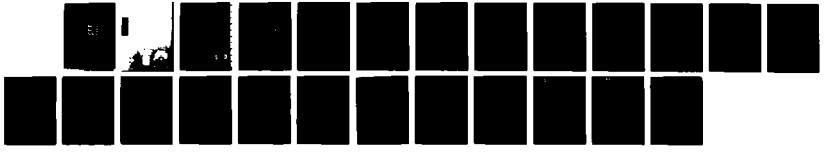
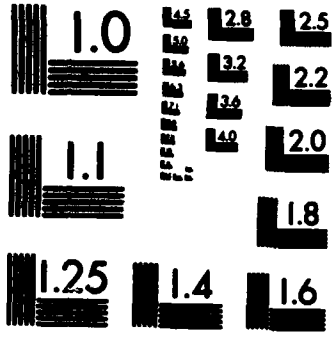
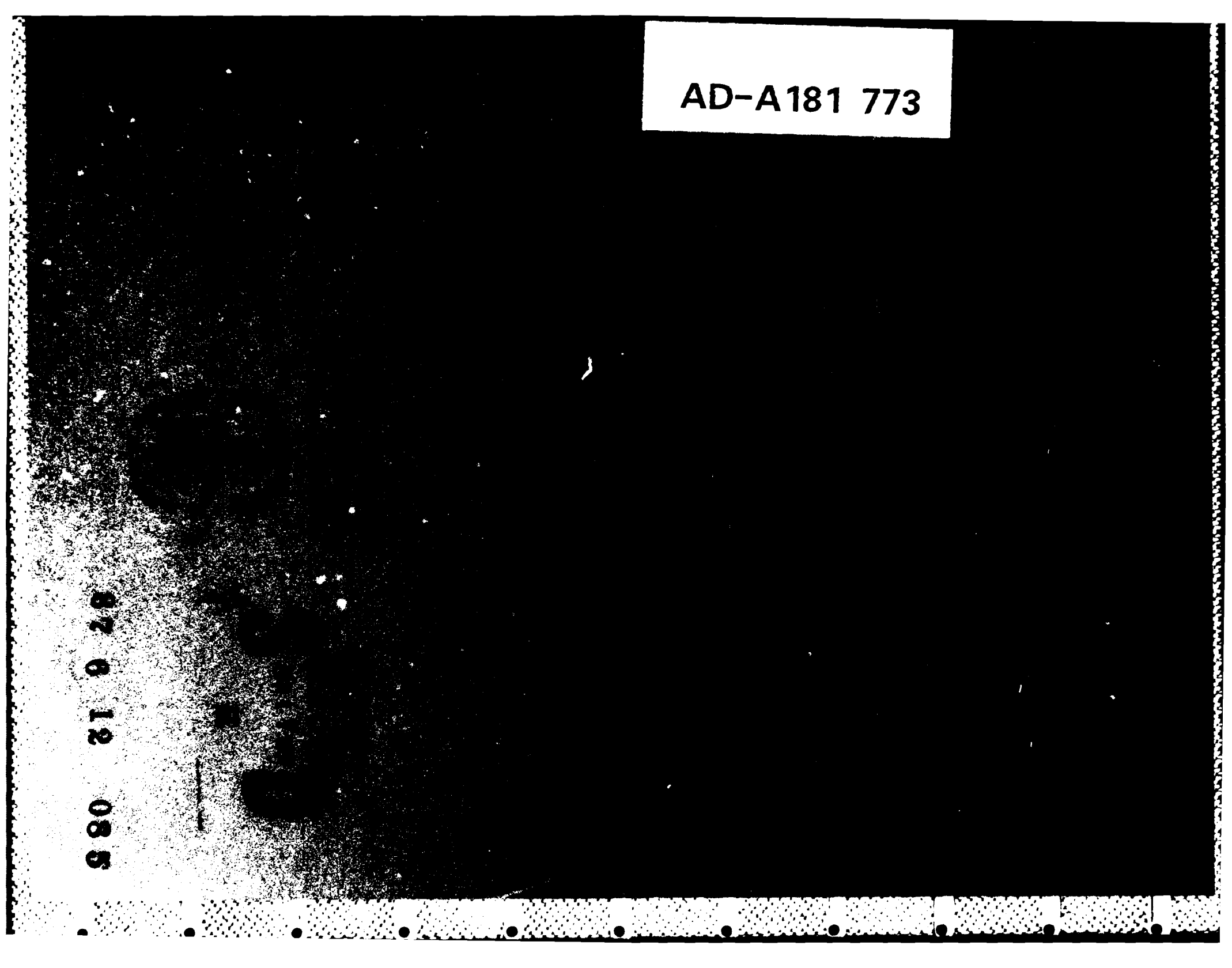
AD-A181 773 GENERALIZED ADDITIVE MODELS CUBIC SPLINES AND PENALIZED 1/1
LIKELIHOOD(U) STANFORD UNIV CA DEPT OF STATISTICS
T HASTIE ET AL. 22 MAY 87 TR-390 N00014-86-K-0156

UNCLASSIFIED F/G 12/3 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A


AD-A181 773

GENERALIZED ADDITIVE MODELS, CUBIC SPLINES AND PENALIZED LIKELIHOOD

BY

TREVOR HASTIE and ROBERT TIBSHIRANI

TECHNICAL REPORT NO. 390
MAY 22, 1987

Prepared Under Contract
N00014-86-K-0156 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Generalized Additive Models, Cubic Splines and Penalized Likelihood

Trevor Hastie

and

Robert Tibshirani

SUMMARY

Generalized additive models (Hastie and Tibshirani, 1986 Statistical Science) extend the class of generalized linear models by allowing an arbitrary smooth function for any or all of the covariates. The functions are estimated by the local scoring procedure, using a smoother as a building block in an iterative algorithm. In this paper we utilize a cubic spline smoother in the algorithm and show how the resultant procedure can be viewed as a method for automatically smoothing a suitably defined "partial residual", and more formally, a method for maximizing a penalized likelihood. The authors also examine convergence of the inner ("backfitting") loop in this case and illustrate these ideas with some binary response data.

Key words: Generalized additive model, spline smoothing, non-parametric regression, partial residual, penalized likelihood.

