# Automatic Smoothing of Regression Functions in Generalized Linear Models

FINBARR O'SULLIVAN, BRIAN S. YANDELL, and WILLIAM J. RAYNOR, JR.*

We consider the penalized likelihood method for estimating nonparametric regression functions in generalized linear models (Nelder and Wedderburn 1972) and present a generalized cross-validation procedure for empirically assessing an appropriate amount of smoothing in these estimates. Asymptotic arguments and numerical simulations are used to show that the generalized cross-validatory procedure preforms well from the point of view of a weighted mean squared error criterion. The methodology adds to the battery of graphical tools for model building and checking within the generalized linear model framework. Included are two examples motivated by medical and horticultural applications.

KEY WORDS: Penalized likelihood; Smoothing splines; IRLS; Cross-validation.

## 1. INTRODUCTION

Nelder and Wedderburn (1972) introduced a collection of statistical regression models known as generalized linear models (GLM's) for the analysis of data from exponential families. With the subsequent development and spread of GLM computer software, the importance of these models in practical data analysis has greatly expanded (see McCullagh and Nelder 1983). As the popularity of these methods has increased, so has the need for more sophisticated model building and diagnostic checking techniques. In this context, nonparametric estimates of the GLM regression surface can be very useful. Recently, Hastie and Tibshirani (1984) studied an approach that combines an additive approximation to the regression surface with the fast one-dimensional smoothing algorithm of Friedman and Stuetzle (1982). In this article we propose a more general multivariate smoothing spline-type estimator and develop an explicit cross-validation score to assess the appropriate correct degree of smoothing. The method adds to the battery of techniques for model building and checking within the GLM framework, and it is particularly suited to the analysis of larger data sets (say $n > 50$ data points).

The basic GLM analysis starts with data of the form

$$(y_i, t_i), \qquad i = 1, 2, \ldots, n,$$

in which the $y_i$ are independent observations, each from a one-parameter exponential family distribution depending on the control variable or covariate $t_i$ (possibly vector-valued). Thus the density of $y_i$ has the form

$$\exp\{(y_i\theta_i - b(\theta_i))/a_i(\phi) + c(y_i, \phi)\}$$

for some appropriate $a_i$, $b$, and $c$. $\phi$ is an unknown scale parameter and often $a_i(\phi) = \phi/w_i$ with the prior weights, $w_i$, known. The mean and variance of $y_i$ are given by

$$E[y_i] = \dot{b}(\theta_i) = \mu_i$$

$$\text{var}[y_i] = \ddot{b}(\theta_i)a_i(\phi) = V_i.$$

Here a dot denotes differentiation. In the usual GLM model the mean is related to the linear predictor or GLM regression surface via the link function transformation $g(\mu_i) = \eta_i$, with $\eta_i = X_i\boldsymbol{\beta}$ (see McCullagh and Nelder 1983, chap. 2). The link function is assumed to be monotonic and differentiable and often transforms the mean value to its natural parameter in the exponential family. Further, it usually has the advantage of removing numerically awkward constraints on the regression function. A wide variety of distributions can be modeled using this approach, including linear regression, with its assumption of additive normal errors; logistic regression, where the response is a binomial random variable; and log-linear models, where the observations are from a Poisson distribution. Many more examples are given in McCullagh and Nelder (1983).

In the case when the predictor, $\eta$, has a known parametric form, there are many powerful techniques available to estimate and assess the fitted model. If the form of the dependence of the predictor on the covariates is not well known, however, it is important to have methods available to indicate where the fitted models fail to capture peculiarities in the data. To this end, we propose to use a nonparametric regression technique to estimate the GLM regression surface. Thus we let $\eta_i = f(t_i)$ and try to estimate $f$ nonparametrically. The results of this technique can then be compared with the usual parametric models to examine the adequacy of the fit and so forth. Like Silverman (1978), Anderson and Blair (1982), and Raynor and Bates (1983), we estimate $f$ by the penalized likelihood method of Good and Gaskins (1971). The nonparametric estimate of $f$ is the function, $\hat{f}_{n\lambda}$, that minimizes the penalized negative logarithm of the likelihood

$$l_{n\lambda}(f) = \sum_{i=1}^{n} l_i(y_i; f(t_i)) + n\lambda J(f), \tag{1.1}$$

where

$$l_i(y_i, f(t_i)) = (b(\theta_i) - y_i\theta_i)/a_i(\phi)$$

and $J(f)$ is a penalty functional designed to incorporate prior notions, such as smoothness, about the behavior of $f$. Similar

* Finbarr O'Sullivan is Assistant Professor, Department of Statistics, University of California, Berkeley, CA 94720. Brian S. Yandell is Assistant Professor, Department of Statistics and Horticulture, University of Wisconsin, Madison, WI 53706. William J. Raynor, Jr., is Group Leader, FIAT Statistics, Kimberly-Clark Corporation, Neenah, WI 54956. The authors acknowledge the advice and encouragement of Dennis Cox, Tom Leonard, Grace Wahba, Richard B. Shekelle, and Philippe Nicot. The research was supported in part by U.S. Department of Agriculture–Cooperative State Research Service Grant 511-100, U.S. Army Grant DAAG29-80-C-0041, and National Science Foundation Grant MCS-8403239.


Journal of the American Statistical Association
March 1986, Vol. 81, No. 393, Theory and Methods

estimators have also been proposed in a Bayesian context by Leonard (1978, 1982). The smoothing parameter, $\lambda$, controls the relative weighting of the penalty function in estimating $f$. When $\lambda \to 0$ the solution will be a function that maximizes the likelihood, whereas when $\lambda \to \infty$ the solution will be determined by the prior. With due care, a large value of $\lambda$ results in the usual GLM estimate for a suitable prior. In most practical settings $\lambda$ is unknown a priori. Below we provide a generalized cross-validation procedure for empirically assessing this parameter.

The article is organized in three parts. Section 2 characterizes the penalized likelihood estimator for fixed $\lambda$ with a generic Laplacian smoothing prior and discusses an algorithm for calculating the estimator. A generalized cross-validation score is introduced in Section 3, and in Section 4 Monte Carlo simulations are used to illustrate the small sample behavior of this score. Some practical examples are presented in Section 5, and the final section discusses our results, pointing out some further areas of applications for these techniques.

