Lecture Notes in Statistics

Edited by D. Brillinger, S. Fienberg, J. Gani,
J. Hartigan, and K. Krickeberg

32

Generalized Linear Models

Proceedings of the GLIM 85 Conference,
held in Lancaster, UK, Sept. 16–19, 1985

Edited by R. Gilchrist, B. Francis and J. Whittaker

Springer-Verlag
Berlin Heidelberg New York Tokyo

Semi-parametric Generalized Linear Models

by

Peter J. Green* and Brian S. Yandell

Department of Statistics
University of Wisconsin-Madison

1. Introduction

When the form of a regression relationship with respect to some but not all of the explanatory variables is unknown, the statistician is caught in a quandary. Should parametric models be abandoned altogether, thus losing the opportunity of estimating parameters of real interest and sacrificing efficiency in estimation and prediction, or should the extraneous variables be forced into a parametric model by imposing a possibly inappropriate functional form without adequate justification?

A compromise is possible using the idea of semi-parametric modelling. This has been considered by several authors in varying degrees of generality; see for example Rice (1981), Green, Jennison and Seheult (1983), Wahba (1984), and Green (1985b). In the present context of generalized linear models we consider replacing the familiar linear predictors $\eta_i = \mathbf{x}_i^T\beta$ by the more general predictors

$$\theta_i = \mathbf{x}_i^T\beta + \gamma(\mathbf{t}_i) \qquad (1)$$

with $\mathbf{x}_i$ a p-vector of explanatory variables for the ith observation, β the corresponding regression coefficients, $\mathbf{t}_i$ the scalar or vector of extraneous variables, and γ a function or curve whose form is not specified. As a simple example, imagine a binomial logistic regression in which the "intercept" term is believed to vary in time or with geographical location. In section 5 we consider in detail an example where a previous investigator has been unsure about the precise dependence of the binary response (presence of tumour) on one of four explanatory variables (age at death).

Straightforward maximization of the log-likelihood function L, which we will write in the composite form $L(\theta(\beta,\gamma))$ to emphasize the roles of predictors, parameters, and unknown curve, is no longer appropriate as a method of estimation. This leads to overfitting in the absence of any constraints on β. Indeed, it typically renders the parameters β unidentifiable. But progress is possible by maximizing instead a penalized version of the log-likelihood, if we are willing to place weak constraints on the form of γ by assuming that it is smooth. Thus we maximize the penalized log-likelihood

$$L(\theta(\beta,\gamma)) - \tfrac{1}{2}\lambda J(\gamma) \qquad (2)$$

where the penalty functional J is some numerical measure of the "roughness" of γ. This might be adopted on ad-hoc grounds (for example, an integrated squared derivative of γ), or follow from a Bayesian argument specifying a prior distribution for γ. The scalar λ is a tuning constant, used to regulate the smoothness of the fitted curve γ. Typically we try a range of values for λ in an exploratory fashion, as well as considering automatic choice based on the data. One ultimate aim may be to discover the form of γ in the hope of modelling it parametrically in future.

2. Maximum Penalized Likelihood Estimates

Here we will only consider maximization of (2) over γ in the span of a set of q prescribed basis functions $\{ \phi_k ; k = 1,2,\ldots,q \}$: we write

$$\gamma = \sum_{k=1}^{q}\xi_k\phi_k$$

and assume in addition that J satisfies

$$J(\sum_{k=1}^{q}\xi_k\phi_k) = \xi^T\mathbf{K}\xi \qquad (3)$$

for some $q \times q$ non-negative definite symmetric $\mathbf{K}$. We thus re-write the penalized log-likelihood in the form

$$L(\theta(\beta,\xi)) - \tfrac{1}{2}\lambda\xi^T\mathbf{K}\xi \qquad (4)$$

to be maximized over choice of the vectors β and ξ.

This finite-dimensional approach is not intended to compromise our non-parametric assumptions about the curve γ. The dimension q, perhaps equal to n, will typically be too large for parametric estimation of ξ to be appropriate, and the basis functions will be chosen so as not to materially constrain the curve, except perhaps in fine detail. With certain penalty functionals, for example those used in spline smoothing, it turns out that with $q = n$ we are not imposing any constraints at all (see section 3).