1. Introduction.

This paper describes a technique for non-parametrically estimating covariate effects in any generalized linear model. In order to motivate the technique we'll consider an example. Risch, Weiss, Lyon, Daling and Liff (1983) analyzed a case-control study of ovarian cancer. Their data consists of 987 observations on the incidence (y) of ovarian cancer (1=case 0=control), and 17 covariates. For illustration here we will focus on two covariates: ovulatory age (OA),

an estimate of the number of years that a woman has ovulated in her life and age at end of ovulatory life (AEO). We shall fit a number of logistic models to these data, with $log[P(Y = 1\,|OA, AEO)/(1-P(Y=1\,|OA, AEO))] = \alpha + f_1(OA) + f_2(AEO)$ for some functions $f_1(OA)$ and $f_2(AEO)$. A linear logistic fit (that is $f_1(OA) = \hat{\beta}_1 OA, f_2(AEO) = \hat{\beta}_2 AEO$) gave $\hat{\alpha} = xx$, $\hat{\beta}_1 = .09$ and $\hat{\beta}_2 = .-07$ both effects being highly significant. Denote the fitted probabilities by p. In order to investigate the adequacy of the linear forms for $f_1(.)$ and $f_2(.)$, Landwehr, Pregibon and Shoemaker (1984) suggested smoothing the partial residual

$$Z_1 = \hat{\alpha} + \hat{\beta}_1 OA + \frac{y - \hat{p}}{\hat{p}(1-\hat{p})} \tag{1}$$

against OA, and similarly for $f_2(.)$. This is justified by the fact that if in reality $log[P(Y = 1\,|OA, AEO)/(1 - P(Y = 1\,|OA, AEO))] = \alpha + \beta_2 AEO + f_1(OA)$, and $f_1[OA)$ is orthogonal to AEO, then $E(Z\,|OA) \approx f_1(OA)$ and likewise for $f_2(AEO)$. A smooth of Z_1 versus OA is shown in Figure 1 and shows some non-linearity. This smooth and all smooths is this paper were computed using a cubic spline smoother, described in the next section.

A shortcoming of this technique is that $E(Z_1\,|OA) \approx f_1(OA)$ only if a linear function for AEO is appropriate, and similarly for $f_2(AEO)$. What is needed is simultaneous estimation of $f_1(.)$ and $f_2(.)$. This can be achieved as follows. We start with initial guesses $\hat{f}_1(.)$ and $\hat{f}_2(.)$ (for example $\hat{\beta}_1 OA$ and $\hat{\beta}_2 AEO$) and corresponding fitted value $\hat{p}$, then construct the quantity

$$Z = \hat{\alpha} + \hat{f}_1(OA) + \hat{f}_2(AEO) + \frac{y - \hat{p}}{\hat{p}(1-\hat{p})} \tag{2}$$

Fixing Z, we alternately smooth $Z - \hat{f}_2(AEO)$ on OA, $Z - \hat{f}_1(OA)$ on AEO, obtaining new $\hat{f}_1(.)$ and $\hat{f}_2(.)$'s, until the smooths don't change much. As a further refinement we can update Z, replacing the old $\hat{f}_i(.)$'s with the new ones (and updating $\hat{p}$), then repeat the first step, and so on until convergence. Note that if this procedure converges, the smooth of $Z - \hat{f}_2(AEO)$ on OA is $\hat{f}_1(OA)$ and similarly for $\hat{f}_2(AEO)$, which is analagous to the use of Z_1 in (1). The resulting function for OA is shown in Figure 2. Notice that the smooth looks quite different from Figure 1– this is due to the fact that the smooth for AEO (not shown) is also non-linear. These data are analyzed more fully in Section 4.

This idea is the heart of the local scoring algorithm, described in this paper. The local

scoring technique is used to fit what we call generalized additive models. In the exponential family, these take the form $g(\mu) = \alpha + \sum_1^p f_i(x_i)$, where the response Y has $EY = \mu$, the x_i's are covariates, and the $f_i(.)$'s are unspecified smooth functions. These models are an extension of generalized linear models (Nelder and Wedderburn 1972) which assume $g(\mu) = \alpha + \sum_1^p x_i\beta_i$. (In generalized linear model terminology, $g(\mu)$ is called the " link function"). The local scoring algorithm estimates the $f_i(.)$'s by repeated smoothing of a suitably defined partial residual. The resulting estimates can be used to suggest transformations of the x's or as a predictive model. One can allow a smooth estimate for all of the covariates or force a linear fit for some of them. Such a semi-parametric model would naturally arise if some of the covariates were categorical in nature but would also be useful if, for reasons specific to the data at hand, a linear fit was desirable for certain specified covariates.

This paper presents an expository view of generalized additive models and the local scoring algorithm, with an emphasis on the use of splines for smoothing. In section 2 we describe the local scoring algorithm for generalized additive models in the exponential family, showing its close relation to the Fisher scoring technique for generalized linear models. Section 3 briefly describes the cubic spline smoother. The ovarian cancer data is analyzed more fully in Section 4, illustrating the use of degrees of freedom and confidence bands for model assessment. In Section 5 we give a justification for the local scoring procedure based on the notion of penalized likelihood. Finally, Section 6 discusses extension of the methods and their relation to other work in the literature.

2. The Local Scoring Method in the Exponential family.

In order to define the local scoring algorithm, it is convenient to first describe the process of repeated smoothing, called backfitting (Friedman and Stuetzle, 1981). Suppose we have data $((x_1, x_{11}, \ldots x_{1p}), \ldots (x_n, x_{n1} \ldots x_{np}))$ and we wish to fit a model of the form $E(Z \mid X_1, \ldots X_p) = \alpha + \sum_1^p f_j(X_j)$. Let $S(Z \mid x)$ represent some estimate of $E(Z \mid x)$, for example a cubic spline smoother. We can estimate the $f_j(.)$'s by repeated smoothing:

Backfitting Algorithm

Initialization: $\{\hat{f}_j(.) = 0, j = 1, \ldots p\}$, $\hat{\alpha} = average(z_i)$.

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

$r_i = z_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ki}), i = 1, \ldots n$

$\hat{f}_j(x_{ji}) = S(r_i \mid x_{ji}), i = 1, \ldots n$

Until $RSS = \sum_{i=1}^{n}(z_i - \hat{\alpha} - \sum \hat{f}_k(x_{ki}))^2$ converges.

