Additive and Generalized Additive Models: a Survey

MICHAEL G. SCHIMEK and BERWIN A. TURLACH[†]
Karl-Franzens-University of Graz, Graz, Austria and
University of Adelaide, Adelaide, Australia

Abstract: This paper is the attempt to summarize the state of art in additive and generalized additive models (GAM). The emphasis is on approaches and numerical procedures which have emerged since the monograph of Hastie and Tibshirani (1990), although reconsidering certain aspects of their work. Apart from GAM, vector GAM (VGAM), alternating conditional expectations (ACE), and additivity and variance stabilization (AVAS) are discussed. Last but not least there are software hints for all these models.

[†]The first author revised this paper during a research visit at Sonderforschungsbereich 373 at the Humboldt-University Berlin. He likes to thank Stefan Sperlich for providing the marginal integration algorithm and for valuable discussions. An extended version will be published in the book *Smoothing and Regression. Approaches, Computation and Application* edited by M. G. Schimek, Wiley, New York, 1999.

1.1 INTRODUCTION

This paper discusses multivariate regression problems evaluated in a nonparametric fashion where the aim is to study the structural relationship between the response variable Y and the vector of d covariates $\mathbf{X} = (X_1, \ldots, X_d)^T$ via

$$m(\mathbf{x}) = \mathcal{E}(Y \mid \mathbf{X} = \mathbf{x})$$

with $\mathbf{x} = (x_1, \ldots, x_d)^T$ and $m(\mathbf{x}) = m(x_1, \ldots, x_d)$.

In the multiple linear regression model one assumption is that the conditional mean relationship between the response and each of the predictors is linear, i.e. that $m(\mathbf{x})$ is linear and additive in the predictors. To gain more flexibility one can drop the linearity assumption but retain additivity. The result is an additive model in which each explanatory variable can be related to the dependent variable via an individual functional form. As long as these functions obey certain smoothness assumptions they can be estimated by scatterplot smoothers in a nonparametric fashion. The case of additive models will be studied in Section 1.2, where we consider penalized least squares and marginal integration methods.

In the parametric class of Generalized Linear Models (GLMs; McCullagh and Nelder, 1989) the unknown regression function $m(\mathbf{x})$ is modelled linearly via a known link function G. An important example is the logistic regression model for binary responses. There, the logit of the regression function is represented linearly. Again we can gain more flexibility by replacing the linear function with some smooth function in a nonparametric setting. Such models have been introduced by Hastie and Tibshirani and summarized in their 1990 monograph. The nonparametric alternative to GLMs is called Generalized Additive Models (GAMs). They allow the conditional mean of the response variable to depend via a fixed link function on a sum of univariate functions, each function having one component of the vector of explanatory variables as argument. The generalized additive case will be discussed in length in Section 1.3, where the emphasis is again on the penalized least squares and the marginal integration methods.

Section 1.3.3 describes Vector Generalized Additive Models (VGAMs), which were recently introduced by Yee and Wild (1996). VGAMs handle multivariate (vector) regression problems.

Two further methods respectively algorithms named Alternating Conditional Expectations (ACE) and Additivity and Variance Stabilization (AVAS) are discussed briefly in Section 1.4 since they are (computationally) closely related to GAMs although they are based on a different working model.

For all models introduced in this paper we provide the numerical background (also pointing out alternative estimation concepts), algorithms and software hints.

Let us now give some motivation for the use of additive models. The most flexible models one could think of do not have any assumptions about the form of the d-variate function $m(\mathbf{x})$. The problem is to fit a d-dimensional surface

to the observed data $\{\mathbf{x}_i^T, Y_i\}$ with $i = (1, \ldots, n)$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^T$. What comes first to our mind is generalizing the univariate smoothing techniques to this multivariate situation. But there is a severe problem: it is necessary to define neighbourhoods in the d-dimensional space, however under the curse of dimensionality we understand that neighborhoods with a fixed number of points become less local as the dimensions increase (Bellman, 1961).

In the spline world the generalization is restricted to certain cases such as thin plate splines. Other problems and thus restrictions arise in the domain of kernel smoothers. The additive model approach is an elegant way around these problems because high-dimensionality is tackled via an additive approximation of simple univariate smooth functions such as cubic smoothing splines or kernel smoothers.

What also remains unsolved is the problem whether the degree of smoothing should be the same in each dimension. This is related to the question whether the variation in the surface is comparable with respect to all covariates (see Scott, 1992, Chapter 7, for further reading). As will be demonstrated in this paper the above problems become less severe when an additive approach is taken, although related questions will emerge. Smoothing parameter respectively bandwidth choice remains a weak point in this class of models. The reason is both a lack of theory and adequate algorithms. In Sections 1.5 and 1.6 we shall discuss the problem of selecting smoothing parameter(s) in (generalized) additive models and conclude with a discussion of model diagnostics, another field still open to research.

1.2 THE ADDITIVE MODEL

Consider the general multiple regression model

$$Y = m(\mathbf{X}) + e, \tag{1.1}$$

where we assume that the error e is independent of the vector of explanatory variables $\mathbf{X}$, $\mathcal{E}(e) = 0$, and $\mathcal{V}(e) = \sigma^2$. Different from linear regression, now only the additivity property of m is required in the following. The dependent variable Y is approximated by the additive model

$$m(\mathbf{X}) \simeq g(\mathbf{X}) = g_0 + \sum_{j=1}^{d} g_j(X_j), \tag{1.2}$$

where g_0 is a constant and the g_js are univariate smooth functions. To avoid free constants in the functions g_j we usually require that $\mathcal{E}[g_j(X_j)] = 0$ for $1 \leq j \leq d$ (centering). These identifiability conditions imply that $\mathcal{E}(Y) = g_0$.

In the following we shall restrict our interest to linear scatterplot smoothers with a $n \times n$ smoother matrix S. Buja *et al.* (1989, p. 453) give the following definition for a linear scatterplot smoother: The function estimate

$\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^T$ can be written as $\hat{\mathbf{y}} = S\mathbf{y}$, where $\mathbf{y} = (y_1, \ldots, y_n)^T$, and S does not depend on $\mathbf{y}$. Important examples of linear smoothers are kernel smoothers and smoothing splines. Not all scatterplot smoothers have this property.

Under identifiability, if the additive model is correct (not just an approximation), we have

$$\mathcal{E}\left[Y - g_0 - \sum_{j \neq k} g_j(X_j) \mid X_k\right] = g_k(X_k)$$

for $k = 1, \ldots, d$. This relationship suggests to adopt an iterative procedure for the estimation of the univariate functions $g_1, \ldots, g_d$ corresponding to the explanatory variables. For a known constant g_0 and given functions $g_j, j \neq k$, the function g_k can be estimated by an univariate regression fit based on the measurements (X_{ik}, Y_i) for $i = 1, \ldots, n$ (n observations). Applying a linear smoother S_j, the centered version of the estimator $\hat{g}_j$ of g_j is

$$\hat{g}_j^* = \hat{g}_j - n^{-1} \sum_{i=1}^{n} \hat{g}_j(X_{ij}).$$

This is the motivation for an iterative scheme that is now known as backfitting. It was introduced and analyzed by Friedman and Stuetzle (1981) and Breiman and Friedman (1985) in the context of projection pursuit regression – the additive model is a special case of the projection pursuit model – and can be described as follows:

Additive Model Algorithm

---

1. *Initialize:* $\hat{g}_0 = n^{-1} \sum_{i=1}^{n} Y_i$, $\hat{g}_j = g_j^0$, $j = 1, \ldots, d$

2. *Find new transformations:* For $j = 1, \ldots, d$
   $\hat{g}_j = S_j \left[Y - \hat{g}_0 - \sum_{j \neq k} \hat{g}_j(X_j) \mid X_k\right]$,
   $\hat{g}_0 = \hat{g}_0 + n^{-1} \sum_{i=1}^{n} \hat{g}_j(X_{ij})$,
   and $\hat{g}_j^* = \hat{g}_j - n^{-1} \sum_{i=1}^{n} \hat{g}_j(X_{ij})$

3. Cycle step 2 until convergence

The idea behind this algorithm is to carry out a fit, calculate partial residuals from that fit and refit again. That is why the iteration scheme is called backfitting. The starting functions $g_1^0, \ldots, g_d^0$ can be obtained in various ways, e.g. from a linear regression fit of Y on the predictors X_k. Technical matters concerning the algorithm such as numerical features, convergence and statistical implications will be discussed later in this paper.

1.2.1 The normal equations and linear scatterplot smoothing

In order to justify smoothing in the additive model we take the L_2 function space view of Hastie and Tibshirani (1990). Let $\mathcal{H}_j$, for $j = 1, \ldots, d$, denote the Hilbert spaces of measurable functions $\phi_j(X_j)$ with $\mathcal{E}\phi_j(X_j) = 0$, $\mathcal{E}\phi_j^2(X_j) < \infty$, and inner product $\langle \phi_j(X_j), \phi_j^*(X_j) \rangle = \mathcal{E}\phi_j(X_j)\phi_j^*(X_j)$. Furthermore, $\mathcal{H}$ is assumed to be the space of arbitrary centered, square integrable functions of $X_1, \ldots, X_d$. We consider the $\mathcal{H}_j$ as subspaces of $\mathcal{H}$ in a canonical way. $\mathcal{H}^{add} \subset \mathcal{H}$ describes a closed linear subspace of the additive functions $\mathcal{H}_1 + \ldots + \mathcal{H}_d$. These are all subspaces of $\mathcal{H}_{YX}$, the space of centered square integrable functions of Y and $X_1, \ldots, X_d$.

The additive model in a population setting amounts to minimizing

$$\mathcal{E}[Y - g(\mathbf{X})]^2 \tag{1.3}$$

over

$$g(\mathbf{X}) = \sum_{j=1}^{d} g_j(X_j) \in \mathcal{H}^{add}.$$

Without the additivity assumption we would simply obtain $\mathcal{E}(Y \mid \mathbf{X})$. Our goal is to find the closest additive approximation to this function. By the definition of $\mathcal{H}^{add}$ the above minimum exists and is unique. This is not true for the individual functions $g_j(X_j)$.

Let P_j denote the conditional expectation operator $\mathcal{E}(\cdot \mid X_j)$ (P_j is the orthogonal projection onto H_j). The minimizer $g(\mathbf{X})$ of (1.3) can be represented by residuals $Y - g(\mathbf{X})$ which are orthogonal to the space of fits $Y - g(\mathbf{X}) \perp \mathcal{H}^{add}$. Since $\mathcal{H}^{add}$ is generated by $\mathcal{H}_j$, we have equivalently $Y - g(\mathbf{X}) \perp \mathcal{H}_j$ for all j or $P_j[Y - g(\mathbf{X})] = 0$ for all j. Considering a single component this can be written as

$$g_k(X_k) = P_k \left[Y - \sum_{\substack{j=1 \\ j \neq k}}^{d} g_j(X_j) \right] = \mathcal{E} \left[Y - \sum_{\substack{j=1 \\ j \neq k}}^{d} g_j(X_j) \mid X_k \right].$$

The following system of estimating equations is necessary and sufficient for $g = (g_1, \ldots, g_d)$ to minimize (1.3):

$$\begin{pmatrix} I & P_1 & P_1 & \ldots & P_1 \\ P_2 & I & P_2 & \ldots & P_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_d & P_d & P_d & \ldots & I \end{pmatrix} \begin{pmatrix} g_1(X_1) \\ g_2(X_2) \\ \vdots \\ g_d(X_d) \end{pmatrix} = \begin{pmatrix} P_1 Y \\ P_2 Y \\ \vdots \\ P_d Y \end{pmatrix} \tag{1.4}$$

or

$$P\mathbf{g} = QY,$$

where P and Q are a matrix and a vector, respectively, of operators. A numerical solution for the additive approximation cannot be obtained from (1.4) since the matrix on the left consists of conditional expectation operators (not real numbers). But the conditional expectations in our population setting are connected to the sample data setting via smoothing.

Let us now take a more formal look at linear scatterplot smoothers in additive models. We assume to compute the fit at design points x_i, in which case we can write a linear smoother as a linear map $S : \mathcal{R}^n \mapsto \mathcal{R}^n$ defined by $\hat{\mathbf{y}} = S\mathbf{y}$. Applying a linear smoothing algorithm we can produce the corresponding smoother matrix S by smoothing unit basis vectors. Smoothing the ith unit vector results in the ith column of S. This cannot be done for a nonlinear smoother such as LO(W)ESS since the estimates depend on y in a nonlinear fashion.

A linear smoother can be written as a smoother matrix S times the response vector $\mathbf{y}$, i.e. $\hat{\mathbf{g}} = S\mathbf{y}$. As pointed out earlier the most prominent linear smoothers are splines and kernel smoothers. Now we can replace the conditional expectation operator P_j by such a smoother with smoother matrix S_j.

Finally a numerical solution for the additive approximation to the regression curves g_j forming a (nd)-dimensional vector $\mathbf{g}$, can be obtained from the data version of the estimating equations (1.4), which forms a $(nd) \times (nd)$ system

$$\begin{pmatrix} I & S_1 & S_1 & \dots & S_1 \\ S_2 & I & S_2 & \dots & S_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_d & S_d & S_d & \dots & I \end{pmatrix} \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \\ \mathbf{g}_d \end{pmatrix} = \begin{pmatrix} S_1\mathbf{y} \\ S_2\mathbf{y} \\ \vdots \\ S_d\mathbf{y} \end{pmatrix}. \tag{1.5}$$

In short form we can write

$$A\mathbf{g} = B\mathbf{y},$$

where A and B are block matrices consisting of identity matrices I and linear smoothing operators S_j. This system of equations is known as the normal equations of the additive regression model.

