smoothers with nonnegative eigenvalues $\operatorname{tr}(SS^t) \le \operatorname{tr}(S) \le \operatorname{tr}(2S - S^tS)$. Since tr($S$) is easiest to compute, it may be the logical choice if a single parameter is desired.

Any of the above degrees-of-freedom measures can be used to determine a value for the smoothing parameter, and will produce a curve using roughly that many degrees of freedom. This provides a reasonable method for calibrating smoothing parameters amongst a class of smoothers, and gives a useful a priori choice in situations where automatic choice is not feasible.

Empirically, we have found that there is a relationship that seems to hold approximately for running-line smoothers (for which the three definitions of degrees of freedom coincide): $1/\text{span} + 1 \le$ degrees of freedom $\le 1/\text{span} + 2$.

3. The additive model.

3.1. *Introduction.* So far we have discussed scatterplot smoothers for a response and a single predictor. When there are two or more predictors, there are a number of possibilities for estimating the regression surface. Probably the most straightforward is through the use of a p-dimensional scatterplot smoother. Cleveland and Devlin (1988) discuss the multidimensional extension of locally weighted running-line smoothers. However, there are a number of problems associated with p-dimensional smoothers:

1. The "curse of dimensionality" [Friedman and Stuetzle (1981)]. When p is large, the neighborhoods are less local for a fixed span than for a single variable smoother and hence large biases can result.

2. When finding neighborhoods in two or more dimensions, there is usually some metric assumption made which can be hard to justify when the variables are measured in different units or are highly correlated.
3. The multivariate versions of all the smoothers mentioned in Section 2 are expensive [$\geq O(n^2)$ operations] to compute.

We take a different approach and use the one-dimensional smoother as a building block for a restricted class of nonparametric multiple regression models. Suppose our data consist of n realizations of random variable Y at p design values, denoted by $\{(y_1, x_{11}, x_{12}, \dots, x_{1p}), \dots, (y_n, x_{n1}, x_{n2}, \dots, x_{np})\}$. Then the *additive model* takes the form

$$E(Y_i | x_{i1}, \dots, x_{ip}) = \sum_{j=1}^{p} f_j(x_{ij}). \tag{15}$$

This model is a special case of both the PPR (projection pursuit regression) model proposed by Friedman and Stuetzle (1981), the ALS (alternating least squares) model of van der Burg and de Leeuw (1983) and the ACE (alternating conditional expectation) model of Breiman and Friedman (1985). Wahba (1986) refers to additive models as main effect partial splines. The additive model avoids all the pitfalls of the p-dimensional smoother listed above—at the cost of approximation errors in using an additive function to model the p-dimensional surface.

The additive model has a stronger motivation as a useful data analytic tool. Since each variable is represented separately in (15), the model retains an important interpretation feature of the linear model: The nature of the effect of a variable on the response surface does not depend on the values of the other variables. In practice this means that once the additive model is fitted to data, we can plot the p coordinate functions separately to examine the roles of the variables in predicting the response.

The additive model can be fitted by an algorithm which consists of estimating each smooth holding all the others fixed, then cycling through this process. Thus if the current estimates are $\hat{f}_k$, $k = 1, \dots, p$, then $\hat{f}_j$ is updated by smoothing the *partial residuals* $r_{ij} = y_i - \sum_{k \neq j} \hat{f}_k(x_{ik})$ against x_{ij}. The procedure implementing this idea is called the *backfitting* algorithm [Friedman and Stuetzle (1981)], described later.

Figure 6 is a scatterplot matrix which shows the relationship between Ozone Concentration and Daggot Pressure Gradient, as well as two additional variables, Inversion Base Temperature and Inversion Base Height discussed earlier. Figures 7(a), (b) and (c) show the functions produced by backfitting Ozone Concentration jointly on these variables. All the curves have been centered at 0. The Daggot Pressure Gradient curve looks much like it did for the univariate smooths (Figure 2). This is not surprising, since it does not exhibit a strong relationship with either of the other two covariates (Figure 6). This is not the case with the other two covariates. Figure 8(a) shows the univariate spline smoother together with the additive model fit for both Inversion Base Temperature and Inversion Base Height. For Inversion Base Height the two fits are quite

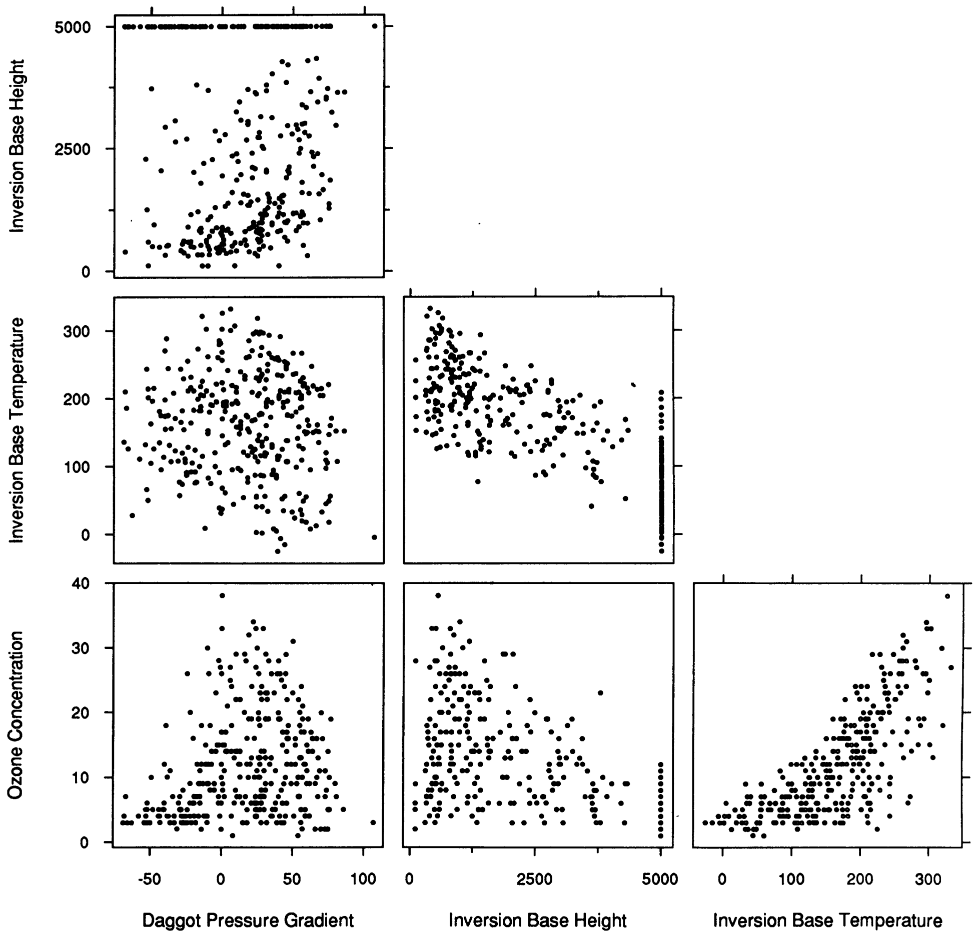


FIG. 6. *Scatterplot matrix of the Ozone Concentration data with three covariates: Daggot Pressure Gradient, Inversion Base Height and Inversion Base Temperature.*

different. Since the two variables are negatively correlated, it seems possible that height is acting as a surrogate for temperature when temperature is not in the model. In Section 4.2 we discuss convergence of backfitting using these data. We deliberately entered Inversion Base Height before temperature; consequently some iteration was needed to change the fitted function. Figure 8(b) shows the fitted functions on the same scale so that the relative strengths of the effects can be compared. We see that Inversion Base Temperature exhibits the strongest effect. It is also possible to perform crude F-tests to judge the importance of variables—see Cleveland and Devlin (1988) and Hastie and Tibshirani (1987). The broken lines in the figures are pointwise $2\times$ estimated standard error curves, based on the estimating equations of the full additive fit. They will reflect high variance regions of the fitted curve, which could be a result of sparse

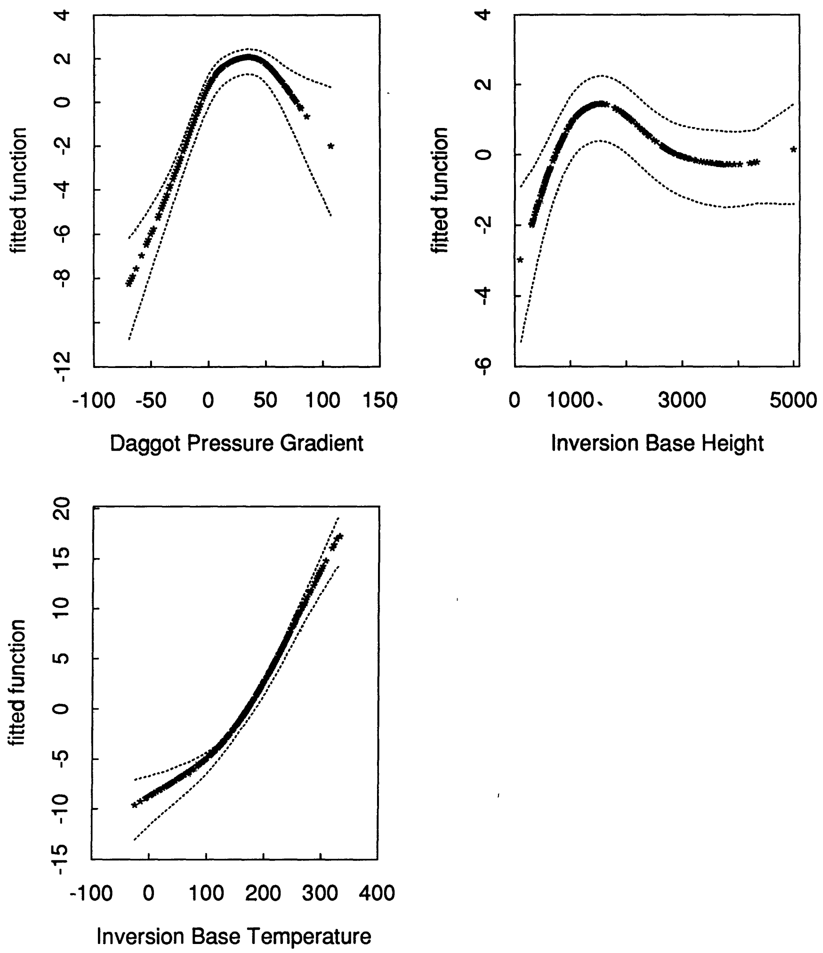


FIG. 7. *Estimated functions for the three covariates in the additive model (solid curves). The broken lines in the plots are curves representing the fitted curve* ±2 × *its estimated standard error. A cubic smoothing spline was used for all three terms, and the smoothing parameters were chosen so that the univariate degrees of freedom* $\mathrm{tr}(S_j^t S_j) \approx 4$.

marginal data there, or interactions with other variables. We discuss these standard errors and methods for calculating them in more detail in Section 5. We will see that backfitting is one of several iterative methods for solving a system of *normal equations* appropriate for estimating the model (15).

3.2. *The additive model as a tool for data analysis.* The additive model provides a logical extension of the standard linear regression model by allowing arbitrary smooth (rather than just linear) functions of the covariates. The backfitting algorithm allows us to use a set of tools or models for summarizing simple x, y data as building blocks in constructing the additive model. We will

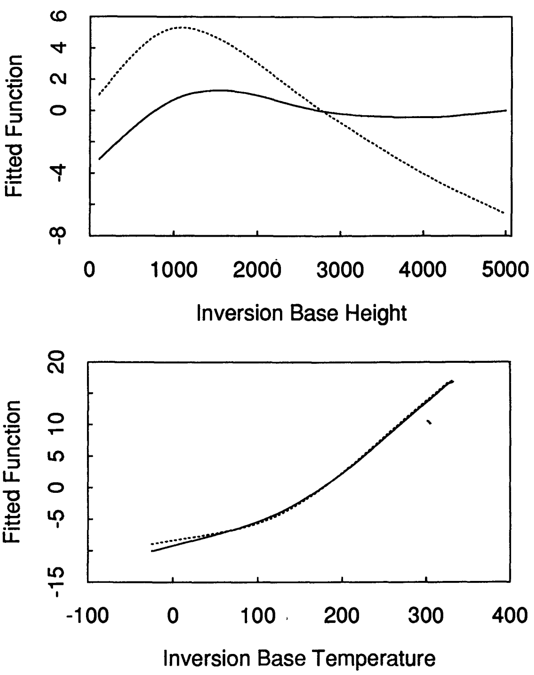


FIG. 8(a). *The solid curves are the additive model fits for Inversion Base Height and Inversion Base Temperature as in Figure 7. The broken curves are the simple spline smooths of the response against the respective variables (with the mean removed).*

touch on some of these here, referring the reader to Hastie and Tibshirani (1986a) and the discussion therein, for further details. For example, if a covariate is categorical in nature, then it does not make sense to consider an arbitrary smooth function of that variable in the additive model. Instead we could model it with a constant for each level. On the other hand, a variable may take on continuous values but for reasons specific to the data at hand we may want to restrict the fit for that variable to be linear or to be of some other parametric form. This semiparametric approach is also advocated by Denby (1984), Green and Yandell (1985), Green, Jennison and Seheult (1985) and Engle, Granger, Rice and Weiss (1986): They allow a single variable to be nonparametric in its effect. (They also allow the nonparametric component to be a function of more than one variable.) More subtly, we may choose a special smoother for a particular variable; for example, if a variable takes on values that are periodic in nature, for example, day of the week, we would want a smoother that "wrapped around" at the endpoints [see Breiman and Friedman (1985)]. We do not want to go into details here: The point is that a mixed strategy may often be used. See also van der Burg and de Leeuw (1983) for a psychometric view.

In our discussion of backfitting below we sometimes assume implicitly that the same smoother is used for each of the variables, but this is only for ease of presentation. The results are general in nature and apply to any backfitting procedure in which any linear smoother is used for any of the variables.

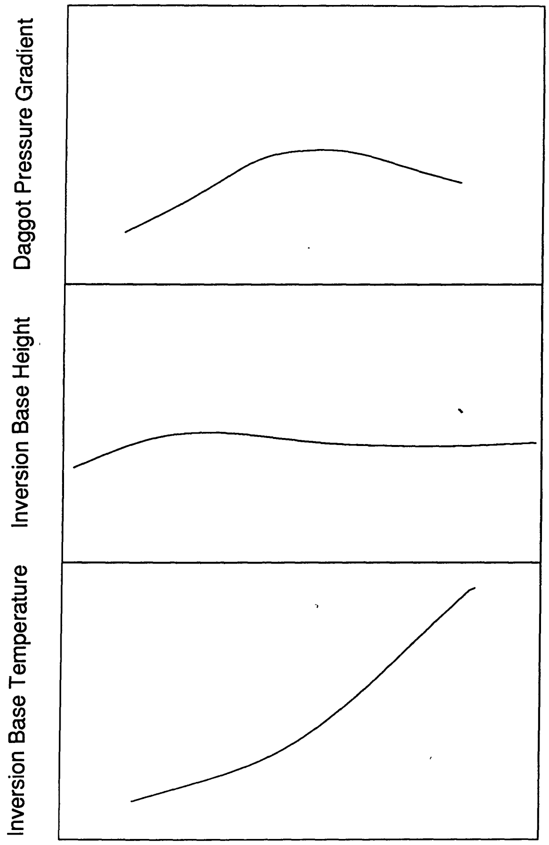


FIG. 8(b). *The fitted functions as in Figure 7, plotted on the same scale.*

The question of interaction is also important. The additive model generalizes the additive structure of the linear model but retains the no-interaction assumption. By defining a new variable to be a function of two or more existing variables (e.g., the product of two variables), one can introduce selected interactions into the additive model. Alternatively, one can model a pairwise interaction quite generally by a two-dimensional surface [e.g., using the surface smoother of Cleveland, Devlin and Grosse (1988)] within the additive model framework. An effective strategy for introducing nonlinearity and interaction into a model is not obvious. These issues are discussed in Hastie and Tibshirani (1986a) and also Wahba (1986).

Finally, in order to use an additive model for data analysis, one needs some tools for inference. We discuss "number of parameters" and standard error bands for the fitted functions in Section 5.

3.3. *The additive model and its normal equations.* The additive model (15) with unrestricted transformations is not meaningful when applied to finite samples; naive least-squares estimation leads to degenerate solutions. The standard parametric approach is to restrict the form of the functions f_j (e.g., to polynomials) and then estimate the parameters by least squares. The approach taken here is a nonparametric one based on smoothers.

We outline two general approaches to motivate the normal equations for the nonparametric case: (a) least squares on populations (as opposed to finite samples) and (b) penalized least squares.

3.3.1. *Least squares on populations.* For a pair of random variables (Y, X) the conditional expectation $f(x) = E(Y|X = x)$ minimizes $E(Y - f(X))^2$ over all L_2 functions f. The idea is to solve the problem in the theoretical setting in terms of conditional expectations, and then estimate the conditional expectations by scatterplot smoothers. We carry this idea a step further for additive models.