Some theoretical results are available for the backfitting algorithm. Breiman and Friedman (1985) show that the theoretical version of backfitting (i.e. with conditional expectations instead of smoothers) converges to the best (in L^2 norm) additive approximation to Z in terms of $X_1, X_2, \ldots X_p$. Furthermore, if $S(Z \mid x)$ represents a least squares fit of Z on x, then the backfitting algorithm can be shown to converge to the least squares fit of Z on $x_1, x_2, \ldots x_p$. In the Appendix, we prove convergence of backfitting with two cubic spline smoothers and discuss the general ($p > 2$) case.

Now consider the class of generalized additive models in the exponential family. We assume that for a fixed value of the scale parameter ϕ, Y has a density $h(y, \mu)$ in the exponential family, with $EY = \mu$, $Var(Y) = V\phi$ and $\eta = g(\mu) = \alpha + \sum_1^p f(X_j)$. The deviance of a model μ is defined by $dev(y, \mu) = 2[\sum \log h(y_i, y_i) - \sum \log h(y_i, \mu_i)]$. The local scoring algorithm uses iterated backfitting of a partial residual to estimate $f_1(.), \ldots f_p(.)$.

The Local Scoring Algorithm

Initialization $\hat{f}_j^0(.) = 0, j = 1, \ldots p$, $\hat{\alpha} = g(average(y))$.

Loop over outer iteration counter m

$\eta_i^m = \hat{\alpha} + \sum_1^p \hat{f}_j^m(x_{ji}), \mu_i^m = g^{-1}(\eta_i^m)$

$z_i = \eta_i^m + (y_i - \mu_i^m)(d\eta_i^m / d\mu_i^m), w_i = (d\mu_i^m / d\mu_i^m)^2 V^{-1}$

Obtain $\hat{f}_j^{(m+1)}, j = 1, \ldots p$ by applying the backfitting algorithm to z_i with weights w_i.

Until $dev(y, \mu)$ fails to decrease

Notice that there are really two nested loops in the algorithm. In the inner (backfitting) loop s is held fixed and the f_j are re-estimated, while in the outer loop, η, μ, z and w are updated. The weights are appropriate because z has variance proportional to $(d\eta/d\mu)^2V$. The quantity z is a general form of Landwehr, Pregibon and Shoemaker's partial residual and is called the adjusted dependent variable or working variate in the generalized linear model literature. In fact, if the smooths in the backfitting algorithm are global least squares fits, then the local scoring algorithm reduces to adjusted dependent variable regression, the GLIM package implementation of maximum likelihood estimation by Fisher scoring (see Nelder and Wedderburn 1972). This would of course be an inefficient implementation because backfitting is an inefficient way to find the least squares fit of Z on $X_1, X_2, \ldots X_p$. Note also that if h is the normal density then $z_i = y_i$ then the algorithm consists only of the inner loop, a backfit of z on $x_1, x_2, \ldots x_p$.

As in generalized linear modelling, one can define degrees of freedom for generalized additive models and obtain an estimate of it. Briefly, the degrees of freedom is defined as the expected decrease in the deviance and is computed from the trace of an appropriate matrix (related to the smoother). The theoretical support for this calculation, however, is not substantial and it is meant to be used only as a rule of thumb. In a similar fashion, pointwise confidence bands can be estimated for the functions. More details on degrees of freedom and confidence bands may be found in Hastie and Tibshirani (1986), and Hastie and Tibshirani (1985a).

3. The Cubic spline smoother.

The local scoring algorithm requires an estimate of a conditional expectation in the backfitting loop. In this paper, we will discuss the use of cubic spline smoothers, which we will review briefly here. Note however that any other reasonable estimate of conditional expectation could be used. In Hastie and Tibshirani (1986) we used running lines smoothers; other candidate smoothers include a kernel smoother (see e.g. Cleveland 1979) , or a smoother such as McDonald and Owen's (1984) " split linear smoother" designed to reproduce discontinuities. One could also use different smoothers for each covariate— for example a " wrap around"

smoother would be appropriate for a periodic variable like month of the year.

Given data $(x_1, z_1), \ldots (x_n, z_n)$, consider the following minimization problem. Find $h(x)$ to minimize

$$\sum_1^n (z_i - h(x_i))^2 + \lambda \int_{-\infty}^{+\infty} [h''(s)]^2 ds \tag{3}$$

where λ is a fixed tuning constant. As shown in Reinsch (1967) (Silverman 1985) the solution $\hat{h}(x)$ is a cubic spline with knots at some of the x_i's, that is, a piecewise cubic function with pieces joined at the x_i's. The parameter λ trades off variance and bias of the solution. When $\lambda = 0$, the solution is any interpolating function, while if $\lambda = +\infty$, the solution is the least squares line. If we consider the value of $h(x)$ only at $x_1, x_2, \ldots x_n$, an equivalent form of this problem is the following. Let $h = (h(x_1), \ldots h(x_n))$, $z = (z_1, \ldots z_n)$ and K be an $n \times n$ "penalty" matrix constructed as follows. Let $q_i = x_{i+1} - x_i, i = 1, \ldots n-1$, Δ be a tri-diagonal $(n-2) \times n$ matrix with $\Delta_{ii} = 1/h_i$, $\Delta_{i,i+1} = -(1/h_i + 1/h_{i+1}), \Delta_{i,i+2} = 1/h_{i+1}$, and let W be a symmetric tri-diagonal matrix of order $n-2$ with $W_{i-1,i} = W_{i,i-1} = h_i/6, W_{ii} = (q_i + q_{i+1})/3$. Finally, let $K = \Delta^t W^{-1} \Delta$. Then the minimizer of (3) also minimizes

$$(z - h)^t (z - h) + \lambda h^t K h \tag{4}$$

Furthermore, one can express the solution $\hat{h}$ as Sz where $S = (I + \lambda K)^{-1}$, I being the $n \times n$ identity matrix. This representation is useful analytically but not very stable computationally. For the latter, an algorithm based on Cholesky decomposition is preferred (see e.g. Yandell 1986).

The parameter λ can either be chosen on subjective grounds, or by cross-validation, generalized cross-validation (Craven and Wahba 1979) or asymptotic generalized cross-validation (Silverman 1985). In "automatic" mode, the local scoring algorithm uses generalized cross-validation to pick λ each time a smooth is computed.

4. Analysis of the ovarian cancer data.

We now analyze the ovarian cancer data discussed in Section 1. Risch and co-workers analyzed a case control study of 987 women in Washington and Utah. They interviewed 284 women with ovarian cancer and 703 controls, and recorded the following covariates: number

of children, number of miscarriages, number of months of lactation, obesity (FAT), oral contraceptive usage and age at end of ovulatory period (AEO). The women were also frequency matched by age category and state of residence. An estimate of ovulatory age (OA) was constructed using all the variables except AEO and the main hypothesis was that OA would be related to the incidence of ovarian cancer.

