UNIVERSITY OF WISCONSIN
DEPARTMENT OF BIOSTATISTICS

Technical Report #76

December 1992

APPLICATION OF INTERACTION SMOOTHING SPLINE TO GENERALIZED LINEAR MODELS

Richard Chappell
Yuedong Wang

Departments of Biostatistics and Statistics

MADISON, WISCONSIN
Application of Interaction Smoothing Spline to Generalized Linear Models

Rick Chappell
Yuedong Wang

Department of Statistics, University of Wisconsin-Madison
Abstract

We consider the penalized likelihood method for estimating nonparametric regression functions in generalized linear models. Interaction splines are used to estimate these functions. Generalized cross-validation or unbiased risk procedure is used to empirically assess the amount of smoothing at each Newton-Raphson iteration. The algorithm is applied to a real data set using binary outcomes. Several simulations are conducted to examine the algorithm and to compare the two methods of assessing smoothness.

KEY WORDS: Generalized linear model; Interaction smoothing spline; Newton-Raphson iteration; GCV estimate; Unbiased risk estimate; Symmetrized Kullback-Leibler discrepancy; Penalized likelihood method.

1 Introduction

Nelder and Wedderburn (1972) introduced a collection of statistical regression models known as generalized linear models (GLIM's) for the analysis of data from exponential families (McCullagh and Nelder (1989)). The applications of GLIM's to practical data analysis has greatly expanded with the subsequent development of theory and computer software. As the popularity of these methods has increased, so has the need for more sophisticated model building and diagnostic checking techniques.

Suppose data have the form \((y_i, \mathbf{x}_i), i = 1, 2, \cdots, n\), where \(y_i\) are independent observations, each from an exponential family distribution depending on the covariate \(\mathbf{x}_i\) (possibly vector-valued). The probability density of \(y_i\) is

\[
f(y_i; \eta_i, \phi) = \exp((y_i g(\eta_i) - b(\eta_i))/a(\phi) + c(y_i, \phi)),
\]

where \(\eta_i = \eta(\mathbf{x}_i)\) is the parameter of interest and \(g(\eta_i)\) is a monotone transformation of \(\eta_i\) known as the canonical parameter. \(\phi\) is an unknown scale parameter. The loglikelihood given \(y_i, \mathbf{x}_i\) is

\[
l_i(\eta_i) = \log f(y_i; \eta_i, \phi) = (y_i g(\eta_i) - b(\eta_i))/a(\phi) + c(y_i, \phi)
\]
The purpose is to investigate the global relationship between \( \eta \) and \( x \). In the usual GLIM model, it is assumed that \( \eta(x) = x^T \beta \), where \( \beta \) consists of parameters to be estimated. The maximum likelihood estimate of \( \beta \), which is equivalent to minimizing \(-\sum_{i=1}^{n} l_i(x_i^T \beta)\), can be obtained by Newton-Raphson procedure. To achieve greater flexibility, many authors proposed nonparametric regression models to relax the rigid constraint: \( \eta(x) = x^T \beta \). In particular, the penalized likelihood smoothing spline regression (Green and Yandell (1985), O'Sullivan, Yandell and Raynor (1986), Gu (1990)) assumes only that \( \eta(x) \) is smooth but imposes a “roughness” penalty \( J(\eta) \), say, to discourage arbitrariness of the estimator. It estimates \( \eta(x) \) with the value which satisfies

\[
\min \left\{ -\sum_{i=1}^{n} l_i(\eta_i) + \lambda J(\eta) \right\},
\]

where \( \lambda > 0 \) is called the smoothing parameter. It controls the trade-off between goodness of fit of the estimator and “smoothness”. Generalized cross-validation (GCV) or unbiased risk estimate of \( \lambda \) for each iteration of Newton-Raphson has been proposed by Yandell (1988) and Gu (1990, 1991).

We consider fitting a multivariate model to response based on ANOVA-like decompositions of functions of several variables,

\[
\eta(x_1, \cdots, x_p) = C + \sum_{i=1}^{p} \eta_i(x_i) + \sum_{i<j}^{p} \eta_{ij}(x_i, x_j) + \cdots + \eta_{1\cdots p}(x_1, \cdots, x_p),
\]

where \( C \) is a constant, \( \eta_i \)'s are the main effects, \( \eta_{ij} \) are i-j 2-interactions, and so on. This decomposition provides a powerful tool for modeling and hypothesis testing (Wahba (1990)). By deleting higher-order interactions, we can get less flexible but more “estimable” models. Especially, the additive model, which contains only main effects, was studied by Hastie and Tibshirani (1990). We will fit (some of) the lower order terms in (2), via a penalized likelihood criterion similar to (1), with multiple smoothing parameters. The exact function to be minimized will be described in Section 2. The Newton-Raphson procedure will be used to estimate \( \eta(x) \). The GCV or unbiased risk method will be used to update smoothing parameters for each iteration.
2 Interaction Smoothing Splines

We will assume that $\eta$ is in a reproducing kernel Hilbert space (RKHS) $\mathcal{M}$ of smooth real-valued functions on $X = X^{(1)} \otimes \cdots \otimes X^{(p)}$. A larger Hilbert space $\mathcal{G}$, in which elements have ANOVA-like decomposition (2), will contain $\eta$. This ANOVA structure will be used to balance the lack of flexibility and the cause of dimensionality by fitting only (some of) the lower order terms in (2). So our model space $\mathcal{M}$ is a subspace of $\mathcal{G}$ containing only the included terms.