The semi-parametric regression problem expressed in the general form (4) is considered in some detail by Green (1985b), who derives the following iterative scheme for the maximum penalized likelihood estimates (MPLEs) $\hat{\beta}$ and $\hat{\xi}$. Suppose we have trial estimates β and ξ. Using these, compute the n-vector of scores $\mathbf{u}$ and the $n \times n$ information matrix $\mathbf{A}$:

$$\mathbf{u} = \frac{\partial L}{\partial\theta}, \qquad \mathbf{A} = E\left[-\frac{\partial^2 L}{\partial\theta\partial\theta^T}\right].$$

In the case of a generalized linear model, $\mathbf{u}$ and $\mathbf{A}$ may be expressed as

$$u_i = \frac{\pi_i(y_i-\mu_i)}{\phi\tau_i^2\delta_i}, \qquad \mathbf{A} = diag\left[\frac{\pi_i}{\phi\tau_i^2\delta_i^2}\right],$$

in the notation of the GLIM3 manual (Baker and Nelder, 1978), where the vectors μ, τ^2, and δ are computed with θ replacing the linear predictor. We also need the $n \times p$ and $n \times q$ matrices of derivatives

$$\mathbf{D} = \frac{\partial\theta}{\partial\beta}, \qquad \mathbf{E} = \frac{\partial\theta}{\partial\xi}.$$

Then updated estimates (β^*, ξ^*) are obtained as the solution to the linear system

* present affiliation: Department of Mathematical Sciences, University of Durham.

$$\begin{bmatrix} \mathbf{D}^T\mathbf{A}\mathbf{D} & \mathbf{D}^T\mathbf{A}\mathbf{E} \\ \mathbf{E}^T\mathbf{A}\mathbf{D} & \mathbf{E}^T\mathbf{A}\mathbf{E} + \lambda\mathbf{K} \end{bmatrix} \begin{bmatrix} \beta^* \\ \xi^* \end{bmatrix} = \begin{bmatrix} \mathbf{D}^T \\ \mathbf{E}^T \end{bmatrix} \mathbf{A}\mathbf{Y}, \tag{5}$$

where

$$\mathbf{Y} = \mathbf{A}^{-1}\mathbf{u} + \mathbf{D}\beta + \mathbf{E}\xi.$$

This scheme is based on the Newton-Raphson method with Fisher scoring, and these updating equations can be seen to combine the iteratively reweighted least squares equations for β used in GLIM (see also Green (1984)), with ridge-regression type equations for ξ (O'Sullivan, Yandell and Raynor, Jr., 1984; Yandell, 1985).

As they stand, the equations (5) are not ideal for practical computation. It is the purpose of this paper to derive various algorithms implementing this scheme, and to illustrate semi-parametric modelling applied to both real and simulated data sets.

3. The One-Dimensional Case, with Cubic Spline Smoothing

For a restrictive but very useful special case, consider a generalized linear model with predictors $\{\theta_i\}$ given by (1), in which $\{ t_i \}$ are one-dimensional, and suppose that the roughness penalty $J(\gamma)$ takes the form $\int (\gamma''(t))^2\, dt$. This allows one additional explanatory variable to enter in a non-parametric fashion; the form of penalty used ensures that the dependence on this variable is "visually smooth". Some aspects of the purely non-parametric version of this problem were discussed by Silverman (1985).

For simplicity, we suppose that the $\{ t_i \}$ are distinct and ordered, $t_1 < t_2 < ... < t_n$, but relaxing this requirement presents no great difficulty. It is well known (Reinsch, 1967) that the γ maximizing (2) for any fixed β and λ is a natural cubic spline with knots at $\{ t_i \}$, that the space of such splines has dimension n, and that we may choose a basis for this space with $\phi_k(t_i) = \delta_{ik}$.

In this case the notation used above simplifies: $q = n$, $\mathbf{D}$ and $\mathbf{E}$ are constant matrices with $\mathbf{D}$ having i th row equal to $\mathbf{x}_i^T$ and $\mathbf{E}$ being the identity, $\mathbf{A}$ is diagonal, and $\mathbf{Y} = \mathbf{A}^{-1}\mathbf{u} + \theta$. Further, it is implicit in Reinsch (1967) that $\mathbf{K}$ may be written $\Delta^T\mathbf{W}^{-1}\Delta$ where Δ is the $(n-2)\times n$ matrix taking second differences:

$$\Delta_{ii} = h_i^{-1}, \quad \Delta_{ii+1} = -(h_i^{-1} + h_{i+1}^{-1}), \quad \Delta_{ii+2} = h_{i+1}^{-1}.$$

and $\mathbf{W}$ is the symmetric tridiagonal matrix of order $(n-2)$:

$$W_{i-1i} = W_{ii-1} = h_i/6, \quad W_{ii} = (h_i + h_{i+1})/3.$$

where $h_i = t_{i+1} - t_i$. The important point about this decomposition is that Δ and $\mathbf{W}$ are banded.