Table 1 shows an analysis of deviance for a number of logistic models fit to these data (see Breslow and Day 1980 for more details on using logistic regression models in case control studies). All models in the table include dummy variable to account for the matching. The first 3 lines of the table indicate that OA is an extremely important factor, after adjusting for the matching variables and obesity. Figure 3 shows the smooth estimate for OA. The risk increases with ovulatory age but levels off at about 35 years. Comparison of lines 2 and 3 of the table confirm that there is a non-linear effect of OA. The upper and lower curves in Figure 3 represent 95% confidence bands mentioned in Section 2.

The middle section of Table 1 shows that even when an adjustment is made for AEO, OA is still extremely important. However, the adjustment for AEO removes the plateau behaviour of OA: the smooth (not shown) looks much like that in Figure 2. The smooth for AEO (Figure 4) is also non-linear. The downturn at about age 45 may be due to the fact that some women who stop ovulating at an early age do so because they have ovarian cancer. The last section of the table examines the effect of entering the remaining variables into the model. The deviance decrease (compared to the second line) is significant indicating that OA may not fully capture the effects of the other variables.

Note that simple interactions can be modelled by taking products of variables and treating the product as a new covariate. Alternatively, we can fit models to subgroups of the data. This might be useful here, for example, in examining a possible interaction between OA and FAT. A more extensive analysis of these data, with an emphasis on the medical aspects, will appear elsewhere.

5. Justification of local scoring through penalized likelihood.

In Hastie and Tibshirani (1986) we viewed the local scoring procedure as an empirical method for minimizing the expected log-likelihood of the data. When a linear smoother such as a cubic spline smoother is used in the algorithm an alternative motivation based on penalized likelihood can be derived.

Let $l(\theta)$ be the log-likelihood, where $\theta = \alpha + \sum_1^p f_j(x_{ji})$, and let $\hat{\alpha} = average(y)$. Let K_i be the $n \times n$ symmetric penalty matrix defined in Section 3, and $f_j = (f_j(x_{j1}), \ldots f_j(x_{jn})), j = 1, 2, \ldots p$ and consider the following problem. Find $f_1, \ldots f_p$ to maximize

$$l(\theta) - \frac{1}{2}\sum_1^p \lambda_i f_i^t K_i f_i \tag{5}$$

Letting $A = E(-d^2l/d\theta\theta^t)$, a diagonal matrix with diagonal elements a_i, we show that a Fisher scoring step is achieved by applying the backfitting algorithm to appropriate adjusted dependent variables. Rather straightforward calculations show that the Fisher scoring step to go from $f_1^{old}, f_2^{old}, \ldots, f_p^{old}$ to $f_1^{new}, f_2^{new}, \ldots, f_p^{new}$ is

$$\begin{pmatrix} A+\lambda_1 K_1 & A & \ldots & A \\ A & A+\lambda_2 K_2 & \ldots & A \\ \vdots & \vdots & \ddots & \vdots \\ A & A & \ldots & A+\lambda_p K_p \end{pmatrix} \begin{pmatrix} f_1^{new} - f_1^{old} \\ f_2^{new} - f_2^{old} \\ \vdots \\ f_p^{new} - f_p^{old} \end{pmatrix} = \begin{pmatrix} u - \lambda_1 K_1 f_1^{old} \\ u - \lambda_2 K_2 f_2^{old} \\ \vdots \\ u - \lambda_p K_p f_p^{old} \end{pmatrix} \tag{6}$$

where $u = \frac{dl}{d\theta}$. Carrying the f_j^{old} terms to the right hand side we get p equations $(A + \lambda_j K_j) f_j^{new} + A\sum_{i \neq j} f_i^{new} = Az, \quad j = 1, p$ where $z = \theta^{old} + A^{-1}u$. These can then be written as

$$\begin{pmatrix} f_1^{new} \\ f_2^{new} \\ \vdots \\ f_p^{new} \end{pmatrix} = \begin{pmatrix} S_1(z - \sum_{j \neq 1} f_j^{new}) \\ S_2(z - \sum_{j \neq 2} f_j^{new}) \\ \vdots \\ S_p(z - \sum_{j \neq p} f_j^{new}) \end{pmatrix} \tag{7}$$

where $S_j = (A + \lambda_j K_j)^{-1} A$. As noted by Green and Yandell (1985), S_j computes a weighted cubic spline smooth, with weights a_i. The backfitting algorithm is a iterative method for solving this system of linear equations. In fact, the backfitting step (7) is exactly a Gauss-

Seidel method solving a linear system. The linear system that it solves can be written as

$$\begin{pmatrix} I & S_1 & S_1 & \cdots & S_1 \\ S_2 & I & S_2 & \cdots & S_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_p & S_p & S_p & \cdots & I \end{pmatrix} \begin{pmatrix} f_1^{new} \\ f_2^{new} \\ \vdots \\ f_p^{new} \end{pmatrix} = \begin{pmatrix} S_1 z \\ S_2 z \\ \vdots \\ S_p z \end{pmatrix} \tag{8}$$

which is just (7) rearranged.

The preceding analysis holds for any penalty matrices K_i; each matrix determines a corresponding smoother by $S_i = (A + \lambda_i K)^{-1} A$. Conversely, given a smoother matrix S_i, the corresponding penalty matrix is given by $K_i = (1/\lambda_i) A(S_i^{-1} - I)$. Following Reinsch (1967), one can also show the the local scoring procedure maximizes $l(\theta) - (1/2)\sum_1^p \lambda_i \int_{-\infty}^{+\infty} [f''(s)]^2 ds$, analogous to the the cubic spline problem (3).

6. Discussion.

Local scoring for generalized additive models provides a flexible method for identifying non-linear covariate effects in a variety of modelling situations; notably the very situations in which it has become popular to use the generalized linear or GLIM models. The additive models can be used in a data analytic fashion to understand the effect of covariates, and to test hypothesis about effects. A more conservative approach is to allow the non-parametric functions to suggest parametric transformations, and then proceed with the usual linear analysis on the transformed variables. The local scoring idea is a very general one, and can be applied in any situation in which the criterion being optimized depends on one or more smooth functions (see Hastie and Tibshirani 1986 for details).