## 2. COMPUTATION OF THE PENALIZED LIKELIHOOD ESTIMATE

### 2.1 The Laplacian Penalty Functional

Let $f$ be a function defined on some design space $\Omega$, which is a subset of $R^d$. In the penalized likelihood framework, the estimation of $f$ is facilitated by employing a Laplacian penalty functional (see Meinguet 1979 and Wahba 1981). The Laplacian penalty functional, denoted by $J_m$, is defined by

$$J_m(f) = \int_\Omega \sum_{i_1,\ldots,i_m=1}^{d} \left[\frac{\partial^m f}{\partial x_{i_1}, \ldots, \partial x_{i_m}}\right]^2 d\underline{x}. \qquad (2.1)$$

Intuitively, $J_m(f)$ measures the visual smoothness of the function $f$. In one dimension, $J_m$ is just the $L_2$ norm of the $m$th derivative of $f$. The Laplacian terminology comes from the fact that if $f$ satisfies so-called "natural" boundary conditions, then integrating by parts, $J_m(f)$ can be written as

$$J_m(f) = (-1)^m \int_\Omega f\Delta^m f \, d\underline{x}, \qquad (2.2)$$

where $\Delta^m$ is the $m$-fold iterated Laplacian. Given $J_m$, we let $S$ be a space of real-valued functions whose derivatives of total order $m$ are square integrable. The penalized likelihood estimator, $\hat{f}_{n\lambda}$, is defined as the minimizer of the penalized likelihood over the space $S$. A simple characterization of the penalized likelihood estimate of $f$ corresponding to the Laplacian choice of penalty functional is available. The characterization says that the minimizer of the penalized likelihood, if it exists, must lie in a finite dimensional space of functions determined by the collection of variables $t_i$. The corresponding result for multivariate smoothing spline estimators can be found in Meinguet (1979) or Wahba (1981).

### 2.2 Characterization and Representation of the Estimate

We begin by looking at the one-dimensional situation. For simplicity let $\Omega = [0, 1]$ and let the penalty functional be $J_2(f) = \int_0^1 [\ddot{f}(t)]^2 \, dt$. Now $S$ is a real Hilbert space, with inner product $\langle \cdot, \cdot \rangle$ given by

$$\langle f, g\rangle = f(0)g(0) + \dot{f}(0)\dot{g}(0) + \int_0^1 \ddot{f}(t)\ddot{g}(t)\, dt. \qquad (2.3)$$

Recall that the penalized likelihood for generalized linear models was written as

$$l_{n\lambda}(f) = \sum_{i=1}^{n} l_i(y_i, f(t_i)) + n\lambda \int_0^1 [\ddot{f}(t)]^2 \, dt. \qquad (2.4)$$

Since evaluation is a continuous linear functional in $S$, by the Riesz representation theorem (see Rudin 1976) there exist functions $e_i$ in $S$, known as representers of evaluation, for which

$$f(t_i) = \langle f, e_i\rangle, \qquad i = 1, 2, \ldots, n. \qquad (2.5)$$

The functions $e_i$, which are piecewise cubic polynomials with continuous second derivative, are given by

$$\begin{aligned} e_i(s) &= 1 + t_i s + t_i s^2/2 - s^3/6, \qquad 0 \le s \le t_i \\ &= 1 + t_i s + t_i^2 s/2 - t_i^3/6, \qquad t_i < s \le 1. \end{aligned} \qquad (2.6)$$

In the Appendix we show that for all $f$ in $S$,

$$l_{n\lambda}(f) \ge l_{n\lambda}(f_1),$$

where $f_1$ is the projection of $f$ onto $S_n$. $S_n = S_0 \oplus \{e_i\}_{i=1}^n$ and $S_0$ is the collection of functions for which $J_2(f) = 0$—that is, linear functions. Thus the minimizer of the penalized likelihood lies in the finite dimensional subspace $S_n$ of $S$.

The elements of the one-dimensional characterization carry over to the multivariate situation. Here we suppose that the design matrix for the regression of the data on the set of polynomials of total degree less than $m$ is of full rank. It follows from this that for the general Laplacian penalty, $J_m(f)$, $S$ is a Hilbert space with inner product given by

$$\langle f, g\rangle = \langle f, g\rangle_0 + \langle f, g\rangle_m,$$

where

$$\langle f, g\rangle_0 = \sum_{i=1}^{n} f(t_i)g(t_i)$$

and

$$\langle f, g\rangle_m = \int_\Omega \sum_{i_1,\ldots,i_m=1}^{d} \frac{\partial^m f}{\partial x_{i_1}, \ldots, \partial x_{i_m}} \frac{\partial^m g}{\partial x_{i_1}, \ldots, \partial x_{i_m}} \, d\underline{x}.$$

Moreover, for $2m - d > 0$ evaluation is a continuous linear functional (see Meinguet 1979). Thus there exist functions $e_i$ in $S$ for which $\langle f, e_i\rangle = f(t_i)$ for all $f$ in $S$. Letting $S_0$ be the collection of functions that are annihilated by $J_m$ (i.e., polynomials of total degree less than $m$), it can be shown that the minimizer of the penalized likelihood over $S$ must now lie in $S_n = \text{span}\,\{e_i\}_{i=1}^n$.

When $\Omega = R^d$, analytic expressions for the representers of evaluation, $e_i$, are available and the functions in $S_n$ can be expressed, for some $\underline{c}$ in $R^n$ and $\underline{d}$ in $R^M$, as

$$\begin{aligned} f(t) &= \sum_{i=1}^{n} c_i \tau_i^{2m-d} \log\tau_i + \sum_{\nu=1}^{M} d_\nu \phi_\nu(t), \quad d \text{ odd} \\ &= \sum_{i=1}^{n} c_i \tau_i^{2m-d} + \sum_{\nu=1}^{M} d_\nu \phi_\nu(t), \qquad d \text{ even}; \end{aligned} \qquad (2.7)$$

$\tau_i = |t_i - t|$, the Euclidean distance between $t_i$ and $t$; $\phi_v$ are polynomials of total degree less than $m$; and $M = \binom{m+d-1}{d-1}$ (see Meinguet 1979 or Wendelberger 1981). If, however, $\Omega$ is a bounded region in $R^d$ and $d > 1$, then analytic expressions for the representers, $e_i$, are no longer available and in this case it is common to use $R^d$-representers to approximate the $\Omega$-representers. Thus one works with an approximate representation of the form (2.7) for the estimate. The computations in this article, like those in Wendelberger (1981) or Wahba and Wendelberger (1980), use this approach. It should be said, however, that this approach typically leads to full $O(n)$ matrices, and as a result the method becomes highly computer intensive with increasing sample size. An alternative approach is to use approximate representations for the functions in $S_n$ that have more desirable numerical properties. For instance, the use of tensor product $B$-splines (see de Boor 1978, chap. 18) leads to matrices with block-banded structure and remarkably efficient algorithms for computing the estimates.