Let $\mathscr{H}_i$, $i = 1, \dots, p$, denote the Hilbert spaces of measurable functions $\phi_i(X_i)$ with $E\phi_i(X_i) = 0$, $E\phi_i^2(X_i) < \infty$ and inner product $\langle \phi_i(X_i), \phi_i'(X_i) \rangle = E(\phi_i(X_i)\phi_i'(X_i))$. In addition, denote by $\mathscr{H}$ the space of arbitrary, centered, square-integrable functions of $X_1, X_2, \dots, X_p$. We consider the $\mathscr{H}_i$ as subspaces of $\mathscr{H}$ in a canonical way. Furthermore, denote by $\mathscr{H}^{\text{add}} \subset \mathscr{H}$ the linear subspace of additive functions: $\mathscr{H}^{\text{add}} = \mathscr{H}_1 + \mathscr{H}_2 + \cdots + \mathscr{H}_p$, which will be closed under some technical assumptions. These are all subspaces of $\mathscr{H}_{YX}$, the space of centered square-integrable functions of Y and $X_1, \dots, X_p$.

The optimization problem in this population setting is to minimize

$$E(Y - g(\mathbf{X}))^2 \tag{16}$$

over $g(\mathbf{X}) = \sum_{j=1}^p f_j(X_j) \in \mathscr{H}^{\text{add}}$. Of course, without the additivity restriction, the solution is simply $E(Y|\mathbf{X})$; we seek the closest additive approximation to this function. Since by assumption $\mathscr{H}^{\text{add}}$ is a closed subspace of $\mathscr{H}$ this minimum exists and is unique; the individual functions $f_i(X_i)$, however, may not be uniquely determined. Denote by P_i the conditional expectation $E(\cdot\,|X_i)$ on $\mathscr{H}_{YX}$; as such P_i is an orthogonal projection onto $\mathscr{H}_i$.

The minimizer $g(\mathbf{X})$ of (16) can be characterized by residuals $Y - g(\mathbf{X})$ which are orthogonal to the space of fits: $Y - g(\mathbf{X}) \perp \mathscr{H}^{\text{add}}$. Since $\mathscr{H}^{\text{add}}$ is generated by $\mathscr{H}_i$ $(\subset \mathscr{H}^{\text{add}})$, we have equivalently: $Y - g(\mathbf{X}) \perp \mathscr{H}_i$, $\forall\ i = 1, \dots, p$, or $P_i(Y - g(\mathbf{X})) = 0$, $\forall\ i = 1, \dots, p$. Componentwise this can be written as

$$f_i(X_i) = P_i\Big(Y - \sum_{j \neq i} f_j(X_j)\Big) = E\Big(Y - \sum_{j \neq i} f_j(X_j)\,|\,X_i\Big). \tag{17}$$

Equivalently, the following system of *normal equations* is necessary and sufficient for $\boldsymbol{f} = (f_1, f_2, \dots, f_p)$ to minimize (16):

$$\begin{pmatrix} I & P_1 & P_1 & \cdots & P_1 \\ P_2 & I & P_2 & \cdots & P_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_p & P_p & P_p & \cdots & I \end{pmatrix} \begin{pmatrix} f_1(X_1) \\ f_2(X_2) \\ \vdots \\ f_p(X_p) \end{pmatrix} = \begin{pmatrix} P_1 Y \\ P_2 Y \\ \vdots \\ P_p Y \end{pmatrix} \tag{18}$$

or

$$\mathbf{P}f = \mathbf{Q}Y,$$

where $\mathbf{P}$ and $\mathbf{Q}$ represent a matrix and vector of operators, respectively, and operator matrix multiplication is defined in the obvious way.

The data version of the normal equations described above is obtained by replacing in (18) the random variables $(Y, \mathbf{X})$ by their realizations $(y_i, \mathbf{x}_i)$, and conditional expectations $P_j = E(\cdot|X_j)$ by smoothers S_j on x_j,

$$(19)\qquad \begin{pmatrix} I & S_1 & S_1 & \cdots & S_1 \\ S_2 & I & S_2 & \cdots & S_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_p & S_p & S_p & \cdots & I \end{pmatrix} \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_p \end{pmatrix} = \begin{pmatrix} S_1\mathbf{y} \\ S_2\mathbf{y} \\ \vdots \\ S_p\mathbf{y} \end{pmatrix}$$

or

$$\hat{\mathbf{P}}\mathbf{f} = \hat{\mathbf{Q}}\mathbf{y}.$$

Recall that the S_j are $n \times n$ smoothing matrices, and $\mathbf{f}_j$ are n vectors, and hence (19) is an $np \times np$ system of equations.

We note that the solutions to (19) automatically satisfy $\mathbf{f}_j \in \mathscr{R}(S_j)$, since $\mathbf{f}_j = S_j(\mathbf{y} - \sum_{k \neq j}\mathbf{f}_k)$. We have assumed here that the components of each $\mathbf{x}_j$ are in the same order as the components of $\mathbf{y}$. As a technical point, this means that S_j will denote here $E_j^{-1}S_jE_j$, where E_j is the permutation matrix that sorts in the order of $\mathbf{x}_j$.

The justification for using (19) as estimates for (18) is not entirely ad hoc. Breiman and Friedman (1985) proved asymptotic consistency results for nearest-neighbor smoothers in this situation. On the other hand, asymptotic results do not solve basic questions such as whether (19) has solutions and under what conditions (consistent equations).

3.3.2. *Penalized least squares.* In the single-smoother case we showed how symmetric linear mappings can be viewed as stationary solutions of penalized constrained least-squares problems. We can extend this approach to additive regression by penalizing the RSS separately for each component function,

$$(20)\qquad Q(\mathbf{f}) = \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{f}_j \right\|^2 + \sum_{j=1}^{p} \mathbf{f}_j^t(S_j^- - I)\mathbf{f}_j,$$

where each S_j is a symmetric mapping, and $\mathbf{f}_j \in \mathscr{R}(S_j)$, $\forall\, j$. It is easily verified that the normal equations (19) are exactly the stationarity conditions for the above criterion. We concentrate on variable j and the eigendecomposition of $S_j = U_jD_jU_j^t$. As before, we partition $U_j = (U_{1j}\colon U_{2j})$ and $D_j = \operatorname{diag}(D_{\theta_j}, 0)$ and write $\mathbf{f}_j = U_{1j}\boldsymbol{\beta}_j$. The stationarity condition for $\boldsymbol{\beta}_j$ is

$$-U_{1j}^t\Big(\mathbf{y} - \sum_{k \neq j}\mathbf{f}_k - U_{1j}\boldsymbol{\beta}_j\Big) + U_{1j}^t\big(U_{1j}D_{\theta_j}^{-1}U_{1j}^t - I\big)U_{1j}\boldsymbol{\beta}_j = \mathbf{0},$$

which simplifies to $\boldsymbol{\beta}_j = D_{\theta_j} U_{1j}^t(\mathbf{y} - \Sigma_{k \neq j} \mathbf{f}_k)$, or equivalently $\mathbf{f}_j = S_j(\mathbf{y} - \Sigma_{k \neq j} \mathbf{f}_k)$ as in (19).

As a leading example, let us look again at cubic smoothing splines. It is natural to generalize the univariate penalized least-squares problem (4) to:

$$\text{minimize}\left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{f}_j \right\|^2 + \sum_{j=1}^{p} \lambda_j \int \left(f_j''(t) \right)^2 dt \tag{21}$$

over all functions $f_j(\cdot) \in W_2$, where for brevity we have written $\mathbf{f}_j$ for $(f_j(x_{1j}), \dots, f_j(x_{nj}))^t$. We prove in Theorem 1 in the Appendix that this system has a solution, that each of the minimizing functions is a natural cubic spline, and that the problem is equivalent to the finite-dimensional problem (20) with each S_j the appropriate cubic spline smoother matrix (and $K_j = S_j^- - I$, the penalty matrix for the jth variable). The theorem, which is an additive extension of the similar single-smoother result in O'Sullivan, Yandell and Raynor (1986), is given, as they did, for the more general problem of penalized likelihoods.

3.4. *Algorithms for solving the normal equations.* For the moment we assume that solutions for the system (19) exist. In subsequent sections we establish conditions for their existence. There are np equations in (19), and although a direct solution is possible, it would be prohibitively expensive except for small data sets. Later we discuss cases in which the effective dimension is substantially less than np and hence a direct solution is feasible.

There are a variety of efficient methods for solving the system (19), which depend on both the number and types of smoothers used. The Gauss–Seidel method, applied to blocks consisting of components $\mathbf{f}_1, \dots, \mathbf{f}_p$, exploits the special structure of (19). It coincides with the backfitting procedure described earlier.

The backfitting or Gauss–Seidel algorithm

Initialize: $\mathbf{f}_i = \mathbf{f}_i^0$, $i = 1, 2, \dots, p$

Cycle: $j = 1, 2, \dots, p, 1, 2, \dots, p, \dots,$

$$\mathbf{f}_j \leftarrow S_j\left(\mathbf{y} - \sum_{k \neq j} \mathbf{f}_k \right) \tag{22}$$

Until: the individual functions do not change.

Convergence of the Gauss–Seidel procedure in this setting is not immediate. In the population setting, Breiman and Friedman (1985) proved convergence for compact projection operators. Bickel, Klaassen, Ritov, and Wellner (1989) prove under milder conditions that the Gauss–Seidel algorithm (22) converges to a solution of (18) in the population setting. The nature of the solution depends on the joint distribution of the X_j's. However, the convergence proof for the population version of (22) does not help much in proving convergence in the data case. The main stumbling block is that most reasonable smoothers S_j are not projections, whereas conditional expectations are. In addition, standard convergence results for Gauss–Seidel and related procedures [cf. Golub and van Loan

(1983), Chapter 10] assume that the matrix of the linear system is symmetric and positive definite. In the present setting $\hat{\mathbf{P}}$ is not symmetric so these results cannot be immediately applied. It is true, however, that in an appropriate coordinate system, $\hat{\mathbf{P}}$ is symmetric, but usually not positive definite. Breiman and Friedman proved convergence for strictly shrinking smoothers. It turns out that this implies positive definiteness, and so convergence is immediate. We derive specialized convergence results for the semidefinite case in Section 4.

The Gauss–Seidel method is only one technique in the large class of iterative schemes called *successive over-relaxation* (SOR) methods. They differ from ordinary Gauss–Seidel procedures by the amount one proceeds in the direction of the Gauss–Seidel updates,

$$\mathbf{f}_j \leftarrow (1-\omega)\mathbf{f}_j + \omega S_j\left(\mathbf{y} - \sum_{k \neq j} \mathbf{f}_k\right). \tag{23}$$

We will see that if the Gauss–Seidel algorithm converges, so do successive over-relaxation iterations for relaxation parameters $0 < \omega < 2$. Experience and some limited theory for special cases indicate that some over-relaxation ($\omega > 1$) can be beneficial, whereas under-relaxation is generally detrimental.

The numerical analysis literature also distinguishes between *successive* and *simultaneous* iterations, usually also referred to as Gauss–Seidel and Jacobi iterations, respectively. The Gauss–Seidel schemes update one component at a time, based on the most recent components available. In contrast, Jacobi schemes form a complete new set of updates from a complete old set. The difference between the two approaches is made explicit in the notation

$$\text{Gauss–Seidel:}\quad \mathbf{f}_j^{\text{new}} \leftarrow S_j\left(\mathbf{y} - \sum_{k<j} \mathbf{f}_k^{\text{new}} - \sum_{k>j} \mathbf{f}_k^{\text{old}}\right),$$

$$\text{Jacobi:}\quad \mathbf{f}_j^{\text{new}} \leftarrow S_j\left(\mathbf{y} - \sum_{k \neq j} \mathbf{f}_k^{\text{old}}\right).$$

However, as written here, Jacobi iterations do not converge; they require so-called "under-relaxation;" see Section 4.3.

If the smoothers are chosen so that the effective dimension of (19) is less than np, then direct solutions become feasible. An important special case is when each of the S_j are orthogonal projections onto a small (relative to n) subspace of $\mathbb{R}^n$. Binning smoothers, linear and polynomial regression and fixed knot regression splines are in this class. The corresponding projection matrices have the form $S_j = \mathbf{X}_j(\mathbf{X}_j^t\mathbf{X}_j)^{-1}\mathbf{X}_j^t$, where the subdesign matrices $\mathbf{X}_j$ are generated by dummy variables, (orthogonal) polynomials of increasing degree or basis spline functions, respectively. It can then be shown that the system (19) is equivalent to the usual least-squares system

$$\mathbf{X}^t\mathbf{X}\mathbf{b} = \mathbf{X}^t\mathbf{y}, \tag{24}$$

where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p)$. This is seen as follows: With $S_j = \mathbf{X}_j(\mathbf{X}_j^t\mathbf{X}_j)^{-1}\mathbf{X}_j^t$ and $\mathbf{f}_j = \mathbf{X}_j\boldsymbol{\beta}_j$ the normal equations (19): $\mathbf{f}_j + S_j(\Sigma_{k \neq j}\mathbf{f}_k) = S_j\mathbf{y}$ become $(\mathbf{X}_j^t\mathbf{X}_j)\boldsymbol{\beta}_j + \mathbf{X}_j^t\Sigma_{k \neq j}\mathbf{X}_k\boldsymbol{\beta}_k = \mathbf{X}_j^t\mathbf{y}$ after left multiplication by $\mathbf{X}_j^t$. This is just $\mathbf{X}_j^t\Sigma_k\mathbf{X}_k\boldsymbol{\beta}_k = \mathbf{X}_j^t\mathbf{y}$, equivalent to (24).

The customary numerical procedures for solving linear least-squares problems are not iterative. They either use a Cholesky or other type of decomposition of $\mathbf{X}^t\mathbf{X}$, or avoid forming the normal equations altogether and decompose X directly either through a Q–R decomposition based on Householder, modified Gram–Schmidt, or Givens transformations, or a singular value decomposition. Avoiding the normal equations is recommended for near degenerate data [Lawson and Hanson (1974)].

In the class of projection smoothers, regression spline smoothing is closer to the present discussion of nonparametric additive models. In this case each X_j has k_j columns consisting of B-splines placed at the k_j judiciously chosen knots for each variable x_j, and evaluated at the data. The numbers k_j and positions of the knots are all parameters of the procedure. Assuming we use $k_j = k$ knots and therefore k parameters per covariate, (24) consists of kp equations. For small to moderate k and p, the problem can thus be solved directly without the use of backfitting; for large k, however, backfitting is a numerically stable alternative to solving a large system of equations. The placement of the knots, however, has remained a problem in this otherwise attractive approach. Recently, however, Friedman and Silverman (1989) proposed a promising stepwise procedure for selecting knots in this context.

Although unnecessary, one can still solve (24) iteratively if all the S_j are projections, and the Gauss–Seidel procedure will converge to the projection of $\mathbf{y}$ onto the column space of X. The real need for iterative schemes, however, arises from problems which cannot be reduced in size by reparametrizations, such as the normal equations (19) with at least some smoothers S_j not of the projection type.

3.5. *A summary of the consistency, degeneracy and convergence results.* It is not a priori clear when the normal equations (19) are consistent, that is, when solutions exist. Nor is it clear when the equations are nondegenerate, that is, when the solutions are unique. This contrasts with the normal equations from ordinary least-squares problems which are always consistent, but possibly degenerate.

In both cases, nondegeneracy implies consistency. However, the normal equations (19) are almost always degenerate, and we need to understand these degeneracies. In the remainder of this section and the next section we derive the following results:

1. For symmetric smoothers with eigenvalues in $[0,1]$, the normal equations $\hat{\mathbf{P}}\mathbf{f} = \hat{\mathbf{Q}}\mathbf{y}$ always have at least one solution.
2. The solution is unique unless there exists a $\mathbf{g} \neq 0$ such that $\hat{\mathbf{P}}\mathbf{g} = \mathbf{0}$, a phenomenon we call "concurvity." This implies that for any solution $\mathbf{f}$ to (19), $\mathbf{f} + \alpha\mathbf{g}$ is also a solution for any α.
3. For this same class of smoothers, (exact) concurvity can only occur if there is a linear dependence among the eigenspaces of the S_j's corresponding to eigenvalue $+1$.
4. For this same class of smoothers, the Gauss–Seidel and related procedures always converge to some solution of $\hat{\mathbf{P}}\mathbf{f} = \hat{\mathbf{Q}}\mathbf{y}$.

In some instances, more general results than these are established, especially in the two-smoother case.

3.6. *Consistency.* To show consistency of the normal equations (19), we need to establish that $\hat{\mathbf{Q}}\mathbf{y} \in \mathscr{R}(\hat{\mathbf{P}})$ for arbitrary data $\mathbf{y} \in \mathbb{R}^n$. In this section we give a number of propositions for establishing consistency. We prove consistency for the general p-smoother case for symmetric smoothers with eigenvalues in $[0, 1]$, and we examine the two-smoother case in detail to develop slightly stronger results for that case.

These and all subsequent results for symmetric smoothers with eigenvalues in $[0, 1]$ apply to, amongst others, cubic spline smoothers, simple linear, polynomial and B-spline regression and bin smoothers. They also can be applied to normal equations comprised of a mixture of these smoothers, that is, a cubic spline smoother for one variable, a simple linear fit for another variable, and so forth. The smoothers need not even be univariate; the results apply to shrinking two- or higher-dimensional surface smoothers as well.

THEOREM 2. *If each S_j is symmetric with eigenvalues in* $[0, 1]$, *the normal equations* (19) *are consistent for every* $\mathbf{y}$.

REMARK. This result generalizes the fact that least squares always leads to consistent normal equations.