When cubic spline smoothers are utilized, local scoring is closely related to recent work by O'Sullivan, Yandell and Raynor (1986), Green (1985) and Green and Yandell (1985). In the first paper, a multidimensional "thin plate spline" is used in a generalized linear model setting. The last two papers describe more general procedures, with an emphasis on semi-parametric models, i.e. models involving a linear component and a single smooth component. In fact, Green derives backfitting equations analogous to (7) in a special case of a model with one smooth. In this special case, the only difference between Green's method and local scoring

with cubic splines is the method for choosing λ. Green uses a quadratic approximation to the generalized cross-validation score, of the form $deviance/\nu^2$, where ν is the estimated degrees of freedom of the model. Since the linearization step in the local scoring is tantamount to a quadratic approximation to the deviance, the two methods are not likely to differ by much. However, the degrees of freedom ν is quite difficult to compute when more than one smooth is present in the model. More generally, the generalized additive models framework (with local scoring) differs from these approaches in that a) it emphasizes additive models b) it can incorporate multiple smooths through the use of backfitting, and c) it can incorporate non-linear smoothers.

A certain amount of theory already exists for these models, notably uniqueness of the best additive approximation at the model and rates of convergence for parametric sub-models(Stone 1985). More theoretical work is needed to refine the degrees of freedom and confidence band computations as well as to understand the effects of collinearity.

SOFTWARE

All the computations in this paper were performed by the Fortran program GAIM (Generalized Additive Interactive Modelling), a package available from the authors upon request. An IBM PC version of GAIM is also available.

ACKNOWLEDGEMENTS

We would like to thank Dr. N. Weiss for making available the ovarian cancer data and Dr. H. Risch for help with the analysis and valuable comments.

Appendix

Convergence of backfitting with two cubic spline smoothers

The cubic spline smoother is constant preserving, i.e. $S\mathbf{1} = \mathbf{1}$. For definitiveness, then, we work with the centered version of each smoother

$$S_j^* = (I - \frac{\mathbf{1}\mathbf{1}^t}{n})S_j \tag{9}$$

In addition we assume that the components of each x_i are in the same order as the components of x. Hence S_i^* will really denote $P_i^{-1}S_i^*P_i$ where P_i is the permutation matrix that sorts x in the order of x_i. (This is fine since $P_i^{-1}S_i^*P_i$ has the same principal value decomposition as S_i^*). Noting that $S_i^*x1 = 0$, the backfitting process consists simply of the alternating steps

$$\begin{aligned} f_1 &= S_1^*(x - f_2) \\ \hat{f}_2 &= S_2^*(x - \hat{f}_1) \end{aligned} \tag{10}$$

Starting with initial values $\hat{f}_1^0$ and $\hat{f}_2^0$, if $\hat{f}_1^m$ and $\hat{f}_2^m$ denote the estimates at the mth stage of the backfitting algorithm, then it is straightforward to show that

$$\begin{aligned} \hat{f}_1^m &= x - \sum_{j=0}^{m-1}(S_1^*S_2^*)^j(I - S_1^*)x \\ \hat{f}_2^m &= x - \sum_{j=0}^{m-1}(S_2^*S_1^*)^j(I - S_2^*)x - (S_2^*S_1^*)^m\hat{f}_2^0 \end{aligned} \tag{11}$$

Let $\|C\| = sup_a[a^tC^tCa]/a^ta$, the natural norm of the matrix C. Then $\hat{f}_1^m$ and $\hat{f}_2^m$ will converge if $\|S_1^*S_2^*\| < 1$ and $\|S_2^*S_1^*\| < 1$. If this is the case, we have

$$\begin{aligned} \hat{f}_1^\infty &= (I - (I - S_1^*S_2^*)^{-1}(I - S_1^*))x \\ \hat{f}_2^\infty &= (I - (I - S_2^*S_1^*)^{-1}(I - S_2^*))x \end{aligned} \tag{12}$$

If S_1^* and S_2^* have principal values values ≤ 1, the conditions $\|S_1^*S_2^*\| < 1$ and $\|S_2^*S_1^*\| < 1$ say that the spaces of of vectors whose length is preserved under each mapping are disjoint. We now show that a cubic spline smoother matrix has real positive eigenvalues less than or equal to one and furthermore, $\|Sx\| < \|x\|$ unless x is a linear function of x. We can verify this through the representation $S = (I + \Delta^tW^{-1}\Delta)^{-1}$ where W and Δ are defined in Section 3. First note that W is positive definite since it is diagonal dominant (i.e. the sum of each row is $\leq$ the diagonal element in that row). Thus W^{-1} exists, $\Delta^tW^{-1}\Delta$ is non-negative definite and hence $(I + \Delta^tW^{-1}\Delta)^{-1}$ has eigenvalues ≤ 1. Now suppose $(I + \Delta^tW^{-1}\Delta)^{-1}z = z$. Then $\Delta^tW^{-1}\Delta z = 0$, $z^t\Delta^tW^{-1}\Delta z = 0$ and thus $\Delta z = 0$ (since W and hence W^{-1} are positive definite). Now Δ takes second differences and hence z hence must be a linear function of x.

Since we have removed the eigenspace corresponding to the constant eigenvector, we see that backfitting will only fail to converge if $x_1 = c_1x_2 + c_2$ for some c_1 and c_2.

In a backfitting algorithm involving p cubic spline smoothers, we have have been unable

to prove convergence. However, we are able to prove convergence for a modified (and more efficient) version of backfitting. Details are in Buja, Hastie and Tibshirani (1986).

Note that one can, in theory, avoid iteration in the backfitting loop through use of formula (12). Unfortunately, these expressions are formidable to compute, requiring the inversion of $n \times n$ matrix. However, Green and Yandell (1985) show than in the special case in which S_1 represents a cubic spline smooth and S_2 a least squares projection, one can compute these expressions explicitly in only $O(n)$ operations.