## 2.3 Minimization of the Penalized Likelihood

Let $S_p = \text{span}_{1\le r\le p}\{h_r\}$ be a finite dimensional subspace of $S$ with dimension $p$. $S_p$ might be equal to $S_n$, but more generally we think of $S_p$ as some collection of functions that are used to approximate elements of $S_n$. Given $S_p$, we now describe how the minimizer of the penalized likelihood over $S_p$ can be carried out. The coefficients, $\boldsymbol{\beta}$, of the function in $S_p$ that minimizes $l_{n\lambda}$ are found by minimizing

$$l^p_{n\lambda}(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(y_i, X_i\boldsymbol{\beta}) + n\lambda J_m\left(\sum_r \beta_r h_r\right), \qquad (2.8)$$

where $X_{ir} = h_r(t_i)$. Dropping "$p$," $l^p_{n\lambda}(\boldsymbol{\beta})$ can be written as

$$l_{n\lambda}(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(y_i, X_i\boldsymbol{\beta}) + n\lambda\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}$$

$$= l_S(\boldsymbol{\beta}) + n\lambda\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}, \qquad (2.9)$$

where $\Sigma_{rs} = \langle h_r, h_s\rangle_m$. In Bayesian terms, $\Sigma^{-1}$ plays the role of a prior covariance for $\boldsymbol{\beta}$. Applying the Newton–Raphson minimization procedure with Fisher's scoring technique, a sequence of approximations, $\{\boldsymbol{\beta}^k\}$, to the minimizer of (2.9) are generated according to

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k - [I_n(\boldsymbol{\beta}^k) + 2n\lambda\Sigma]^{-1}\nabla l_{n\lambda}(\boldsymbol{\beta}^k), \qquad (2.10)$$

where

$$I_n(\boldsymbol{\beta}^k) = E\{[\partial^2 l_s(\boldsymbol{\beta}^k)/\partial\beta_r\partial\beta_s] \mid \eta_i = X_i\boldsymbol{\beta}^k\}$$

estimates the sample Fisher information matrix. The derivation in McCullagh and Nelder (1983, pp. 32–33) can be modified to obtain that $\boldsymbol{\beta}^{k+1}$ is equivalently the minimizer of

$$\sum_{i=1}^{n} w_i[z_i - X_i\boldsymbol{\beta}']^2 + n\lambda\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}, \qquad (2.11)$$

where

$$z_i = X_i\boldsymbol{\beta}^k + (y_i - \mu_i)\frac{\partial g(\mu_i)}{\partial\mu_i}, \qquad w_i^{-1} = 2V_i\left[\frac{\partial g(\mu_i)}{\partial\mu_i}\right]^2.$$

The mean and variance functions, $\mu_i$ and $V_i$, are both evaluated as though $X_i\boldsymbol{\beta}^k$ were the true value of $f(t_i)$. It follows that the Newton–Raphson iteration, with Fisher's scoring technique, is equivalent to an iteratively reweighted "ridge" regression procedure.

*Remark.* The existence and uniqueness of minimizers of $l_{n\lambda}$ is a separate issue. If the log-likelihood is convex and has a unique minimizer over $S_0$, then it can be shown that there is a unique minimizer of $l_{n\lambda}$ over $S$ (see O'Sullivan 1983). In other words, when the penalized likelihood is convex, then the existence of a unique maximum likelihood estimator (MLE) on $S_0$ guarantees the existence of the penalized likelihood estimator over the whole space.

# 3. ASSESSING THE SMOOTHING PARAMETER

## 3.1 Cross-Validation Scores for Smoothing Spline Estimators

The penalized likelihood estimator of the GLM regression surface depends on the value assigned to the smoothing parameter $\lambda$. For the Laplacian penalty, $J_m$, small values of the smoothing parameter produce rougher-looking estimates. Although a strong case can be made for visually inspecting estimators corresponding to a variety of $\lambda$ values, there are situations in which an automatic procedure for isolating a single "ball park" value is convenient. One of the standard procedures for assessing model adequacy is cross-validation. The use of cross-validation to choose the smoothing parameter in smoothing spline-type estimators was developed by Wahba and Wold (1975) and Craven and Wahba (1979). To describe these methods, consider the usual smoothing spline setup in which we observe a smooth function with error

$$y_i = f(t_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

The $\varepsilon_i$'s are mean zero uncorrelated, and let $\text{var}(\varepsilon_i) = w_i^{-1}\phi$ with $\phi$ unknown. The smoothing spline estimator, $\hat{f}_{n\lambda}$, is the minimizer over the space of functions $S$ of

$$\frac{1}{n}\sum_{i=1}^{n} w_i[y_i - f(t_i)]^2 + \lambda J_m(f). \qquad (3.1)$$

Expressing $\hat{f}_{n\lambda}$ as a linear combination of representers of evaluation, we have $\hat{f}_{n\lambda} = \Sigma_{j=1}^{p}\beta_j e_j$, with

$$\hat{\boldsymbol{\beta}} = [X'WX + n\lambda\Sigma]^{-1}X'W\mathbf{y},$$

where $X_{ij} = \eta_j(t_i)$, $\Sigma_{jk} = \langle\eta_j, \eta_k\rangle_m$, and $W = \text{diag}(w_1, \ldots, w_n)$. Thus the vector of predictions, $\hat{\mathbf{y}} = (\hat{f}_{n\lambda}(t_1), \ldots, \hat{f}_{n\lambda}(t_n))$, can be written as

$$\hat{\mathbf{y}} = X[X'WX + n\lambda\Sigma]^{-1}X'W\mathbf{y} \equiv H(\lambda)\mathbf{y}, \qquad (3.2)$$

where $H(\lambda)$ is the smoothing spline hat matrix. With this the ordinary cross-validation score of Wahba and Wold (1975) is given by

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n} w_i\left(\frac{y_i - \hat{y}_i}{1 - H_{ii}(\lambda)}\right)^2. \qquad (3.3)$$

The generalized cross-validation (GCV) score, proposed by Craven and Wahba (1979), replaces the individual leverages,

$H_{ii}(\lambda)$, by their average value. This score is usually written as

$$V(\lambda) = \frac{\sum_{i=1}^{n} w_i[y_i - \hat{y}_i]^2}{[n - \text{trace } H(\lambda)]^2}. \quad (3.4)$$

Interestingly, in the usual linear model framework, $V_0(\lambda)$ is equivalent to Allen's prediction sum of squares (PRESS), and $V(\lambda)$ becomes the residual mean square divided by the degrees of freedom for error, which was proposed by Anscombe (1967) as a variable selection criterion (see Mosteller and Tukey 1977, pp. 385–387).