PROOF. We use the penalized least-squares approach of Section 3.3.2. The penalized least-squares criterion (20) $Q(\mathbf{f}) = \|\mathbf{y} - \mathbf{f}_+\|^2 + \Sigma \mathbf{f}_j^t(S_j^- - I)\mathbf{f}_j$ is a quadratic function in $\mathbf{f}$ for $\mathbf{f}_j \in \mathscr{R}(S_j)$, that is, it is the sum of a quadratic form, a linear form and a constant. Under the stated conditions, $Q(\mathbf{f})$ is nonnegative for all $\mathbf{f}$ such that $\mathbf{f}_j \in \mathscr{R}(S_j)$, since each of the penalizations $\mathbf{f}_j^t(S_j^- - I)\mathbf{f}_j$ is nonnegative on $\mathscr{R}(S_j)$ ($S_j^- - I$ has eigenvalues $1/\theta - 1$ where $0 < \theta \leq 1$). The stationarity conditions (19) then characterize the minima of $Q(\mathbf{f})$, which must exist since every multivariate quadratic function bounded below has at least one minimum. □

For the case in which the symmetric smoothers have eigenvalues in $[0, 1)$ only (i.e., $+1$ excluded), we can write down closed formulas for the solutions.

PROPOSITION 3 [Breiman and Friedman (1985)]. *If the smoothers S_j are symmetric with eigenvalues in* $[0, 1)$, *the solutions of the normal equations* (19) *can be written* $\mathbf{f}_j = A_j(I + A)^{-1}\mathbf{y}$, *where* $A_j = (I - S_j)^{-1}S_j$ *and* $A = \Sigma_j A_j$.

PROOF. The normal equations are equivalent to $(I - S_j)\mathbf{f}_j = S_j(\mathbf{y} - \mathbf{f}_+)$, $j = 1, \dots, p$. Thus $\mathbf{f}_j = A_j(\mathbf{y} - \mathbf{f}_+)$ and $\mathbf{f}_+ = A(\mathbf{y} - \mathbf{f}_+)$. Since A_j is symmetric and nonnegative definitive under the assumptions on S_j, the same holds for A. It follows that $\mathbf{f}_+ = (I + A)^{-1}A\mathbf{y}$ exists, and $\mathbf{f}_j = A_j[I - (I + A)^{-1}A]\mathbf{y} = A_j(I + A)^{-1}\mathbf{y}$. □

Below is a less stringent necessary and sufficient condition on the S_j for consistency of the normal equations which leads to a more general result for the two-smoother case.

THEOREM 4. *For arbitrary linear mappings S_j, the normal equations* (19) *are consistent for arbitrary* $\mathbf{y}$ *iff one of the following two equivalent conditions hold*:

1. $\mathbf{f}_+ = \mathbf{0}$ *whenever* $\hat{\mathbf{P}}^t\mathbf{f} = \mathbf{0}$.
2. $\mathbf{f}_j \in \mathscr{M}_1(S_j^t)$ *for at least one and hence all* j *whenever* $\hat{\mathbf{P}}^t\mathbf{f} = \mathbf{0}$.

We prove Theorem 4 in the Appendix. The two-smoother case is of special interest, and simplified conditions can be given.

COROLLARY 4.1. *For arbitrary linear mappings S_1 and S_2, the two-smoother normal equations are consistent for arbitrary* $\mathbf{y}$ *iff* $\mathbf{f}_1 = S_1^t\mathbf{f}_1$ *whenever* $\mathbf{f}_1 = (S_1S_2)^t\mathbf{f}_1$.

PROOF. The second condition of Theorem 4 becomes $S_1^t\mathbf{f}_1 = \mathbf{f}_1$ whenever $\mathbf{f}_1 = -S_2^t\mathbf{f}_2$ and $\mathbf{f}_2 = -S_2^t\mathbf{f}_1$, which is equivalent to the condition above. □

COROLLARY 4.2. *If in the two-smoother case both smoothers are symmetric with eigenvalues in* $(-1, 1]$, *then the normal equations are consistent for arbitrary* $\mathbf{y}$.

PROOF. The proof follows from the fact that such smoothers are shrinking: $\|S_j\mathbf{f}_j\| \le \|\mathbf{f}_j\|$, and equality holds iff $S_j\mathbf{f}_j = \mathbf{f}_j$. We verify the conditions of Corollary 4.1: If $\mathbf{f}_1 = S_2S_1\mathbf{f}_1$, we get $\|\mathbf{f}_1\| = \|S_2S_1\mathbf{f}_1\| \le \|S_1\mathbf{f}_1\|$, hence $S_1\mathbf{f}_1 = \mathbf{f}_1$. □

The conditions of Theorem 4 and Corollary 4.1 above are difficult to establish for arbitrary smoothers. As we have seen, some commonly used smoothers (running lines and locally weighted running lines) may have singular values larger than 1. We have empirical evidence, however, that:

1. The spectral radius $\rho(S)$ equals one for both these smoothers, with the constant and linear terms belonging to the eigenvalue 1,
2. for the two-smoother case, the constant is the only eigenvector of $(S_1S_2)^t$ with eigenvalue 1 (unless $\mathbf{x}_1 = \mathbf{x}_2$), and
3. no complex eigenvalue of absolute value 1 other than 1 exists.

Thus the conditions of Corollary 4.1 may be empirically verified in most cases. Figure 9 shows (a) the eigendecomposition of S_1 for a running-line smoother based on 50 pseudorandom Gaussian observations $\mathbf{x}_1$ with span 50%, and (b) the decomposition of $(S_1S_2)^t$ for this variable and a similar (correlated) vector $\mathbf{x}_2$. The eigenvalue 1 has multiplicity 1 for the constant term. If $\mathbf{x}_1$ and $\mathbf{x}_2$ were exactly collinear, it would have multiplicity 2.

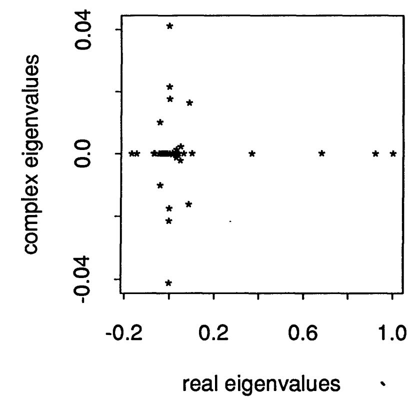


FIG. 9(a). *The (complex) eigendecomposition for a running-line smoother matrix S_1 based on* 50 *pseudorandom Gaussian observations, using a span of* 50%. *The eigenvalue* 1 *has multiplicity* 2, *corresponding to the constant and linear terms.*

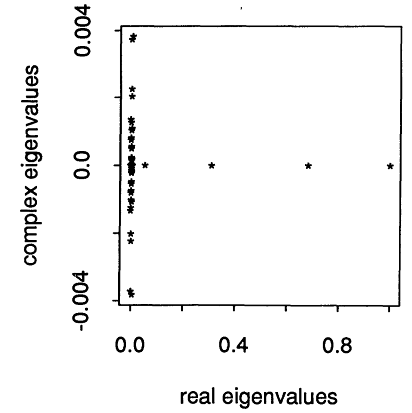


FIG. 9(b). *The eigendecomposition of $(S_1S_2)^t$ for S_2 similar to S_1. Here* 1 *has multiplicity* 1 *for the constant term. This is an empirical demonstration that the conditions for Corollary* 4.1 *are satisfied for these data.*

A condition like $\|S_1\| < 1$, $\|S_2\| < 1$ for some matrix norm would also be sufficient to guarantee consistency, since any matrix norm dominates the spectral radius [Householder, (1964), Section 2.2], and the condition of Corollary 4.1 would be vacuous. Since most smoothers reproduce constants and linear functions of the predictor on which they are based, this condition would seldom apply. A less stringent condition is given by Corollary 4.3.

COROLLARY 4.3. *If in the two-smoother case, the smoothers are arbitrary but $\|S_1S_2\| < 1$ for an arbitrary matrix norm, then the normal equations are consistent for all* $\mathbf{y}$.

PROOF. If we use the two-norm $\|C\|_2 = \sup_{\mathbf{a}\neq 0}\|C\mathbf{a}\|/\|\mathbf{a}\|$, then the proof is simple, since $\|S_1S_2\|_2 = \|(S_1S_2)^t\|_2 \geq \rho((S_1S_2)^t)$. Thus 1 is not an eigenvalue of $(S_1S_2)^t$ and the conditions of Corollary 4.1 are vacuous. In fact, for any other matrix norm one can show [Householder (1964), Section 2.2] that $\|S_1S_2\| \geq \rho((S_1S_2)^t)$ and so the result is true in general. □

3.7. *Degeneracy of smoother-based normal equations*: *collinearity and concurvity*. The problem of nonunique solutions of normal equations is a standard topic in the teaching of multiple linear regression. Collinearity detection as part of regression diagnostics is a must in every good regression analysis. Practioners are usually concerned with approximate collinearity and its inflationary effects on standard errors of regression coefficients. Exact (up to numerical precision) collinearity is rare and usually results from "underdetermined models" with too many or redundant variables included. Just the same, exact degeneracy of normal equations is an extreme case worth exploring. Its structure should be fully understood before approximate degeneracies are tackled. In this paper we deal only with exact degeneracy of smoother-based normal equations.

While the term "collinearity" refers to linear dependencies among predictors as the cause of degeneracy, the term "concurvity" has been used [Buja, Donnell and Stuetzle (1986)] to describe nonlinear dependencies which lead to degeneracy in additive models. In a technical sense, concurvity boils down to collinearity of (nonlinear) transforms of predictors. Consider, for example, additive regression where we allow linear and quadratic transformations of the predictors. This amounts to multiple linear regression including the square of each predictor. Although collinearity among raw and squared predictors describes degeneracy in a technical sense, it is more intuitive to think of it as an additive dependence $\mathbf{f}_+ = \mathbf{0}$, where each component $\mathbf{f}_j$ is a quadratic polynomial in the predictor $\mathbf{x}_j$. Similarly, we can describe degeneracy in terms of general polynomial transforms, B-splines and others, by associating with predictor $\mathbf{x}_j$ a linear space V_j of transformations, and defining concurvity to hold if there exist nontrivial $\mathbf{f}_j \in V_j$ such that $\mathbf{f}_+ = \mathbf{0}$. This covers at least the situation of additive models where each smoother is an orthogonal projection with $\mathscr{R}(S_j) = V_j$.

For general smoother-based normal equations, exact concurvity is defined as the existence of a nonzero solution of the corresponding homogeneous equations

$$\hat{\mathbf{P}}\mathbf{g} = \mathbf{0}. \tag{25}$$

It is clear that if such a $\mathbf{g}$ exists, and if $\mathbf{f}$ is a solution to $\hat{\mathbf{P}}\mathbf{f} = \hat{\mathbf{Q}}\mathbf{y}$, then so is $\mathbf{f} + \omega\mathbf{g}$ for any ω, and thus infinitely many solutions exist. The set of all nonzero solutions to the homogeneous equations $\hat{\mathbf{P}}\mathbf{g} = \mathbf{0}$ will be called concurvity space for the normal equations and the additive model defined by them. It is easy to check that

$$\mathbf{g} = \begin{pmatrix} \alpha\mathbf{1} \\ -\alpha\mathbf{1} \end{pmatrix}$$

lies in the concurvity space of the two-smoother problem if they both reproduce constants. Similarly, for p such smoothers, the concurvity space has dimension at least $p - 1$.

For the general p-smoother case, we again must restrict attention to symmetric matrices with eigenvalues in $[0, 1]$. For its formulation, we use the quadratic form of the penalized least-squares criterion obtained by setting $\mathbf{y} = \mathbf{0}$,

$$Q(\mathbf{g}) = \|\mathbf{g}_+\|^2 + \sum_{j=1}^{p} \mathbf{g}_j^t(S_j^- - I)\mathbf{g}_j,$$

defined for $\mathbf{g}_j \in \mathscr{R}(S_j)$.

THEOREM 5. *If the smoothers S_j are all symmetric with eigenvalues in* $[0, 1]$, *a vector $\mathbf{g} \neq \mathbf{0}$ with $\mathbf{g}_j \in \mathscr{R}(S_j)$ represents a concurvity* ($\hat{\mathbf{P}}\mathbf{g} = \mathbf{0}$) *iff one of the following equivalent conditions is satisfied*:

1. $Q(\mathbf{g}) = 0$, *that is*, $\mathbf{g}$ *minimizes* Q.
2. $\mathbf{g}_j \in \mathscr{M}_1(S_j)$, $j = 1, \dots, p$, *and* $\mathbf{g}_+ = \mathbf{0}$.

PROOF. Under the given assumptions, Q is minimized iff the normal equations are satisfied. Setting $\mathbf{y} = \mathbf{0}$ in $\hat{\mathbf{P}}\mathbf{g} = \hat{\mathbf{Q}}\mathbf{y}$, this shows the equivalence of concurvity and condition (1). The quadratic form $Q(\mathbf{g}) = \|\mathbf{g}_+\|^2 + \Sigma_j \mathbf{g}_j^t(S_j^- - I)\mathbf{g}_j$ is minimized iff all its summands are minimized, that is, $\|\mathbf{g}_+\|^2 = 0$ and $\mathbf{g}_j^t(S_j^- - I)\mathbf{g}_j = 0$, $j = 1, \dots, p$. Since the eigenvalues of S_j^- are $1/\theta$ for eigenvalues θ of S_j, the equivalence of conditions 1 and 2 follows. □

As mentioned above, condition 2 implies that exact concurvity is exact collinearity if, for example, all smoothers are cubic spline smoothers. Approximate concurvity, however, can be described by approximate minimizers of $Q(\mathbf{g})$, which leads to approximate nonlinear additive relations among the predictors. The authors and D. Donnell are working on a theory of approximate concurvity.

REMARK. If S_j, $j = 1, \dots, p$, are symmetric with eigenvalues in $[0, 1)$, then $\hat{\mathbf{P}}$ is nonsingular. We had this result implicitly in Proposition 3, where an explicit solution was given.

This remark seems irrelevant if most smoothers reproduce constants. However, in practice we usually separate the constant term in the additive model, and adjust each of the smooth terms to have mean 0. This means that implicitly we have redefined our smoothers to $S^* = S - \mathbf{1}\mathbf{1}^t/n$, and S^* has eigenvalue 0 for the vector of constants. Many smoothers also reproduce linear functions as well, so more adjustments would be needed.

The work of O'Sullivan (1983) can be extended to complement the results found here. He established (for more general loss and penalty functions) for the single function version of (21) that existence and uniqueness of a minimizer of (21) depend on the existence and uniqueness of the *linear* minimizer of the least-squares part. We have thus extended these results to the additive case; uniqueness of the additive minimizer is guaranteed if there is no collinearity!

We now examine in more detail the two-smoother case.

PROPOSITION 6. *For two smoothers, there exists exact concurvity iff* $\mathbf{f}_1 = (S_1S_2)\mathbf{f}_1$ *for some* $\mathbf{f}_1 \neq \mathbf{0}$.

PROOF. The homogeneous equations are $\mathbf{f}_1 = -S_1\mathbf{f}_2$ and $\mathbf{f}_2 = -S_2\mathbf{f}_1$. It follows that $\mathbf{f}_1 = (S_1S_2)\mathbf{f}_1$. On the other hand, if $\mathbf{f}_1 = (S_1S_2)\mathbf{f}_1$, set $\mathbf{f}_2 = -S_2\mathbf{f}_1$, which will satisfy the homogeneous equations. □

COROLLARY 6.1. *If* $\|S_1S_2\| < 1$ *for some matrix norm, concurvity does not exist.*

PROOF. As mentioned earlier, any matrix norm is a bound on the spectral radius. Thus $+1$ cannot be an eigenvalue. □

COROLLARY 6.2. *For two symmetric smoothers with eigenvalues in* $(-1, +1]$, *concurvity exists iff* $\mathscr{M}_1(S_1) \cap \mathscr{M}_1(S_2) \neq 0$.

PROOF. The condition $\mathbf{f}_1 = (S_1S_2)\mathbf{f}_1$ of Proposition 6 is satisfied under the given assumptions iff $\mathbf{f}_1 = S_2\mathbf{f}_1$ and $\mathbf{f}_1 = S_1\mathbf{f}_1$. □

Corollary 6.2 has again the consequence that exact concurvity, for example, for a pair of cubic spline smoothers, can only be an exact collinearity between the untransformed predictors, since cubic splines preserve constant and linear fits. Such results have to be taken with a grain of salt when it comes to approximate concurvity, which can be generated by eigenvalues close to, but not exactly equal to 1. Cubic splines, especially in large samples with suitably small bandwidths, tend to have a good number of eigenvalues near 1. The outcome of all this is that even though the covariates may lie exactly on a lower-dimensional manifold (i.e., a curve for two predictors), this will not constitute an exact degeneracy unless the components of the additive function defining the manifold are preserved by the respective smoothers.

The definition of concurvity carries over immediately to function space, that is $\mathbf{P}\mathbf{g} = \mathbf{0}$. If the operators in $\mathbf{P}$ are all conditional expectations, exact concurvity may be defined as the existence of a set of p functions $g_1, \dots, g_p$, not all zero, such that

$$\sum_{j=1}^{p} g_j(X_j) = 0 \quad \text{a.s.} \tag{26}$$

If the covariates are real-valued, and if the functions g_j are smooth, such a relationship means that the covariates are contained in a $p-1$-dimensional manifold of $\mathbb{R}^p$.