Table 1

Analysis of deviance for ovarian cancer data

Model	Residual deviance	Degrees of Freedom
FAT	1144.8	977.0
FAT, OA(λ=+1.25)	1129.0	974.7
FAT, OA(linear)	1134.8	976.0
FAT, AEO(λ=+.65)	1133.7	973.2
FAT, AEO(λ=+.65), OA(λ=+1.25)	1115.3	970.9
all covariates including OA	1104.3	967.3

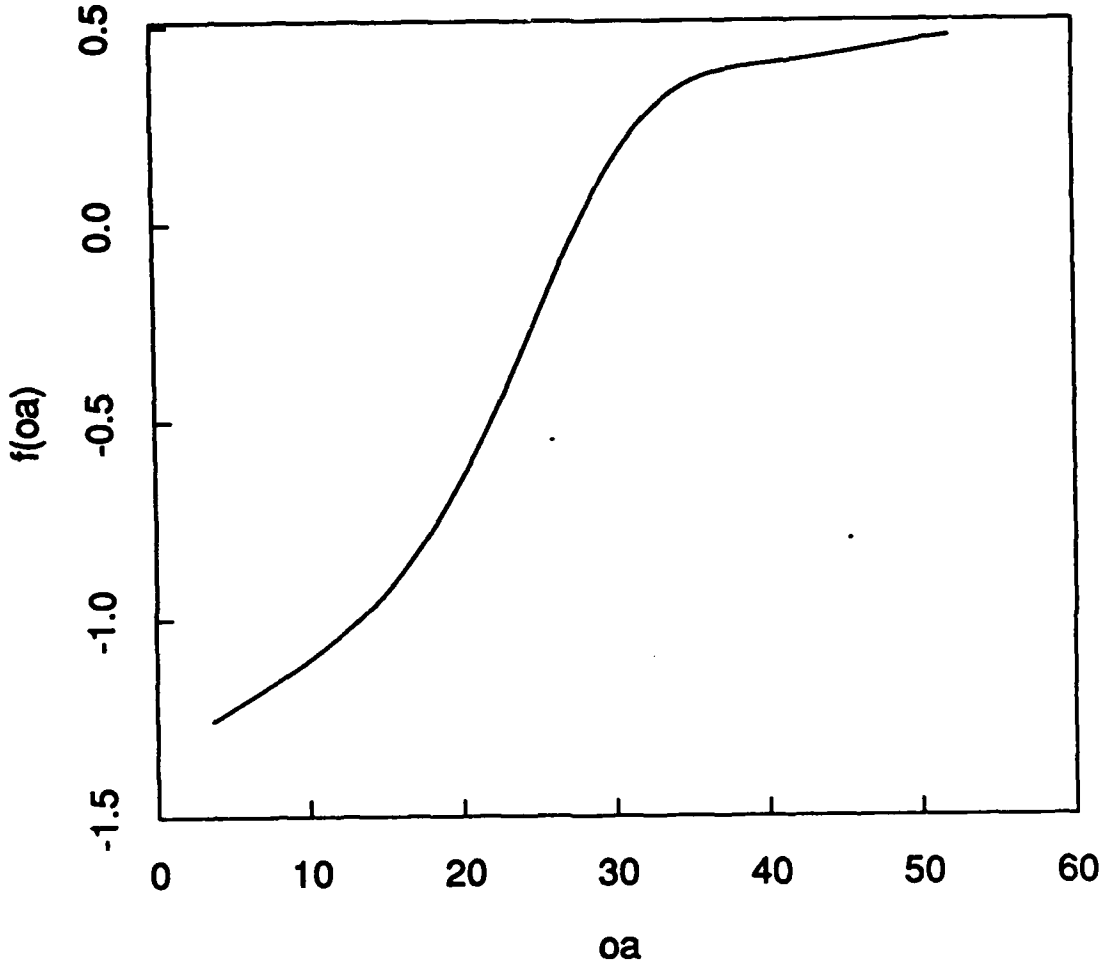