The above system can be adapted to handle tied observations, an important modification for applied data analysis. Schimek *et al.* (1993) proposed a weighting scheme (based on permutation and reduction matrix operators) which does not change the structural features of the system matrix (below this approach is discussed in detail for smoothing splines).

1.2.2 Penalized least squares

In this section we shall concentrate on smoothing splines which have the additional property (not shared, for instance, by kernel smoothers) that the

smoothing matrix S is symmetric. As a direct consequence of this symmetry and that splines can reproduce straight lines we have

$$HS\mathbf{y} = (S^T H^T)^T \mathbf{y} = (SH)^T \mathbf{y} = H^T \mathbf{y} = H\mathbf{y},$$

where H is the so-called hat matrix (often written $\hat{H}$) for linear least squares regression. This property will prove useful when studying the formal structure of the backfitting algorithm. Due to symmetry is also the fact that the eigenvectors of the smoother matrix S of a spline resemble those of polynomials of increasing degree (Eubank, 1984). The first two eigenvalues are one and correspond to linear functions of the design variable x. The eigendecomposition of S allows us to analyse the smoother, analogously to linear stochastic processes for time series, by means of the spectrum.

For convenience let us study the popular case of cubic smoothing splines. We have to minimize with respect to g the penalized least squares criterion

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int_{-\infty}^{+\infty} [g''(z)]^2 dz, \tag{1.6}$$

where λ ($\lambda > 0$) is a fixed smoothing parameter. The solution $\hat{g}$ is a cubic Reinsch spline (Reinsch, 1967) with knots at each distinct x_i.

Following Green and Yandell (1985, p. 46) we introduce $h_i = x_{i+1} - x_i$ for $i = 1, 2, \ldots, n-1$, a tridiagonal $(n-2) \times n$ matrix Δ 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 a symmetric tridiagonal $(n-2) \times (n-2)$ matrix C with $c_{i-1,i} = c_{i,i-1} = h_i/6$, $c_{ii} = (h_i + h_{i+1})/3$. Then the minimization problem in (1.6) can be equivalently expressed by

$$\|\mathbf{y} - \mathbf{g}\|^2 + \lambda \mathbf{g}^T K \mathbf{g} \to min, \tag{1.7}$$

where K denotes a quadratic penalty matrix with

$$K = \Delta^T C^{-1} \Delta.$$

The solution is now

$$\hat{\mathbf{g}} = S\mathbf{y},$$

where the linear smoother matrix S is given by

$$S = (I + \lambda K)^{-1}. \tag{1.8}$$

The concept of penalized least squares can be also applied to the additive model itself when smoothing splines are used as scatterplot smoothers. Further, it provides another motivation apart from that of function spaces for the application of smoothing in additive models.

Above we introduced the cubic smoothing spline as the minimizer of the penalized least squares criterion (1.6) over all twice continuously differentiable

functions g. In order to extend this idea to the additive model, we generalize the criterion (1.6) in a straightforward manner. We seek to minimize

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{d} g(x_{ij}))^2 + \sum_{j=1}^{d} \lambda_j \int_{-\infty}^{+\infty} [g_j''(z)]^2 dz, \tag{1.9}$$

over all twice continuously differentiable functions g_j. Each function in (1.9) is penalized by a separate fixed smoothing parameter λ_j. This in turn determines the smoothness of that function in the solution. There are two extreme cases: if all the λ_js take the value zero the solution to (1.6) is any interpolating set of functions satisfying $y_i = \sum_{j=1}^{d} g(x_{ij})$ for $i = 1, 2, \ldots, n$. The other extreme is that each λ_j goes to infinity, resulting in a penalty term which itself goes to infinity unless $g_j''(z) = 0$ for all j (i.e. each g_j is linear as in linear least squares regression).

The solution of (1.9) is a cubic smoothing spline in each of the predictors. Evaluation at all n observations leads to

$$(\mathbf{y} - \sum_{j=1}^{d} \mathbf{g}_j)^T (\mathbf{y} - \sum_{j=1}^{d} \mathbf{g}_j) + \sum_{j=1}^{d} \lambda_j \mathbf{g}_j^T K_j \mathbf{g}_j, \tag{1.10}$$

where the K_js are penalty matrices for each predictor as in the univariate setting of (1.7). Differentiating (1.10) with respect to the function $\mathbf{g}_k$ yields

$$\hat{\mathbf{g}}_k = S_k(\mathbf{y} - \sum_{j \neq k} \hat{\mathbf{g}}_j) \tag{1.11}$$

where S_k is a smoother matrix given by (compare with (1.8))

$$S_k = (I + \lambda_k K_k)^{-1} \tag{1.12}$$

and K_k, the individual penalty matrices. Writing equation (1.11) for $k = 1, 2, \ldots, d$ produces the same $(nd) \times (nd)$ system of normal equations as obtained from the function space considerations. Again we end up with

$$\begin{pmatrix} I & S_1 & S_1 & \ldots & S_1 \\ S_2 & I & S_2 & \ldots & S_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_d & S_d & S_d & \ldots & I \end{pmatrix} \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \\ \mathbf{g}_d \end{pmatrix} = \begin{pmatrix} S_1\mathbf{y} \\ S_2\mathbf{y} \\ \vdots \\ S_d\mathbf{y} \end{pmatrix} \tag{1.13}$$

with d different cubic spline smoother matrices S_j. According to Hastie and Tibshirani (1990, p. 111) we can interpret the penalty terms in (1.10) as a down-weighting of each of the components of $\mathbf{g}_j$, the down-weighting determined by the corresponding eigenvalue of that component and λ_j.

Next we have to cope with the technical problem of tied predictor values, mentioned in Hastie and Tibshirani (1990, p. 74). There is almost no practical

situation in multivariate regression where the data within one variable are all different from each other.

The problem of tied observations in additive models is twofold: (i) the predictor variables may require different orderings (relevant for instance when smoothing splines are adopted), and (ii) in each predictor variable values may occur with frequency greater than one (tied values). Schimek *et al.* (1993) put forward a computationally efficient modification of the normal equations in (1.13) which accommodates for problems (i) and (ii) at the same time.

We apply again the penalized least squares criterion for cubic smoothing splines as well as for the additive model. Let us define $x_i = (x_{i1}, \ldots, x_{id})^T$ as the design points (i.e. explanatory observations) for $i = 1, \ldots, n$.

The estimation of model (1.2) requires to determine the $\mathbf{g}_j$s for each coordinate of the design points x_{ij}. For the distinct design points we consider the minimization problem (with respect to $\mathbf{g}_j$)

$$(\mathbf{y} - \sum_{j=1}^{d} \mathbf{g}_j)^T (\mathbf{y} - \sum_{j=1}^{d} \mathbf{g}_j) + \sum_{j=1}^{d} \lambda_j \mathbf{g}_j^{(l)T} K_j \mathbf{g}_j^{(l)} \tag{1.14}$$

where the K_j are penalty matrices and the λ_j ($\lambda_j > 0$) smoothing parameters of the individual cubic smoothing splines as before.

The vectors $\mathbf{g}_j^{(l)}$ are of equal or smaller (hence notation "(l)") dimension than the original $\mathbf{g}$ vectors of estimators (i.e. $n_j \leq n$) can be obtained as follows. Let O_j be a permutation operator such that $x_{ij}^{'} = O_j x_{ij}$ is ordered ascending, $j = 1, \ldots, d$; and R_j be a reduction operator defined by a $n_j \times n$ matrix with elements r_{ts}, where $n_j = \#\{x_{1j}, \ldots, x_{nj}\}$; $r_{ts} := 1$ for $s = 1, \ldots, n$ and $t = t(s) = \#\{x_{1j}^{'}, \ldots, x_{sj}^{'}\}$; $r_{ts} := 0$ elsewhere. Let $D_j := R_j R_j^T$ and $\mathbf{g}_j^{(l)} := D_j^{-1} R_j O_j \mathbf{g}_j$. The $\mathbf{g}_j^{(l)}$s consist of those remaining values $\mathbf{g}_{ij}$ corresponding to the distinct and ordered design points $x_{ij}^{'}$, $i = 1, \ldots, n_j$. Hence we can write

$$\mathbf{g}_j = O_j^{-1} R_j^T \mathbf{g}_j^{(l)}.$$

The least squares term in (1.14) is equivalently expressed as the square of

$$O_r \mathbf{y} - (R_r^T \mathbf{g}_r^{(l)} + \sum_{j \neq r} O_r O_j^{-1} R_j^T \mathbf{g}_j^{(l)}), \tag{1.15}$$

for $r = 1, \ldots, d$. Differentiation of (1.14) using (1.15) with respect to $\mathbf{g}_r^{(l)}$ yields

$$\lambda_r K_r \mathbf{g}_r^{(l)} - R_r O_r \mathbf{y} + R_r R_r^T \mathbf{g}_r^{(l)} + \sum_{j \neq r} R_r O_r O_j^{-1} R_j^T \mathbf{g}_j^{(l)} = 0.$$

$T_{j,l} := S_j R_j O_j O_l^{-1} R_l^T$, where $S_j = (D_j + \lambda_j K_j)^{-1}$ describes a modified smoother matrix and $\mathbf{z}_j := S_j R_j O_j \mathbf{y}$ for $j = 1, \ldots, d$. Then the normal

equations can be expressed by

$$\begin{pmatrix} I_1 & T_{1,2} & \dots & T_{1,d-1} & T_{1,d} \\ T_{2,1} & I_2 & \dots & T_{2,d-1} & T_{2,d} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ T_{d-1,1} & T_{d-1,2} & \dots & I_{d-1} & T_{d-1,d} \\ T_{d,1} & T_{d,2} & \dots & T_{d,d-1} & I_d \end{pmatrix} \begin{pmatrix} \mathbf{g}_1^{(l)} \\ \mathbf{g}_2^{(l)} \\ \vdots \\ \mathbf{g}_{d-1}^{(l)} \\ \mathbf{g}_d^{(l)} \end{pmatrix} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_{d-1} \\ \mathbf{z}_d \end{pmatrix}, \tag{1.16}$$

where I_j is the $n_j \times n_j$ identity matrix. The $\mathbf{g}_j^{(l)}$ are the new estimators of the additive model in (1.2). The system in (1.16) can be solved in the same way as the usual normal equations. In the next section we learn about the problems and numerical techniques associated with the solution of such systems of the additive model.

1.2.3 Solution of the normal equations

1.2.3.1 Statistical considerations First we consider consistency of the normal equations in (1.5). Consistency requires that $B\mathbf{y} \in \mathcal{R}(A)$ for arbitrary data $\mathbf{y} \in R^n$ ($\mathcal{R}$ is the range). For the d-dimensional smoother case the following results are known for linear scatterplot smoothers with smoother matrices S_j.

Theorem 1. If each S_j is symmetric with eigenvalues in $[0, 1]$, the normal equations are consistent for every $\mathbf{y}$ (for the proof see Buja *et al.*, 1989, p. 481).

Further, for symmetric smoothers with eigenvalues in $[0, 1)$ closed formulas for the solutions can be given:

Proposition 1. If the S_j are symmetric with eigenvalues in $[0, 1)$, the solutions of the normal equations can be written as $\mathbf{g}_j = A_j(I + A)^{-1}\mathbf{y}$, where $A_j = (I - S_j)^{-1}S_j$ and $A = \sum_j A_j$ (for the proof see Breiman and Friedman, 1985).

Less stringent necessary and sufficient conditions on S_j for consistency could be derived by Buja *et al.* (1989, p. 482f) solely for the case of two smoothers which is of limited practical value.

Next we discuss reasons for degeneracy of the normal equations which results in non-unique solutions. The first reason is collinearity or concurvity. Collinearity is well-known from linear regression and describes a situation in which predictors are linearly dependent. If non-linear dependencies are concerned, the term "concurvity" has been established. Both collinearity and concurvity have the same effect on the additive model, degeneracy of the normal equations. In practice exact singularity of the system matrix is unlikely, however, some degeneration is quite common. There is a second reason for degeneracy of the normal equations. Linear scatterplot smoothers impose certain weighting schemes on the data which can cause near-singularity in some instances, especially in combination with concurvity.

For smoother-based normal equations, exact singularity (concurvity) is de-

fined as the existence of a non-zero solution of the corresponding homogeneous equations

$$A\mathbf{g}^* = \mathbf{0}.$$

If such a $\mathbf{g}^*$ exists, and if $\mathbf{g}$ is a solution to $A\mathbf{g} = B\mathbf{y}$, then so is $\mathbf{g} + \gamma\mathbf{g}^*$ for arbitrary γ. Hence there are infinitely many solutions.

Theorem 2. If the smoothers S_j are all symmetric with eigenvalues in $[0, 1]$, then a vector $\mathbf{g} \neq \mathbf{0}$ with $\mathbf{g} \in \mathcal{R}(S_j)$ represents a concurvity ($A\mathbf{g}^* = \mathbf{0}$) *iff* one of the following conditions is satisfied: $B\mathbf{g}^* = \mathbf{0}$, i.e. $\mathbf{g}$ minimizes B or $\mathbf{g}_j^* \in \mathcal{M}_1(S_j)$ for $j = 1, 2, \ldots, d$, and $\mathbf{g}_+^* = \mathbf{0}$. $\mathcal{M}_1$ denotes the eigenspace corresponding to eigenvalue one and $\mathbf{g}_+^* = \sum_j \mathbf{g}_j^*$ (for the proof see Buja *et al.*, 1989, p. 485).