One possible algorithm implementing (5) involves an inner iteration between the pair of equivalent equations

$$\beta^* = (\mathbf{D}^T\mathbf{A}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{A}(\mathbf{Y}-\xi^*)$$

$$\xi^* = \mathbf{S}\,(\mathbf{Y}-\mathbf{D}\beta^*) \tag{6}$$

where $\mathbf{S} = (\mathbf{A} + \lambda\mathbf{K})^{-1}\mathbf{A}$. This will always converge (Green, 1985a). But further iteration can be avoided by eliminating ξ^* from (5) to give

$$\beta^* = (\mathbf{D}^T\mathbf{A}(\mathbf{I}-\mathbf{S})\mathbf{D})^{-1}\mathbf{D}^T\mathbf{A}(\mathbf{I}-\mathbf{S})\mathbf{Y}. \tag{7}$$

Solution of this small $(p\times p)$ system for β^* is followed by use of (6) to obtain ξ^*. From the updated (β^*, ξ^*) we recompute θ, thence $\mathbf{u}$ and $\mathbf{A}$, and the cycle is repeated to convergence.

This approach is highly practicable, and very economical, since apart from solving the linear equations (7), and some matrix multiplications, we only need to apply the "smoothing operator" $\mathbf{S} = (\mathbf{A} + \lambda\mathbf{K})^{-1}\mathbf{A}$ to form $\mathbf{SY}$ and $\mathbf{SD}$. But a consequence of the special structure of $\mathbf{K}$ mentioned above is that $\mathbf{S}$ can be applied to a vector in only $O(n)$ operations. We use a minor modification of the version of Reinsch's algorithm given by De Boor (1978) to obtain a very fast implementation.

An almost identical approach may be adopted more generally, with any penalty functional for which $\mathbf{S} = (\mathbf{A} + \lambda\mathbf{K})^{-1}\mathbf{A}$ may pre-multiply a vector in $O(n)$ time. This could include splines of different orders, penalties based on discrete differences, and "moving average" smoothers.

4. Goodness-of-fit, Standard Errors, and Choice of λ

Goodness-of-fit can be assessed globally, as in generalized linear models, by the deviance $\Delta = 2\{sup_\theta\, L(\theta) - L(\theta(\hat\beta,\hat\xi))\}$ where $(\hat\beta,\hat\xi)$ are the MPLEs. Locally, it is measured by residuals: either the deviance residuals, the signed square roots of the individual contributions to Δ, or in GLIM fashion as

$$z_i = u_i \,/\, A_{ii}^{1/2} = \left\{\frac{\pi_i}{\phi}\right\}^{1/2} \left\{\frac{y_i-\mu_i}{\tau_i}\right\}. \tag{8}$$

This is all standard, but we do need a new concept of degrees-of-freedom to assign to Δ. It turns out (Green, 1985b) that the appropriate value, not in general an integer, is given by

$$\nu = n - tr(\mathbf{S}) - tr\left[(\mathbf{D}^T\mathbf{A}(\mathbf{I}-\mathbf{S})\mathbf{D})^{-1}\mathbf{D}^T\mathbf{A}(\mathbf{I}-\mathbf{S})^2\mathbf{D}\right]. \tag{9}$$

This is an approximation to the asymptotic expectation of Δ, and reduces to the usual $n - rank(\mathbf{D})$ when the non-parametric part of the model is omitted. In the non-parametric case, this ν has been used informally for linear models (Eubank, 1984; Eubank, 1985) and generalized linear models (O'Sullivan, Yandell and Raynor, Jr., 1984; O'Sullivan, 1985; Yandell, 1985).