Asymptotically, the minimizer of the GCV also minimizes the weighted mean squared error between the estimate and the truth; that is,

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n} w_i[\hat{f}_{n\lambda}(t_i) - f(t_i)]^2. \quad (3.5)$$

More precisely, a slight modification to the result of Craven and Wahba (1979) gives the following: If $\lambda_n^*$ is the minimizer of $EV(\lambda)$, then

$$ER(\lambda_n^*)/\min_{\lambda} ER(\lambda) \to 0 \quad \text{as} \quad n \to \infty.$$

Further results are given in Li (1983) and Speckman (1982).

### 3.2 Extension of the Cross-Validation Score

The GCV score in (3.4) has the form of a ratio of the residual sum of squares divided by the square of "effective" degrees of freedom for error ($[n - \text{trace } H(\lambda)]^2$ term). An analog of the residual sum of squares in the GLM context is the generalized Pearson $\chi^2$ statistic (see McCullagh and Nelder 1983, p. 26). Replacing the residual sum of squares by the Pearson $\chi^2$ leads to the GCV score

$$V(\lambda) = \left[\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2/\hat{V}_i\right] \Big/ [n - \text{trace } H(\lambda)]^2, \quad (3.6)$$

where $\hat{\mu}_i$, $\hat{V}_i$, and the linearized hat matrix, $H(\lambda) = X[X'WX + 2n\lambda\Sigma]^{-1}X'W$, are all computed at the final stage of the iteratively reweighted algorithm in Section 2.

The GCV score can be given a more theoretical motivation, and from this we are led to conjecture that the $\lambda$ minimizing the GCV score will approximately minimize the weighted mean squared error criterion

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n} w_i[\hat{f}_{n\lambda}(t_i) - f(t_i)]^2, \quad (3.7)$$

where $w_i = V_i^{-1}[\partial g(\mu_i)/\partial\mu_i]^2$. $R(\lambda)$ is a mean squared error criterion weighted by the expected Fisher information (for the predictor) at the design points. Thus $R(\lambda)$ reflects the attitude that the experimenter is most interested in understanding the regression surface, $f$, in regions where the experiment has been made most informative.

### 3.3 Theoretical Motivation for the GCV Score

Cox and O'Sullivan (1985) gave results on the asymptotic behavior of general penalized likelihood estimators. From these results it can be shown that our GLM regression surface smoother is first-order asymptotically equivalent to $\bar{f}_{n\lambda}$, where $\bar{f}_{n\lambda}$ is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n} \bar{w}_i[\bar{z}_i - f(t_i)]^2 + \lambda J_m(f), \quad (3.8)$$

$$\bar{z}_i = f_\lambda(t_i) + (y_i - \bar{\mu}_i)[\partial g(\bar{\mu}_i)/\partial\mu_i],$$

and

$$\bar{w}_i^{-1} = 2\bar{V}_i[\partial g(\bar{\mu}_i)/\partial\mu_i]^2.$$

The mean and variance functions are now evaluated at $f_\lambda$, which is a root of the variational equation corresponding to the limiting version of the penalized likelihood (i.e., the sample penalized likelihood, $l_{n\lambda}$, has a limit, $l_\lambda$, as $n$ tends to infinity, and $f_\lambda$ is a solution to $\nabla l_\lambda = 0$).

The asymptotic equivalence between $\hat{f}_{n\lambda}$ and $\bar{f}_{n\lambda}$ holds for all $\lambda$'s in some interval $[\lambda_n, \lambda_0]$ with $\lambda_n \to 0$. The rate of convergence of $\lambda_n$ is such that $\lambda_n \ll \lambda_n^s$, where $\lambda_n^s$ minimizes the expected value of the weighted mean squared error, $R(\lambda)$. Thus the interval $[\lambda_n, \lambda_0]$ contains all of the most interesting values of $\lambda$. Given that $\bar{f}_{n\lambda}$ is defined as a smoothing spline-type estimator of the form (3.1), we are led to consider the corresponding GCV score in (3.4):

$$\bar{V}(\lambda) = \sum_{i=1}^{n} \bar{w}_i[\bar{z}_i - \bar{f}_{n\lambda}(ti)]^2 \Big/ [n \text{ trace } \bar{H}(\lambda)]^2, \quad (3.9)$$

where $\bar{H}(\lambda)$ is the relevant hat matrix. With some analysis, we can show that $\bar{V}(\lambda)$ has the following property: If $\bar{\lambda}_n$ is the minimizer in $[\lambda_n, \lambda_0]$ of $E\bar{V}(\lambda)$, then

$$ER(\bar{\lambda}_n)/\min_{[\lambda_n,\lambda_0]} ER(\lambda) \to 0 \quad \text{as} \quad n \to \infty.$$

Thus, in the Craven and Wahba sense, the minimizer of $\bar{V}(\lambda)$ minimizes $R(\lambda)$ in large sample sizes. Unfortunately, because of the dependence on $f_\lambda$, $\bar{V}(\lambda)$ is uncomputable; however, we can reasonably approximate $\bar{V}(\lambda)$ by replacing $f_\lambda$ by its sample analog, $\hat{f}_{n\lambda}$. Carrying out this substitution we obtain

$$V(\lambda) = \left[\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2/\hat{V}_i\right] \Big/ [n - \text{trace } H(\lambda)]^2, \quad (3.10)$$

which is the form given in (3.6).

Given that $\bar{V}(\lambda)$ is so closely related to $V(\lambda)$, one would suspect that $V(\lambda)$ ought to do a good job of picking the minimizer of the weighted mean squared error loss, and our simulations bear this out.

## 4. MONTE CARLO EXPERIMENTS

Simulations with binomial and Poisson data in one and two dimensions were carried out for a moderate sample size (100 distinct data points). Only 25 Monte Carlo runs were performed, but, even so, one can see that the modified GCV function (3.6) does a good job of picking the minimizer of the weighted mean squared error (WMSE). The efficacy of the GCV estimate of $\lambda$ is measured by computing the ratio of the minimum WMSE to the value of WMSE attained at the GCV estimate $\hat{\lambda}$. Table 1 summarizes these efficacy comparisons. In nearly all cases the efficacy was close to one.