The last condition implies that exact concurvity is exact collinearity if, for instance, all smoothers are of the cubic spline type. Approximate concurvity, however, can be characterized by approximate minimizers of $B(\mathbf{g}^*)$, which leads to approximate nonlinear additive relationships beween the predictors. Another aspect of interest is the following: If the S_j, $j = 1, 2, \ldots, d$, are symmetric with eigenvalues in $[0, 1)$, then A is non-singular. This remark is of practical relevance as the constant term is usually separated in the additive model and a zero-mean adjustment made for each of the smooth terms. As a matter of fact doing this we redefine our smoother matrices S to S^* with an eigenvalue of zero for the constant.

Finally we consider the asymptotic rate of convergence of the nonparametric additive model. Stone (1985) derived the interesting result that the d-dimensional additive model can be estimated with the optimal rate of convergence of one-dimensional smoothing. Under a number of technical conditions and the following definitions, i.e. $\mathbf{g}^*$ is the best additive approximation to the true response function $\mathbf{g}$, o is the assumed measure of smoothness of $\mathbf{g}^*$, $\hat{\mathbf{g}}_n$ is the additive (spline) estimator, $\hat{\mathbf{g}}_{n1}, \ldots, \hat{\mathbf{g}}_{nd}$ are the component functions of $\hat{\mathbf{g}}_n$, and $r = \frac{o}{2o+1}$, we have:

Theorem 3. $E[\|\hat{\mathbf{g}}_{nj} - \mathbf{g}^*\|_j^2 | X_1, \ldots, X_n] = O_p(n^{-2r})$ for $1 \leq j \leq d$.

For the technical aspects and the proof see Stone (1985, p. 693ff). In a corollary he showed that the rate of convergence does not depend on the number of dimensions d, another surprising result which means that the curse of dimensionality does not effect the asymptotic convergence rate. These results hold for projection smoothers (e.g. regression splines) and smoothing splines but not for general linear smoothers. Opsomer and Ruppert (1997) and Opsomer (1996) derived similar results for local linear regression, a non-projection smoother, under rather strong conditions on the smoothing matrices.

1.2.3.2 Numerical procedures The normal equations in (1.5) are a linear system of the form

$$A\mathbf{g} = \mathbf{b}$$

where $\mathbf{b} = B\mathbf{y}$. Let the elements of the matrix A be a_{ij} and of the vector $\mathbf{b}$ be b_i. Because of the size of the equation system it is usually solved by iterative numerical techniques.

Here we relate backfitting to standard iterative procedures, i.e. Jacobi and Gauss-Seidel, developed for solving linear equation systems with non-singular system matrices. After a brief overview we discuss their shortcomings in the context of fitting additive regression models nonparametrically by linear scatterplot smoothers. As pointed out ealier in this paper, concurvity is a main source of concern, causing ill-posed system matrices. However, standard iterative procedures are not designed to cope with degeneracy.

Backfitting

Let us first define the so-called Jacobi procedure

$$v_i^{(m)} = [b_i - \sum_{\substack{j=1 \\ j \neq i}}^{N} a_{ij} v_j^{(m-1)}]/a_{ii},$$

with iterative solutions $v_i^{(m)}$, $i = 1, \ldots, N$, an iteration counter m and starting values $v_i^{(0)}$ (usually $v_i^{(0)} = 0$). On the other hand, the Gauss-Seidel procedure is defined by

$$v_i^{(m)} = [b_i - \sum_{j=1}^{i-1} a_{ij} v_j^{(m)} - \sum_{j=i+1}^{N} a_{ij} v_j^{(m-1)}]/a_{ii}.$$

How do they differ numerically? Jacobi is a complete step and Gauss-Seidel a single step procedure. This can be best seen, when the information update is studied. It describes the way how currently available results (estimates $v^{(m-1)}$) are used to obtain the successive results (estimates $v^{(m)}$). The Jacobi algorithm computes the estimates in iteration m only on the basis of estimates from iteration $m-1$. All estimates from step m are stored and remain unused within this step. The information update takes place when moving from step m to $m+1$. Thus the information flow is low. By way of contrast, Gauss-Seidel makes use of all the information currently available, no matter where it comes from (m or $m-1$). There is permanent information update and maximal information flow again.

Backfitting was developed in the context of nonparametric multidimensional regression (Friedman and Stuetzle, 1981; Hastie and Tibshirani, 1986). To serve that purpose it was built upon the idea of determining estimates for the covariates in a successive manner, taking advantage of the specific structural features of the estimation problem.

It uses the currently available information from all covariates, except the covariate of which estimates are just computed. This leads to a splitting of the system matrix into d blocks P_j of size $N/d \times N$, where each block corresponds to one of the predictor variables X_j, $j = 1, \ldots, d$. Then an iterative procedure has to be applied to these blocks resulting in d vectors $\mathbf{v}_j^{(l)}$ (l denoting the last iteration), each of length N/d. The common choice is Gauss-Seidel iterations.

Information update takes place within the iterations, but with a certain lag. Therefore the information flow is higher compared to the Jacobi procedure and lower than in the Gauss-Seidel procedure. As a consequence, given the block structure imposed by the additive or generalized additive model, backfitting is a special case of the Gauss-Seidel algorithm. Moreover, backfitting is most effective in combination with Reinsch splines, because then we can avoid the explicit calculation of the S_k in equation (1.11) (see Green and Silverman, 1994, p. 19ff for details). This explains why many software implementers resort to backfitting and the Reinsch algorithm.

Convergence results for the above mentioned iterative procedures found in the numerical literature were obtained under the assumption of a non-singular system matrix A. Convergence depends on the eigenvalues λ_i of the iteration matrix (for details see Hämmerlin and Hoffmann, 1992, p. 372 and p. 376). Acceleration techniques were developed to improve the speed of convergence. We can write the iterative (Jacobi) scheme as

$$\mathbf{v}^{(m)} = (I - A)\mathbf{v}^{(m-1)} + \mathbf{b} = \mathbf{v}^{(m-1)} - D^{(m-1)}, \tag{1.17}$$

where $D^{(m-1)} = A\mathbf{v}^{(m-1)} - \mathbf{b}$ is the deficiency of step $m-1$. Equation (1.17) reveals that the procedure may be viewed in the light of correcting the m-th estimate for deficiency. A relaxation parameter ω can be introduced which allows us to control the amount of correction, i.e.

$$\mathbf{v}^{(m)} = \mathbf{v}^{(m-1)} - \omega D^{(m-1)}. \tag{1.18}$$

For relaxed Gauss-Seidel iteration the term *successive overrelaxation (SOR)* is established (Golub and van Loan, 1989, p. 510). After rewriting (1.17) the following relationship between the relaxed $\mathbf{v}_{rel}^{(m)}$ and unrelaxed $\mathbf{v}^{(m)}$ holds:

$$\mathbf{v}_{rel}^{(m)} = (1 - \omega)\mathbf{v}^{(m-1)} + \omega\mathbf{v}^{(m)}. \tag{1.19}$$

For $\omega < 1$ we have underrelaxationq, for $\omega > 1$ we have overrelaxation, and for $\omega = 1$ we obtain the unrelaxed algorithm. Theoretical results about the admissible range of ω on one hand and the optimal choice of ω on the other are available in the numerical literature. Under non-singularity the optimal choice of ω is as follows: (i) Assuming the Jacobi procedure converges, we have convergence for $0 < \omega \leq 1$. The optimal parameter is given by $\omega^* = 2/(2 - \lambda_1 - \lambda_n)$ with $1/2 < \omega^* < \infty$. (ii) Assuming the Gauss-Seidel procedure converges, convergence for *SOR* iteration can only be achieved with $0 < \omega < 2$. There is evidence that this result holds at least for special matrices e.g. symmetric positive definite matrices (see Gander and Golub, 1989, p. 530). However, the optimal choice ω^* is not known.

For more details we refer to Schimek *et al.* (1994). We should not close this discussion without mentioning that relaxation concepts have not found their way into commercial software such as S-Plus or XploRe. One obstacle is certainly the need for interaction between the user and the program (activation and selection of a relaxation parameter). Recently, numerical alternatives have been studied.

Such an alternative to backfitting and related procedures is a technique based on relaxed iterative projections, proposed by Schimek (1996). Especially in cases where backfitting fails totally to provide appropriate estimation results in an additive or generalized additive model (e.g. when the covariates are substantially correlated), a numerical approach which accommodates for degeneracy of the normal equations is desirable. In the following this new technique is described in essence.

Relaxed iterative projections

Again, let us consider the linear equation systems $A\mathbf{x} = \mathbf{b}$ (with $\mathbf{x} = \mathbf{g}$ and $\mathbf{b} = B\mathbf{y}$ in the additive model). Further let us have a square $n \times n$ system matrix A (usually large and sparse) and n-dimensional vectors $\mathbf{x}$ and $\mathbf{b}$. In contrast to standard iterative procedures which are equation-oriented, the iterative projection method is column-oriented.

We assume the matrix $A = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ to consist of column vectors $\mathbf{a}_i$ for $i = 1, 2, \ldots, n$. Let us have a linear space $sp(A)$ generated by the columns of A and

$$\mathbf{b} \in sp(A).$$

We define two real sequences, one is (μ_j) with

$$\left(\mathbf{b} - \sum_{i=1}^{j} \mu_i \mathbf{a}_i, \mathbf{a}_j \right) = 0, \qquad j = 1, 2 \ldots,$$

where $\mathbf{a}_i$ are the column vectors of A as defined above. The other sequence is (s_{ik}) defined by

$$s_{ik} = \sum_{j} \mu_j, \qquad j = i + nk, \qquad k = 1, 2, \ldots \tag{1.20}$$

In the j-th iteration step μ_j is determined by the orthogonal (perpendicular) projection of the previous "unexplained" residual component $\mathbf{u}_{j-1}$ onto the dimension $\mathbf{a}_j$. This means that the coefficients μ_j can be calculated by dot (inner) products. Hence

$$\mu_j = \frac{(\mathbf{u}_{j-1}, \mathbf{a}_j)}{(\mathbf{a}_j, \mathbf{a}_j)} \tag{1.21}$$

where

$$\mathbf{u}_{j-1} = \mathbf{b} - \sum_{i=1}^{j-1} \mu_i \mathbf{a}_i,$$

which makes the geometric interpretation clear. The norm (length) of μ is usually shrinking and convergence can be expected. Because the s_{ik} from equation (1.20) tend to the x_i for $k \to \infty$ each element x_i of the solution vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ can be calculated by (for k sufficiently large)

$$x_i = \sum_{j} \mu_j, \qquad j = i + nk, \qquad k = 1, 2, \ldots$$

The necessary condition is that the residual components $\mathbf{u}_j$ tend to zero, which is always true. Indeed convergence does not depend on special features of the system matrix A, such as positive definiteness or diagonal dominance. Even for singular systems a solution can be obtained (for the details see Schimek, 1996, and the references therein). Hence we can cope with ill-conditioned systems which cannot be ruled out in additive models.

However, the iterative projection method converges slower than does backfitting. This drawback can be compensated through relaxation (Schimek, 1996, p. 456). For that purpose a relaxation parameter ω is introduced in equation (1.21), leading to

$$\mu'_j = \frac{(\omega \mathbf{u}_{j-1}, \mathbf{a}_j)}{(\mathbf{a}_j, \mathbf{a}_j)}$$

where

$$\mathbf{u}_{j-1} = \mathbf{b} - \sum_{i=1}^{j-1} \mu_i \mathbf{a}_i .$$

As a matter of fact one can show that the relaxed version of the iterative projection method maintains all its desirable characteristics. In addition there is some theory as well as evidence from simulations concerning the choice of the relaxation parameter ω.

First numerical experiences for non-singular as well as near-singular system matrices are quite promising. The computational burden can be significantly reduced by relaxation in all instances. Degeneracy requires larger ω values. As pointed out earlier, ill-posed linear equation systems should not be solved with classical techniques as regularity of the system matrix is required throughout. The relaxed iterative projection algorithm has the potential to bridge this gap in numerical methodology. But there is always a price to be paid: the explicit calculation of the S_k when applied to the system of normal equations in (1.13).

Some final remarks on iterative solutions

When an additive model is evaluated via an iterative numerical procedure in the sample world of scatterplot smoothers some questions still remain open. Obviously small sample behaviour is different from asymptotic behaviour. But apart from that, the convergence behaviour is not even fully understood for backfitting, although being a well established technique. Statistical considerations in these iterative schemes become really tedious. It is only recently that findings were published for certain smoother types.

Until recently the only findings for backfitting with arbitrary linear scatterplot smoothers have been limited to the case of two dimensions. Buja *et al.* (1989, p. 489ff) and Hastie and Tibshirani (1990, p. 118ff) showed that both the convergence of the backfitting algorithm and the uniqueness of its solution depend on the behaviour of the product of the two smoother matrices S_1 and S_2, which have to be symmetric and shrinking. We call a smoother matrix S shrinking if $\|S\mathbf{y}\| \leq \|\mathbf{y}\|$ for all $\mathbf{y}$ (Euclidean or any other matrix

norm). This will be the case if all its singular values are less or equal one. Smoothers fulfilling this requirement are projection smoothers (expressed in parametric terms such as regression splines) and smoothing splines. Opsomer (1996) could extend the above result to more than two dimensions. He derived recursive expressions for the backfitting estimator. Also for $d > 2$ the existence and the uniqueness of this estimator depend on the characteristics of the pairwise products of the smoother matrices: A d-variate additive model with smoother matrices $S_1, \ldots, S_d$ will converge to an unique solution, if

$$\max_{\delta \in [2,d]} \| \sum_{j=1}^{\delta-1} S_\delta S_j \| < 1 \tag{1.22}$$

for some matrix norm where δ denotes the δs block consisting of a $n \times n$ identity matrix and zero elements otherwise in a matrix E_δ, so that $\hat{g}_\delta = E_\delta A^{-1} B y$ (provided the inverse of A for equation system (1.5) exists; for more details see Opsomer, 1996). Thus the index δ corresponds to the δ's block of the solution vector $\hat{g}$. One can prove that the system matrix A is invertible when the spectral radius of $T_\delta = \| \sum_{j=1}^{\delta-1} S_\delta^* S_j^* \|$ is less than one for all $\delta \in [2, d]$. Because of the fact that the ordering of the predictor functions is arbitrary, it is sufficient to show that the condition (1.23) holds for one particular ordering.