One of the most important cases of concurvity, however, arises from nonsmooth functions g_j, which may indicate the presence of multivariate clusters. As a simple example, consider random variables X_1, X_2 with a joint distribution which satisfies $P(X_1 < 0,\ X_2 < 0) = \frac{1}{2}$, $P(X_1 \geq 0,\ X_2 \geq 0) = \frac{1}{2}$, that is, the values lie in only two of the four quadrants of the plane. The step functions

$g_1(x_1) = 1_{(x_1 \geq 0)} - \frac{1}{2}$ and $g_2(x_2) = 1_{(x_2 < 0)} - \frac{1}{2}$ lead to $g_1(X_1) + g_2(X_2) = 0$ a.s., a nontrivial degeneracy which qualifies as concurvity like any other. For a development of such phenomena in the context of ACE, see Buja (1989). In finite samples rather than distributions, we observed that sufficiently flexible smoothers will try to approximate step functions in such situations.

4. Convergence of the Gauss–Seidel (backfitting) algorithm and related procedures. In this section we prove that the Gauss–Seidel and related algorithms converge for the normal equations $\hat{\mathbf{P}}\mathbf{f} = \hat{\mathbf{Q}}\mathbf{y}$ if suitable conditions are imposed. The general result establishes convergence for symmetric smoother matrices having eigenvalues in $[0, 1]$. Slightly stronger results are derived for the two-smoother case.

Consistency is obviously a necessary condition for convergence. We show, however, that degeneracy (concurvity) does not need to be avoided to assure convergence. This is in contrast to the treatment of Gauss–Seidel and Jacobi iterations in the literature, where linear systems are usually assumed nondegenerate. An exception is Keller (1965) who deals expressly with degeneracy and produces results of great generality for symmetric nonnegative definite systems. These results would not cover indefinite systems in the two-smoother situation, where we can prove sharper results than Keller's. Rather than mechanically verifying Keller's conditions in the p-smoother case, we prefer to derive convergence of backfitting from a more intuitive descent principle for seminorms which is of interest in itself. Interestingly, most of Keller's results follow from it.

4.1. *The convergence of backfitting*: *p smoothers*. In this section we show that for symmetric, smoothers with eigenvalues in $[0, 1]$, the backfitting algorithm always converges.

We begin by centering the normal equations at an arbitrary solution $\tilde{\mathbf{f}}$ to reduce the problem to that of solving the homogeneous equations, a common tactic in problems of this sort. Thus $\hat{\mathbf{P}}\tilde{\mathbf{f}} = \hat{\mathbf{Q}}\mathbf{y}$ and we need to find $\mathbf{f}$ such that $\hat{\mathbf{P}}(\mathbf{f} - \tilde{\mathbf{f}}) = \mathbf{0}$. We have to show that for $\mathbf{y} = \mathbf{0}$, backfitting converges to some solution of $\hat{\mathbf{P}}\mathbf{f} = \mathbf{0}$. If the normal equations are nonsingular, this means convergence to $\mathbf{f} = \mathbf{0}$.

To describe the effect of updating the jth component under Gauss–Seidel on the homogeneous equations, we define the linear map

$$\hat{T}_j\colon \mathbf{f} \mapsto \begin{pmatrix} \mathbf{f}_1 \\ \vdots \\ S_j\left(-\Sigma_{k \neq j}\mathbf{f}_k\right) \\ \vdots \\ \mathbf{f}_p \end{pmatrix}, \qquad \mathbb{R}^{np} \mapsto \mathbb{R}^{np}. \tag{27}$$

A full cycle of backfitting is described by the map $\hat{\mathbf{T}} = \hat{\mathbf{T}}_p\hat{\mathbf{T}}_{p-1} \cdots \hat{\mathbf{T}}_1$, while m full cycles correspond to $\hat{\mathbf{T}}^m$. The problem is to show that $\mathbf{f}^m = \hat{\mathbf{T}}^m\mathbf{f}^0$ converges to a solution $\mathbf{f}^{\infty}$ of the homogeneous equations $\hat{\mathbf{P}}\mathbf{f} = \mathbf{0}$, for arbitrary initialization $\mathbf{f}^0$. A relation between $\hat{\mathbf{T}}$ and $\hat{\mathbf{P}}$ is given by Proposition 7.

PROPOSITION 7. *For arbitrary smoothers*, $\hat{\mathbf{P}}\mathbf{f} = \mathbf{0}$ *iff* $\hat{\mathbf{T}}\mathbf{f} = \mathbf{f}$.

The proposition is proved in the Appendix. In formal terms it says $\mathscr{N}(\hat{\mathbf{P}}) = \mathscr{M}_1(\hat{\mathbf{T}})$. The solutions of the homogeneous system are exactly the fixed points under Gauss–Seidel iterations, and $\hat{\mathbf{P}}$ is nonsingular iff there are no fixed points other than $\mathbf{0}$.

The convergence proof is then complete if we can show that all (complex) eigenvalues λ of $\hat{\mathbf{T}}$ are either $+1$ or in the interior of the unit disk ($|\lambda| < 1$), and that the Jordan blocks of $\hat{\mathbf{T}}$ for $\lambda = 1$ do not contain off-diagonal (nilpotent) components. In other words, the geometric and algebraic multiplicity of the eigenvalue $\lambda = 1$ are the same. This is formalized by Lemma 8.1 in the Appendix. Rather than verify the conditions of this lemma directly, we take an intermediate step. Exploiting the fact that the S_j's are symmetric and have eigenvalues in $[0,1]$, we can interpret $\hat{\mathbf{T}}$ as a descent method for the corresponding penalized least-squares problem. In particular, the following theorem gives a sufficient condition for convergence and is a consequence of Lemma 8.1.

THEOREM 8 (Seminorm descent principle). *If* $|\mathbf{f}|$ *is a complex seminorm and* $\hat{\mathbf{T}}$ *a linear mapping on* $\mathbb{C}^N$ *satisfying* $|\hat{\mathbf{T}}\mathbf{f}| < |\mathbf{f}|$ *unless* $|\mathbf{f}| = 0$, *and* $\hat{\mathbf{T}}\mathbf{f} = \mathbf{f}$ *for* $|\mathbf{f}| = 0$, *then* $\hat{\mathbf{T}}^m$ *converges to a limit* $\hat{\mathbf{T}}^{\infty}$ *with the properties* $|\hat{\mathbf{T}}^{\infty}\mathbf{f}| = 0$ *for all* $\mathbf{f}$, $(\hat{\mathbf{T}}^{\infty})^2 = \hat{\mathbf{T}}^{\infty}$ *and* $\hat{\mathbf{T}}\hat{\mathbf{T}}^{\infty} = \hat{\mathbf{T}}^{\infty}\hat{\mathbf{T}} = \hat{\mathbf{T}}^{\infty}$.

The theorem is proved in the Appendix. It is easily applied to the Gauss–Seidel iteration $\hat{\mathbf{T}}$ under the assumptions that all smoothers are symmetric with eigenvalues in $[0,1]$. In this case, the complex quadratic form

$$Q(\mathbf{f}) = \mathbf{f}_+^t \overline{\mathbf{f}_+} + \sum_{j=1}^{p} \mathbf{f}_j^t (S_j^- - I)\overline{\mathbf{f}_j}$$

is nonnegative for $\mathbf{f}_j \in \mathscr{R}(S_j)$, and $|\mathbf{f}| = \sqrt{Q(\mathbf{f})}$ defines a complex seminorm. Its space of degeneracy $\{\mathbf{f} \mid \mathbf{f}_j \in \mathscr{R}(S_j),\ Q(\mathbf{f}) = 0\}$ coincides with $\mathscr{N}(\hat{\mathbf{P}})$ (by Theorem 5) and $\mathscr{M}_1(\hat{\mathbf{T}})$ by Proposition 7. Thus the condition $\hat{\mathbf{T}}\mathbf{f} = \mathbf{f}$ for $|\mathbf{f}| = 0$ is verified. To show that $|\hat{\mathbf{T}}\mathbf{f}| < |\mathbf{f}|$ unless $|\mathbf{f}| = 0$, we notice that $\hat{\mathbf{T}}_j\mathbf{f}$ is the minimizer of $Q(\mathbf{f})$ over the jth component of $\mathbf{f}$. This ensures that $|\hat{\mathbf{T}}\mathbf{f}| \le |\mathbf{f}|$. If $|\hat{\mathbf{T}}\mathbf{f}| = |\mathbf{f}|$, no strict descent is possible along any component, hence $\hat{\mathbf{T}}_j\mathbf{f} = \mathbf{f}$, $j = 1, \dots, p$. This implies $\hat{\mathbf{T}}\mathbf{f} = \mathbf{f}$, and $|\mathbf{f}| = 0$ follows. We have thus proved

THEOREM 9. *If all the smoothers* S_j *are symmetric with eigenvalues in* $[0,1]$, *then the backfitting algorithm converges to some solution of the normal equations.*

If $\tilde{\mathbf{f}}$ is an arbitrary solution of (19): $\hat{\mathbf{P}}\tilde{\mathbf{f}} = \hat{\mathbf{Q}}\mathbf{y}$, the effect of the iterations can be described by

$$\mathbf{f}^m - \tilde{\mathbf{f}} = \hat{\mathbf{T}}^m(\mathbf{f}^0 - \tilde{\mathbf{f}}).$$

Letting $m \to \infty$, this becomes

$$\mathbf{f}^\infty - \tilde{\mathbf{f}} = \hat{\mathbf{T}}^\infty(\mathbf{f}^0 - \tilde{\mathbf{f}}),$$

where $\mathbf{f}^\infty$ is the solution of $\hat{\mathbf{P}}\mathbf{f} = \hat{\mathbf{Q}}\mathbf{y}$ which backfitting produces from the initialization $\mathbf{f}^0$. The mapping $\hat{\mathbf{T}}^\infty$ is an oblique projection (i.e., idempotent linear transformation) which maps $\mathbb{R}^{np}$ onto concurvity space $\mathscr{R}(\hat{\mathbf{T}}^\infty) = \mathscr{N}(\hat{\mathbf{P}})$. If the normal equations are nonsingular, that is, $\mathscr{N}(\hat{\mathbf{P}}) = 0$, the limiting map $\hat{\mathbf{T}}^\infty$ is 0, and the backfitting iterates converge to the unique solution $\tilde{\mathbf{f}} = \hat{\mathbf{P}}^{-1}\hat{\mathbf{Q}}\mathbf{y}$.

The speed of convergence of the sequence $\hat{\mathbf{T}}^m$ to $\hat{\mathbf{T}}^\infty$ depends on the largest absolute eigenvalue $|\lambda| < 1$, which can also be described as the spectral radius $\rho(\hat{\mathbf{T}} - \hat{\mathbf{T}}^\infty)$. The reason is that $\hat{\mathbf{T}}^\infty$ is exactly the part of the Jordan decomposition of $\hat{\mathbf{T}}$ which belongs to $\lambda = 1$. The asymptotic rate of convergence for $\hat{\mathbf{T}}^m$ is $\rho(\hat{\mathbf{T}} - \hat{\mathbf{T}}^\infty)^m$ [Householder (1964), Section 7.4].

If, for instance, $\hat{\mathbf{T}} - \hat{\mathbf{T}}^\infty$ has an eigenvalue very close to 1 ($1\text{–}10^{-3}$ say), it might as well be considered equal to 1 for all practical purposes, because the 2 to 100 iterations one allows in practice will be unable to power it down far below 1 ($1\text{–}2 \cdot 10^{-3}$ to $1\text{–}10^{-1}$ approximately). These simple considerations suggest that there is a link between the total number of iterations applied and the strength of the effect of approximate concurvity.

Later we describe modified algorithms that partially account for these situations by *removing* $\mathscr{M}_1(S_j)$ from each of the smoothers.

4.2. *Convergence of backfitting for two smoothers.* We examine the two-smoother case separately because we can show stronger results than for p smoothers, and the effect of concurvity on the behavior of the backfitting algorithm can be worked out explicitly.

The normal equations

$$\begin{aligned}\mathbf{f}_1 &= S_1(\mathbf{y} - \mathbf{f}_2),\\ \mathbf{f}_2 &= S_2(\mathbf{y} - \mathbf{f}_1)\end{aligned} \tag{28}$$

can be formally solved in closed form in some cases:

$$\begin{aligned}(I - S_1S_2)\mathbf{f}_1 &= S_1(I - S_2)\mathbf{y},\\ (I - S_2S_1)\mathbf{f}_2 &= S_2(I - S_1)\mathbf{y},\end{aligned}$$

which leads to the unique solutions

$$\begin{aligned}\mathbf{f}_1 &= (I - S_1S_2)^{-1}S_1(I - S_2)\mathbf{y},\\ \mathbf{f}_2 &= (I - S_2S_1)^{-1}S_2(I - S_1)\mathbf{y}\end{aligned} \tag{29}$$

if the inverse exists. A common sufficient condition is $\|S_1S_2\| < 1$ for an arbitrary matrix norm [Householder (1964), Section 2.5]. In Section 3.5 we saw that this is sufficient for consistency and nondegeneracy.

We wish to obtain a formal solution of the normal equations (19) and to exhibit the convergence points of backfitting in the face of exact concurvity. We will do so under the assumption of symmetric smoothers with eigenvalues in the half-open interval $(-1, 1]$. The results are therefore more general than for p smoothers.

We decompose $S_1 = \tilde{S}_1 + H_U$ and $S_2 = \tilde{S}_2 + H_U$, where H_U is the orthogonal projection onto $U = \mathscr{M}_1(S_1) \cap \mathscr{M}_1(S_2)$. We have $\tilde{S}_j H_U = H_U \tilde{S}_j = 0$ and $\|\tilde{S}_1 \tilde{S}_2\|_2 < 1$. This latter inequality is immediate from the fact that $\mathscr{M}_1(S_1 S_2) = \mathscr{M}_1(S_2 S_1) = U$ under the present assumptions. Invariance of U and $U^\perp$ under S_1 and S_2 allows us to examine separately the Gauss–Seidel process on the two subspaces.

Consider first $\mathbf{y}$ and $\mathbf{f}_2^0$ in $U^\perp$: For such data and initialization, the normal equations have only one solution which is the convergence point of Gauss–Seidel,

$$\mathbf{f}_1^\infty = \left(I - \tilde{S}_1 \tilde{S}_2\right)^{-1} \tilde{S}_1 \left(I - \tilde{S}_2\right) \mathbf{y},$$

$$\mathbf{f}_2^\infty = \left(I - \tilde{S}_2 \tilde{S}_1\right)^{-1} \tilde{S}_2 \left(I - \tilde{S}_1\right) \mathbf{y}.$$

Second, for $\mathbf{y}$ and $\mathbf{f}_2^0$ in U, the Gauss–Seidel process comes to rest after one cycle of updates,

$$\mathbf{f}_1^1 = H_U\left(\mathbf{y} - \mathbf{f}_2^0\right) = H_U \mathbf{y} - H_U \mathbf{f}_2^0,$$

$$\mathbf{f}_2^1 = H_U\left(\mathbf{y} - \mathbf{f}_1^1\right) = H_U \mathbf{f}_2^0$$

and $\mathbf{f}_1^1 = \mathbf{f}_1^m = \mathbf{f}_1^\infty$, $\mathbf{f}_2^2 = \mathbf{f}_2^m = \mathbf{f}_2^\infty$. Putting pieces together, we get Theorem 10:

THEOREM 10. *If S_1 and S_2 are symmetric with eigenvalues in $(-1, 1]$, then the Gauss–Seidel algorithm converges to a solution of the normal equations, and*

$$\begin{aligned} \mathbf{f}_1^\infty &= \left(I - \tilde{S}_1 \tilde{S}_2\right)^{-1} \tilde{S}_1 \left(I - \tilde{S}_2\right) \mathbf{y} + H_U \mathbf{y} - H_U \mathbf{f}_2^0, \\ \mathbf{f}_2^\infty &= \left(I - \tilde{S}_2 \tilde{S}_1\right)^{-1} \tilde{S}_2 \left(I - \tilde{S}_1\right) \mathbf{y} + H_U \mathbf{f}_2^0, \end{aligned} \tag{30}$$

where $U = \mathscr{M}_1(S_1) \cap \mathscr{M}_1(S_2)$, $\tilde{S}_j = S_j - H_U$ and $\mathbf{f}_2^0$ is the initialization.