Similar somewhat approximate asymptotics lead to an estimated variance matrix for $\hat\beta$ of the form

$$(\mathbf{D}^T\mathbf{A}(\mathbf{I}-\mathbf{S})\mathbf{D})^{-1}\mathbf{D}^T\mathbf{A}(\mathbf{I}-\mathbf{S})^2\mathbf{D}(\mathbf{D}^T\mathbf{A}(\mathbf{I}-\mathbf{S})\mathbf{D})^{-1} \tag{10}$$

from which standard errors may be calculated. In the absence of the appropriate distribution theory, neither the deviance nor the standard errors should be used in formal significance tests, at present, but they do seem to provide adequate guidelines for model selection.

Computation of these quantities follows naturally from the algorithm outlined in section 3, and consists of solving $p\times p$ linear systems following the repeated application of $\mathbf{S}$ to $\mathbf{D}$. The only part of this that is not simple to implement in $O(n)$ time is the first trace term, $tr(\mathbf{S})$, which in our present program takes about $7n^2$ multiplications or divisions. However an $O(n)$ algorithm for this computation in linear spline smoothing has recently been announced by O'Sullivan (1985), and we will adapt this to the present context.

As for automatic choice of λ, Wahba's generalized cross-validation (GCV) method (Wahba, 1977), which uses an invariant modification of a predictive mean-squared error criterion, may be adapted to this situation. A quadratic approximation to the quantity to be minimized (over λ) is simply Δ/v^2, so no further computation is involved. We use a simple one-dimensional search over λ to find the minimum. Other approaches to the automatic choice of λ would be possible, for example the empirical Bayesian methods proposed by Leonard (1982).

5. Examples

Logistic Regression, and Tumour Prevalence Data

Dinse and Lagakos (1983) consider logistic regression models for data from a U.S. National Toxicology Program bioassay of a flame retardant. Data on 127 male and 192 female rats exposed to various doses of the agent consist of a binary response variable (y) indicating presence or absence of bile duct hyperplasia at death, and four explanatory variables: log dose (x_1), initial weight (x_2), cage position (x_3), and age at death (t). Dinse and Lagakos express some doubts as to whether the fourth of these variables enters the model linearly, so they consider fitting higher-order polynomials, or step functions based on age intervals. A reasonable alternative seemed to be the semi-parametric approach described here, which allows age at death (t) to enter the binomial logistic model in a non-parametric fashion, whilst still allowing estimation of the log dose regression coefficient.

The results of our analyses are presented graphically (see Figure 1) for various values of the tuning constant λ. In each plot, the upper two traces display on the same scale the fitted parametric and non-parametric components of the predictor, $\sum D_{ij}\hat{\beta}_j$ and $\hat{\gamma}(t) = \sum \hat{\xi}_k \phi_k(t)$, plotted against t_i. The lower panel displays the corresponding residuals from (8). (The slightly bizarre appearance of the residual plot is due to the binary nature of the response variable.) In both data sets there were a considerable number of ties in values of $\{ t_i \}$, which we have broken arbitrarily with small deterministic displacements. This seems not to lead to any numerical instability in our program, it avoided some modifications to the coding to handle coincident $\{ t_i \}$, and had the incidental advantage of clarifying the plots so that each case can be distinguished.

For the data on the male rats, Figures 1(a) and 1(b) demonstrate the effect of using very small and large values of λ respectively, suggesting under- and over-smoothing. These values are respectively 0.01 and 100 times the automatic GCV choice of $\lambda = 1380$, for which the relevant plot is Figure 1(c). This indicates a nonlinear dependence on t, but note that the left-hand part of the curve, up to the first turning value, is based on rather little data. The parameter estimates, with approximate standard errors, are –0.139±0.148, –0.012±0.022, and 0.068±0.151, and we thus agree with Dinse and Lagakos about the lack of significance of the regression coefficients.

In the case of the female rats, the GCV method for choice of λ is not so well behaved. Whilst there is a turning value of Δ/v^2 at $\lambda = 6.24$, this is only a local minimum, and the GCV criterion seems to decrease to 0 for very small λ. Figure 1(d) displays the fitted values and residuals for $\lambda = 6.24$, suggestive of a complicated nonlinear dependence on t. Our parameter estimates are