*Table 1. Efficacies of Generalized Cross-Validation, Relative to Weighted Mean Squared Error*

| Model | Mean | Median | Minimum | Maximum |
|---|---|---|---|---|
| Binomial | .86 | .95 | .52 | 1.00 |
| Poisson (1-D) | .82 | .87 | .19 | 1.00 |
| Poisson (2-D) | .93 | 1.00 | .62 | 1.00 |

Random numbers used in the simulations were obtained using the pseudo-random number generator of Marsaglia (see Gross 1978), which is publicly available as part of the Portable Statistical Library from Bell Laboratories' Computing Information Service (in Murray Hill, New Jersey). This library is incorporated in the *S* system (Becker and Chambers 1984), which served as the computing environment for our analyses. The computation of estimates was done using the iteratively reweighted ridge regression procedure described in Section 2. The regression algorithm, though similar to that of Wendelberger (1981), avoided singular value decompositions and found $\hat{\lambda}$ by evaluating the GCV on a course grid (in log scale). The source code, as either an *S* function or a stand-alone rational FORTRAN program, can be obtained from us. Users will also need the Linpack subroutine package (Dongarra, Bunch, Moler, and Stewart 1979).

## 4.1 Binomial Simulations

The risk curve used in the binomial simulation has a long plateau at .25, increases like a logistic function until it reaches its maximum at .75, and levels off. This type of risk function could occur in a situation in which there was a baseline hazard that increased with the levels of the independent variable until it reached a maximum observable level of risk. A motivation for looking at this kind of risk comes from epidemiological data discussed in Section 5.

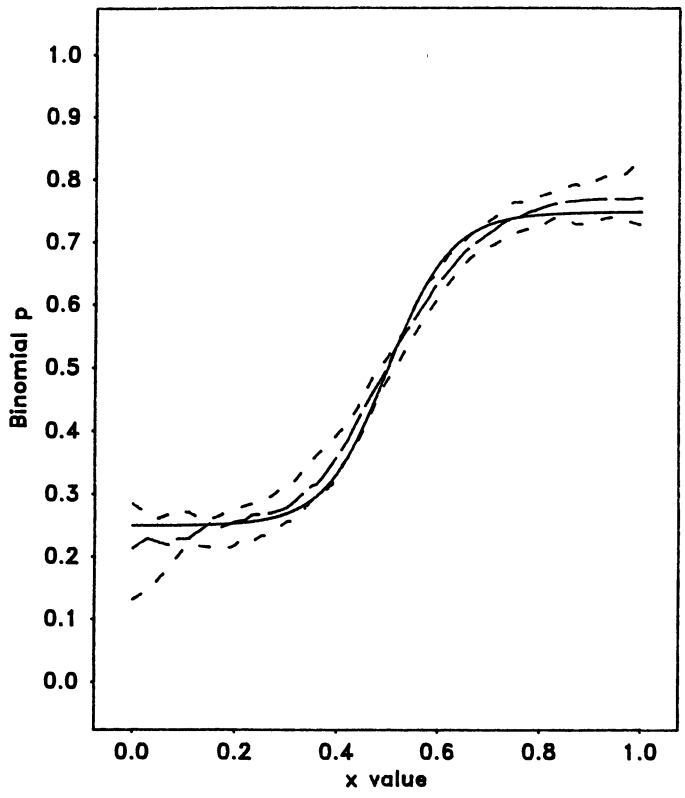


*Figure 1. Binomial Test Function and Quartiles of Simulation Estimates: ——, true function; — —, median of estimates (25 simulations); - - -, lower and upper quartiles (25 simulations).*

In each simulation observations were generated according to

$$y_i \sim B(5, p(t_i)), \qquad i = 1, \ldots, 100, \tag{4.1}$$

where $t_i$ are equispaced and the risk, $p(\cdot)$, is given in Figure 1. Given these data, the smoothing procedure is used to estimate the underlying logit curve; that is,

$$f(t) = \log[p(t)/(1 - p(t))], \tag{4.2}$$

with the smoothing parameter being chosen by minimizing the GCV score. The minimum of $V(\lambda)$ and $R(\lambda)$, the GCV and WMSE, were found by computing the functions at values of $\log(\lambda)$ from 6 to -16 in steps of 2. Both functions tend to attain their minima at similar values of $\lambda$.

The median estimate for each $t$ value and the outer quartiles are shown with the true function in Figure 1. The smoothed estimates do quite well in finding the general shape of the curve, although they tend to overshoot in areas of high curvature, such as the beginning and end of the plateau areas. Even there, the behavior of the smoothed estimate does a credible job of indicating the trend of the underlying curve.

## 4.2 Poisson Simulations

A second set of simulations with Poisson data was performed. The motivation for this came from work with plant pathologists at the University of Wisconsin who were interested in describing the distribution of virus activity in potato fields. See Section 5 for more discussion.

Simulations were carried out in one and two dimensions. The Poisson rate parameter, $\mu(\cdot)$, ranged from 1 at the edge of the "field" to 20 at the center. The one-dimensional runs had 100 equispaced data points whereas the two-dimensional runs had 100 points on a regular 10 × 10 grid. For each simulated data set, the GLM smoother was used to estimate the logarithm of the Poisson rate

$$f(t) = \log \mu(t), \tag{4.3}$$

with the smoothing parameter again being chosen by minimizing the GCV score. (Here $\lambda$ was stepped through in natural logs from 5 to -16.)

Figure 2 shows the true function and the median of the 25 simulations, along with the outer quartiles, for the one-dimensional runs. The same problem of overshooting occurs as in the binomial, presumably due to the high curvature at certain points and to low expected counts at the boundary.

## 4.3 Efficacy

For both the binomial and Poisson examples, the minimizer, $\hat{\lambda}$, of the GCV function on the observed lattice of values has a weighted mean squared error close to the minimum possible value. The efficacy of GCV, defined as the ratio of the minimum value of WMSE to its value at $\hat{\lambda}$, ranged in the binomial simulation from .52 to 1.00, with a mean of .86 and a median of .95. The Poisson simulations showed the same general pattern.