Interpretation. The components $\mathbf{f}_1^\infty$ and $\mathbf{f}_2^\infty$ can be decomposed into (a) the part within $U^\perp$ which is uniquely determined and depends on the data $\mathbf{y}$ only; (b) the part within U which depends on the sequence of iteration (we started updating $\mathbf{f}$ from $\mathbf{f}_2^0$) and the initialization $\mathbf{f}_2^0$. The corollary shows that the concurvity component $H_U \mathbf{y}$ of the response $\mathbf{y}$ gets absorbed into the first component $\mathbf{f}_1^\infty$. Indeed, the absorption of $H_U \mathbf{y}$ occurs at the very first update $\mathbf{f}_2^0 \mapsto \mathbf{f}_1^1$ and does not change any more as we see from (30). The situation is opposite for the concurvity component $H_U \mathbf{f}_2^0$ of the initialization $\mathbf{f}_2^0$. Since it is part of $\mathbf{f}_2^0$, it stays part of all iterates $\mathbf{f}_2^m$ [see (30)], while the iteration $\mathbf{f}_1^m$ gets suitably adjusted by subtracting out $H_U \mathbf{f}_2^0$. The terms $H_U \mathbf{y}$ and $H_U \mathbf{f}_2^0$ represent the arbitrariness of choices in decomposing $\hat{\mathbf{y}}$ into the two components $\mathbf{f}_1^\infty$ and $\mathbf{f}_2^\infty$ in the presence of concurvity. The backfitting algorithm imposes a choice by

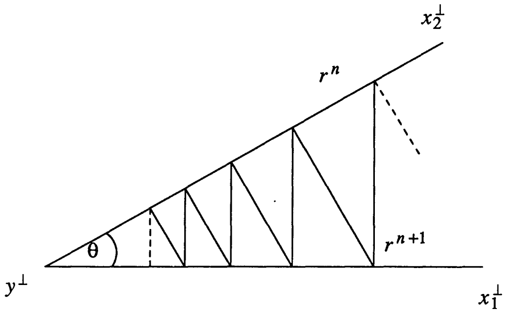


FIG. 10. *Backfitting to the least-squares linear fit with two covariates. The residual* $\mathbf{r}^m$ *converges to the least-squares residual* $\mathbf{y}^\perp$ *in a zigzag fashion. In the figure, all vectors are projected onto the* $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2$ *plane.*

forcing a user to make a decision on the sequence of iteration and the initialization. This corresponds to picking a specific generalized inverse of $\hat{\mathbf{P}}$ in the normal equations (19).

It is worthwhile to examine some special cases in detail.

(a) *Linear regression.* Suppose both S_1 and S_2 produce least-squares fits with an intercept, and we start with both functions $\mathbf{0}$. Then $\mathbf{f}_1$ absorbs the means, and the algorithm converges to the least-squares projection $\hat{\mathbf{y}}$ onto the space spanned by $\mathbf{1}$, $\mathbf{x}_1$ and $\mathbf{x}_2$. It can be shown that $\|\tilde{S}_1\tilde{S}_2\|_2 = \cos(\theta)$, where θ is the angle between $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ (see Figure 10), where $\tilde{\mathbf{x}}_j$ denotes centered $\mathbf{x}_j$. The matrices $\tilde{S}_1$ and $\tilde{S}_2$ correspond to simple regressions through the origin on $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$.

Since in addition the S_j are projections, it can be shown that at the mth iteration, the residual $\mathbf{r}^m$ is given by

$$\mathbf{r}^m = \mathbf{y}^\perp + \left[(I - \tilde{S}_2)(I - \tilde{S}_1)\right]^m \hat{\mathbf{y}}, \tag{31}$$

where $\mathbf{y}^\perp$ is the true least-squares residual. Thus convergence is geometric with rate equal to the cosine of the smallest angle between the two spaces [see Deutsch (1983) for this and more general results]. When $\|\tilde{S}_1\tilde{S}_2\|_2 = 1$ $(\theta = 0)$, then $\tilde{\mathbf{x}}_1 = c\tilde{\mathbf{x}}_2$, and the algorithm converges in one step with $\mathbf{f}_1 = \hat{\mathbf{y}}$ and $\mathbf{f}_2 = \mathbf{0}$.

(b) *Running-line smoothers.* One can show numerically that the backfitting algorithm may not converge in this case. For example if $n = 5$ and $\mathbf{x}_1 = (1, 2, 3, 4, 5)^t$, $\mathbf{x}_2 = (1, 2, 4, 3, 5)$ and both spans are 0.5, then $\|S_1S_2\|_2 \approx 1.07$. The algorithm will not converge if, for example, $\mathbf{y} = (0.687, 0.230, -0.003, -0.287, -0.626)^t$. In a later section we propose an improved algorithm that converges empirically for this example.

(c) *Analysis of variance.* Consider a two-way design with unequal numbers of replications in each cell. As an (inefficient) alternative to the usual least-squares estimation, one can estimate the row and column effects alternately, iterating until convergence. This is a case of backfitting with two bin smoothers (see Section 2.1, Example 3). When only means are used, this approach has little value. However, if the means are replaced by medians or some other robust measure of location, a resistant method for analysis of variance can be produced. This is discussed by Mosteller and Tukey (1977) who call it "median polish." Its convergence is studied by Siegel (1983), Kemperman (1984) and Light and Cheney (1985). Because of the nonlinear nature of medians, this case is outside the scope of this paper.

4.3. *Successive over-relaxation.* In its simplest form, successive over-relaxation is a modification of the Gauss–Seidel procedure in the following sense: The update of component j is a linear combination

$$\mathbf{f} \leftarrow (1 - \omega_j)\mathbf{f} + \omega_j \begin{pmatrix} \mathbf{f}_1 \\ \vdots \\ S_j(\mathbf{y} - \Sigma_{k \neq j}\mathbf{f}_k) \\ \vdots \\ \mathbf{f}_p \end{pmatrix}$$

of the old vector and the Gauss–Seidel update. For the homogeneous problem we therefore consider the update mappings

$$\tilde{\mathbf{T}}_j\mathbf{f} = (1 - \omega_j)\mathbf{f} + \omega_j \begin{pmatrix} \mathbf{f}_1 \\ \vdots \\ S_j(-\Sigma_{k \neq j}\mathbf{f}_k) \\ \vdots \\ \mathbf{f}_p \end{pmatrix}.$$

Under the condition of Theorem 9 the quadratic form $Q(\tilde{\mathbf{T}}_j\mathbf{f})$ as a function of the relaxation parameter ω_j is a parabola symmetric about its minimum at $\omega_j = +1$. Therefore, $Q(\tilde{\mathbf{T}}_j\mathbf{f}) < Q(\mathbf{f})$ for all values $0 < \omega_j < 2$ iff $Q(\hat{\mathbf{T}}_j\mathbf{f}) < Q(\mathbf{f})$. As a consequence, the reasoning which lead to Theorem 9 applies to $\tilde{\mathbf{T}} = \tilde{\mathbf{T}}_p\tilde{\mathbf{T}}_{p-1} \cdots \tilde{\mathbf{T}}_1$ as well as for arbitrary fixed values $\omega_j \in (0, 2)$.

PROPOSITION 11. *Under the assumptions of Theorem* 9, *the successive over-relaxation modification of the backfitting algorithm for* $\omega_j \in (0, 2)$ *converges to some solution of the normal equations.*

PROPOSITION 12. *Make the assumptions of Theorem* 9, *and consider the two-smoother case. If we allow only one nontrivial relaxation parameter* ω_1, *while* $\omega_2 = 1$, *then the value of* ω_1 *that decreases* Q *the most, for a given* $\mathbf{f} \in \mathscr{R}(\hat{\mathbf{T}}_2)$, *is*

$$\omega_1 = \frac{Q((I - \hat{\mathbf{T}}_1)\mathbf{f})}{Q(\hat{\mathbf{T}}_2(I - \hat{\mathbf{T}}_1)\mathbf{f})} \geq 1.$$

Note that considering only $\mathbf{f} \in \mathscr{R}(\hat{\mathbf{T}}_2)$ is no essential restriction: This is always satisfied after the first iteration

$$\hat{\mathbf{T}}_2\big((1-\omega_1)\mathbf{I} + \omega_1\hat{\mathbf{T}}_1\big)\mathbf{f} \in \mathscr{R}\big(\hat{\mathbf{T}}_2\big)$$

since we leave $\omega_2 = 1$ unrelaxed.

For more than two smoothers, we cannot expect a general result on speed-up for over-relaxation. The analytical reason is that the product of more than two orthogonal projections can have negative eigenvalues, while a product of two cannot. This fact could be used to construct examples where over-relaxation is detrimental. This may be an atypical case, however, and some over-relaxation may still result in speed-up in most cases.

We conclude this section with a remark on Jacobi or simultaneous iterations. The naive (generally divergent) version is $\mathbf{f}_j^{\text{new}} = S_j(-\sum_{k \neq j}\mathbf{f}_k^{\text{old}})$ for the homogeneous normal equations. This can also be written as $\mathbf{f}^{\text{new}} = (\mathbf{I} - \hat{\mathbf{P}})\mathbf{f}^{\text{old}}$. We mentioned in Section 3 that a certain amount of under-relaxation is necessary to achieve convergence:

$$\mathbf{f}^{\text{new}} = (1-\omega)\mathbf{f}^{\text{old}} + \omega(\mathbf{I} - \hat{\mathbf{P}})\mathbf{f}^{\text{old}} = (\mathbf{I} - \omega\hat{\mathbf{P}})\mathbf{f}^{\text{old}}$$

As for the Gauss–Seidel iterator $\hat{\mathbf{T}}$ it is immediate that $\mathscr{M}_1(\mathbf{I} - \omega\hat{\mathbf{P}}) = \mathscr{N}(\hat{\mathbf{P}})$ for $\omega \neq 0$. We wish to examine for which $\omega \neq 0$ the iteration $\mathbf{I} - \omega\hat{\mathbf{P}}$ has no absolute eigenvalue ≥ 1 other than possibly $+1$. Lemma 8.1 in the Appendix will assure convergence of $(\mathbf{I} - \omega\hat{\mathbf{P}})^m$ to an oblique projection onto $\mathscr{N}(\hat{\mathbf{P}})$.

One can show that $\hat{\mathbf{P}}$ is diagonalizable, that its eigenvalues are real, nonnegative and bounded by p [Buja, Donnell and Stuetzle (1986)]. If λ is an eigenvalue of $\hat{\mathbf{P}}$, so is $1 - \omega\lambda$ for $\mathbf{I} - \omega\hat{\mathbf{P}}$, and we recognize that $(\mathbf{I} - \omega\hat{\mathbf{P}})^m$ converges to the eigenprojection of $\hat{\mathbf{P}}$ for the eigenvalue $\lambda = 0$ iff $0 < \omega < 2/\rho(\hat{\mathbf{P}})$. A conservative choice would therefore be $\omega = 2/p$, since generally $\rho(\hat{\mathbf{P}}) < p$.

4.4. *An improved backfitting algorithm.* Practical experience with the backfitting algorithm has shown that for correlated covariates, a great many iterations can be required to get the correct average slope of the functions. We have seen (Section 4.2) that for linear fitting, the rate of convergence for two variables is equal to their correlation. Since smoothers typically contain eigenspaces with eigenvalue $+1$, it makes sense to extract these "projection parts" from the smoothers and perform the projection in one multiple linear regression step. Denote by H_i the orthogonal projection onto $\mathscr{M}_1(S_i)$. Define the modified smoother $S_j^* = H_j + (I - H_j)S_j = H_j + \tilde{S}_j$ (if S_j is symmetric, then $S_j^* = S_j$). We now define the modified backfitting algorithm.

0. Initialize $\tilde{\mathbf{f}}_1, \dots, \tilde{\mathbf{f}}_p$ and set $\tilde{\mathbf{f}}_+ = \tilde{\mathbf{f}}_1 + \tilde{\mathbf{f}}_2 + \cdots + \tilde{\mathbf{f}}_p$.
1. Regress $\mathbf{y} - \tilde{\mathbf{f}}_+$ onto the space spanned by $\mathscr{M}_1(S_1), \dots, \mathscr{M}_1(S_p)$, that is, set $\mathbf{g} = H(\mathbf{y} - \tilde{\mathbf{f}}_+)$, where H is the orthogonal projection onto $\mathscr{M}_1(S_1) + \cdots + \mathscr{M}_1(S_p)$ in $\mathbb{R}^n$.
2. Fit an additive model to $\mathbf{y} - \mathbf{g}$ using smoothers $\tilde{S}_i$; this step yields an additive fit $\tilde{\mathbf{f}}_+ = \tilde{\mathbf{f}}_1 + \cdots + \tilde{\mathbf{f}}_p$.
3. Repeat steps 1 and 2 until convergence. The final estimate for the fit is $\mathbf{f}_+ = \mathbf{g} + \tilde{\mathbf{f}}_+$.

Step 2 is deliberately vague, since a number of possibilities exist; we consider two:

2(a) The additive fit is obtained by one backfitting loop based on $\tilde{S}_1, \dots, \tilde{S}_p$, that is, p smooths in all;

2(b) as in (a), but the backfitting loop is iterated until convergence.

THEOREM 13. *If S_j, $j = 1, 2, \dots, p$, are symmetric with eigehvalues in $[0,1]$, then the modified backfitting algorithm converges in the sense that $\mathbf{g}, \tilde{\mathbf{f}}_1, \dots, \tilde{\mathbf{f}}_p$ converge.*

PROOF. We prove convergence for both versions corresponding to steps 2(a) and 2(b) given above.

(a) The convergence with step 2(a) follows immediately from Theorem 9. It is a backfitting algorithm with $p + 1$ smoothers: $H, \tilde{S}_1, \dots, \tilde{S}_p$.

(b) For the algorithm with step 2(b), the inner loop converges (each time) once again by Theorem 9. In fact, since all the smoothers $\tilde{S}_j$ are strictly shrinking, the inner loop converges to a unique solution $\tilde{\mathbf{f}}_1, \dots, \tilde{\mathbf{f}}_p$. To show that the outer loop converges, we can apply the result from Section 4.2 on backfitting with two smoothers, H: the least-squares projection matrix, and B, the linear operator resulting from the converged inner loop. Notice that neither H nor B are univariate smoothers! We will show that $\|B\|_2 < 1$ which implies $\|HB\|_2 < 1$ since $\|H\|_2 = 1$. Let $A_j = (I - \tilde{S}_j)^{-1}\tilde{S}_j$ and $A = \sum_{j=1}^{p} A_j$ as in Section 3.6, Proposition 3; thus $B = (I + A)^{-1}A$. From the proof of Proposition 3, A is symmetric and nonnegative definite, and thus B is symmetric with nonnegative eigenvalues less than 1, as eigenvalues θ of A translate into eigenvalues $\theta/(1 + \theta)$ of B with the same eigenvectors. □

One might compromise between steps 2(a) and 2(b) and perform a fixed number $q > 1$ of inner loops. In practice, none of these alternatives appears to dominate, and all dominate the original backfitting algorithm. We discuss convergence issues further in Section 5.3.

If the S_j are symmetric with eigenvalues in $[0,1]$, then $\tilde{S}_j = S_j - H_j$, and $\|\tilde{S}_j\|_2 < 1$. Cubic smoothing splines belong to this class, and hence the algorithm always converges for them. Running-line smoothers are asymmetric and may have a singular value > 1, so the result cannot be applied. If cubic smoothing splines are used for all predictors, H is the projection matrix corresponding to the least-squares regression on $(\mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_p)$. The nonlinear functions $\tilde{\mathbf{f}}$ are uniquely determined. Exact concurvity (collinearity) can show up only in the H step, where it is dealt with in the standard linear least-squares fashion. At convergence, one may then decompose $\mathbf{g} = \sum \mathbf{g}_j$, $\mathbf{g}_j \in \mathscr{M}_1(S_j)$, and reconstruct final components $\mathbf{f}_j = \mathbf{g}_j + \tilde{\mathbf{f}}_j$. If S_j is a cubic spline smoother and if $\mathbf{y}$ was centered initially, then $\mathbf{g}_j = \hat{\beta}_j \cdot \mathbf{x}_j$, where $\hat{\beta}_1, \dots, \hat{\beta}_p$ are the coefficients from the multiple linear regression of $\mathbf{y} - \sum \tilde{\mathbf{f}}_k$ on $\mathbf{x}_1, \dots, \mathbf{x}_p$.

THEOREM 14. *Suppose the modified backfitting algorithm has converged with smoothers $\tilde{S}_j$ and projection H, yielding functions $\tilde{\mathbf{f}}_j$ and $\mathbf{g}_j \in \mathscr{M}_1(S_j)$. Then the components $\mathbf{f}_j = \mathbf{g}_j + \tilde{\mathbf{f}}_j$ are solutions to the normal equations with smoothers $S_j^* = H_j + (I - H_j)S_j$.*

We prove Theorem 14 in Appendix A.2. Notice that we are not assuming symmetry in contrast to most previous results. For symmetric smoothers ($S_j^* = S_j$), the theorem says that the modified backfitting algorithm solves the original system of equations. Using this result, together with the closed form expression in Proposition 3, it is easy to arrive at Proposition 15.

PROPOSITION 15. *With H and B defined above, the closed form expressions for the normal equation solutions are $\tilde{\mathbf{f}}_+ = (I - BH)^{-1}B(I - H)\mathbf{y}$ and $\mathbf{g} = H(\mathbf{y} - \tilde{\mathbf{f}}_+)$. These can be combined to form*

$$\begin{aligned} \mathbf{f}_+ &= S_+\mathbf{y} \\ &= \big(H + (I - H)(I - BH)^{-1}B(I - H)\big)\mathbf{y}, \\ \tilde{\mathbf{f}}_j &= \big(I - \tilde{S}_j\big)^{-1}\big(\mathbf{y} - \mathbf{g} - \tilde{\mathbf{f}}_+\big) \end{aligned}$$

and

$$\mathbf{f}_j = \mathbf{g}_j + \tilde{\mathbf{f}}_j,$$

where the $\mathbf{g}_j$'s are any vectors $\mathbf{g}_j \in \mathscr{M}_1(S_j)$ such that $\sum \mathbf{g}_j = \mathbf{g}$.