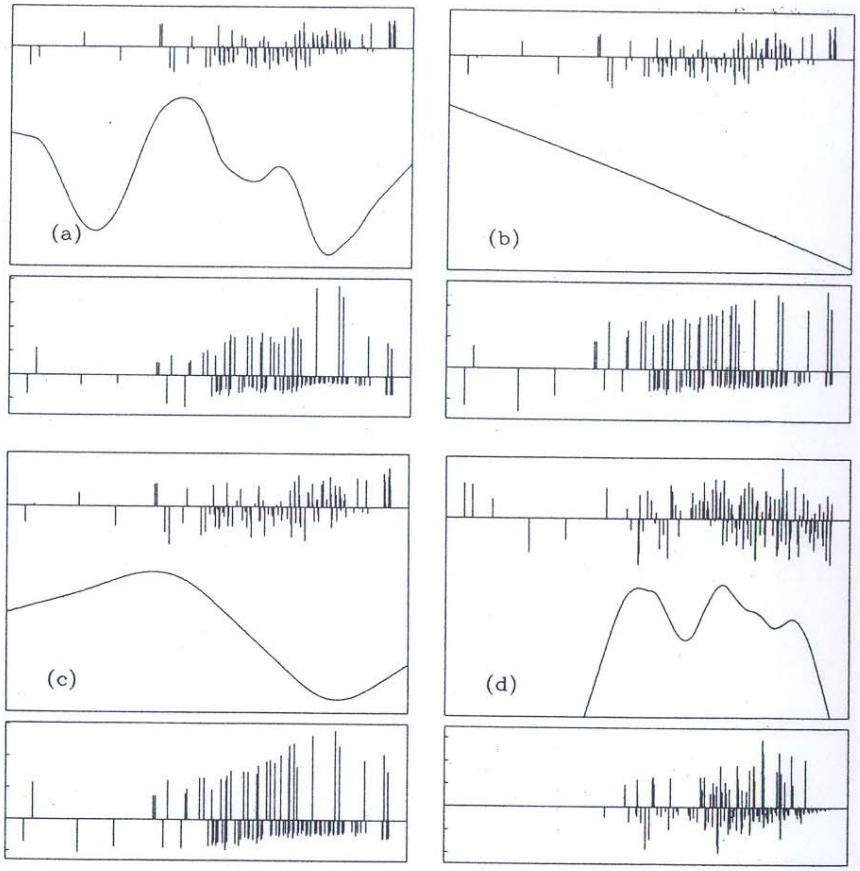

Figure 1. Fitted Logits and Residuals for Tumour Prevalence in Rats

0.492±0.137, 0.040±0.015 and 0.270±0.158. Dinse and Lagakos obtain $\beta_1 = 0.554 \pm 0.20$ using a model linear in t, implying that if we have correctly identified the form of $\gamma(t)$ then their estimate is both slightly biased and inefficient.

It may be of interest to report some details of the performance of our algorithm applied to these data. For the male rats, with the GCV choice of λ, four iterations were needed to converge from initial estimates corresponding to empirical logits to a point where neither β nor the deviance changed by more than 10^{-4}. Excluding the calculation of $tr(S)$ (see section 4), the computation time was 1.26 seconds on a VAX 11/780. For the female rats, 10 iterations were required, and the time was 3.77 seconds.

Poisson Log-linear Regression, with Simulated Data

As a demonstration that our approach can properly identify a smooth $\gamma(t)$, we analyzed a simulated data set, with $\{ y_i \, ; i = 1,2,\ldots,200 \}$ distributed independently with Poisson distributions with means $\{ \exp(\theta_i) \}$, where

$$\theta_i = \sum_{j=1}^{5} D_{ij}\beta_j + 1 + \sin(t_i).$$

The $\{ t_i \}$ were chosen independently and uniformly on the interval $(0,3\pi)$, so that the true curve $\gamma(t) = 1 + \sin t$ has three turning values in the range of the data. The design matrix D was taken as that for 40 replicates of a randomized complete blocks design on 5 treatments, each successive 5 ordered $\{ t_i \}$ forming a block. The true β was (1, 0.5, 0, –0.5, –1), so that the Poisson means varied between about 0.4 and 20.

Figure 2(a) corresponds to the GCV choice of $\lambda = 3.13$: the fitted $\gamma(t)$ has the correct form, and the parameter estimates are (1.031, 0.517, 0.020, –0.583, –0.984) with standard errors (0.058, 0.066, 0.078, 0.098, 0.116). (Clearly for identifiability a constraint must be placed on β : we used $\sum\beta_j = 1$). Figure 2(b) demonstrates that when the tuning constant λ is set much too high, in this case 100 times the GCV value, the fitted curve $\gamma(t)$ cannot match the structure in the data, which is therefore forced into the residuals.