Figure 1. Partial residual smooth for OA with AEO linear

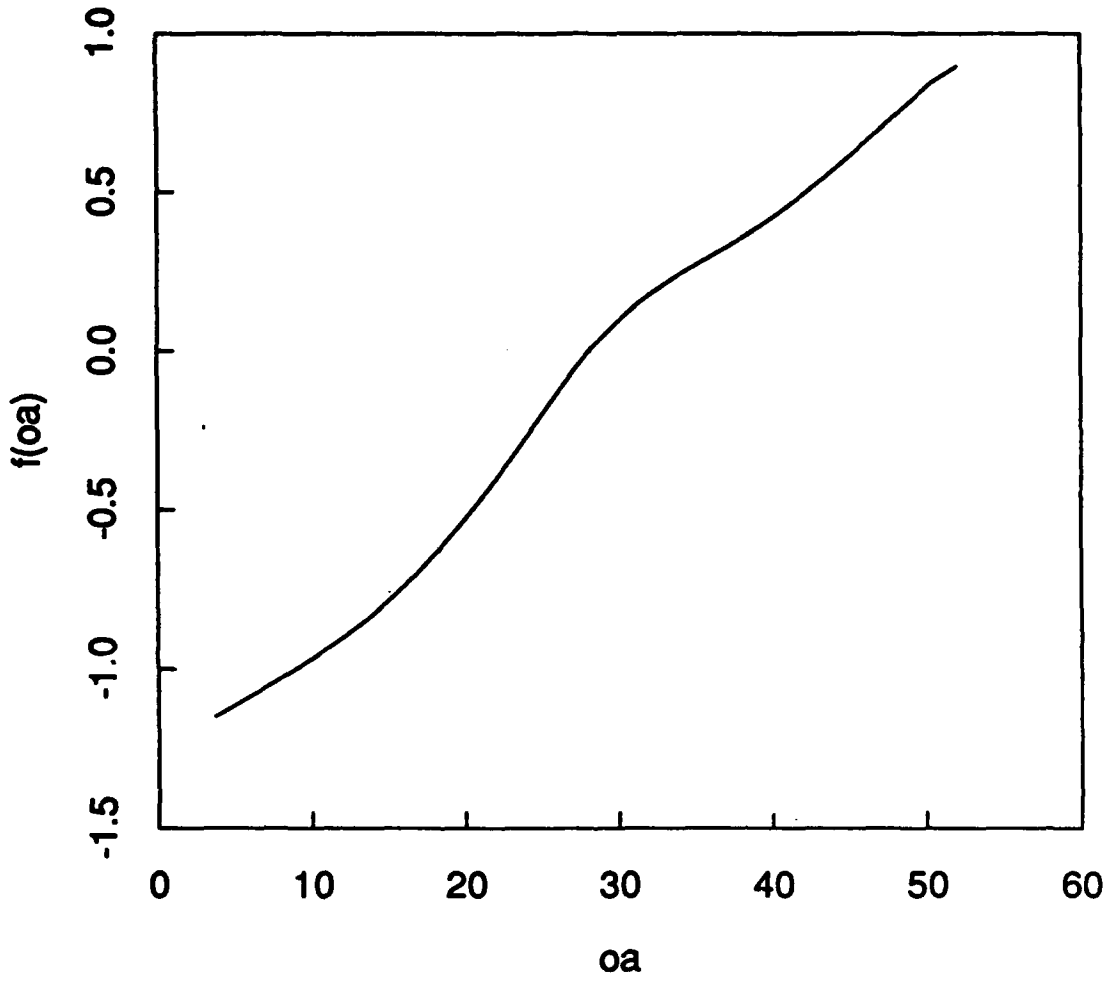


Figure 2. Partial residual smooth for OA with AEO smooth

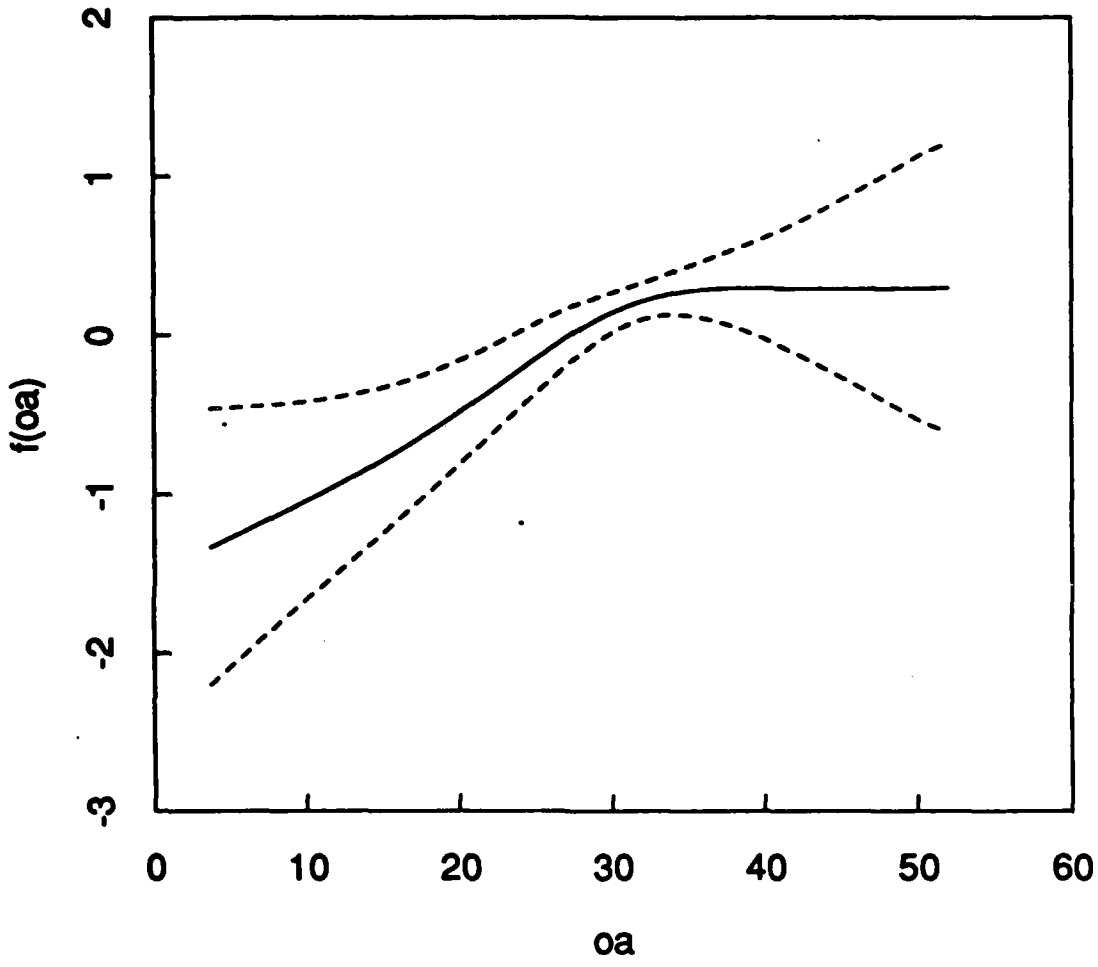


Figure 3. Partial residual smooth for OA, adjusting for FAT. Broken curves are 95% confidence bands