5. Further features of additive models.

5.1. *Weighted penalized least squares.* In many situations, one prefers weighted least squares. This is the case if the observations have known but unequal relative precisions, or when the least-squares problem is part of another iterative procedure, for example, the local scoring procedure of Hastie and Tibshirani (1986a, b). An additional complication may arise when the data are correlated with a covariance matrix which is known up to a constant. All these cases are covered by the penalized least-squares criterion

$$Q(\mathbf{f}) = \left(\mathbf{y} - \sum_{j=1}^{p} \mathbf{f}_j\right)^t W\left(\mathbf{y} - \sum_{j=1}^{p} \mathbf{f}_j\right) + \sum_{j=1}^{p} \mathbf{f}_j^t K_j \mathbf{f}_j, \tag{32}$$

which is to be minimized under the constraints $\mathbf{f}_j \in U_j$. The matrix W is a symmetric positive-definite matrix (generally the inverse of the covariance matrix up to a scale factor), K_j is the positive semidefinite penalization matrix and U_j is the constraint space for the jth predictor. The customary coordinate transformation $\tilde{\mathbf{y}} = W^{1/2}\mathbf{y}$, $\tilde{\mathbf{f}}_j = W^{1/2}\mathbf{f}_j$, $\tilde{K}_j = W^{-1/2}K_jW^{-1/2}$, $\tilde{U}_j = W^{1/2}U_j$, brings us back to unweighted penalized constrained least squares. Thus, all results on existence, degeneracy and convergence of algorithms apply to this situation as well. We may just add that in the unconstrained case the smoother matrix associated with the jth predictor can be written $S_j = (I + W^{-1}K_j)^{-1}$.

This is obviously not symmetric in the canonical coordinate system, but it is so after the above coordinate transformation.

5.2. *Remarks on inference for additive models.* For the remainder of this section we will assume that $y_i = g(\mathbf{x}_i) + \varepsilon_i$, where $g(\mathbf{x}_i)$ is the true regression function and the errors ε_i are uncorrelated with zero expectation and common variance σ^2.

5.2.1. *Bias and consistency.* As in the univariate case, a fitted additive model will typically be biased, unless rigid assumptions are made. For example, if we assume that $g(\mathbf{x})$ is a polynomial of fixed degree, then the appropriate least-squares fit will be unbiased. If we assume that g is additive but otherwise arbitrary, the additive fits will be biased just as in the univariate case. Typically investigations of finite sample bias involve simulation and bootstrap methods.

Asymptotic consistency is a more manageable issue that has been studied in the literature. Either we assume the additive model is correct, or study consistency for the projection of g onto the space of additive fits. Breiman and Friedman (1985) discuss consistency using simple running-mean smoothers. Stone (1985) gives a detailed study of rates of convergence for additive model fits using regression splines, and shows that they have the same rate as a univariate fit. The details are beyond the scope of this paper.

5.2.2. *Variance.* From the previous sections we note that each estimated smooth from the backfitting algorithm is the result of a linear mapping or smoother applied to $\mathbf{y}$. This means that the variance and degrees-of-freedom formulas developed earlier can be applied to the backfitting algorithm. At convergence, we can express the np vector of fits as $\mathbf{f} = \hat{\mathbf{P}}^{-}\hat{\mathbf{Q}}\mathbf{y} \equiv \mathbf{R}\mathbf{y}$. If the observations have i.i.d. errors, then $\operatorname{cov}(\mathbf{f}) = \mathbf{R}\mathbf{R}^t\sigma^2$, where $\sigma^2 = \operatorname{var}(y_i)$. As in the least-squares case, if $\hat{\mathbf{P}}$ has eigenvalues close to 0, this will be reflected in $\operatorname{cov}(\mathbf{f})$ as large variances and covariances. In this setting eigenvalues equal to 0 do not result in infinite variances; since they are ignored in the generalized inverse they do not contribute to the covariance matrix. The covariance matrix is restricted to the subspace for which the estimates are unique. Rather the infinite variances are reflected in our freedom of choice of the starting values and thus the solutions found by Gauss–Seidel.

Direct computation of $\mathbf{R}$ is formidable; instead we apply the backfitting procedure to the unit vectors. The result of backfitting the ith unit vector is the ith column of $\mathbf{R}$. The confidence bands in Figure 7 were constructed using $\pm$twice the square root of the diagonal elements of $\mathbf{R}\mathbf{R}^t$. Hastie (1988) has developed parametric approximations to the additive model fit; amongst other uses, they provide useful approximations to such second-order information as is sought here, without any iterations. They are also useful for demonstrating the effects of approximate concurvity on the standard errors.

5.2.3. *Degrees of freedom.* It would be convenient if the degrees of freedom of the additive model were additive; that is, if the total degrees of freedom were simply $\Sigma_1^p \operatorname{tr}(2S_i - S_i^t S_i)$. This is the case if each S_i is an orthogonal projection. If this held true, the computation of degrees of freedom would be much easier, for one would have only to compute the degrees of freedom of each individual smoother. We will briefly investigate the additivity of degrees of freedom in an additive model.

Consider first a single-smoother situation. We assume that all smoothers are centered, shrinking and symmetric, with nonnegative eigenvalues. Denote the eigenvalues of the smoother by $1 \geq \tau_1 \geq \tau_2 \geq \cdots \geq \tau_n \geq 0$. The expression $\operatorname{tr}(2S - S^t S)$ equals $\Sigma_1^n \tau_i(2 - \tau_i)$. Now consider a multiple predictor backfitting algorithm with smoothing matrices $S_1, S_2, \dots, S_p$. It turns out that the exact degrees of freedom of the fitted model is difficult to compute analytically; in order to get an upper bound we consider the extreme case of concurvity in which the p-smoother matrices are identical. A fairly straightforward calculation using the expression for the fitted model given in Appendix A.2 shows that the degrees of freedom of the resultant fit is $\Sigma_1^n \theta_i(2 - \theta_i)$, where $\theta_i = p\tau_i/(1 + (p - 1)\tau_i)$. Now if we simply were to add up the degrees of freedom of the p smoothers, the estimated degrees of freedom would be $p\Sigma_1^n \tau_i(2 - \tau_i)$. How do these quantities relate? It is easy to show that

$$\sum_1^n \tau_i(2 - \tau_i) \leq \sum_1^n \theta_i(2 - \theta_i) \leq p\sum_1^n \tau_i(2 - \tau_i). \tag{33}$$

We see that (1) smoothing on the same covariate increases the degrees of freedom and (2) adding up the degrees of freedom of the individual fits provides an upper bound on the true degrees of freedom of the fitted model.

Does this relationship hold in intermediate situations, that is, if the smoother matrices are not identical? We ran a small experiment to investigate this further. Two samples of size 20 were generated from the standard normal distribution and the second sample was adjusted so that its sample correlation with the first sample took on the values 0, 0.5, 0.9 and 1. Then we computed the quantities $\Sigma_1^n \theta_i(2 - \theta_i)$ and $2\Sigma_1^n \tau_i(2 - \tau_i)$ for a cubic spline smoother with various values of the smoothing parameter. The results are shown in Figure 11. (When we repeated this experiment so that different sets of design values were generated, the results changed very little.)

We see that adding up the degrees of freedom of the two smoothers is quite a good approximation, and is only inaccurate when a small smoothing parameter is used or the covariates are very highly correlated.

5.3. *A closer look at convergence.* In this section we demonstrate the convergence patterns for the algorithms presented so far, as well as for other variants and algorithms for solving the system (19).

Consider the case of extreme collinearity, where we have two identical covariates and a cubic spline smoother matrix S. Some simple calculations show that starting the backfitting algorithm from $\mathbf{0}$, the residual after m smooths is given

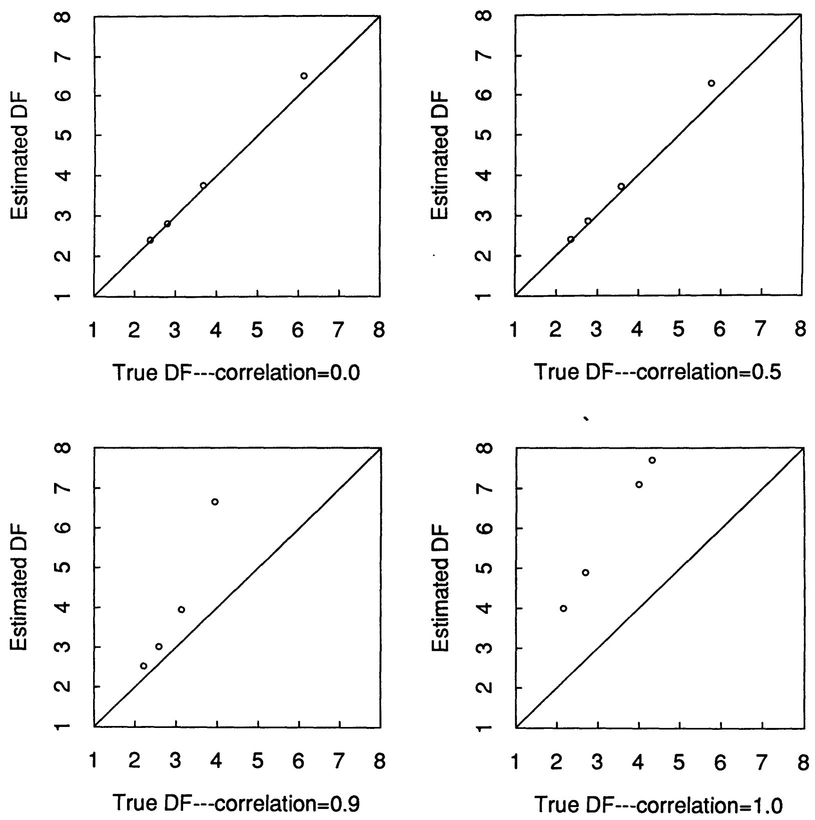


FIG. 11. *The "true" DF of the additive fit, based on the formula* $\operatorname{tr}(2R - R^tR)$ *vs. the DF obtained by adding* DF_i *for the individual smoothers.*

by

$$
\begin{aligned}
\mathbf{r}^m &= \left[I - S + S^2 - S^3 + \cdots + (-1)^{m-1}S^{m-1}\right](I - S)\mathbf{y} \\
&\xrightarrow[\infty]{m} (I + S)^{-1}(I - S)\mathbf{y}.
\end{aligned}
\tag{34}
$$

This shows us a number of things:

1. The residuals (and their norm) oscillate as they converge.
2. The converged model is *rougher* than a single smoother. This is true since the eigenvalues of $(I + S)^{-1}(I - S)$ are at most those of $I - S$, so the residuals are "smaller."
3. By looking at every other iteration,

$$
\mathbf{r}^{2m} = (I + S^2 + S^4 + \cdots + S^{2(m-1)})(I - S)^2\mathbf{y}, \tag{35}
$$

it is clear that the norm of the residuals converges *upwards*, after every even number of steps.

4. $\mathbf{r}^2$ is the same as the "twicing" [Tukey (1977)] residual, where twicing enhances a smooth by adding in the smooth of the residual. If twicing is continued, however, the "$n$ing" residual is $(I - S)^n$ which converges to 0 for shrinking smoothers.

We have seen similar behavior in real data examples where the variables are strongly correlated. One point is clear: It is not appropriate to track the residual sum of squares to test for convergence when backfitting with smoothers. This should not surprise us, since we are minimizing a penalized residual sum of squares.

Figure 12(a) shows the convergence patterns for the ozone data analyzed in Section 3.1. Spline smoothers were used, standardized to have the same degrees of freedom of approximately 4. The upper curve (dots) is log(RSS − 5800) for the ordinary backfitting algorithm, plotted against the number of smoothers. The middle curve labeled "m" is the same for the modified backfitting algorithm, with one inner backfitting iterations per least-squares fit. We have counted the least-squares fit as one smooth in this comparison. The modified algorithm is clearly an improvement. The lower curve is once again the regular algorithm, but with special initial functions. If $\mathbf{s}_j$ is the vector of fitted values obtained from the simple spline smooth against variable j, then the initial function for this variable is $\alpha_{j1}\mathbf{x}_j + \alpha_{j2}\mathbf{s}_j$, where the α_{ji} are chosen *globally* by least squares. This has the same flavor as the augmented partial residual plots of Mallows (1986), and in this example does very well. It is interesting to note again that for the "s" curve the RSS increases initially!

We use as convergence criterion $\Delta_m = \Sigma_{j=1}^{p}\|\mathbf{f}_j^m - \mathbf{f}_j^{m-1}\|^2$, the sum of squares of the changes in the functions after each inner loop. An alternative would be to

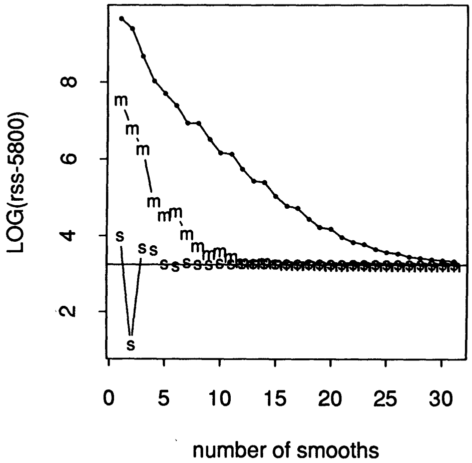


FIG. 12(a). *Convergence patterns for three different backfitting algorithms. The dotted curve is the standard algorithm starting from* **0**, *the curve labeled* "m" *the modified algorithm, and the curve labeled* "s" *the modified algorithm with special initial functions.*

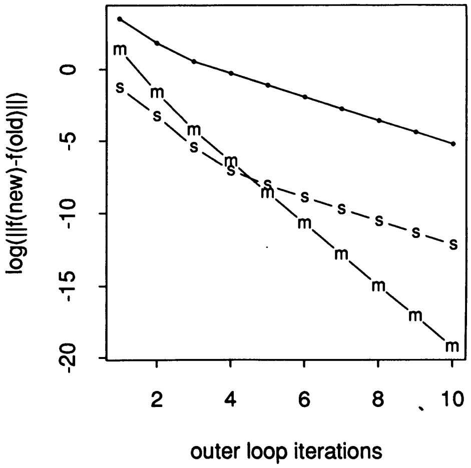


FIG. 12(b). *The change in the squared norm Δ_m for the three algorithms, plotted on the logarithmic scale. Each point corresponds to a complete cycle of the inner loop.*

use the penalized residual sum of squares itself. Figure 12(b) plots $\log(\Delta_m)$ vs. m for the three algorithms above. The modified algorithm appears to have a higher convergence rate than the unmodified algorithm. Our experience to date indicates that it is hard to improve dramatically over the regular algorithm, provided good starting values are used.

5.4. *Related methods*: *linear models with a single smooth term.* Consider an additive model in which all but one term is assumed to be linear. The corresponding backfitting algorithm can be thought of as having two smoothers, one representing a least-squares fit $X\hat{\boldsymbol{\beta}}$ on one or more covariates (represented by the design matrix X) and the other a smoother S_2 producing an estimate $\mathbf{f}_2^\infty$. The backfitting steps are $\mathbf{f}_1 = S_1(\mathbf{y} - \mathbf{f}_2) = X(X^tX)^{-1}X^t(\mathbf{y} - \mathbf{f}_2) \equiv X\hat{\boldsymbol{\beta}}$, and $\mathbf{f}_2 = S_2(\mathbf{y} - X\hat{\boldsymbol{\beta}})$. It turns out that we can solve for $\hat{\boldsymbol{\beta}}$ and $\mathbf{f}_2^\infty$ explicitly,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(X^t(I - S_2)X\right)^{-1}X^t(I - S_2)\mathbf{y},\\ \mathbf{f}_2^\infty &= S_2(\mathbf{y} - X\hat{\boldsymbol{\beta}}).\end{aligned} \tag{36}$$

Green and Yandell (1985) derived (36) and a more general version of it in their work on semiparametric generalized linear models. Within a general likelihood model they allow a smooth function of one or more variables, and base its estimation on a penalized likelihood approach.

Denby (1984) derived (36) as a method for discovering nonlinearity for a single covariate in regression. Her starting point was not the backfitting algorithm; instead, she considered the penalized least-squares criterion

$$\|\mathbf{y} - X\boldsymbol{\beta} - \mathbf{f}_2\|^2 + \lambda\mathbf{f}_2^tK\mathbf{f}_2, \tag{37}$$

where K is the penalty matrix discussed earlier. Clearly (37) is a special case of (20). Denby also considered an alternative criterion. She chose $\hat{\beta}$ to minimize

$$\|\mathbf{y} - X\beta - S_2(\mathbf{y} - X\beta)\|^2. \tag{38}$$

Surprisingly this has a solution different from (36): $\mathbf{f}_2^\infty = S_2(\mathbf{y} - X\hat{\beta})$ but

$$\hat{\beta} = \left(X^t(I - S_2^t)(I - S_2)X\right)^{-1}X^t(I - S_2^t)(I - S_2)\mathbf{y}. \tag{39}$$

Denby's investigation suggests that in many practical cases the differences will not be substantial.

The solution (36) requires the condition that $(I - S_2)\mathbf{X}$ has full rank; this is met if the covariate in the smoother is not collinear with any of the columns of $\mathbf{X}$. The estimates require $O(n)$ operations to compute, since we apply S successively to each of the columns of $\mathbf{X}$, and this operation is $O(n)$. In general, such explicit expressions for the solution are cumbersome and very expensive to compute. Even in the two-smoother case the expression given by (29) is $O(n^3)$ to compute. In that case backfitting is a much more efficient method for obtaining the solutions.

Hastie and Tibshirani (1987) discuss this special case of the backfitting algorithm as a technique for nonlinear analysis of covariance, and contrast it with other methods in the literature.