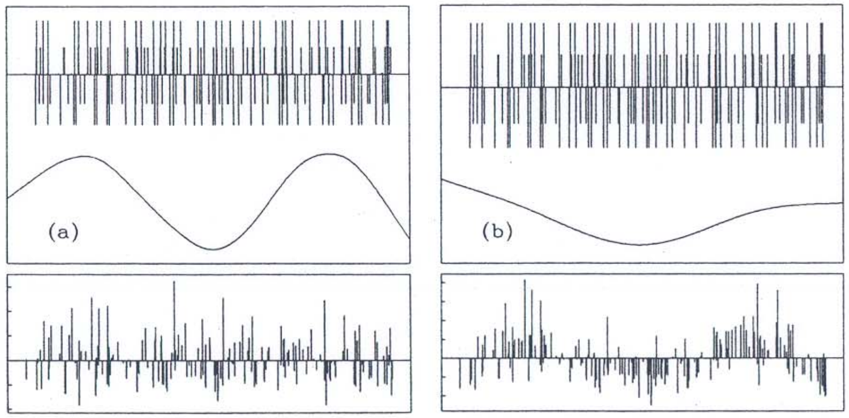

Figure 2. Fitted Log Rates and Residuals for Simulated Poisson Data

6. Using B-splines

For an application of spline smoothing to the penalized log-likelihood (4) more general than that of section 3, suppose that the $\{t_i\}$ remain one-dimensional, and that the roughness penalty takes the form

$$J(\gamma) = \int[\gamma^{(m)}(t)]^2 dt.$$

In this case, an alternative set of basis splines is useful, namely the natural B-splines of order $2m$ (De Boor, 1978; Schumaker, 1981). These are defined on any sequence of knots $s_1 < \cdots < s_q$ as piecewise polynomials of degree $(2m-1)$ between the knots, of degree $(m-1)$ outside (s_1, s_q), and with $(2m-2)$ continuous derivatives. This basis leads to stable, economical computing as B-splines are non-negative and have limited support. In the cubic spline case ($m = 2$), $\{\phi_k \, ; k=3, \cdots ,q-2\}$ are each non-negative only on (s_{k-2}, s_{k+2}), whilst ϕ_1 , ϕ_2 , ϕ_{q-1} and ϕ_q are linear outside (s_1, s_q). The matrix $\mathbf{K}$, which has the form

$$K_{ij} = \int \phi_i^{(m)}(t)\phi_j^{(m)}(t)dt,$$

is banded, and so an algorithm based on the equations (7) and (6) can be implemented in $O(n)$ time; see also Silverman (1985).

This approach is particularly useful in the case where n is very large and it is desirable that the number q of basis functions be much smaller. We are then only attempting the restricted minimization of (4), but with say $q = 50$ or 100 knots equally spaced to cover the range of $\{t_i\}$, this restriction is not of practical importance.

7. An Algorithm for the General Case

The linear system (5) can be expressed as finding β and ξ to minimize

$$||\mathbf{B}^T(\mathbf{Y}-\mathbf{D}\beta-\mathbf{E}\xi)||^2 + \lambda\xi^T\mathbf{K}\xi \tag{11}$$

in which $\mathbf{A}=\mathbf{B}\mathbf{B}^T$. Suppose that $\mathbf{K}$ is of rank $r < q$. Two matrices, $\mathbf{J}$ and $\mathbf{T}$, with q rows and full column ranks r and $q-r$, respectively, can be formed such that $\mathbf{J}^T\mathbf{K}\mathbf{J} = \mathbf{I}$, $\mathbf{T}^T\mathbf{K}\mathbf{T} = 0$, and $\mathbf{J}^T\mathbf{T} = 0$ (see below). Rewriting ξ as

$$\xi = \mathbf{T}\delta + \mathbf{J}\varepsilon, \tag{12}$$

with ε and δ of lengths r and $q-r$, respectively, (11) becomes

$$||\mathbf{B}^T(\mathbf{Y} - [\mathbf{D}{:}\mathbf{E}\mathbf{T}]\begin{pmatrix}\beta\\ \delta\end{pmatrix} - \mathbf{E}\mathbf{J}\varepsilon)||^2 + \lambda\varepsilon^T\varepsilon.$$