Ansley and Kohn (1994) studied the statistical features of d-dimensional backfitting for the special case of smoothing splines. Opsomer and Ruppert (1997) analysed backfitting for the bivariate additive model and Opsomer (1996) for the d-variate additive model when local polynomial regression is applied. Linton, Mammen and Nielsen (1997) derived the asymptotic properties of a backfitting projection algorithm under weaker conditions for the local linear and a special Nadaraya–Watson estimator in an additive model of dimension d. Even more research is desirable as far as the convergence behaviour of these and other scatterplot smoothers is concerned. This is specially true for the iterative projection algorithm.

Recently there has been considerable effort to avoid iterative techniques in the evaluation of additive models. These developments are introduced in the next section.

1.2.4 The marginal integration method

Here we discuss a direct method, introduced independently by Newey (1994), Tjøstheim and Auestad (1994), and Linton and Nielsen (1995), labelled marginal integration. It is based on the fact that up to a constant, $g_j(x_j)$ is equal to

$$\mathcal{E}\{m(X_1, \ldots, X_{j-1}, x_j, X_{j+1}, \ldots, X_d)\},$$

where $m(x) = E(Y|\mathbf{X} = \mathbf{x})$. The estimate of g_j is obtained by marginal integration of an estimate of m. Therefore no iterations are necessary. The explicit definition of the estimation procedure allows a detailed asymptotic analysis which is easier to carry out compared to the backfitting algorithm.

Linton and Nielsen (1995) studied the additive regression model

$$g(x_1, x_2) = g_1(x_1) + g_2(x_2),$$

where $g(x_1, x_2)$ is a bivariate regression function. Let Q be a deterministic (continuous or discrete) weight function with $dQ(x_2) = 1$. Assume a density q of Q with respect to either a Lebesgue or a counting measure. The contrast

$$\alpha_Q(x_1) = \int g(x_1, x_2) dQ(x_2) \tag{1.23}$$

is considered, where $\alpha_Q(x_1) = g_1(x_1) + c_1$ with $c_1 = \int g_2(x_2) dQ(x_2)$. Thus $\alpha_Q(x_1)$ is identifiable up to a constant, the univariate component of the additive structure we are interested in.

Let us have observed data $\{x_{1i}, x_{2i}, y_i\}$ for $i = 1, \ldots, n$. Linton and Nielsen (1995) propose for the additive fit a local linear smoother with product kernels. Thus $\hat{g}(x_1, x_2) = \sum w_j(x_1, x_2) y_j$ (the w_j denoting some weights) is the first element of

$$(D^T K D)^{-1} D^T K y,$$

where $y = (y_1, \ldots, y_n)^T$ and $D = (d_1, \ldots, d_n)^T$ with $d_j = (1, x_{1j} - x_1, x_{2j} - x_2)^T$, while K is a diagonal $n \times n$ matrix with typical diagonal elements $k_{b_1}(x_{1j} - x_1) k_{b_2}(x_{2j} - x_2)$. The b_1 and b_2 are scalar bandwidths, and $k_b(\cdot) = b^{-1} k(b^{-1})$ for any b. k denotes a univariate differentiable density function, symmetric about zero. Then $\alpha_Q(x_1)$ is estimated by the sample version of (1.23), i.e.

$$\hat{\alpha}_Q(x_1) = \int \hat{g}(x_1, x_2) dQ(x_2) = \sum_{j=1}^{n} w_{Qj}(x_1) y_j,$$

where $w_{Qj}(x_1) = \int w_j(x_1, x_2) dQ(x_2)$.

We already know from Stone (1985) that both g_1 and g_2 can be estimated with the one-dimensional convergence rate $n^{2/5}$. Under the usual assumptions such as independent, identically distributed observations, a joint density $f(x_1, x_2)$, marginals $f_{x_1}(x_1)$ and $f_{x_2}(x_2)$ with marginal cumulative distribution functions $F_{x_1}(x_1)$ and $F_{x_2}(x_2)$, Linton and Nielsen (1995, p. 95) proved the following theorem:

Theorem 4. Assume that g possesses two continuous partial derivatives in each direction, while f is continuously differentiable. Suppose also that $b_1, b_2 \to 0$ and $n b_1 b_2^2 \to \infty$. Then conditional on $\{x_{1i}, x_{2i}\}_{i=1}^n$,

$$(n b_1)^{1/2} [\hat{\alpha}_Q(x_1) - \mathcal{E}\{\hat{\alpha}_Q(x_1)\}] \to N(0, s^2(x_1)),$$

in distribution, where $s^2(x_1) = v(k) \sigma^2 \int f^{-1}(x_1, x_2) q^2(x_2) dx_2$, with $v(k) = \int k(t)^2 dt$, while

$$\mathcal{E}[\hat{\alpha}_Q(x_1)] - \alpha_Q(x_1) = \{b_1^2 \beta_1(x_1) + b_2^2 \beta_2(x_1)\}(1 + o(1)),$$

where

$$\beta_1(x_1) = \frac{\mu(k)}{2} \int \frac{\partial^2 g}{\partial x_1^2}(x_1, x_2) dQ(x_2), \quad \beta_2(x_1) = \frac{\mu(k)}{2} \int \frac{\partial^2 g}{\partial x_2^2}(x_1, x_2) dQ(x_2),$$

with $\mu(k) = \int t^2 k(t) dt$. Hence, when $b_1, b_2 = O(n^{-1/5})$, then $\hat{\alpha}_Q(x_1)$ of the marginal integration method converges at $n^{2/5}$, the optimal rate.

Further Q has to be chosen. Linton and Nielsen (1995) recommend an empirical distribution function as an approximation of Q which is integrated mean squared error optimal. Although in general there is a dependence on both f and g. If g is exactly additive and $b_2 = o(n^{-1/5})$, the bias does not depend on Q.

A main disadvantage compared to the iterative backfitting procedure is the fact that the marginal integration method as outlined above does not hold true for dimensions (explanatory variables) $d > 2$. Theorem 4 cannot be extended to cope with higher dimensions. Moreover, the local linear pilot estimator is a non-optimal choice. There have been several yet unpublished attempts to introduce a marginal integration estimator for $d > 2$. Among those I would like to mention Severance-Lossin and Sperlich (1997), who also propose a methodology to estimate the derivatives for additive separable models, important for certain applications in economics and biology. They also suggest the application of a local polynomial pilot estimator restricted to the direction of interest, with the effect that more information remains in the constant. This leads to a faster algorithm with little loss in optimality.

In the marginal integration method the scatterplot smoothers are restricted to local polynomial and kernel fits. In the following we sketch the marginal integration algorithm for local polynomials with kernel weights.

The algorithm is represented in the GAUSS language. The input parameters are: x the (n x d) matrix of the explanatory variables, y the response variable, xg the design points, respectively the grid $(ng \times d)$ on which we want to estimate the additive components, and the bandwidths $h, hs \in I\!R^d$ for the directions of interest, respectively not of interest. The result is written in fh.

Marginal Integration Algorithm

```
fh = zeros(ng,d);
p = 1;     @ degree of local polynomial @
pick = 1~zeros(1,p);     @ pick function estimate @
j = 1;
     do until (j>d);
     i = 1;
     do until (i>ng);     @ for dimension of interest @
          hv = g;
          hv[j,.] = h;
          xest = x;
          xest[.,j] = xg[i,j].*ones(n,1);
```

```
        weight = zeros(1,n);
        Z = ones(rows(x),1);
        k = p;
        do while (k);     @ create polynomial design @
             Z = ones(n,1)~(Z*~(x[.,j]-xest[i,j]));
             k = k-1;
        endo;
        l = 1;
        do until (l>n);     @ kernel weights @
             dx = kernel((x-xest[l,.])./(hv'));
             zw = Z'.*dx';
             weight = pick*inv(zw*Z)*zw+weight;
             l = l+1;
        endo;
        fh[i,j] = (weight./n)*y;     @ result @
        i = i+1;
    endo;
    j = j+1;
endo;
```

Fan *et al.* (1996) extend the marginal integration method to the estimation of semiparametric additive partial linear models (for an overview of semiparametric regression models see Schimek, 1997). A problem in practice is certainly the necessary $O(n^3)$ algorithm. Sperlich, Tjøstheim and Yang (1998) propose a concept which not only allows for interaction in additive models, but also estimating and testing such interactions.

Finally let us make some comments about the marginal integration method compared to backfitting. For the case of two explanatory variables with smoother matrices S_1 and S_2, respectively, the backfitting procedure converges to the n-dimensional vectors

$$\hat{\mathbf{g}}_1^{\infty} = [I - (I - S_1 S_2)^{-1}(I - S_1)]\mathbf{y}$$

and

$$\hat{\mathbf{g}}_2^{\infty} = [I - (I - S_2 S_1)^{-1}(I - S_2)]\mathbf{y}$$

provided $\|S_1 S_2\| < 1$ (see Hastie and Tibshirani, 1990, p. 118f). These expressions are quite intractable for general linear smoothers – although their linearity can be exploited to construct pointwise confidence bounds, etc. – and as a consequence the bias and variance cannot be derived apart from special cases (see the final remarks on iterative solutions). This is certainly a drawback compared to marginal integration with its result of Theorem 4.

It is also a well-known fact that the marginal integration method can be very inefficient with respect to the regression function m. Let us study the marginal integration approach in terms of L_2 function space. As in Section 1.2.1 assume $\mathcal{H}_j$, for $j = 1, \ldots, d$, denoting the Hilbert spaces of measur-

able functions $\phi_j(X_j)$ with $\mathcal{E}\phi_j(X_j) = 0$, $\mathcal{E}\phi_j^2(X_j) < \infty$, and inner product $\langle \phi_j(X_j)\phi_j^*(X_j)\rangle = \mathcal{E}\phi_j(X_j)\phi_j^*(X_j)$. Moreover let $\mathcal{H}^{add} = \sum_j^d \mathcal{H}_j$ and all $\mathcal{H}_j$ be subspaces of $\mathcal{H}$, the space of measurable functions of $X_1, \ldots, X_d$. We remember that solving the normal equations means finding that member of $\mathcal{H}^{add}$ which is closest to the regression function $m \in \mathcal{H}$, corresponding to equation (1.1).

The empirical marginal integration map $\pi : \mathcal{H} \to \mathcal{H}^{add}$ is of the form

$$\begin{aligned} \pi(g)(x) &= \int g(x_1, x_2) f_{x_2}(x_2) dx_2 + \int g(x_1, x_2) f_{x_1}(x_1) dx_1 \\ &\quad - \int g(x) f(x_1, x_2) dx, \end{aligned}$$

where

$$g(x) = g_0 + \sum_{j=1}^{2} g_j(X_j). \tag{1.24}$$

The additive functions are fixed points of π. That is the reason why the marginal integration method consistently estimates g in equation (1.24), hence π is idempotent. It is also linear, i.e. for any $a, b \in \mathcal{H}$, $\pi(a + b) = \pi(a) + \pi(b)$. However, π is not self-adjoined and thus not an orthogonal projection (Linton, 1997, p. 470). This explains why $\mathcal{E}[m(\mathbf{X})] \in \mathcal{H}^{add}$ in solving the normal equations, provides a better mean squared error approximation to the regression function m. This also means that the marginal integration method is not that efficient in estimating g, respectively its components.

To overcome this drawback of the marginal integration method Linton (1997) suggested to calculate starting values via marginal integration and then to apply a single-step backfitting iteration. He could prove efficiency of this technique in the sense of being equivalent to a procedure based on knowing the other components of the regression function. The interpretation of these estimation results is not clear, especially when the additivity assumption is violated. Another yet unpublished proposal to improve efficiency, also from a computational point of view, is due to Hengartner (1996). He tackles the problem of pilot estimator choice, favouring a so-called internalized estimator.

Apart from the fact that there is no agreement about appropriate algorithms for marginal integration there are only two studies (one published) comparing it with the backfitting procedure. The one is Nielsen and Linton (1998), and the other Sperlich, Linton and Härdle (1997). According to Nielsen and Linton (1998) both marginal integration and backfitting can be seen with respect to minimizing an integrated mean squared error criterion. Marginal integration optimizes the criterion with weighting given by an independent product measure. This is correct independently of whether additivity holds or not. Backfitting achieves the same goal with weighting given a joint empirical measure (joint density). The latter makes sense without prior knowledge of the situation the data come from. The definite advantage of the marginal integration estimator is that it is explicitly defined. However,

there is a loss of efficiency for non-independent designs. In conclusion Nielsen and Linton (1998, p.221) write "Perhaps the more significant disadvantage of integration, which is specific to this nonparametric setting, is that the curse of dimensionality is completely eliminated. Thus we must use bias reduction arguments to achieve the optimal rate in high dimensions and we might expect poor small sample performance relative to the asymptotics".

Sperlich, Linton and Härdle (1997) have undertaken the most extensive simulation study till now, in which they tried to trace down performance differences between the iterative backfitting procedure and the direct marginal integration method for small samples and $d = 2$, and in one instance $d = 4$. They applied local polynomial fitting and Nadaraya–Watson kernel smoothing to a number of simple additive functions. The error assumptions were uniform and Normal with constant variance and correlation $\rho = \{0.0, 0.4, 0.8\}$ between the covariates (case $d = 2$). The bandwidth was chosen by a rule of thumb due to Linton and Nielsen (1995) and by the plug-in method of Severance-Lossin and Sperlich (1997).