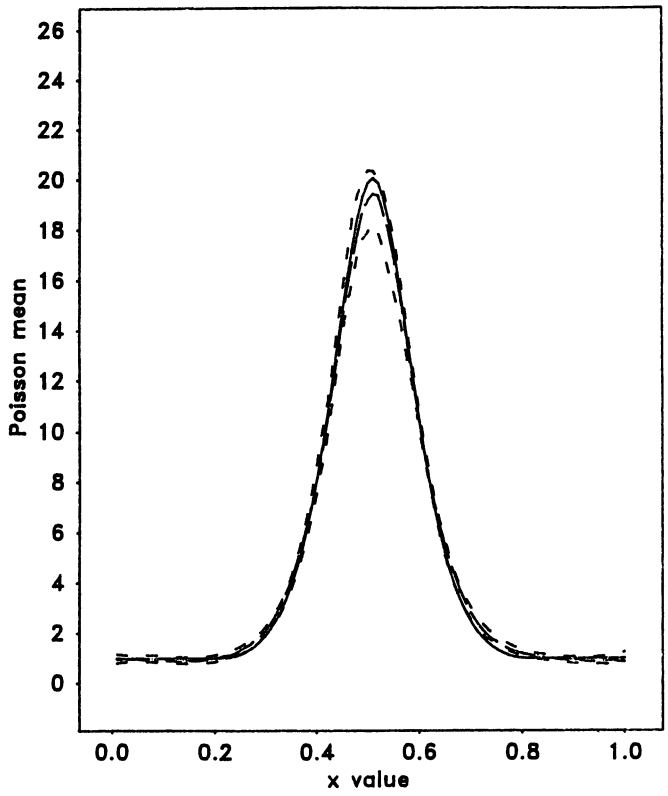


*Figure 2. Poisson Test Function and Quartiles of Simulation Estimates: ——, true function; — —, median of estimates (25 simulations); - - -, lower and upper quartiles (25 simulations).*

The efficacies of $\hat{\lambda}$ for these simulations are summarized in Table 1. The table suggests that the GCV function might be used to guide the analyst toward a reasonable starting value for the smoothing parameter. A careful researcher would do well to try some smaller $\lambda$ values in the neighborhood of the GCV minimizer in order to get a better understanding of the regression surface.

## 5. SOME PRACTICAL EXAMPLES

### 5.1 Heart Disease and Concomitant Variables

This analysis concerns 19-year death rates in 1,665 men from the Western Electric Health Study (Raynor, Shekelle, Rossof, Maliza, and Paul 1981). The data include men who were alive at the end of the follow-up period and those who had died from heart disease. Participants dying from other causes were excluded. A natural way to model these data is via the standard logistic regression approach, which allows one to estimate the 19-year probability of death after adjusting for concomitant factors. The results of studies such as these help determine the risks associated with various factors and are used by physicians and others to provide medical advice to their patients. We fit a logistic regression using two factors, diastolic blood pressure (DBP) and cholesterol, and compare this model with a smoothed logistic regression in the same variables (see Figure 3). (A first guess for the smoothing parameter was obtained by the GCV, although a slightly smaller value of $\lambda$ was used for the figure.) The most outstanding difference between the two surfaces is

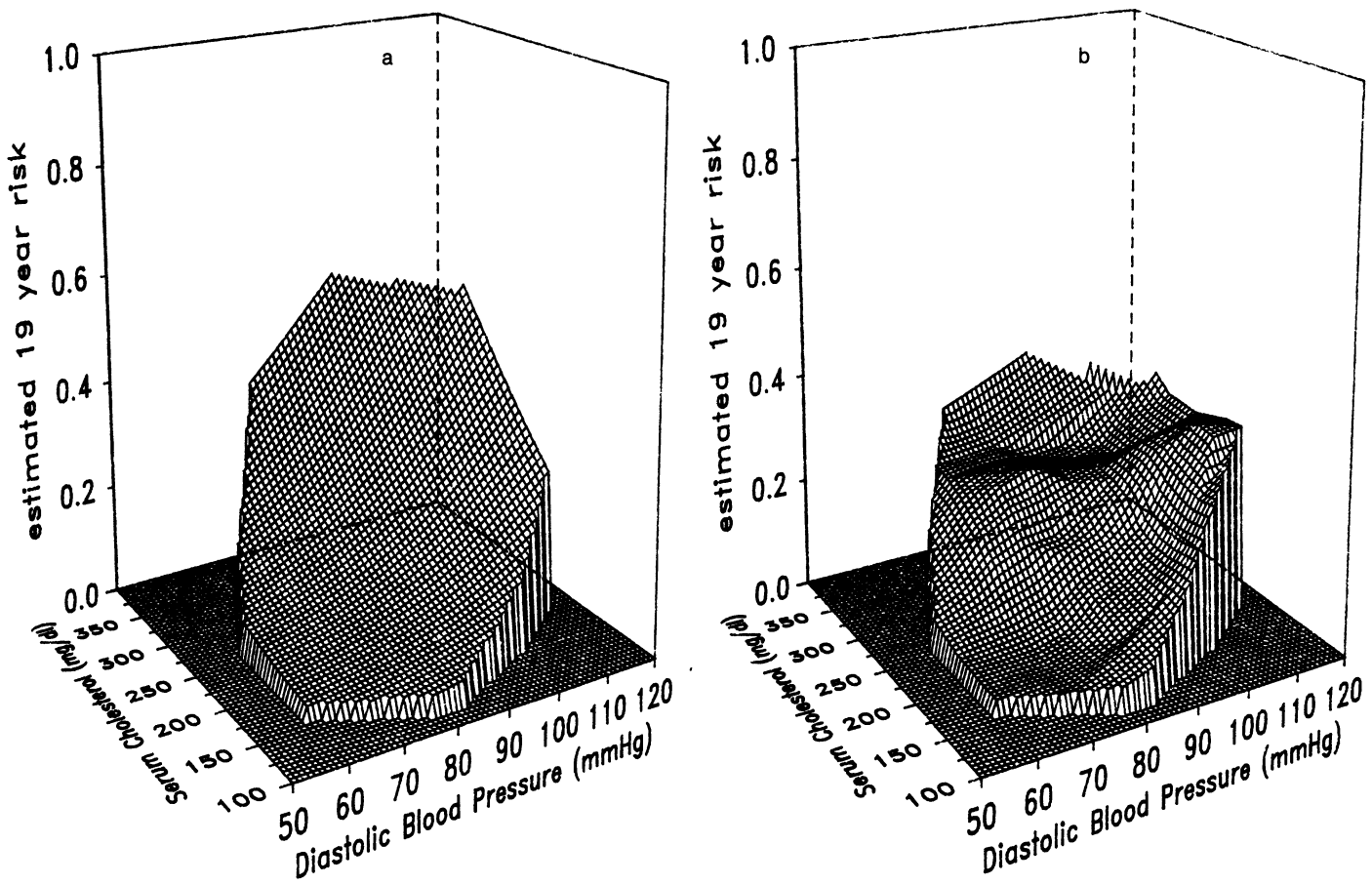


*Figure 3. Risk Estimates: (a) from linear generalized linear model; (b) smoothed.*