A Householder decomposition (Dongarra et al., 1979) allows one to separate the solution of β and δ from that of ε . In other words, we decompose

$$\mathbf{Q}_1^T\mathbf{B}^T[\mathbf{D}{:}\mathbf{E}\mathbf{T}] = \mathbf{R}, \ \ \mathbf{Q}_2^T\mathbf{B}^T[\mathbf{D}{:}\mathbf{E}\mathbf{T}] = 0,$$

in which $\mathbf{Q} = [\mathbf{Q}_1{:}\mathbf{Q}_2]$ is orthogonal and $\mathbf{R}$ is nonsingular, upper triangular, and of full rank $p+q-r$. The linear problem becomes: minimize the sum of

$$\|Q_1^T B^T Y - R\begin{bmatrix}\beta\\ \delta\end{bmatrix} - Q_1^T B^T E J \varepsilon\|^2 \tag{13}$$

and

$$\|Q_2^T B^T Y - Q_2^T B^T E J \varepsilon\|^2 + \lambda \varepsilon^T \varepsilon. \tag{14}$$

The first term (13) can be set to zero by appropriate choice of β and δ given ε. If we define $Y^* = Q_2^T B^T Y$ and $Z = Q_2^T B^T E J$, (14) becomes a problem of minimizing

$$\|Y^* - Z\varepsilon\|^2 + \lambda \varepsilon^T \varepsilon$$

which is an ordinary ridge regression problem. See Bates and Wahba (1983); Golub, Heath and Wahba (1979). The solution is

$$\varepsilon^* = (Z^T Z + \lambda I)^{-1} Z^T Y^*.$$

The other parameters are solved as

$$\begin{bmatrix}\beta^*\\ \delta^*\end{bmatrix} = R^{-1} Q_1^T B^T (Y - E J \varepsilon^*).$$

One then computes ξ^* using (12) and proceeds with the nonlinear iteration discussed in section 2.

Decomposition of K

One need only compute J and T once as K depends on the model only through $\{ t_i \}$. Using a pivoted Cholesky decomposition (Dongarra et al., 1979),

$$P^T K P = L^T L,$$

with L of dimension $r \times q$ and P a permutation matrix such that the first r columns of KP are linearly independent. A Householder decomposition of L^T yields

$$F_1^T L^T = G, \ F_2^T L^T = 0,$$

in which $F = [F_1 : F_2]$ is orthogonal and G is nonsingular, upper triangular, and of full rank r. It is known (Dongarra et al., 1979) that $G^{-1} F_1^T$ is the Moore-Penrose inverse of L^T. Therefore, one can construct the matrices J and T as

$$T = PF_2 \text{ and } J = PF_1 G^{-T}.$$

A further refinement is possible if the partial derivative matrices E and D can be written as $E = E^* M$ and $D = D^* M$, in which M depends only on $\{ t_i \}$. For instance, with a B-spline basis with model (1), $M_{ik} = \phi_k(t_i)$. One can left multiply by M exactly once, forming $T^* = MT$ and $J^* = MJ$ each with n rows, and replace E and D by E^* and D^* in subsequent computations.

Auxiliary Statistics

Auxiliary statistics can be constructed in the general case in a similar fashion to Section 4, with S in equations (9) and (10) replaced by

$$S = E(E^T A E + \lambda K)^{-1} E^T A.$$

Alternatively, using the notation of this section, one can show that (9) becomes

$$\nu = n - p - q + r - tr\left[Z^T Z (Z^T Z + \lambda I)^{-1}\right].$$

The variance (10) can be reexpressed as

$$var\begin{bmatrix}\hat\beta\\ \hat\delta\end{bmatrix} = R^{-1} Q_1^T \left[I + B^T E J (Z^T Z + \lambda I)^{-1} Z^T Z (Z^T Z + \lambda I)^{-1} J^T E^T B\right] Q_1 R^{-T}.$$

O'Sullivan (1985) observed that the trace and the diagonal "leverage" elements of the hat matrix for the B-spline basis can be computed in $O(r)$ multiplications/divisions by using a Cholesky decomposition of ($Z^T Z + \lambda I$) as in Silverman (1985).

8. Related Work in Progress

We have nearly completed an implementation of the algorithm for the general case, which will allow specification of models through subroutine evaluation of the forms of A, D, E, K and u. This software will be in the public domain and uses LINPACK (Dongarra et al., 1979) This work is related to a larger programming effort involving Douglas Bates, Grace Wahba, and Mary Lindstrom.