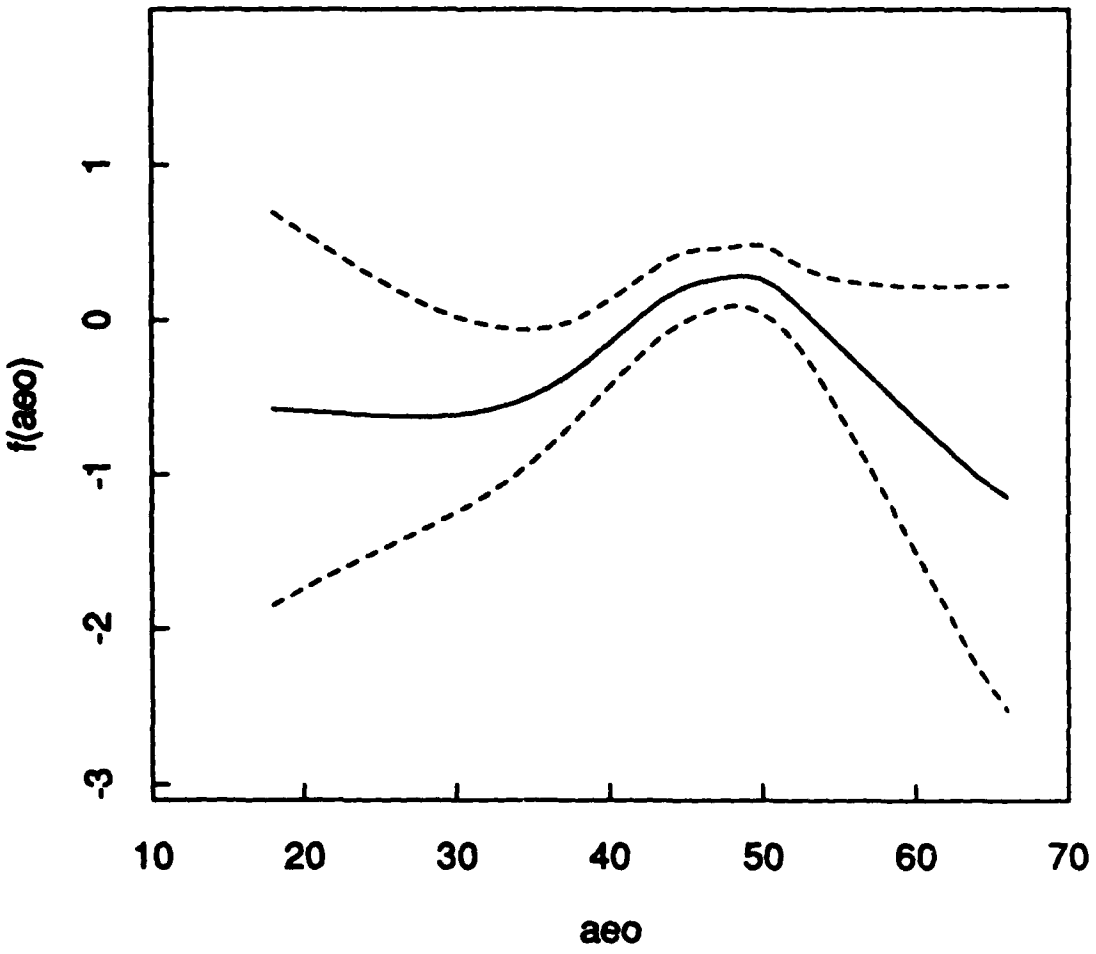


Figure 4. Partial residual smooth for AEO, adjusting for FAT and OA. Broken curves are 95% confidence bands.

REFERENCES

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. To appear, J. Amer .Stat. Assoc.

Breslow, N. and Day, N. (1980). Statistical methods in cancer research, volume 1- the analysis of case-control studies. International agency for research on cancer, IARC 32, Lyon.

Buja, A., Hastie, T. and Tibshirani, R. (1986). Linear smoother and addtive models. In preparation.

Cleveland, W. (1979). Robust locally weighted regression and smoothing of scatterplots, J. Amer. Statist. Assoc., 74, 829-836.

Cox, D.R. (1972). Regression models and life tables (with discussion). J. Roy. Stat. Soc B.,34, 187-202.

Friedman, J. and Stuetzle, W. (1981) Projection pursuit regression, J. Amer. Statist. Assoc., 76, 817-823.

Friedman, J. and Stuetzle, W. (1982) Smoothing of scatterplots,Tech. rep. Orion 3, Dept. of Statistics, Stanford University.

Green, P. (1985). Penalized likelihood for general semi-parametric likelihood models. Tech. rep. 2819, Math Research Ctr., U of Wisc.

Green, P. and Yandell, B. (1985). Semi parametric generalized linear models. Proc. 2nd Intern. Glim Conference. Springer-Verlag lecture notes in Statistics #32, Berlin Heidelberg.

Hastie, T. (1982). Comment on " Graphical methods for assessing logistic regression models" by Landwehr et al. J. Amer. Statist. Assoc., 79, 61-63.

Hastie and Tibshirani (1984). Generalized additive models. Tech. rep. 98, Dept. of Statistics, Stanford University.

Hastie, T. and Tibshirani, R (1985a). Generalized additive models- some applications. Tech. rep. 14, Dept. of Statistics, Univ. of Toronto.

Hastie, T. and Tibshirani, R. (1985b). Generalized additive models; some Applications, Proc. 2nd Intern. Glim Conference. Springer-Verlag lecture notes in Statistics #32, Berlin Heidelberg.

Hastie, T. and Tibshirani, R. (1985c). Non-parametric logistic and proportional odds regression. Tech. rep. 1986-001, Biostatistics group, University of Toronto.

Hastie and Tibshirani (1986). Generalized additive models. To appear, Statistical Science.

Landwehr, J., Pregibon, D., and Shoemaker, A. (1982) Graphical methods for assessing logistic regression models, J. Amer. Statist. Assoc., 79, 61-81.

McCullagh, P. (1980). Regression models for ordinal data. J. R. Statist. Soc B.,42,1, 109-142.

McDonald, J. and Owen, A. (1984). A split linear smoother. Tech. rep, Dept. of Stat., Stanford University.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models, J. R. Statist. Soc. A, 135, 370-384.

O'Sullivan, F. Yandell, B., and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models. J. Amer. Statist. Assoc. 81, 393, p 96-103.

Reinsch, C. (1967). Smoothing by spline functions, Numer. Math., 10, 177-183.

Risch, H., Weiss, N., Lyon, J., Daling, J. and Liff, J. (1983). Events of reproductive life and the incidence of ovarian cancer. Amer. J. Epidem. 117, 128-139.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with Discussion). J. R. Statist. Soc. B, 36, 111-147.

Stone. C. (1985). The dimensionality principle for generalized additive models. Tech. rep., Dept. of Statistics, U. Calif., Berkeley.

Tibshirani, R. (1984). Local likelihood estimation. Unpublished Ph.D thesis, and tech. rep., Dept. of Statistics, Stanford University.

Wahba, G., and Wold, S. (1975). A Completely automatic French curve: fitting spline functions by cross-validation, Comm. Statistics, 4, 1-7.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 390	2. GOVT ACCESSION NO.: AD-A181773
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Generalized Additive Models, Cubic Splines and Penalized Likelihood	5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL REPORT
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Trevor Hastie and Robert Tibshirani	8. CONTRACT OR GRANT NUMBER(s): N00014-86-K-0156
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, Stanford University, Stanford, CA 94305	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Statistics & Probability Program Code 1111	12. REPORT DATE: May 22, 1987 13. NUMBER OF PAGES: 22
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report): APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED	
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)	
18. SUPPLEMENTARY NOTES	
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Generalized additive model, spline smoothing, non-parametric regression, partial residual, penalized likelihood	
20. ABSTRACT (Continue on reverse side if necessary and identify by block number): PLEASE SEE FOLLOWING PAGE.	

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TECHNICAL REPORT NO. 390

20. ABSTRACT

Generalized additive models (Hastie and Tibshirani, 1986 Statistical Science) extend the class of generalized linear models by allowing an arbitrary smooth function for any or all of the covariates. The functions are estimated by the local scoring procedure, using a smoother as a building block in an iterative algorithm. In this paper we utilize a cubic spline smoother in the algorithm and show how the resultant procedure can be viewed as a method for automatically smoothing a suitably defined "partial residual", and more formally, a method for maximizing a penalized likelihood. We also examine convergence of the inner ("backfitting") loop in this case and illustrate these ideas with some binary response data.

END

7-87

DTIC