The results can be summed up as follows: There are many similarities between the backfitting algorithm and the marginal integration algorithm with respect to their statistical performance. Both algorithms run into severe problems in designs with increasing correlation (see remarks on ill-posed systems in 1.2.3.2), although backfitting does perform slightly better. At least for $d = 2$ asymptotics hold empirically. The adopted smoothing method does generally not matter much. Backfitting works better at boundary points and under data sparseness while the integration method is more capable of estimating the components as opposed to the function itself (marginal influences). This is specially true for $d = 4$, i.e. higher dimensions. The obtained results are not conclusive in that sense of generally favouring one method.

Implementations of the marginal integration procedure are rare, more or less of prototype nature. Publically available is solely the macro `intest` in XploRe 4.0 (for Windows 95, Windows NT and UNIX) similar to the algorithm presented above. XploRe can be downloaded at `http://www.xplore-stat.de/`.

Finally we should mention that bandwidth and smoothing parameter choice (see later in this paper) remain troublesome issues in the context of additive models, whatever the method of estimation is, and are likely to hamper the interpretation of comparative simulation studies (controlled for in the experiment described above).

1.3 GENERALIZED ADDITIVE MODELS

In this section we extend the additive model to the class of Generalized Additive Models (GAMs). GAMs were introduced in a series of papers by Hastie and Tibshirani (1986, 1987a, 1987b) and Stone (1986). They are described in detail in Hastie and Tibshirani (1990).

Their purpose is to allow for even more flexibility than in additive models. On the other hand they retain an important feature of GLMs, additivity of the predictors. However, the predictor effects are generally nonlinear due to arbitrary functions g_j.

A special case occurs if only one predictor function, say $g_1(x_1)$, is evaluated nonparametrically, while the remaining explanatory variables still enter as a linear combination, say $\tilde{X}^T\beta = x_2\beta_2 + \cdots + x_d\beta_d$. Such semi-parametric models were first considered by Green and Yandell (1985). New developments in this area are reviewed in Schimek (1997).

Our generalization of the additive model becomes

$$\mathcal{E}[Y|\mathbf{X} = x] = G\left(g_0 + \sum_{j=1}^{d} g_j(x_j)\right),$$

where $G(\cdot)$ is a fixed link function and the distribution of Y is assumed to belong to the exponential family as in GLMs. The assumptions concerning identifiability of the functions g_j remain the same (see Section 1.2).

The fitting of a GAM consists of two parts: Estimating the additive predictor and linking it to the function $G(\cdot)$ in an iterative manner. The first part requires solving the system of normal equations as already discussed. For the second part the so-called local scoring algorithm is applied.

The local scoring algorithm is practically identical with the Fisher scoring algorithm used in GLMs, except that the least squares step is replaced by the solution step of the normal equations. In GLMs the least squares step is used to update the estimate $\hat{\beta}$ for the linear predictor $X^T\beta$. Here we apply the backfitting or iterative projection algorithm to update the estimates for g_0 and the g_js.

In Section 1.2.1 smoothing in the additive model was motivated in L_2 function space. In Section 1.2.4 the Hilbert space interpretation was used to interpret its estimation through marginal integration. This view also helps to understand GAMs (Hastie and Tibshirani, 1990, p. 148f).

Given $\mathbf{X} = (X_1, \ldots, X_d)$, the response Y has conditional density $h(y, \zeta)$, where $\zeta = \zeta(\mathbf{X})$ is the true regression parameter fulfilling $\zeta \in \mathcal{H}$. The corresponding log-likelihood for a single observation is denoted by l. For $\zeta(\mathbf{X})$ we try to obtain the best approximation by maximizing the expected log-likelihood

$$\mathcal{E}l[\eta(\mathbf{X}), Y] \tag{1.25}$$

over $\eta(\mathbf{X}) = \sum_j^d g_j(X_j) \in \mathcal{H}^{add}$. Stone (1986) gave conditions for the existence and uniqueness of the best additive approximation. The maximum of equation (1.25) is characterized by a score function $\partial l/\partial\eta$ orthogonal to the space of fits, or equivalently

$$\mathcal{E}\left(\frac{\partial l}{\partial \eta} \mid X_j\right) = 0$$

for all j. A solution for these nonlinear equations in η and g_j can be found by a linearization about an approximate η_0. The final result is

$$g_j(X_j) = \frac{\mathcal{E}\left[W_0(\mathbf{X})\left\{Z_0 - \sum_{k\neq j} g_k(X_k)\right\} \mid X_j\right]}{\mathcal{E}[W_0(\mathbf{X} \mid X_j]}, \tag{1.26}$$

where

$$Z_0 = \eta_0 + (\partial l)/(\partial \eta_0)/(-\partial^2 l)/(\partial \eta_0^2)$$

and

$$W_0(\mathbf{X}) = (-\partial^2 l)/(\partial \eta_0^2)$$

The two operators $\mathcal{E}$ in equation (1.26) are weighted conditional expectations which can be evaluated by scatterplot smoothers.

1.3.1 Penalized least squares

Let us consider again the penalized least squares concept introduced in Section 1.2.2, which results in cubic smoothing splines. The estimation of the additive model (representing regression with a continuous, usually Gaussian response) was based on equation (1.10). Now, for the generalized case (representing regression with a non-continuous response variable) we can apply the penalized log-likelihood criterion (Hastie and Tibshirani, 1986; Fahrmeir and Tutz, 1994)

$$PL(g_1, \ldots, g_d) = \sum_{i=1}^{n} l_i(y_i; \eta_i) - \frac{1}{2}\sum_{j=1}^{d} \lambda_j \mathbf{g}_j^T K_j \mathbf{g}_j,$$

where l denotes the log-likelihood (as in GLMs), η_i the additive predictor values

$$\eta_i = \sum_{j=1}^{d} g_j(x_{ij}),$$

and $\mathbf{g}_j = (g_j(x_{1j}), \ldots, g_j(x_{nj}))^T$, $\mathbf{g}_j \in \mathcal{H}_j$, $j = 1, \ldots, d$, the vector of smooth spline functions. The penalty matrices K_j for each predictor X_j are defined as in equation (1.14) for the additive model.

The Fisher scoring algorithm was originally designed to maximize a (penalized) log-likelihood criterion (see McCullagh and Nelder, 1989 for details). Here we study the Fisher scoring algorithm first and modify it later to what is now known as local scoring procedure in nonparametric regression.

1.3.1.1 Fisher scoring and local scoring Let η be the n-dimensional vector of additive predictor values η_i. Differentiation of $\partial PL/\partial \mathbf{g}_j$ for $j = 1, \ldots, d$, yields the likelihood equations

$$s_1 = \lambda_1 K_1 \mathbf{g}_1, \ldots, s_d = \lambda_d K_d \mathbf{g}_d,$$

where the derivative $s = (s_1, \ldots, s_n)$ of the log-likelihood is given by

$$s_i = \frac{D_i}{\sigma_i^2}(y_i - \mu_i),$$

with $D_i = \partial G / \partial \eta_i$ as the first derivative of the response function and σ_i^2 the variance function evaluated at $\mu_i = G(\eta_i)$.

Let $W = \text{ diag}(w_1, \ldots, w_n)$ with $w_i = D_i^2/\sigma_i^2$ be the expected information matrix. Then the Fisher scoring iterations are given by

$$\begin{pmatrix} W^{(k)} + \lambda_1 K_1 & W^{(k)} & \ldots & W^{(k)} \\ W^{(k)} & W^{(k)} + \lambda_2 K_2 & \ldots & W^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ W^{(k)} & W^{(k)} & \ldots & W^{(k)} + \lambda_d K_d \end{pmatrix} \begin{pmatrix} \mathbf{g}_1^{(k+1)} - \mathbf{g}_1^{(k)} \\ \mathbf{g}_2^{(k+1)} - \mathbf{g}_2^{(k)} \\ \vdots \\ \mathbf{g}_d^{(k+1)} - \mathbf{g}_d^{(k)} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{s}^{(k)} - \lambda_1 K_1 \mathbf{g}_1^{(k)} \\ \mathbf{s}^{(k)} - \lambda_2 K_2 \mathbf{g}_2^{(k)} \\ \vdots \\ \mathbf{s}^{(k)} - \lambda_p K_d \mathbf{g}_d^{(k)} \end{pmatrix},$$

where $W^{(k)}$ and $\mathbf{s}^{(k)}$ are W and s evaluated at $\eta^{(k)} = \eta(\mathbf{g}_1^{(k)}, \ldots, \mathbf{g}_d^{(k)})$. Defining the working observation vector

$$\tilde{\mathbf{y}}^{(k)} = \eta^{(k)} + (W^{(k)})^{-1}\mathbf{s}^{(k)}$$

and the smoother matrices

$$S_j^{(k)} = (W^{(k)} + \lambda_j K_j)^{-1} W^{(k)}$$

the iterations can be expressed in the form

$$\begin{pmatrix} I & S_1^{(k)} & \ldots & S_1^{(k)} \\ S_2^{(k)} & I & \ldots & S_2^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ S_d^{(k)} & S_d^{(k)} & \ldots & I \end{pmatrix} \begin{pmatrix} \mathbf{g}_1^{(k+1)} \\ \mathbf{g}_2^{(k+1)} \\ \vdots \\ \mathbf{g}_d^{(k+1)} \end{pmatrix} = \begin{pmatrix} S_1^{(k)} \tilde{\mathbf{y}}^{(k)} \\ S_2^{(k)} \tilde{\mathbf{y}}^{(k)} \\ \vdots \\ S_d^{(k)} \tilde{\mathbf{y}}^{(k)} \end{pmatrix}. \tag{1.27}$$

This is a $(n \times d)$-dimensional linear system comparable to the normal equations in (1.5). For the same reasons as pointed out there a direct solution is not feasible, apart from special instances, to obtain the next iterate $\mathbf{g}^{(k+1)} = (\mathbf{g}_1^{(k+1)}, \ldots, \mathbf{g}_d^{(k+1)})$. Instead (1.27) is rewritten as

$$\begin{pmatrix} \mathbf{g}_1^{(k+1)} \\ \mathbf{g}_2^{(k+1)} \\ \vdots \\ \mathbf{g}_d^{(k+1)} \end{pmatrix} = \begin{pmatrix} S_1(\tilde{y}^{(k)} - \sum_{j \neq 1} \mathbf{g}_j^{(k+1)}) \\ S_2(\tilde{y}^{(k)} - \sum_{j \neq 2} \mathbf{g}_j^{(k+1)}) \\ \vdots \\ S_d(\tilde{y}^{(k)} - \sum_{j \neq d} \mathbf{g}_j^{(k+1)}) \end{pmatrix}$$

and solved iteratively by the backfitting or iterative projection algorithm in the inner loop of the local scoring algorithm. As mentioned earlier, solving the linear system in the core of Fisher scoring iteratively, produces a new algorithm, known as local scoring. It can be characterized as follows:

Generalized Additive Model Algorithm

---

1. *Initialization of outer loop*: $\mathbf{g}_0^{(0)} = F\left(n^{-1}\sum_{i=1}^n y_i\right), \mathbf{g}_1^{(0)} = \ldots = \mathbf{g}_d^{(0)} = 0, F = G^{-1}$

2. *Scoring steps*: For $k = 0, 1, 2, \ldots$:
   Compute the current working observations

$$\tilde{\mathbf{y}}_i^{(k)} = \eta_i^{(k)} + \frac{\mathbf{y}_i - G\left(\eta^{(k)}\right)}{D_i^{(k)}}$$

   and the weights $w_i^{(k)} = \left(D_i^{(k)}/\sigma_i^{(k)}\right)^2$ with $\eta_i^{(k)} = \mathbf{g}_0^{(k)} + \sum_{j=1}^d \mathbf{g}_j^{(k)}(x_{ij})$, $i = 1, \ldots, n$

3. *Backfitting or iterative projection steps*: Solve (1.27) for $\mathbf{g}^{(k+1)}$

   (a) *Initialization of inner loop*:

$$\mathbf{g}_0^{(k)} = n^{-1}\sum_{i=1}^n \tilde{\mathbf{y}}_i^{(k)}, \mathbf{g}_j^0 := \mathbf{g}_j^{(k)}, \quad j, p, c = 1, \ldots, d$$

   (b) Compute updates $\mathbf{g}_j^0 \to \mathbf{g}_j^1, \quad j = 1, \ldots, d$ in each backfitting iteration (index p for "previous" and c for "current")

$$\mathbf{g}_j = S_j^1\left[\tilde{\mathbf{y}} - \mathbf{g}_0 - \sum_{c<j}\mathbf{g}_c^1 - \sum_{p>j}\mathbf{g}_p^0\right]$$

   and apply weighted cubic spline fits to the "working residuals" $\tilde{\mathbf{y}} - \mathbf{g}_0 - \sum_{c<j}\mathbf{g}_c^1 - \sum_{p>j}\mathbf{g}_p^0$
   Set $\mathbf{g}_j^0 := \mathbf{g}_j^1$ for $j = 1, \ldots, d$ after each iteration

   (c) Stop backfitting or iterative projections when $\|\mathbf{g}_j^0 - \mathbf{g}_j^1\|$ for $j = 1, \ldots, d$ is very small
   Set $\mathbf{g}_j^{(k+1)} := \mathbf{g}_j^1$ after the final iterate

4. Iterate steps 2 and 3 until some termination criterion is met

As termination criterion one could use for instance

$$\frac{\sum_{j=1}^d \|\mathbf{g}_j^{(k+1)} - \mathbf{g}_j^{(k)}\|}{\sum_{j=1}^d \|\mathbf{g}_j^{(k)}\|} \leq \varepsilon.$$

Finally we should point out again that the constant g_0 was introduced for the purpose of identifiability of the additive approximation. Hence the local scoring algorithm is applied to centered working observations, i.e. $\tilde{\mathbf{y}}$ is centered to have mean zero which ensures that throughout the iterations the $\mathbf{g}_j s$ have mean zero.