the presence of two plateaus in the smoothed surface. The logistic model exhibits a constant rate of increase in logit risk, completely missing the plateaus. If these plateaus can be verified by other analyses, they would suggest that with some combinations of risk factors a slight modification in status can result in a dramatic shift in risk, whereas with other combinations only a change in a certain direction results in any appreciable change in risk. For example, a person at the far edge of the upper plateau needs a large change in DBP and cholesterol in order to show an appreciable change in risk, whereas a person at the near edge can have a substantial lowering of risk through a small change in these factors.

### 5.2 Potato Early Dying Disease in a Potato Field

*Verticilium* is a virus that invades potato roots and seems to induce the "potato early dying" disease. Farmers and researchers would like to determine the spatial pattern of virus in the soil in order to decide whether and where to fumigate. Although little is known about the movement of the virus, it seems to invade potato plants in the spring and remains in the decaying stem material in the fall, when the farmer plows for the next season. The virus can be counted by diluting a soil sample and plating the solution on a petri dish in the laboratory. The petri dish is left for several days, after which the visible colonies are counted. The counts seem to be distributed Poisson-wise. A regular grid of soil samples and corresponding dish counts yields a two-dimensional grid that can be fit by smoothing the logarithm of the underlying Poisson rate to obtain a surface map of virus activity in the field. Such a surface was constructed (see Figure 4), based on a 20 × 20 grid with 30 meters between centers. Data were collected by Philippe Nicot of the Plant Pathology Department at the University of Wisconsin as part of his dissertation research. The large ridges in this picture occur along the rows of the field. This suggests that the virus is spread primarily along rows as the farmer tills his field and diffusion across rows is rather slow, lending some support to the beliefs of plant pathologists. Again, more careful experimentation would be required to validate these findings.

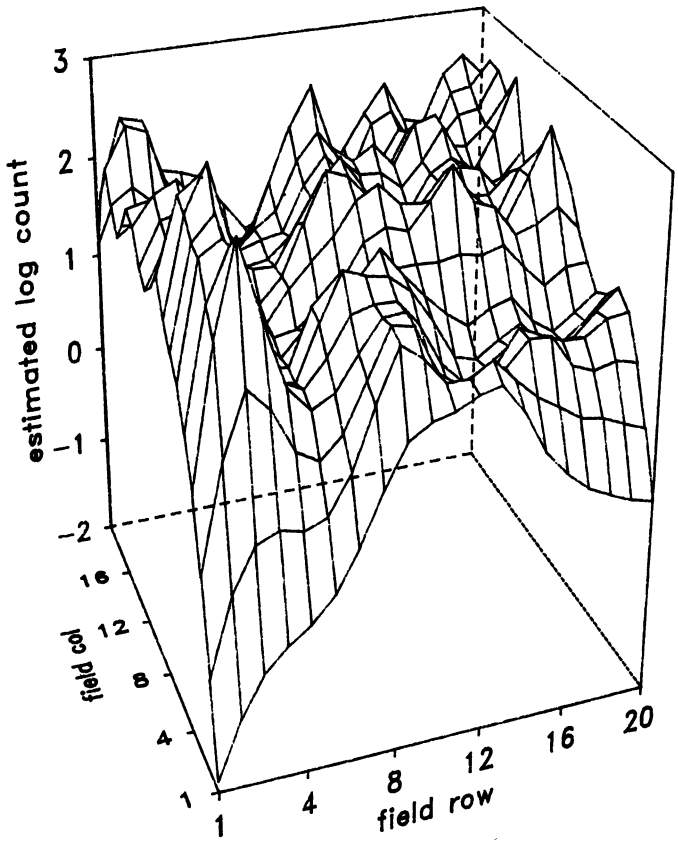


*Figure 4. Map of Smoothed Verticulum Counts.*

## 6. DISCUSSION

The goal of this article has been to develop a methodology for obtaining and assessing smooth pictures for complicated data, such as those from binomial and Poisson responses. Many authors have recognized that the penalized likelihood formulism can be applied in a natural way to smoothing logistic surfaces (see Anderson and Blair 1982, Raynor and Bates 1983, and Silverman 1978). We have shown that this extends naturally to the class of GLM models. In addition a generalized cross-validation method can be used to assess the appropriate degree of smoothing in such estimates. A number of variations on the cross-validation score are possible. For example, a plausible ordinary cross-validation assessment might be the sum of squares of the individual Pearson residuals divided by $(1 - H_{ii}(\lambda))$. Deviance or Anscombe residuals could also be used.

The cross-validation score is motivated by a linearization of the penalized likelihood estimator. Perhaps such linearizations can be employed more generally to obtain cross-validation scores for other penalized likelihood estimators. Of course, an alternative technique, which we have not explored, would be to apply a straightforward leave-one-out cross-validation directly to the log-likelihood. In a density estimation this generates the well-known Kullbach–Leibler-type estimator of Boneva, Kendall, and Stefanov (1971). (Also see Bowman 1984 and Bowman, Hall, and Titterington 1984.)

Our limited experience with our proposed smoothing technique suggests that the cross-validated regression surface loses sight of some structure, especially where there is high localized curvature. This is a natural consequence of both the loss and the penalty function used. In practice, it may be wise to examine the model with the cross-validated fit and with a slightly rougher fit. The results of the smoothing can then be examined intrinsically or compared with other models to determine features that are best reflected in a parametric model and those special features that would be missed by standard approaches. Confidence intervals for the fitted surfaces could probably be obtained by the methods introduced by Wahba (1983). In the examples presented in the last section, the smoothed models suggest that there are features in the data that would probably be missed by a "standard" analysis. The results need to be confirmed by other analyses, but the techniques certainly provide a flexible exploratory and diagnostic tool for use by data analysts.