Notice that the first equation in (36) is the solution to a generalized least-squares problem with variance $(I - S_2)^{-1}$. Denby (1984) and Green (1987) explore this connection between generalized least squares and smoothing.

5.5. *Convergence of the ACE algorithm.* The ACE algorithm [Breiman and Friedman (1985)] estimates an additive model with a transformation of the response as well as each of the predictors. It does so in an alternating fashion, iterating between a backfitting algorithm applied to the current estimate of the response transformation, and a smoother of the current additive fit on the response. Let $S_1, \dots, S_p$ be the linear smoothers for the predictors, symmetric with eigenvalues in $[0, 1]$ to ensure the convergence of backfitting. Let S_y be the linear smoother for the response. Further, let S_+ be the linear operator defined in Proposition 15 that produces the additive fits. Then Breiman and Friedman show that the ACE procedure converges if the largest eigenvalue of the product operator S_yS_+ is real and positive. Using this, we can show that ACE converges if S_y is symmetric and nonnegative definite. From Proposition 15, $S_+ = (H + (I - H)(I - BH)^{-1}B(I - H)$. One can check that S_+ is symmetric and nonnegative definite. Hence if S_y is symmetric and nonnegative definite, then S_yS_+ is the product of two symmetric, nonnegative definite matrices and therefore has real, nonnegative eigenvalues.

It is interesting that we do not require S_y to be shrinking. The reason is that in the ACE algorithm the transformation for y is rescaled after each iteration, and thus S_y is effectively scaled so that its eigenvalues are in $[0, 1]$.

6. Discussion. In this paper we have looked at linear smoothers and their use in additive models. We summarize the main points.

1. Many useful smoothers, in particular the running-line and cubic spline smoothers, are linear and hence are easily accessible to analysis through the corresponding smoother matrix.
2. Smoother matrix plots, singular value decompositions and self-influence plots are useful ways of investigating the operating characteristics of smoothers. In our limited experience locally weighted running lines and the cubic spline smoother seem to be quite similar in the way they smooth data.
3. The cubic spline smoother matrix is particularly tractable because it is symmetric and has eigenvalues ≤ 1. Only constant and linear functions are passed through a cubic spline smoother unchanged.
4. The additive model is a useful nonparametric regression model that is more flexible than the standard linear model and at the same time much more interpretable than a general high-dimensional regression surface.
5. Estimation of the additive model with linear smoothers leads to a linear system of equations for the unknown functions. The backfitting algorithm provides an efficient method for solving this system and is equivalent to the well-known Gauss–Seidel procedure.
6. We have established consistency of the system of equations and convergence of the Gauss–Seidel procedure (and related methods) when symmetric shrinking smoothers are used. Nonuniqueness occurs when "concurvity" exists and we have studied this phenomenon in some detail.
7. A penalized least-squares criterion has been derived, whose minimum is given by this same system of equations. This connection was exploited in establishing the consistency, degeneracy and convergence results.
8. We have developed modified backfitting algorithms that separate out the eigenspaces of eigenvalue 1. The resultant procedure is faster than the usual algorithm and we have proven its convergence for symmetric, shrinking smoothers.
9. We have described some inferential tools for linear smoothers and additive models including estimation of the number of parameters of the fitted model and standard error bands for the functions.

This work leaves open a number of issues for further study. Many of these have been mentioned already. We raise some additional questions below.

(i) How do the various smoothers perform with real data? This is a very complex question that might be addressed with a large-scale simulation study.

(ii) Can the results for additive models be extended to algorithms that use nonlinear smoothers? Many simple smoothers are nonlinear, for example, the running-median smoother. Also, as mentioned earlier, if a data-based criterion is used to pick the smoothness parameter the resultant smoother is nonlinear. More complicated smoothers, such as the "supersmoother" [Friedman and Stuetzle (1982)] which is used in the ACE algorithm, are also nonlinear. Nonlinear smoothers are difficult to analyze theoretically but seem to work well in practice. In the experience of Breiman and Friedman (1985), the backfitting algorithm, using supersmoother, rarely fails to converge. Degrees of freedom and variance of the fit are difficult to compute analytically but simulation can instead be used.

(iii) Can the results for additive models be extended to more complicated models? A number of extensions of additive models have been proposed, all using the backfitting algorithm as part of the estimation process. These include generalized additive models [Hastie and Tibshirani (1986a)] and semiparametric generalized linear models [Green and Yandell (1985)], which extend the class of generalized linear models and other nonlinear models, and projection pursuit regression [Friedman and Stuetzle (1981)] which allows arbitrary linear combinations of the covariates. Hastie and Tibshirani (1986b) have used the results of this paper to establish convergence of the local scoring algorithm for generalized additive models, under suitable conditions.

(iv) How do we assess and deal with approximate concurvity?

(v) Can computable measures for influence and lack of fit be developed for additive models?

(vi) What are the operating characteristics of the standard error curves? Are they approximate confidence curves, in a pointwise or uniform sense?

APPENDIX

A.1. Proofs of results in Section 3.

Existence of penalized likelihood solutions. O'Sullivan, Yandell and Raynor (1985) proved an important result which makes existence and uniqueness proofs for certain penalized likelihood models simpler—they show that a finite-dimensional approximation always does better for the criterion. We first state their result for the univariate cubic smoothing spline.

Let $\mathscr{S}$ be the Sobolev space of real-valued functions f defined, for simplicity, on $\Omega = [0,1]$, with penalty functional $J_2(f) = \int_0^1 [\ddot{f}\,]^2\,dt$, and inner product $\langle f, g\rangle = f(0)g(0) + \dot{f}(0)\dot{g}(0) + \int_0^1 \ddot{f}(t)\ddot{g}(t)\,dt$.

Consider penalized likelihoods of the form

$$l_{n\lambda}(f) = \sum_{i=1}^{n} l_i(y_i, f(t_i)) - \lambda J_2(f).$$

Let e_i be the piecewise cubic polynomial *representers of evaluation*, such that $f(t_i) = \langle f, e_i\rangle$. Finally, let $\mathscr{S}^n = \mathscr{S}^0 \oplus \{e_i\}_{i=1}^n$, where $\mathscr{S}^0$ is the class of linear functions [for which $J_2(f^0) = 0$]. Then for all $f \in \mathscr{S}$,

$$l_{n\lambda}(f) \le l_{n\lambda}(f^1),$$

where f^1 is the projection of f onto $\mathscr{S}^n$. This shows that the maximizer of $l_{n\lambda}$, if it exists, lies in the finite-dimensional subspace $\mathscr{S}^n$ of $\mathscr{S}$.

The proof of O'Sullivan, Yandell and Raynor has two steps:

1. $f(t_i) = f^1(t_i)$, and
2. $J_2(f) = J_2(f^1) + J_2(f^2)$, where $f = f^1 + f^2$ with $f^2 \in \mathscr{S}^{n\perp}$, and this establishes the result.

Our penalized likelihoods, which include penalized residual sums of squares as a special case, have the form

$$l_{n\lambda}(f) = \sum_{i=1}^{n} l_i\big(y_i, f_1(t_{i1}) + f_2(t_{i2}) + \cdots + f_p(t_{ip})\big) - \sum_{j=1}^{p} \lambda_j J_2(f_j), \tag{40}$$

where $\boldsymbol{f}$ is the vector of functions f_j.

Let $\mathscr{S}_j$ denote the Sobolev space (defined above) for functions of variable t_j, and define the Cartesian product space $\mathscr{S}_{\mathrm{prod}} = \mathscr{S}_1 \times \mathscr{S}_2 \times \cdots \times \mathscr{S}_p$. A natural inner product on $\mathscr{S}_{\mathrm{prod}}$ is

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \sum_{j=1}^{p} \lambda_j \langle f_j, g_j \rangle. \tag{41}$$

$\mathscr{S}_{\mathrm{prod}}$ is a Hilbert space for which the natural imbeddings of $\mathscr{S}_j$ are all closed linear subspaces. Also, the norm topology of $\mathscr{S}_{\mathrm{prod}}$ coincides with the product topology inherited from the factors $\mathscr{S}_j$.

The representers have the form $\boldsymbol{e}_{ij} = (0, \dots, e_{ij}/\lambda_j, 0, \dots, 0)$, where e_{ij} is the ith representer for $\mathscr{S}_j$, and $\langle \boldsymbol{f}, \boldsymbol{e}_{ij} \rangle = f_j(t_{ij})$.

THEOREM 1. *Denote by $\mathscr{S}_j^n$ the version of $\mathscr{S}^n$ for $\mathscr{S}_j$, and let $\mathscr{S}_{\mathrm{prod}}^n = \mathscr{S}_1^n \times \cdots \times \mathscr{S}_p^n$ be the appropriate subspace of $\mathscr{S}_{\mathrm{prod}}$. For any function $\boldsymbol{f} \in \mathscr{S}_{\mathrm{prod}}$, let $\boldsymbol{f}^1 \in \mathscr{S}_{\mathrm{prod}}^n$ denote the vector of functions whose elements are the* coordinate-wise *projections onto $\mathscr{S}_j^n$. Then $l_{n\lambda}(\boldsymbol{f}) \le l_{n\lambda}(\boldsymbol{f}^1)$ for all $\boldsymbol{f} \in \mathscr{S}_{\mathrm{prod}}$.*

PROOF. The proof follows very closely that of O'Sullivan, Yandell and Raynor. For each j,

$$\begin{aligned} f_j(t_{ij}) &= \langle \boldsymbol{f}, \boldsymbol{e}_{ij} \rangle \\ &= \langle \boldsymbol{f}^1, \boldsymbol{e}_{ij} \rangle + \langle \boldsymbol{f}^2, \boldsymbol{e}_{ij} \rangle \\ &= \langle f_j^1, e_{ij} \rangle \\ &= f_j^1(t_{ij}). \end{aligned} \tag{42}$$

Also $\Sigma_j \lambda_j J_{2j}(f_j) = \langle \boldsymbol{f} - \boldsymbol{f}^0, \boldsymbol{f} - \boldsymbol{f}^0 \rangle$, where $\boldsymbol{f}^0$ is the vector of functions with elements the coordinate-wise projections onto $\mathscr{S}_j^0$. This implies that $\Sigma_j \lambda_j J_{2j}(f_j) = \Sigma_j \lambda_j J_{2j}(f_j^1) + \Sigma_j \lambda_j J_{2j}(f_j^2)$. □

THEOREM 4. *For arbitrary linear mappings, the normal equations* (19) *are consistent for arbitrary* $\mathbf{y}$ *iff one of the following two equivalent conditions holds*:

1. $\mathbf{f}_+ = \mathbf{0}$ *whenever* $\hat{\mathbf{P}}^t\mathbf{f} = \mathbf{0}$.
2. $\mathbf{f}_j \in \mathscr{M}_1(S_j^t)$ *for at least one and hence all* j *whenever* $\hat{\mathbf{P}}^t\mathbf{f} = \mathbf{0}$.

PROOF. We wish to show that $\hat{\mathbf{Q}}\mathbf{y} \in \mathscr{R}(\hat{\mathbf{P}})$ iff the conditions of this proposition are met. The proof consists of a specialization of the general fact $\mathscr{R}(\hat{\mathbf{P}}) = \mathscr{N}(\hat{\mathbf{P}}^t)^{\perp}$. Consistency is thus equivalent to $\hat{\mathbf{Q}}\mathbf{y} \perp \mathscr{N}(\hat{\mathbf{P}}^t)$, or $\Sigma_k\langle S_k\mathbf{y}, \mathbf{f}_k\rangle = 0$ for $\mathbf{f} \in \mathscr{N}(\hat{\mathbf{P}}^t)$, that is, $\mathbf{P}^t\mathbf{f} = \mathbf{0}$. The condition $\Sigma_k\langle S_k\mathbf{y}, \mathbf{f}_k\rangle = 0$, if rewritten as $\langle \mathbf{y}, \Sigma_k S_k^t\mathbf{f}_k\rangle = 0$, is seen to be equivalent to $\Sigma_k S_k^t\mathbf{f}_k = \mathbf{0}$ since we require this condition to hold for arbitrary $\mathbf{y} \in \mathbb{R}^n$. Thus consistency is equivalent so far to

$$\sum_{k=1}^{p} S_k^t\mathbf{f}_k = \mathbf{0} \quad \text{whenever } \hat{\mathbf{P}}^t\mathbf{f} = \mathbf{0}.$$

An additional simplification occurs by noticing that the if-part is equivalent to

$$S_j^t\mathbf{f}_j - \mathbf{f}_j = \mathbf{c}, \qquad j = 1,\dots,p \text{ for } \mathbf{c} = \sum_{k=1}^{p} S_k^t\mathbf{f}_k.$$

Summing up, we obtain $\mathbf{c} - \mathbf{f}_+ = p\mathbf{c}$, or $(p-1)\Sigma_k S_k^t\mathbf{f}_k = -\mathbf{f}_+$. Thus for $p > 1$, $\Sigma_k S_k^t\mathbf{f}_k = \mathbf{0}$ iff $\mathbf{f}_+ = \mathbf{0}$. To show equivalence with condition 2, we observe that $\mathbf{c} = \Sigma_k S_k^t\mathbf{f}_k = \mathbf{0}$ iff $S_j^t\mathbf{f}_j - \mathbf{f}_j = \mathbf{c} = \mathbf{0}$ for some (and hence all) j. □

A.2. Proofs of results in Section 4.

PROPOSITION 7. *For arbitrary smoothers*, $\hat{\mathbf{P}}\mathbf{f} = \mathbf{0}$ *iff* $\hat{\mathbf{T}}\mathbf{f} = \mathbf{f}$.

PROOF. Observing that each factor $\hat{\mathbf{T}}_j$ of $\hat{\mathbf{T}} = \hat{\mathbf{T}}_p,\dots,\hat{\mathbf{T}}_1$ modifies only one component $\mathbf{f}_j$ of $\mathbf{f}$ at a time, we conclude that $\hat{\mathbf{T}}\mathbf{f} = \mathbf{f}$ is equivalent to $\hat{\mathbf{T}}_1\mathbf{f} = \mathbf{f}$, $\hat{\mathbf{T}}_2\mathbf{f} = \mathbf{f},\dots,\hat{\mathbf{T}}_p\mathbf{f} = \mathbf{f}$. This is the same as $\mathbf{f}_j = -S_j(\Sigma_{i\neq j}\mathbf{f}_i)$, $j = 1,\dots,p$, that is, $\hat{P}\mathbf{f} = 0$. □

THEOREM 8 (Seminorm descent principle). *If* $|\mathbf{f}|$ *is a complex seminorm and* $\hat{\mathbf{T}}$ *a linear mapping on* $\mathbb{C}^N$ *satisfying* $|\hat{\mathbf{T}}\mathbf{f}| < |\mathbf{f}|$ *unless* $|\mathbf{f}| = 0$, *and* $\hat{\mathbf{T}}\mathbf{f} = \mathbf{f}$ *for* $|\mathbf{f}| = 0$, *then* $\hat{\mathbf{T}}^m$ *converges to a limit* $\hat{\mathbf{T}}^{\infty}$ *with the properties* $|\hat{\mathbf{T}}^{\infty}| = 0$ *for all* $\mathbf{f}$, $(\hat{\mathbf{T}}^{\infty})^2 = \hat{\mathbf{T}}^{\infty}$ *and* $\hat{\mathbf{T}}\hat{\mathbf{T}}^{\infty} = \hat{\mathbf{T}}^{\infty}\hat{\mathbf{T}} = \hat{\mathbf{T}}^{\infty}$.

PROOF. We wish to apply Lemma 8.1 below.

(a) We first show that the eigenvalues of $\hat{\mathbf{T}}$ lie in $\{\lambda \mid |\lambda| < 1 \text{ or } \lambda = 1\}$: If $\hat{\mathbf{T}}\mathbf{f} = \lambda\mathbf{f}$, then $|\hat{\mathbf{T}}\mathbf{f}| = |\lambda|\,|\mathbf{f}|$. However, we have either $|\hat{\mathbf{T}}\mathbf{f}| < |\mathbf{f}|$ or $|\mathbf{f}| = 0$. Thus $|\lambda| < 1$ or $\lambda = 1$.

(b) We show $\mathscr{R}(I - \hat{\mathbf{T}}) \cap \mathscr{M}_1(\hat{\mathbf{T}}) = \{\mathbf{0}\}$: Let $\hat{\mathbf{T}}\mathbf{f} = \mathbf{f}$ and $\mathbf{f} = \mathbf{g} - \hat{\mathbf{T}}\mathbf{g}$ for some $\mathbf{g}$. From $\hat{\mathbf{T}}\mathbf{f} = \mathbf{f}$ follows $|\mathbf{f}| = 0$, hence $|\hat{\mathbf{T}}\mathbf{g} - \mathbf{g}| = 0$. For any seminorm we have $|\,|\hat{\mathbf{T}}\mathbf{g}| - |\mathbf{g}|\,| \leq |\hat{\mathbf{T}}\mathbf{g} - \mathbf{g}|$, hence $|\hat{\mathbf{T}}\mathbf{g}| = |\mathbf{g}|$. It follows that $\hat{\mathbf{T}}\mathbf{g} = \mathbf{g}$, hence $\mathbf{f} = \mathbf{0}$. □

LEMMA 8.1. *The powers* $\hat{\mathbf{T}}^m$ *of a linear mapping* $\hat{\mathbf{T}}$ *converge to a limit* $\hat{\mathbf{T}}^{\infty}$ *iff the following two conditions are satisfied*:

(a) *All (complex) eigenvalues* λ *of* $\hat{\mathbf{T}}$ *lie in* $\{\lambda \mid |\lambda| < 1$ *or* $\lambda = 1\}$.