As far as convergence is concerned, the local scoring algorithm does not require special attention. Local scoring iterations correspond to Newton-Raphson steps. Convergence results and ways of step optimization, usually not needed for the estimation of GAMs, are considered in Ortega and Rheinboldt (1970). Crucial in the fitting of GAMs is the inner loop providing the solution of the normal equations, not the outer loop carrying out the local scoring iterations.

In this section we emphasized cubic smoothing splines because of their numerical practicability and software availability. Alternatively any other symmetric linear smoother could be applied. Such smoothers can also be represented by matrices $S_1, \ldots, S_d$ of similar shape as those of smoothing splines. The only difference is that one has to define the penalty matrices $K_1, \ldots, K_d$ in the penalized log-likelihood criterion by

$$K_j = S_j^- - I,$$

where S_j^- is any generalized inverse of S_j (see Golub and van Loan, 1989, for generalized inverses). Weighted cubic spline smoothing in the inner loop of the algorithm is then substituted by a corresponding symmetric linear smoother, for instance a running line smoother.

Further we have to remark that the local scoring algorithm is not limited to the use of backfitting or iterative projections. Any iterative procedure which solves a linear system would serve this purpose.

GAM can be fitted in the statistical environment S-Plus for UNIX and DOS (Windows 95 or NT) platforms. Alternative sofware is XploRe (Härdle *et al.*, 1995) for both UNIX and DOS (Windows 95 or NT) computers. Furthermore there is a function `addreg` in Douglas Nychka's FUNFITS software, available at `http://www.stat.ncsu.edu/~nychka/funfits/index.html` for UNIX platforms with a S-Plus environment (S/S-Plus functions are required to execute FUNFITS). The most widely used implementation is actually that of S-Plus. It is based on backfitting in combination with local scoring as described above. Very useful for GAM fitting in S-Plus is Chambers and Hastie (1992, especially Chapter 7) but also Venables and Ripley (1997, Chapter 10), apart from the manual (MathSoft, 1997, Chapter 8).

Finally we would like to mention that the additive fitting concept can also be applied to the estimation of autoregressive time series (beyond the scope of this paper). Härdle and Chen (1995, p. 379ff) discuss the so-called Nonlinear Additive AR models (NAAR).

Let us now switch from spline smoothing in combination with backfitting and local scoring to kernel smoothing. There we discuss recent developments in the marginal integration method which make it possible to apply it to

GAMs.

1.3.2 The marginal integration method for generalized additive models

marginal integration, for generalized additive models

In Section 1.2.4 we described a marginal integration method for the evaluation of additive models, introduced by Linton and Nielsen (1995). Now this direct method (no iterations) is based on the fact that up to a constant, $g_j(x_j)$ is equal to

$$\mathcal{E}\{\eta(X_1, \ldots, X_{j-1}, x_j, X_{j+1}, \ldots, X_d)\},$$

where $\eta(x) = \mathcal{E}(Y|\mathbf{X} = \mathbf{x})$. The estimate of g_j is obtained by marginal integration of an estimate of η. Linton and Härdle (1996) extended this idea to the estimation of

$$G\{\eta(x)\} = g_0 + \sum_{j=1}^{d} g_j(x_j), \tag{1.28}$$

where G is a known link function. They differentiate between a full and a partial model specification. Here we are interested in the first kind of specification which represents generalized additive models with errors following an exponential family distribution. This means that the variance is functionally related to the mean. In the partial model specification the variance function is unrestricted. When G is the identity function we have exactly the situation examined in Section 1.2.4.

Let us assume a partition $X = (X_1, X_2)$, where X_1 is the one-dimensional direction of interest and X_2 a $(d-1)$-dimensional nuisance direction ($d \geq 2$ predictors). Further let $x = (x_1, x_2)$. For any predictor function η and link function G, a functional

$$\gamma_1(x_1) = \int G\{\eta(x_1, x_2)\} p_2(x_2) dx_2 \tag{1.29}$$

is defined, where $p_2(x_2)$ is the joint density of x_2. Because of the additive structure of (1.28) γ_1 is g_1 up to a constant g_0. Linton and Härdle (1996) proposed to replace both η and p_2 in (1.29) by estimates. For that purpose they use the Nadaraya–Watson kernel estimator

$$\hat{\eta}(x_1, x_2) := \frac{n^{-1} \sum_{i=1}^{n} K_h(x_1 - X_{1i}) L_l(x_2 - X_{2i}) Y_i}{n^{-1} \sum_{i=1}^{n} K_h(x_1 - X_{1i}) L_l(x_2 - X_{2i})},$$

where K and L are compactly supported Lipschitz continuous kernels integrating to one. We have $K_h(\cdot) = h^{-1} K(h^{-1} \cdot)$ and $L_l(\cdot) = l^{-(d-1)} L(l^{-1} \cdot)$ and take K to be a second-order kernel and L to be product of univariate kernels of order q, i.e. $\int L(u) u^k du = 0$ for $k = 1, \ldots, q-1$. For high dimensions d it is necessary to reduce the bias in the nuisance directions to achieve the optimal one-dimensional rate of convergence for the direction of interest.

The functional $\gamma_1(x_1)$ can be estimated by the sample version of (1.29), that is

$$\tilde{\gamma}_1(x_1) := \sum_{i=1}^{n} G\{\eta(x_1, X_{2i})\}.$$

Apart from the case where G is the identity function, $\tilde{\gamma}_1(x_1)$ is a non-linear function of y_i. For the estimation of the regression surface the above procedure is applied to each direction by redefining in turn the jth covariate to be X_1 and the remainder to be X_2. Estimates are obtained for each γ_j at all the design points. Finally $\eta(x)$ is reestimated by

$$\tilde{\eta}(x) := F\left\{\sum_{j=1}^{d} \tilde{g}_j(x_j) + \tilde{g}_0\right\},$$

where $F = G^{-1}$,

$$\tilde{g}_0 = d^{-1}n^{-1}\sum_{j=1}^{d}\sum_{i=1}^{n} \tilde{\gamma}_j(X_{ji})$$

and $\tilde{g}_1(x_1) = \tilde{\gamma}_1(x_1) - \tilde{g}_0$.

In practice it is often a problem to identify relevant covariates with respect to a response variable. Related work in an unpublished research report by Härdle and Korostelev (1994) addresses the problem of searching for significant variables.

Different from the estimation concept for GAMs worked out in the previous section, the marginal integration method allows for asymptotic considerations. Linton and Härdle (1996, p. 534ff) proved the following theorem.

Theorem 5. Let the order q of L satisfy $q > d - 1$. Let $h = \beta n^{-1/5}$. Assume that $n^{2/5}l^q \to \infty$, that $n^{2/5}l^{d-1} \to \infty$, that F is twice continuously differentiable, and that the additivity assumption holds. Then

$$n^{2/5}\{\tilde{\eta}(x) - \eta(x)\} \to N\{b(x), v(x)\}$$

in distribution, where

$$b(x) = F'\left\{\sum_{j=1}^{d} g_j(x_j) + g_0\right\}\sum_{j=1}^{d} b_j(x_j)$$

and

$$v(x) = F'\left\{\sum_{j=1}^{d} g_j(x_j) + g_0\right\}\sum_{j=1}^{d} v_j(x_j).$$

The main consequence is that the rate of convergence of $\tilde{\eta}$ is not influenced by the curse of dimensionality. The obtained rate of $n^{2/5}$ is still that derived by Stone (1986) for one-dimensional regression functions. This is a remarkable result.

Despite asymptotic optimality one has to be cautious with respect to the algorithm, especially for dimensions $d > 2$. In practice covariates are more or less correlated, data sparse and sample sizes small. Extensive simulations for variable sample sizes and different combinations of regression and link functions are necessary before any conclusions can be drawn. Another problem is the lack of software apart from the macro `gintest` in XploRe 4.0, available at `http://www.xplore-stat.de/` based on a prototype algorithm.

In conclusion, the marginal integration method for generalized additive models is certainly interesting from a theoretical point of view but too new, to comment its value for real data analysis.

1.3.3 Vector generalized additive models

Yee and Wild (1996) extended the class of GAMs to handle multivariate (vector) regression problems. Vector Generalized Additive Models (VGAMs) enhance the idea of vector GLMs (not explicitly dealt with in McCullagh and Nelder, 1989).

Suppose that for each measurement unit under study a q-dimensional response vector $\mathbf{y} = (y_1, \ldots, y_q)^T$ and a d-dimensional covariate vector $\mathbf{x} = (x_1, \ldots, x_d)^T$ are observed. A VGAM is any model for which the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is of the form

$$f(\mathbf{y} \mid \mathbf{x}) = f^*(\mathbf{y}, \eta_1(\mathbf{x}), \ldots, \eta_M(\mathbf{x}))$$

for some function f^*, where (for $j = 1, \ldots, M$)

$$\eta_j(\mathbf{x}) = g_{(j)0} + g_{(j)1}(x_1) + \ldots + g_{(j)d}(x_d).$$

For the evaluation of a VGAM simultaneous smoothing is required. This can be achieved by an interesting generalization of cubic smoothing splines due to Fessler (1991). His vector spline technique has some nice features, also from a numerical point of view. As a matter of fact it can be seen as a generalization of univariate Reinsch splines which are computationally highly efficient. It should be pointed out that the underlying system matrix of such a VGAM is very large. Another feature of vector splines is that they allow for correlated errors ϵ_i as long as their covariance matrix Σ_i is known and the errors are independent between samples.

We suppose to have a vector response $\mathbf{y}_i$ of dimension M at each value of a scalar x_i, assumed to be a realization from the vector measurement model in $\mathcal{R}^M$

$$\mathbf{y}_i = \mathbf{g}(x_i) + \epsilon_i$$

for $i = 1, \ldots, n$ with error assumptions $\mathcal{E}(\epsilon_i) = 0$ and $\mathcal{E}(\epsilon_i \epsilon_i^T) = \delta_{ij}\Sigma_i$, where the Σ_i are known symmetric and positive definite error covariances. The smooth vector function $\mathbf{g}(x) = (g_1(x), \ldots, g_M(x))^T$ can be estimated by min-

imizing a penalized least squares criterion

$$\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{g}(x_i))^T \Sigma_i^{-1}(\mathbf{y}_i - \mathbf{g}(x_i)) + \sum_{j=1}^{M}\lambda_j \int_{a_j}^{b_j}[g_j''(t)]^2 dt. \qquad (1.30)$$

The roughness penalty term penalizes for lack of smoothness in the component functions. Each component demands independent smoothing, not necessarily of the same degree (measured on arbitrary scales). Hence M fixed smoothing parameters $\lambda_j > 0$ are required. We refer to Fessler (1991, p. 855) for the choice of the λ_j in practice, which is non-trivial.

In the special case of a covariance matrix $\mathbf{\Sigma} = I$ the minimization problem of equation (1.30) is reduced to the unweighted version of an univariate (scalar) spline evaluation.

The solution $\hat{\mathbf{g}}$ can be obtained similarly to that of univariate cubic smoothing splines (see Section 1.1). For ordered design values $x_1 < x_2 < \ldots < x_n$ and (block) vectors $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$,

$$\mathbf{g} = (f_1(x_1), \ldots, f_M(x_1), \ldots, f_1(x_n), \ldots, f_M(x_n))^T,$$

and $\mathbf{\Sigma} = diag(\Sigma_1, \ldots, \Sigma_n)$ the penalized least squares criterion in (1.30) is equivalent to

$$(\mathbf{y} - \mathbf{g})^T \mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{g}) + \mathbf{g}^T K \mathbf{g}$$

for a matrix K of the same structure (not depending on $\mathbf{y}$) as in Section **??**. To be correct, there is one difference: here the smoothing parameters λ are part of K. The quadratic in $\mathbf{g}$ is minimized when

$$\hat{\mathbf{g}} = S(\lambda)\mathbf{y}$$

with the smoother matrix

$$S(\lambda) = (I + \mathbf{\Sigma}K)^{-1}.$$

As a result the vector spline is a linear smoother. In combination with VGAMs all the findings for additive models due to Buja *et al.* (1989) hold.

The estimation procedure for VGAMs can be developed in the same way as was demonstrated in the section about Fisher and local scoring for GAMs. The starting point is again a penalized likelihood criterion. Let us consider solely the kth covariate. Then the minimization problem is

$$\sum_{i=1}^{n}(\mathbf{z}_i - \mathbf{g}_k(x_{ik}))^T W_i(\mathbf{z}_i - \mathbf{g}_k(x_{ik})) + \sum_{j=1}^{M}\lambda_{(j)k}\int_a^b g_{(j)k}''(t)^2 dt, \qquad (1.31)$$

where the M-dimensional vector $\mathbf{z}_i$ represents the adjusted dependent variable (Yee and Wild, 1996, p. 483ff, give a detailed study for the identity link) subject to smoothing, $\mathbf{g}_k(x_{ik}) = (g_{(1)k}(x_{ik}), \ldots, g_{(M)k}(x_{ik}))^T$, and a $M \times M$ matrix W_i with elements

$$(w_i)_{jk} = -(\partial^2 l_i)/(\partial\eta_j \partial\eta_k),$$

where l_i is the likelihood of the ith measurement unit. The solution of (1.31) is a vector spline with $\Sigma_i^{-1} = W_i$. For the regression problem with d predictors we have to apply the backfitting or the iterative projection algorithm to $\mathbf{z}_i$ with vector spline smoothing. The obtained results are for the vector additive model

$$\mathcal{E}(\mathbf{y}_i) = \mathbf{g}_0 + \sum_{j=1}^{d} \mathbf{g}_j(x_{ij}),$$

which is fitted to the vector response $\mathbf{y}$. The backfitting algorithm adapted for vector smoothing is given in Yee and Wild (1996, p. 485). The underlying large linear (block matrix) equation system would in fact be better solved by the relaxed iterative projection algorithm (Schimek, 1996). VGAMs are an application where ill-posed estimation problems, depending primarily on the features of the covariance structure and the vector smoothing operator, are likely. The slow convergence of backfitting iterations in such applications is indicative of these problems.