Yandell and Bates (unpublished) have also observed that one can use the singular value decomposition for ridge regression problems (Bates and Wahba, 1983; Golub, Heath and Wahba, 1979) to iterate on the "optimal" choice of tuning constant λ (in the sense of minimizing the generalized cross validation function) within each linear step. Thus λ changes with each nonlinear iteration, as do β and ξ. Empirically, convergence takes about the same number of steps as it would for a fixed λ near the optimal value. Unfortunately, the singular value decomposition is expensive, taking $O(r^3)$ multiplications/divisions. However, it may be possible to combine this approach with the Cholesky decomposition approach used by O'Sullivan, Yandell and Raynor, Jr. (1984) and Silverman (1985) to strike a healthy compromise between heavy computation and finding the optimal amount of smoothing.

Acknowledgements

This research has been supported in part by United States Department of Agriculture CSRS grant 511-100, National Sciences Foundation grant DMS-8404970, and United States Army contract DAAG29-80-C-0041. Computing was performed on the UW-Madison Statistics VAX 11/750 Research Computer and the UW-Madison MRC VAX 11/780 Computer.

References

Baker, R. J. and Nelder, J. A. (1978) *The GLIM System Release 3*. Oxford: Numerical Algorithms Group.

Bates, D. M. and Wahba, G. (1983) A truncated singular value decomposition and other methods for generalized cross-validation. Technical Report#715, Dept. of Statistics, U. of Wisconsin.

De Boor, C. (1978) *A Practical Guide to Splines.* New York: Springer.

Dinse, G. E. and Lagakos, S. W. (1983) Regression analysis of tumour prevalence data. *Appl. Statist.*, **32**, 236-248.

Dongarra, J. J., Bunch, J. R., Moler, C. B. and Stewart, G. W. (1979) *Linpack User's Guide.* Philadelphia: SIAM.

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

Eubank, R. L. (1985) Diagnostics for smoothing splines. *J. R. Statist. Soc. B*, **47**. (to appear)

Golub, G. H., Heath, M. and Wahba, G. (1979) Generalised cross validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215-224.

Green, P. J., Jennison, C. and Seheult, A. H. (1983) Contribution to the discussion of the paper by Wilkinson et al. *J. R. Statist. Soc. B*, **45**, 193-195.

Green, P. J. (1984) Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives (with discussion). *J. R. Statist. Soc. B*, **46**, 149-192.

Green, P. J. (1985a) Linear models for field trials, smoothing, and cross-validation. *Biometrika*, **72**. (to appear)

Green, P. J. (1985b) Penalized likelihood for general semi-parametric regression models. Technical Report#2819, Math. Research Center, U. of Wisconsin.

Leonard, T. (1982) An empirical Bayesian approach to the smooth estimation of unknown functions. Technical Report#2339, Math. Research Center, U. of Wisconsin.

O'Sullivan, F., Yandell, B. S. and Raynor, Jr., W. J. (1984) Automatic smoothing of regression functions in generalized linear models. Technical Report#734, Dept. of Statistics, U. of Wisconsin.

O'Sullivan, F. (1985) Contribution to the discussion of the paper by Silverman. *J. R. Statist. Soc. B*, **47**. (to appear)

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

Rice, J. R. (1981) An approach to peak area estimation. *J. Res. Nat. Bur. Stand.*, **87**, 53-65.

Schumaker, L. L. (1981) *Spline Functions: Basic Theory.* New York: Wiley.

Silverman, B. W. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. R. Statist. Soc. B*, **47**. (to appear)

Wahba, G. (1977) A survey of some smoothing problems and the method of generalized cross-validation for solving them.. In *Applications of Statistics* (P. R. Krishnaiah, ed.), pp.507-523. Amsterdam: North Holland.

Wahba, G. (1984) Cross validated spline methods for the estimation of multivariate functions from data on functionals. In *Statistics: An Appraisal, Proceedings 50th Anniversary Conference Iowa State Statistical Laboratory* (H. A. David,, ed.) Iowa State U. Press.

Yandell, B. S. (1985) Graphical analysis of proportional Poisson rates. *Proceedings of the 17th Symposium on the Interface*, Lexington, Kentucky 17-19 March 1985.