(b) $\mathscr{R}(I - \hat{\mathbf{T}}) \cap \mathscr{M}_1(\hat{\mathbf{T}}) = \{\mathbf{0}\}$.

PROOF. Let $\hat{\mathbf{T}} = \Sigma_\lambda(\lambda P_\lambda - D_\lambda)$ be the Jordan decomposition [Kato (1984), Section 1.4] of $\hat{\mathbf{T}}$ with eigenprojection P_λ and nilpotent D_λ for the eigenvalues λ. We get $\hat{\mathbf{T}}^m = \Sigma_\lambda(\lambda P_\lambda - D_\lambda)^m$. Now a power $(\lambda P_\lambda - D_\lambda)^m$ of a Jordan block converges iff: Either $|\lambda| < 1$, or $\lambda = 1$ and $D_1 = 0$. In the former case the limit is 0, in the latter the sequence is fixed equal to $P_{\lambda=1}$. This can be shown along the lines of Householder (1964), Section 7.3.

To finish the proof, it is easy to show that condition (b) of the lemma is equivalent to $D_{\lambda=1} = 0$. □

PROPOSITION 12. *Make the assumptions of Theorem 9, and consider the two-smoother case. If we allow only one nontrivial relaxation parameter ω_1, while $\omega_2 = 1$, then the value of ω_1 that decreases Q the most, for a given $\mathbf{f} \in \mathscr{R}(\hat{\mathbf{T}}_2)$, is*

$$\omega_1 = \frac{Q\big((I - \hat{\mathbf{T}}_1)\mathbf{f}\big)}{Q\big(\hat{\mathbf{T}}_2(I - \hat{\mathbf{T}}_1)\mathbf{f}\big)} \geq 1.$$

PROOF. We make use of the bilinear form

$$B(\mathbf{f}, \mathbf{g}) = \langle \mathbf{f}_+, \mathbf{g}_+ \rangle + \sum_{j=1}^{2} \langle \mathbf{f}_j, (S_j^- - I)\mathbf{g}_j \rangle.$$

Since $B(\mathbf{f}, \mathbf{f}) = Q(\mathbf{f}) \geq 0$ under the given assumptions, the form B is symmetric and nonnegative definite. Thus B has all the properties of a scalar product, except there may exist $\mathbf{f} \neq \mathbf{0}$ with $B(\mathbf{f}, \mathbf{f}) = 0$. Furthermore, the Gauss–Seidel updates $\hat{\mathbf{T}}_j$ are orthogonal projections w.r.t. B: $\hat{\mathbf{T}}_j^2 = \hat{\mathbf{T}}_j$ and $B(\hat{\mathbf{T}}_j\mathbf{f}, \mathbf{g}) = B(\mathbf{f}, \hat{\mathbf{T}}_j\mathbf{g})$. We wish to examine how much $\hat{\mathbf{T}}_\omega = \hat{\mathbf{T}}_2(I - (I - \hat{\mathbf{T}}_1)\omega)$ decreases Q:

$$Q\big(\hat{\mathbf{T}}_\omega\mathbf{f}\big) = Q\big(\hat{\mathbf{T}}_2(I - \hat{\mathbf{T}}_1)\mathbf{f}\big)\omega^2 - 2B\big(\hat{\mathbf{T}}_2\mathbf{f}, (I - \hat{\mathbf{T}})\mathbf{f}\big)\omega + Q\big(\hat{\mathbf{T}}_2\mathbf{f}\big).$$

For the second term, we made use of symmetry w.r.t. B and idempotence of $\hat{\mathbf{T}}_2$. Clearly, if the coefficient of ω^2 vanishes, so does the coefficient of ω, and the criterion stays flat as a function of ω. Otherwise, the coefficient of ω^2 will be positive, and the minimizing ω is

$$\omega_{\min} = \frac{B\big(\hat{\mathbf{T}}_2\mathbf{f}, (I - \hat{\mathbf{T}}_1)\mathbf{f}\big)}{Q\big(\hat{\mathbf{T}}_2(I - \hat{\mathbf{T}}_1)\mathbf{f}\big)}.$$

The assertion of Proposition 12 follows under the assumption $\mathbf{f} \in \mathscr{R}(\hat{\mathbf{T}}_2)$: In this case $\hat{\mathbf{T}}_2\mathbf{f} = \mathbf{f}$ since $\hat{\mathbf{T}}_2$ is a projection, and since $I - \hat{\mathbf{T}}_1$ is an orthogonal projection for B, we get

$$B\big(\hat{\mathbf{T}}_2\mathbf{f}, (I - \hat{\mathbf{T}}_1)\mathbf{f}\big) = Q\big((I - \hat{\mathbf{T}}_1)\mathbf{f}\big).$$

Clearly, $Q((I - \hat{\mathbf{T}}_1)\mathbf{f}) \geq Q(\hat{\mathbf{T}}_2(I - \hat{\mathbf{T}}_1)\mathbf{f})$ since again $\hat{\mathbf{T}}_2$ is an orthogonal projection under B. □

The modified backfitting algorithm as a solution to the original problem. Let S_j be the smoother matrix for the jth variable. Define the modified smoother

$S_j^* = H_j + (I - H_j)S_j$ as before. Let $\tilde{S}_j = (I - H_j)S_j$, and note that if S_j is symmetric and shrinking, $\tilde{S}_j$ is strictly shrinking. For such smoothers, we have proved that the modified backfitting algorithm always converges whether we iterate the inner loop or not.

A question remaining is whether the solution to the modified algorithm solves the original normal equations. This is indeed the case as we show below.

THEOREM 14. *Suppose the modified backfitting algorithm has converged with smoothers $\tilde{S}_j$ and projection H, yielding functions $\tilde{\mathbf{f}}_j$ and $\mathbf{g}_j \in \mathcal{M}_1(S_j)$. Then the components $\mathbf{f}_j = \mathbf{g}_j + \tilde{\mathbf{f}}_j$ are solutions to the normal equations with smoothers $S_j^* = H_j + (I + H_j)S_j$.*

REMARK. If S_j is symmetric, we have $S_j^* = S_j$ and thus the solutions to the modified algorithm solve the normal equations with smoothers S_j. These correspond to the original backfitting algorithm.

PROOF. Convergence implies that we have stationarity conditions for both $\tilde{\mathbf{f}}_j$ and $\mathbf{g}_j$,

$$\tilde{\mathbf{f}}_j = \tilde{S}_j\left(\mathbf{y} - \sum_{i=1}^{p} \mathbf{g}_i - \sum_{i \neq j} \tilde{\mathbf{f}}_i\right), \qquad j = 1, \dots, p, \tag{43}$$

$$\mathbf{g}_j = H_j\left(\mathbf{y} - \sum_{i \neq j} \mathbf{g}_i - \sum_{i=1}^{p} \tilde{\mathbf{f}}_i\right), \qquad j = 1, \dots, p. \tag{44}$$

Now, since H_j projects onto $\mathcal{M}_1(S_j)$, we have $S_j H_j = H_j$, hence $\tilde{S}_j H_j = (I - H_j)S_j H_j = (I - H_j)H_j = 0$. From this follows $\tilde{S}_j \mathbf{g}_j = \mathbf{0}$. Since $H_j \tilde{S}_j = 0$, we also have $H_j \tilde{\mathbf{f}}_j = \mathbf{0}$. Hence $\mathbf{g}_j$ may be dropped from (43), and $\tilde{\mathbf{f}}_j$ from (44), and we get

$$\begin{aligned} \mathbf{f}_j &= \mathbf{g}_j + \tilde{\mathbf{f}}_j \\ &= H_j\left(\mathbf{y} - \sum_{i \neq j} \mathbf{g}_i - \sum_{i \neq j} \tilde{\mathbf{f}}_i\right) + \tilde{S}_j\left(\mathbf{y} - \sum_{i \neq j} \mathbf{g}_i - \sum_{i \neq j} \tilde{\mathbf{f}}_i\right) \\ &= S_j^*\left(\mathbf{y} - \sum_{i \neq j} \mathbf{f}_i\right). \end{aligned}$$

□

Acknowledgments. The authors thank Peter Bickel for his proof of an important special case of Theorem 5 and other helpful comments, an Associate Editor and two referees for valuable comments, and Tashkin Atilgan, John Chambers, Bill Cleveland, Walter Gander, Gene Golub, Peter Green, Jeff Lagarias, Colin Mallows, Brad Murphey, Finbarr O'Sullivan, Daryl Pregibon, Jim Reeds, Werner Stuetzle and Jon Wellner for useful suggestions, comments and references.

REFERENCES

BECKER, R. and CHAMBERS, J. (1984). *S: An Interactive Language for Data Analysis and Graphics.* Wadsworth, Belmont, Calif.

BICKEL, P., KLAASSEN, C., RITOV, Y. and WELLNER, J. (1989). *Efficient and Adaptive Estimation for Semiparametric Models.* To appear.

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.* **80** 580–619.

BUJA, A. (1989). Remarks on functional canonical variates, alternating least squares methods, and ACE. Technical Memorandum, Bellcore.

BUJA, A., DONNELL, D. and STUETZLE, W. (1986). Additive principal components. Technical Report, Dept. Statistics, Univ. Washington.

BURMAN, P. (1988). Estimation of generalized additive models. Technical Report, Dept. Statistics, Rutgers Univ.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

CLEVELAND, W. S. (1983). Seasonal and calendar adjustment. In *Handbook of Statistics* (D. R. Brillinger and P. R. Krishnaiah, eds.) **3** 39–72. North-Holland, Amsterdam.

CLEVELAND, W. S. and DEVLIN, S. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83** 596–610.

CLEVELAND, W. S., DEVLIN, S. and GROSSE, E. (1988). Regression by local fitting: Methods, properties and computational algorithms. *J. Econometrics* **37** 87–114.

COX, D. D. (1983). Asymptotics for *M*-type smoothing splines. *Ann. Statist.* **11** 530–551.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.

DE BOOR, C. (1978). *A Practical Guide to Splines.* Springer, New York.

DEMMLER, A. and REINSCH, C. (1975). Oscillation matrices and spline smoothing. *Numer. Math.* **24** 375–382.

DENBY, L. (1984). Smooth regression functions. Ph.D. dissertation, Univ. Michigan.

DEUTSCH, F. (1983). von Neumann's alternating method: The rate of convergence. In *Approximation Theory IV* (C. K. Chui, L. L. Schumaker and J. D. Ward, eds.) 427–434. Academic, New York.

DEVLIN, S. J. (1986). Locally-weighted multiple regression: Statistical properties and a test of linearity. Technical Memorandum, Bellcore.

ENGLE, R. F., GRANGER, C. W. J., RICE, J. and WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

EUBANK, R. L. (1984). The hat matrix for smoothing splines. *Statist. Probab. Lett.* **2** 9–14.

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics.* To appear.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

FRIEDMAN, J. H. and STUETZLE, W. (1982). Smoothing of scatterplots. Technical Report, Orion 3, Dept. Statistics, Stanford Univ.

FRIEDMAN, J. H., GROSSE, E. and STUETZLE, W. (1983). Multidimensional additive spline approximation. *SIAM J. Sci. Statist. Comput.* **4** 291–301.

GASSER, TH. and MÜLLER, H.-G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 23–68. Springer, Berlin.

GOLUB, G. H. and VAN LOAN, C. F. (1983). *Matrix Computations.* Johns Hopkins Univ. Press, Baltimore, Md.

GREEN, P. and YANDELL, B. (1985). Semi-parametric generalized linear models. *Generalized Linear Models. Lecture Notes in Statist.* **32** 44–55. Springer, Berlin.

GREEN, P., JENNISON, C. and SEHEULT, A. (1985). Analysis of field experiments by least squares smoothing. *J. Roy. Statist. Soc. Ser. B* **47** 299–315.

GREEN, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Internat. Statist. Rev.* **55** 245–260.

HÄRDLE, W. (1987). Resistant smoothing using the fast Fourier transform, AS222. *Appl. Statist.* **36** 104–111.

HASTIE, T. (1988). Pseudo-smoothers and additive model approximations. Technical Memorandum, AT&T Bell Laboratories.

HASTIE, T. and TIBSHIRANI, R. (1986a). Generalized additive models (with discussion). *Statist. Sci.* **1** 297–318.

HASTIE, T. and TIBSHIRANI, R. (1986b). Generalized additive models, cubic splines and penalized likelihood. Technical Report, Div. Biostatistics, Univ. Toronto.

HASTIE, T. and TIBSHIRANI, R. (1987). Generalized additive models: Some applications. *J. Amer. Statist. Assoc.* **82** 371–386.

HASTIE, T. and TIBSHIRANI, R. (1988). Comment on "Monotone regression splines in action" by J. O. Ramsay. *Statist. Sci.* **3** 450–456.

HOUSEHOLDER, A. S. (1964). *The Theory of Matrices in Numerical Analysis*. Dover, New York.

KATO, T. (1984). *Perturbation Theory for Linear Operators*. Springer, New York.

KELLER, H. B. (1965). On the solution of singular and semidefinite linear systems by iteration. *SIAM J. Numer. Anal.* **B2** 281–290.

KEMPERMAN, J. (1984). Least absolute value and median polish. In *Inequalities in Statistics and Probability* (Y. L. Tong, ed.) 84–103. IMS, Hayward, Calif.

LAWSON, C. L. and HANSON, R. J. (1974). *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, N.J.

LIGHT, W. A. and CHENEY, E. W. (1985). *Approximation Theory in Tensor Product Spaces. Lecture Notes in Math.* **1169**. Springer, Berlin.

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

MALLOWS, C. L. (1980). Some theory of nonlinear smoothers. *Ann. Statist.* **8** 695–715.

MALLOWS, C. L. (1986). Augmented partial residuals. *Technometrics* **28** 313–320.

MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, Mass.

O'SULLIVAN, F. (1983). The analysis of some penalized likelihood estimation schemes. Technical Report 726, Dept. Statistics, Univ. Wisconsin, Madison.

O'SULLIVAN, F. (1986a). Estimation of densities and hazards by the method of penalized likelihood. Technical Report 58, Dept. Statistics, Univ. California, Berkeley.

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

O'SULLIVAN, F., YANDELL, B. and RAYNOR, W. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–103.

PRIESTLEY, M. B. and CHAO, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34** 385–392.

RAMSAY, J. O. (1988). Monotone regression splines in action (with discussion). *Statist. Sci.* **3** 425–461.

REINSCH, C. (1967). Smoothing by spline functions. *Numer. Math.* **10** 177–183.

RICE, J. A. (1986). Comment on "A statistical perspective on ill-posed inverse problems" by F. O'Sullivan. *Statist. Sci.* **1** 522–523.

RICE, J. and ROSENBLATT, M. (1983). Smoothing splines, regression derivatives and convolution. *Ann. Statist.* **11** 141–156.

ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

SIEGEL, A. F. (1983). Low median and least absolute residual analysis of two-way tables. *J. Amer. Statist. Assoc.* **78** 371–374.

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

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–620.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

STONE, C. J. and KOO, C.-Y. (1985). Additive splines in statistics. *Proc. Statist. Comp. Sec.* 45–48. Amer. Statist. Assoc., Washington.

TIBSHIRANI, R. and HASTIE, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82** 559–568.

TUKEY, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, Mass.

UTRERAS, F. D. (1979). Cross-validation techniques for smoothing spline functions in one or two dimensions. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 196–232. Springer, Berlin.

VAN DER BURG, E. and DE LEEUW, J. (1983). Non-linear canonical correlation. *British J. Math. Statist. Psych.* **36** 54–80.

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150.

WAHBA, G. (1986). Partial and interaction splines for the semiparametric estimation of functions of several variables. Technical Report 784, Dept. Statistics, Univ. Wisconsin, Madison.

WATSON, G. S. (1964). Smooth regression analysis. *Sankyā Ser. A* **26** 359–372.

WHITTAKER, E. (1923). On a new method of graduation. *Proc. Edinburgh Math. Soc.* **41** 63–75.

ANDREAS BUJA
BELL COMMUNICATIONS RESEARCH
MORRISTOWN, NEW JERSEY 07960-1910

TREVOR HASTIE
AT&T BELL LABORATORIES
600 MOUNTAIN AVENUE
MURRAY HILL, NEW JERSEY 07974-2070

ROBERT TIBSHIRANI
DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO
CANADA M5S 1A8

DISCUSSION

LEO BREIMAN

University of California, Berkeley

After finishing the ACE paper [Breiman and Friedman (1985)] I hoped that others would tie up some of the significant loose ends. The work under discussion does a good part of that admirably.

But is it interesting that since that time both Friedman and myself have veered off in the direction of using splines for additive and more general models, thus circumventing the problem of convergence of iterated smooths which occupies much of the present paper.

I think it would be useful, in the context of the present paper, to give the itinerary of my journey from smoothers to splines. In addition, another problem that has occupied me is the incorporation of bivariate interaction into the model and I will also comment on that below.

Bivariate smoothers, in and of themselves are not of undying statistical interest. The interest in them developed because of realization, in the ACE paper, that additive models could be fitted through an iterated sequence of bivariate smooths. Now additive models are very interesting, since they form a useful and often revealing extension to linear models.