It should be pointed out that "generalized" in VGAMs is not used in the sense of directly evaluating arbitrary link functions G (within the exponential family distribution framework). That is why they do not have to develop an alternative to Fisher scoring (see the previous section on GAMs for comparison). A generalization, apart from the important concept of vector variables, can be seen in the fact that Yee and Wild (1996) introduced constraints with respect to the covariates. They not only allow for different sets of predictors to be used for each predictor function but also constraints on how they act (for details see Yee and Wild, 1996, p. 486ff).

Several types of VGAM models have been implemented in the program VGAM, an S-Plus/ANSI-C implementation written by Thomas Yee. The software is available at `http://www.stat.auckland.ac.nz/yee`.

Important models that can be fitted with the VGAM software are the vector additive model (including seemingly unrelated regression), the proportional odds model, bivariate logistic models, and multinomial logit models.

1.4 ALTERNATING CONDITIONAL EXPECTATIONS, ADDITIVITY AND VARIANCE STABILIZATION

There are two other methods that we wish to discuss in this paper, the Alternating Conditional Expectation algorithm (ACE; Breiman and Friedman, 1985) and the Additivity and Variance Stabilization algorithm (AVAS; Tibshirani, 1988). Although these methods are based on a different working model than (generalized) additive models, the estimation process is very similar. Specifically, the working model for these two algorithms is

$$\theta(Y) = \alpha + \sum_{j=1}^{d} \psi_j(X_j) + e, \tag{1.32}$$

where e has mean zero and is independent of the X_js and $\theta(\cdot)$, $\psi_j(\cdot)$, $j = 1, \ldots, d$ are unknown (smooth) functions with $\mathcal{E}[\psi_j(X_j)] = 0$ for identifiability reasons.

Note that now we assume that the conditional expectation of a transformation of the Y-variable is given by an additive model of the X-variables. By way of contrast, in a GAM, see Section 1.3, we assume that a transformation of the conditional expectation of Y is given by an additive approximation of the X-variables. Hence, models of the form (1.32) are also know as *transformation models* or *transform both sides models* (TBS; see, among others, Carroll and Ruppert, 1988; Nychka and Ruppert, 1995).

1.4.1 Alternating Conditional Expectations

We assumed in (1.32) that $\mathcal{E}[\psi_j(X_j)] = 0$ for $j = 1, \ldots, d$. From this it follows that $\mathcal{E}[\theta(Y)] = \alpha$ and hence, without loss of generality, we can incorporate the constant α into the function $\theta(\cdot)$ and assume also that $\mathcal{E}[\theta(Y)] = 0$. Our working model is now

$$\theta(Y) = \sum_{j=1}^{d} \psi_j(X_j) + e.$$

The ACE algorithm estimates $\theta(\cdot)$ and $\psi_j(\cdot)$, $j = 1, \ldots, d$ by minimizing

$$\mathcal{E}\left[\theta(Y) - \sum_{j=1}^{d} \psi_j(X_j)\right]^2.$$

Obviously, this criterion is trivially minimized by choosing $\theta(\cdot) \equiv \psi_j(\cdot) \equiv 0$. Hence, a further normalization is needed, e.g. $\mathcal{V}[\theta(Y)] = 1$. Note, that if (1.32) holds we have similarly to the additive model in (1.2)

$$\mathcal{E}\left[\theta(Y) - \sum_{j \neq k}^{d} \psi_j(X_j) \mid X_k\right] = \psi_k(X_k)$$

and

$$\mathcal{E}\left[\sum_{j=1}^{d} \psi_j(X_j) \mid Y\right] = \theta(Y).$$

This suggests to estimate $\theta(\cdot)$ and $\psi_j(\cdot)$, $j = 1, \ldots, d$ by calculating alternatively these conditional expectations. Hence, using a univariate scatterplot smoother to approximate these conditional expectations leads to the basic ACE algorithm.

ACE Algorithm

---

1. *Initialize:* Set $\hat{\theta}(\cdot) = \left(\cdot - \bar{Y}\right)/\hat{\sigma}_Y$ where $\overline{Y} = n^{-1}\sum Y_i$ and $\hat{\sigma}_Y = n^{-1}\sum (Y_i - \bar{Y})^2$

2. *Find new transformations of Xs:* Fit an additive model to $\hat{\theta}(\cdot)$ to obtain new estimates $\hat{\psi}_j(\cdot)$, $j = 1, \ldots, d$ (see page iv)

3. *Find new transformation of Y:* Obtain a new estimate $\hat{\theta}(\cdot)$ by smoothing $\sum_j \hat{\psi}(X_j)$ against Y and standardize such that $\sum_i \hat{\theta}(Y_i) = 0$ and $\sum_i \hat{\theta}(Y_i)^2 = 1$

4. Alternate between step 2 and 3 until convergence is reached

It can be shown (see Breiman and Friedman, 1985) that the solutions of the ACE algorithm are closely related to the solutions of minimizing the correlation between $\theta(Y)$ and $\sum_j \psi_j(X_j)$ under the condition that $\mathcal{V}[\theta(Y)] = \mathcal{V}[\sum_j \psi_j(X_j)]$. This implies that ACE is more suitable as a correlation tool than a regression tool. Indeed, if viewed as a regression tool ACE has several disturbing features (see the discussion of Breiman's and Friedman's article or Hastie and Tibshirani, 1990, p. 184ff). One of these features is that even if (1.32) holds, the ACE algorithm may not reconstruct the functions $\theta(\cdot)$ and $\psi(\cdot)$ since the optimal transformations (that minimize the above criterion) depend on the joint distribution of the Xs and e. In the next section we shall describe a modification that tries to overcome these anomalies.

On the other hand, the original FORTRAN77 implementation of the ACE algorithm of Breiman and Friedman (available from the archive `lib.stat.cmu.edu`) allows the user to specify that some of the transformations should have certain features like, e.g. being monotone or linear (the same is true for the `ace` function in S-Plus). Indeed, by restricting $\theta(\cdot)$ to be linear, the ACE algorithm can be used as an effective exploratory tool (see Raftery and Richardon, 1996, and the references therein).

1.4.2 Additivity and Variance Stabilization

Tibshirani (1988) proposed a modification of the ACE algorithm to make it more suitable as a regression tool. Essentially he proposed that instead of calculating alternatively conditional expectations to identify the transformations $\theta(\cdot)$ and $\psi_j(\cdot)$, $j = 1, \ldots, d$, a variance stabilizing transformation should be used to estimate $\theta(\cdot)$. Specifically, we try find transformation such that

$$\mathcal{E}[\theta(Y) \mid X_1, \ldots, X_d] = \sum_{j=1}^{d} \psi_j(X_j)$$

and

$$\mathcal{V}\left[\theta(Y) \mid \sum_{j=1}^{d} \psi_j(X_j)\right] = c,$$

where c is an arbitrary constant. Here, we assume additionally that $\theta(\cdot)$ is strictly monotone and without loss of generality we may assume that $\theta(\cdot)$ is strictly increasing.

Now, if a random variable Z has mean μ and variance $\mathcal{V}(\mu)$ then the variance stabilizing transformation for Z is given by (see, Serfling, 1980, p. 120f; Sen and Singer, 1993, p. 139f):

$$h(t) = \int_0^t \frac{1}{\sqrt{\mathcal{V}(u)}} du.$$

That is, $h(Z)$ has approximately constant variance. The proposed algorithm is now as follows:

AVAS Algorithm

---

1. *Initialize:* Set $\hat{\theta}(\cdot) = \left(\cdot - \bar{Y}\right)/\hat{\sigma}_Y$ where $\bar{Y} = n^{-1}\sum Y_i$ and $\hat{\sigma}_Y = n^{-1}\sum (Y_i - \bar{Y})^2$.

2. *Find new transformations of Xs:* Fit an additive model to $\hat{\theta}(\cdot)$ to obtain new functions $\hat{\psi}_j(\cdot)$, $j = 1, \ldots, d$ (see page iv).

3. *Find new transformation of Y:* Set $m(\mathbf{X}) = \sum_{j=1}^{d} \psi_j(X_j)$ and compute the variance function $V(u) = \mathcal{V}[\theta(Y) \mid m(\mathbf{X}) = u]$. Then calculate the variance stabilizing transformation

$$h(t) = \int_0^t \frac{1}{\sqrt{V(u)}} du.$$

   and define the new $\hat{\theta}(\cdot)$ as the transformation $h(\hat{\theta}(\cdot))$ of the old estimate for $\theta(\cdot)$. Finally, renormalize $\hat{\theta}(\cdot)$ such that $\mathcal{E}[\theta(Y) = 0]$ and $\mathcal{V}[\theta(Y)] = 1$.

4. Alternate step 2 and 3 until convergence is reached.

When implementing this algorithm, the variance stabilizing transformation has to be calculated by numerical quadrature and $V(u)$ is calculated via an appropriate smoothing operation. FORTRAN77 and RATFOR code implementing this algorithm was submitted by Robert Tibshirani to the archive at `lib.stat.cmu.edu` and is available from there. S-Plus has an implementation of this algorithm too, called `avas`.

Banks *et al.* (1995) compare several methods for (high-dimensional) nonparametric regression including ACE and AVAS. Their results seem to indicate that both methods perform similarly and that ACE is often slightly better in mean integrated squared error sense. This is somewhat surprising since AVAS was designed to be a regression tool whereas ACE is rather a correlation tool. Although some theoretical results regarding AVAS are available, several questions remain unsolved, e.g. its global convergence has not been established yet.

1.5 SMOOTHING PARAMETER AND BANDWIDTH DETERMINATION FOR GAMS

All the estimation methods for additive models and GAMs discussed so far solely work when smoothing parameter or bandwidth values are chosen for each dimension beforehand. Although univariate in nature the choice of the degree of smoothing remains a multivariate problem. As a result of that, the usual generalized cross-validation criterion cannot be applied any more. However, a data-driven choice of the degree of smoothing, i.e. for $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_d)$ based on generalized cross-validation is possible in principle. For additive models the criterion is

$$GCV_{add}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{(y_i - \hat{\mu}_i)/\hat{\sigma}_i}{1 - tr(R_\lambda)/n} \right\}^2 ,$$

where R_λ is the smoother (hat) matrix that generates the additive predictor $\hat{\eta} = R_\lambda \tilde{\mathbf{y}}$ in the last iteration step. R_λ can be interpreted as a weighted additive fit operator. However, optimization of the criterion would require efficient computation of $tr(R_\lambda)$ in each step of a multidimensional search algorithm. Further it would be very difficult to find a global minimum in this multidimensional setting. For the generalized case the problem gets even more complicated because the link function also plays a role in the search for the optimal degree of smoothing. Taking into account popular link functions, Burman (1990) made a proposal which works for B-splines with equispaced knots. His proposal addresses the question of choosing the correct number of knots using cross-validation as criterion. For a pure additive spline model there is an $O(n^3)$ algorithm available (Gu and Wahba, 1988). Using the marginal integration method instead of iterative procedures does not circumvent this messy and computational expensive situation. The standard choice are simple plug–in techniques for bandwidth selection (e.g. in Severance-Lossin and Sperlich, 1997). There is also a proposal for a plug–in technique for the backfitting procedure when local linear fitting is applied (Opsomer and Ruppert, 1998). All these proposals are limited because there is neither theoretical justification, when the predictors are correlated, nor for dimensions $d > 2$.

We know from univariate function fitting that the degrees of freedom *df* are simply $df = tr(S)$, where S is a linear smoother matrix. As a matter of fact *df* amounts to be the sum of eigenvalues of S, indicating the extent of smoothing. Moreover, $tr(S)$ is easy to compute because only the diagonal elements of S are required. For certain linear smoothers there are $O(n)$ algorithms (e.g. Hutchinson and de Hoog, 1985, for smoothing splines).

Just like in univariate function fitting, the overall degrees of freedom are $df = tr(R_\lambda)$ in the additive model (Hastie and Tibshirani, 1990, p. 128f). The problem is that $tr(R_\lambda)$ is not exactly the sum the individual R_{λ_j} for $j = 1, \ldots, d$. Buja *et al.* (1989, p. 496f.) provided some evidence that under the assumptions of not too small smoothing parameters λ_j and not too heavily

correlated predictors x_j, adding up the individual degrees of freedom forms an upper bound for the overall degree of freedom of R_λ. The degrees of freedom obtained in this way are sometimes called the equivalent degrees of freedom (Hastie, 1992, p. 251). Their prespecification is often used to determine the amount of smoothing in an additive model or GAM. Apart from the fact that a prior guess can be wrong, it is implicitly assumed that the necessary amount of smoothing in each dimension is the same. But this is not true in many instances.

A different approach is feasible based on an improved (correcting for the tendency to undersmooth) Akaike information criterion due to Hurvich, Simonoff and Tsai (1998). Their proposal is for univariate smoothing problems but the idea could be extended to multivariate situations (e.g. for the additive model). Its advantage is that unlike plug–in techniques there is less limitation with respect to the type of linear smoothers.