Finally, it would be interesting to extend the methodology in various ways. For instance, McCullagh and Nelder (1983) demonstrated that the GLM method could be used to model censored survival data, using either a parametric model, such as a Weibull function, or a semiparametric model, such as proportional hazards. The techniques outlined in the article

might be useful in fitting such hazards. They may also apply to assessing the lack of fit of autoregressive moving average models.

## APPENDIX

*Theorem A.* Let $S_0$ be the collection of functions in $S$ for which $J_2(f) = 0$ (i.e., liner functions) and $S_n = S_0 \oplus \{e_i\}_{i=1}^n$. Then for all $f$ in $S$, $l_{n\lambda}(f) \geq l_{n\lambda}(f_1)$, where $f_1$ is the projection of $f$ onto $S_n$.

*Proof.* Any $f$ in $S$ can be written as $f = f_1 + f_2$, with $f_1$ in $S_n$ and $f_2$ in $S_n$, the orthogonal complement of $S_n$. By definition of $e_i$ and $S_n$,

$$\begin{aligned} f(t_i) &= \langle f, e_i \rangle \\ &= \langle f_1, e_i \rangle + \langle f_2, e_i \rangle \\ &= \langle f_1, e_i \rangle \\ &= f_1(t_i). \end{aligned}$$

Thus

$$\sum_{i=1}^{n} l_i(y_i, f(t_i)) = \sum_{i=1}^{n} l_i(y_i, f_1(t_i)). \tag{A.1}$$

Now $J_1(f) = \langle f - f_0, f - f_0 \rangle$, where $f_0$ is the component of $f$ lying in $S_0$. Therefore, by straightforward algebra, $J_2(f_1 + f_2) = J_2(f_1) + J_2(f_2)$. Combining this with (A.1), we have

$$\begin{aligned} l_{n\lambda}(f) &= \sum_{i=1}^{n} l_i(y_i, f_1(t_i)) + n\lambda J_2(f_1) + n\lambda J_2(f_2) \\ &\geq l_{n\lambda}(f_1), \end{aligned}$$

which proves the result.

[*Received March 1984. Revised August 1985.*]

## REFERENCES


Anderson, J. A., and Blair, V. (1982), "Penalized Maximum Likelihood Estimation in Logistic Regression and Discrimination," *Biometrika*, 69, 123–136.

Anscombe, F. J. (1967), "Topics in the Investigation of Liner Relations Fitted by Least Squares," *Journal of the Royal Statistical Society*, Ser. B, 29, 1–52.

Becker, R. A., and Chambers, J. M. (1984), *S—An Interactive Environment for Data Analysis and Graphics*, Belmont, CA: Wadsworth.

Boneva, L. T., Kendall, D. G., and Stefanov, I. (1971), "Spline Transformations: Three New Diagnostic Aids for the Data Analyst" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 30, 1–70.

Bowman, A. W. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika*, 71, 353–360.

Bowman, A. W., Hall, P., and Titterington, D. M. (1984), "Cross-Validation in Nonparametric Estimation of Probabilities and Probability Densities," *Biometrika*, 71, 341–351.

Cox, D. D., and O'Sullivan, F. (1985), "Analysis of Penalized Likelihood-Type Estimators With Application to Generalized Smoothing in Sobolev Spaces," unpublished manuscript.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, 31, 377–403.

DeBoor, C. (1978), *A Practical Guide to B-Splines*, New York: Springer-Verlag.

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

Friedman, J., and Stuetzle, W. (1982), "Smoothing of Scatterplots," Technical Report Orion 3, Stanford University, Statistics Dept.

Good, I. J., and Gaskins, R. A. (1971), "Non-parametric Roughness Penalties for Probability Densities," *Biometrika*, 58, 255–277.

Gross, A. M. (1978), "Portable Random Number Generation," technical report, AT & T Bell Laboratories, Murray Hill, NJ 07974.

Hastie, T. J., and Tibshirani, R. J. (1984), "Generalized Additive Models," Technical Report 98, Stanford University, Division of Biostatistics.

Leonard, T. (1978), "Density Estimation, Stochastic Processes, and Prior Information," *Journal of the Royal Statistical Society*, Ser. B, 40, 113–146.

——— (1982), "An Empirical Bayesian Approach to the Smooth Estimation of Unknown Functions," Technical Report 2339, University of Wisconsin, Mathematics Research Center.

Li, K. C. (1983), "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross Validation," Technical Report 83-34, Purdue University, Statistics Dept.

McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models*, London: Chapman & Hall.

Meinguet, J. (1979), "Multivariate Interpolation at Arbitrary Knots Made Simple," *Journal of Applied Mathematics and Physics*, 30, 292–304.

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley.

Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalised Linear Interactive Models," *Journal of the Royal Statistical Society*, Ser. A, 135, 370.

O'Sullivan, F. (1983), "On the Analysis of Some Penalized Likelihood Estimation Schemes," Technical Report 726, University of Wisconsin, Dept. of Statistics.

Raynor, W. J., and Bates, D. M. (1983), "Spline Smoothing of Binary Regressions in Large Data Sets," University of Wisconsin, Dept. of Statistics.

Raynor, W. J., Shekelle, R. B., Rossof, A. H., Maliza, C., and Paul, O. (1981), "High Blood Pressure and 17-Year Cancer Mortality in the Western Electric Health Study," *American Journal of Epidemiology*, 113, 371–377.

Rudin, W. (1976), *Principles of Mathematical Analysis*, New York: McGraw-Hill.

Silverman, B. W. (1978), "Density Ratios, Empirical Likelihood and Cot Death," *Applied Statistics*, 27, 26–33.

Speckman, P. (1982), "Efficient Nonparametric Regression With Cross-Validated Smoothing Splines," Technical Report 45, University of Missouri, Statistics Dept.

Wahba, G. (1981), "Spline Interpolation and Smoothing on the Sphere," *SIAM Journal on Scientific and Statistical Computing*, 2, 5–16.

——— (1983), "Bayesian Confidence Intervals for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society*, 45, 133–150.

Wahba, G., and Wendelberger, J. (1980), "Some New Mathematical Methods for Variational Objective Analysis Using Splines and Cross Validation," *Monthly Weather Review*, 108, 1122–1143.

Wahba, G., and Wold, S. (1975), "A Completely Automatic French Curve: Fitting Spline Functions by Cross Validation," *Communications in Statistics*, 4, 1–17.

Wendelberger, J. (1981), "The Computation of Laplacian Smoothing Splines With Examples," Technical Report 648, University of Wisconsin, Dept. of Statistics.