In summary we cannot help saying that smoothing parameter or bandwidth choice remains a week point in any kind of additive model discussed so far. It seems that additional research is required for the data–driven selection of the necessary amount of smoothing.

1.6 MODEL DIAGNOSTICS FOR GAMS

At convergence of an additive fit we can express $\hat{\mathbf{g}}_j$ as $R_{\lambda_j}\mathbf{y}$ for some matrix R_{λ_j} of dimension $n \times n$ because the $\hat{\mathbf{g}}_j$ results from a linear mapping applied to $\mathbf{y}$. Suppose errors are independent and identically distributed, then the covariance of the estimator is

$$\mathcal{C}(\hat{\mathbf{g}}_j) = R_{\lambda_j} R_{\lambda_j}^T \sigma^2,$$

where $\sigma^2 = \mathcal{V}(Y_i)$. If the system matrix A of the normal equations $A\mathbf{g} = B\mathbf{y}$ has singular values close to zero, i.e. highly correlated covariates, then $\mathcal{C}(\hat{\mathbf{g}}_j)$ will be large.

The direct computation of R_{λ_j} is very expensive. Hastie and Tibshirani (1990, p. 127f) propose an $O(kn^2)$ backfitting procedure, where $k = dMc$ with d the number of covariates, M the number of backfitting iterations, and c a linear smoother-specific constant.

Hastie and Tibshirani (1990, p. 128 and p. 156f) also discuss the construction of confidence bands for the estimator. Today it is believed that resampling techniques are the better choice.

As in GLMs the analysis of deviance plays an important role for inference in GAMs. The deviance or likelihood-statistic for a fitted model represented by $\hat{\eta}$ is defined by

$$D(y; \hat{\eta}) = 2 \left\{ l(\eta_{max}; y) - l(\hat{\eta}; y) \right\},$$

where η_{max} is the parameter value that maximizes the log-likelihood $l(\eta; y)$

over all η (the saturated model). The deviance replaces the residual sum of squares $RSS = \mathbf{y}^T(I - R_\lambda)^T(I - R_\lambda)\mathbf{y}$ used in the simple additive model.

For GLMs there is an asymptotic distribution theory. Consider two linear models, η_1 and η_2, with η_1 nested within η_2. If η_1 is correct and some regularity conditions are fulfilled, then $D(\eta_2;\eta_1) = D(y;\eta_1) - D(y;\eta_2)$ has an asymptotic χ^2-distribution with degrees of freedom equal to the difference in the dimensions of the models. The result is usually summarized in an analysis of deviance table. For GAMs the deviance remains a sensible means of model assessment. The problem is that it is not even asymptotically χ^2-distributed, although empirical evidence due to Hastie and Tibshirani (1990, p. 155f) supports the use of the χ^2-distribution.

Finally we consider the degrees of freedom for error df_{err}. In the additive model df_{err} is derived in terms of the expected value of the residual sum of squares RSS, and defined by

$$df_{err} = n - tr(2R_\lambda - R_\lambda R_\lambda^T).$$

The analogous quantity to RSS in the additive model is the deviance D in the GAM. Starting from an asymptotic approximation to the deviance

$$D(y;\eta) \approx (\tilde{\mathbf{y}} - \hat{\eta})^T \tilde{W} (\tilde{\mathbf{y}} - \hat{\eta}),$$

we end up with degrees of freedom for error

$$df_{err} = n - tr(2R_\lambda - R_\lambda^T \tilde{W} R_\lambda \tilde{W}^{-1})$$

for the GAM obtained from local scoring in (1.25). For the use of df_{err} see Hastie and Tibshirani (1990, p. 157f) and Hastie (1992, p. 302f).

In conclusion one must say that since the publication of Hastie's and Tibshirani's 1990 monograph, there have not been substantial new developments in the field of model critique for additive models or GAMs (for testing in nonparametric models see also Bowman and Azzalini, 1997, Chapter 5). The same is true for model selection. Current research emphasizes the application of resampling techniques.

Other most recent developments address the estimation of additive models via Markov chain Monte Carlo techniques. This is truly a change of paradigm away from solving the normal equation system or equivalent concepts. Wong and Kohn (1996) not only present a promising Bayesian estimation technique but can at the same time handle the tedious problem of smoothing parameter selection (for the special case of regression splines). This Bayesian approach also facilitates the estimation of diagnostic quantities.

REFERENCES

1. Ansley, C. F. and Kohn, R. (1994). Convergence of the backfitting algorithm for additive models. *Journal of the Australian Mathematical Society, A, 57*, 316–329.

2. Banks, D., Maxion, R. and Olszewski, R. (1995). Comparing methods for nonparametric regression. *ASA Proceedings of the Statistical Computing Section*, 136–141.

3. Bellman, R. E. (1961). *Adaptive control processes.* Princeton University Press, Princeton.

4. Bowman, A. W. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations.* Oxford University Press, Oxford.

5. Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc., 80*, 580–619.

6. Buja, A. (1990). Remarks on functional canonical variates, alternating least squares methods, and ACE. *Ann. Statist., 18*, 1032–1069.

7. Buja, A. (1996). What criterion for a power algorithm? In Rieder, H. (ed.) *Robust statistics, data analysis, and computer intensive methods.* Springer-Verlag, New York, 49–61.

8. Buja, A., Hastie, T. J. and Tibshirani, R.J. (1989), Linear smoothers and additive models. *Ann. Statist., 17*, 453–510.

9. Burman, B. (1990). Estimation of generalized additive models. *Journal of Multivariate Analysis, 32*, 230–255.

10. Carroll, R. J. and Ruppert, D. (1988). *Transformation and weighting in regression.* Chapman and Hall, London.

11. Chambers, J. M. and Hastie, T. J. (eds., 1992). *Statistical models in S.* Wadsworth & Brooks, Pacific Grove, CA.

12. Eubank, R. L. (1984). The hat matrix for smoothing splines. *Statist. and Prob. Letters, 2*, 9–14.

13. Fahrmeir, L. and Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models.* Springer-Verlag, New York.

14. Fan, J., Härdle, W. and Mammen, E. (1996). Direct estimation of additive and linear components for high dimensional data. *Technical Report 96.1, Department of Statistics*, Chinese University of Hong Kong.

15. Fessler, J. A. (1991). Nonparametric fixed-interval smoothing with vector splines. *IEEE Trans. Signal Proces., 39*, 852–859.

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

17. Gander, W. and Golub, G. H. (1989). Discussion of 'Buja, A., Hastie, T. and Tibshirani, R.: Linear smoothers and additive models.' *Ann. Statist., 17*, 529–532.

18. Golub, G. H. and van Loan, C. F. (1989) *Matrix computations.* John Hopkins University Press, Baltimore.

19. Green, P. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models. A roughness penalty approach.* Chapman and Hall, London.

20. Green, P. and Yandell, B. (1985). *Semi-parametric generalized linear models.* In Gilchrist, R., Francis, B. J. and Whittaker, J. (eds.). Generalized linear models. Springer-Verlag, Berlin, 44–55.

21. Gu, C. and Wahba, G. (1988). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *Technical Report 847*, University of Wisconsin, Madison.

22. Hastie, T. J. (1992). Generalized additive models. In Chambers, J. M. and Hastie, T. J. (eds.). *Statistical models in S.* Wadsworth & Brooks, Pacific Grove, CA, 249–307.

23. Hastie, T. J. and Tibshirani, R. J. (1986). Generalized additive models. *Statist. Scien., 1*, 297–318.

24. Hastie, T. J. and Tibshirani, R. J. (1987a). Generalized additive models: Some applications. *J. Amer. Statist. Assoc., 82*, 371–386.

25. Hastie, T. J. and Tibshirani, R. J. (1987b). Non-parametric logistic and proportional odds regression. *Appl. Statist., 36*, 260–276.

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

27. Härdle, W. and Chen, R. (1995). Nonparametric time series analysis, a selective review with examples. *Bulletin of the International Statistical Institute, Proceedings of the 50th Session, Book 1*, 375–394.

28. Härdle, W., Klinke, S. and Turlach, B. A. (eds., 1995). *XploRe: An interactive statistical computing environment.* Springer-Verlag, New York.

29. Härdle, W. and Korostelev, A. (1994). Search of significant variables in nonparametric additive regression. *Discussion Paper 42, SFB 373*, Humboldt University Berlin.

30. Hämmerlin, G. and Hoffmann, K.-H. (1995). *Numerical mathematics.* Springer-Verlag, New York.

31. Hengartner, N. (1996). Rate optimal estimation of additive regression via the integration method in presence of many covariates. *Statistics Preprint 96oct-1*, Yale University.

32. Hurvich, C. M., Simonoff, J. S. and Tsai, C.–L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc. B*, 60, 271–293.

33. Hutchinson, M. F. and de Hoog, F. R. (1985). Smoothing noisy data with splines functions. *Numer. Math., 47*, 99–106.

34. Linton, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika, 84*, 469–473.

35. Linton, O. B. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika, 82*, 93–100.

36. Linton, O. B. and Härdle, W. (1996). Estimation of additive regression models with known links. *Biometrika, 83*, 529–540.

37. Linton, O. B., Mammen, E. and Nielsen, J. P. (1997). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Cowles Foundation Discussion Paper No. 1160*, Yale University.

38. MathSoft (1997). S-PLUS 4. Guide to statistics. MathSoft, Inc., Seattle, Washington.

39. McCullagh, P. and Nelder, J. A. (1989, 2nd ed.). *Generalized linear models*. Chapman and Hall, London.

40. Neubauer, G. P. and Schimek, M. G. (1996). A note on cross-validation for smoothing splines. In Härdle, W. and Schimek, M. (eds.), *Statistical theory and computational aspects of smoothing*. Physica-Verlag, Heidelberg, 165–177.

41. Newey, W. K. (1994). Kernel estimation of partial means. *Econometric Theory, 10*, 233-253.

42. Nielsen, J. P. and Linton, O. B. (1998). An optimization interpretation of integration and back-fitting estimators for separable nonparametric models. *J. R. Statist. Soc. B, 60*, 217-222.

43. Nychka, D. and Ruppert, D. (1995). Nonparametric transformation for both sides of a regression model. *J. R. Statist. Soc. B, 57*, 519–532.

44. Opsomer, J. D. (1996). On the existence and asymptotic properties of backfitting estimators. *Preprint 96-12, Statistical Laboratory*, Iowa State University.

45. Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist., 25*, 186–211.

46. Opsomer, J. D. and Ruppert, D. (1998). A fully automatic bandwidth selector method for fitting additive models. *J. Amer. Statist. Assoc., 93*, 605–619.

47. Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables.* Academic Press, New York.

48. Raftery, A. E. and Richardon, S. (1996). Model selection for generalized linear models via GLIB: Application to nutrition and breast cancer. In Berry, D. A. and Stangl, D. K. (eds.) *Bayesian biostatistics.* Marcel Dekker, New York, 321–354.

49. Reinsch, C. (1967). Smoothing by spline functions. *Numerische Mathematik, 10*, 177-183.

50. Schimek, M. G. (1997). Non- and semiparametric alternatives to generalized linear models. *Computational Statistics, 12*, 173–191.

51. Schimek, M. G. (1996). An iterative projection algorithm and some simulation results. In Prat, A. (ed.) *COMPSTAT '96. Proceedings in Computational Statistics.* Physica-Verlag, Heidelberg, 453–458.

52. Schimek, M. G., Neubauer, G. and Stettner, H. (1994). Backfitting and related procedures for non-parametric smoothing regression: A comparative view. In Grossmann, W. and Dutter, R. (eds.) *COMPSTAT' 94 Proceedings in Computational Statistics.* Physica-Verlag, Heidelberg, 63–68.

53. Schimek, M. G., Stettner, H., Haberl, J. and Neubauer, G. P. (1993). Approaches for fitting additive models with cubic smoothing splines and associated problems. *Bulletin of the International Statistical Institute. Contributed Papers, 49th Session, Book 2*, 381–382.

54. Scott, D. W. (1992). *Multivariate density estimation. Theory, practice and visualization.* Wiley, New York.

55. Sen, P. K. and Singer, J. M. (1993). *Large sample methods in statistics: an introduction with applications.* Chapman and Hall, New York.

56. Serfling, R. J. (1980). *Approximation theorems of mathematical statistics.* Wiley, New York.

57. Severance-Lossin, E. and Sperlich, S. (1997). Estimation of derivatives for additive separable models. *Discussion Paper 30, SFB 373.* Humboldt University Berlin.

58. Sperlich, S., Linton, O. B. and Härdle, W. (1997). A simulation comparison between integration and backfitting methods of estimating separable nonparametric regression models. *Discussion Paper 66, SFB 373.* Humboldt University Berlin.

59. Sperlich, S., Tjøstheim, D. and Yang, L. (1998). Nonparametric estimation and testing of interaction in additive models. *Discussion Paper 14, SFB 373.* Humboldt University Berlin.

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

61. Stone, C. J. (1986). The dimension reduction principle for generalized additive models. *Ann. Statist., 14*, 590–606.

62. Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization, *J. Amer. Statist. Assoc., 83*, 394–405.

63. Tjøstheim, D. and Auestad, B. (1994). Nonparametric identification of nonlinear time series: Projections. *J. Amer. Statist. Assoc., 89*, 1398–1409.

64. Venables, W. N. and Ripley, B. D. (1997, second edition). *Modern applied statistics with S-Plus.* Springer–Verlag, New York.

65. Wong, C. and Kohn, R. (1996). A Bayesian approach to additive semiparametric regression. *Journal of Econometrics, 74*, 209–235.

66. Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models. *J. R. Statist. Soc. B, 58*, 481–493.