For the special case $X^{(i)} = [0, 1], i = 1, \cdots, p$, we will take the component space on $[0, 1]$ as the RKHS

$$W_2^m = \{ f : f^{(v)} \text{ a\ö. cont.}, \; v = 0, \cdots, m - 1, \; \int_0^1 (f^{(m)})^2 < \infty \}$$

with norm

$$f^2 = \sum_{v=0}^{m-1} (\int_0^1 f^{(v)})^2 + \int_0^1 (f^{(m)})^2.$$  \hfill (4)

We decompose $W_2^m = \mathcal{N} \oplus \mathcal{P}_{m-1} \oplus \mathcal{S}_m$, where $\mathcal{N}$ is the space of constants, with the square norm $(\int_0^1 f)^2$; $\mathcal{P}_{m-1}$ is the space of polynomials of degree less than $m$ which integrate to zero, with the square norm $\sum_{v=0}^{m-1} (\int_0^1 f^{(v)})^2$; and $\mathcal{S}_m$ is the space of functions with square integrable $m$th derivative and satisfy $\int_0^1 f^{(v)} = 0, \; v = 0, \cdots, m - 1$, with the square norm $\int_0^1 (f^{(m)})^2$. So the polynomial spline penalty functional $J(f) = \int_0^1 (f^{(m)})^2$ is the norm of the projection of $f$ to $\mathcal{S}_m$. The space $\mathcal{P}_0$ vanishes when $m = 1$. Let $\mathcal{G}$ be the tensor products of component spaces

$$\mathcal{G} = \bigotimes_{i=1}^p (\mathcal{N}_i \oplus \mathcal{P}_m \oplus \mathcal{S}_m)$$

$$= \{ 1 \} \oplus \{ (\bigoplus_{i=1}^p \mathcal{P}_m) \oplus (\bigoplus_{i=1}^p \mathcal{S}_m) \} \oplus (\bigoplus_{i<j} (\mathcal{P}_m \oplus \mathcal{P}_m))$$

$$\oplus (\bigoplus_{i\neq j} (\mathcal{P}_m \otimes \mathcal{S}_m)) \oplus (\bigoplus_{i<j} (\mathcal{S}_m \otimes \mathcal{S}_m)) \oplus \cdots,$$

where with some abuse of notation, we are omitting factors of the form $\otimes \{ \mathcal{N}_i \}; \{ 1 \}$ is the space of constant functions on $[0, 1]^p$; $\mathcal{P}_m = N^1 \otimes \cdots \otimes \mathcal{P}_m \otimes \cdots \otimes N^p$ is the space of
functions that is a constant on \( x_j \), \( j \neq i \) and a polynomial of degree less than \( m \) on \( x_i \). Others have similar interpretation. The terms in the three brackets are the spaces for the constant, the main effects and the 2-interactions respectively. We can select a model from simple to complex by including the constants, or the constants and some of the main effects, or the constants plus some of the main effects and 2-interactions, and so on. When a model is chosen, we can write the model space as

\[
\mathcal{M} = \mathcal{H}^0 \oplus \sum_{j=1}^{q} \mathcal{H}^j,
\]

where \( \mathcal{H}^0 \) is a finite dimensional space of polynomials which are not going to be penalized, and the \( \mathcal{H}^j \)'s are orthogonal subspaces of the form \( S_m^i \) or composite spaces of tensor products of two or more spaces of the form \( S_m^i \) and \( P_m^k \), at least one of these spaces being \( S_m^i \). The norms on the composite \( \mathcal{H}^j \) are the tensor product norms induced by the norms on the component subspaces. So \( \| \eta \|^2 = \| P_0 \eta \|^2 + \sum_{j=1}^{q} \| P_j \eta \|^2 \), where \( P_j \) is the orthogonal projector in \( \mathcal{M} \) onto \( \mathcal{H}^j \).

A direct generalization of (1) is

\[
\min - \sum_{i=1}^{n} l_i(\eta_i) + \frac{n}{2} \sum_{j=1}^{q} \lambda_j \| P_j \eta \|^2,
\]

in \( \mathcal{M} \), where \( \lambda_j \)'s are a set of smoothing parameters. Writing \( \lambda_j = \lambda / \theta_j \), (5) becomes

\[
\min - \sum_{i=1}^{n} l_i(\eta_i) + \frac{n}{2} \lambda \| P_* \eta \|^2_{\Theta},
\]

where \( P_* = \sum_{j=1}^{q} P_j \) is the orthogonal projection in \( \mathcal{M} \) onto \( \mathcal{H}_* = \sum_{j=1}^{q} \mathcal{H}^j \) and

\[
\| \eta \|^2_{\Theta} = \| P_0 \eta \|^2 + \sum_{j=1}^{q} \theta_j^{-1} \| P_j \eta \|^2,
\]

is a modified norm indexed by \( \Theta = (\theta_1, \cdots, \theta_q) \). We denote \( R_j \) as the reproducing kernel (RK) for \( \mathcal{H}^j \); it can be shown that \( \theta_j R_j \) is the RK under the norm \( \| \cdot \|_{\Theta} \). So the RK for \( \sum_{j=1}^{q} \mathcal{H}^j \) under \( \| \cdot \|_{\Theta} \) is \( R_{\Theta} = \sum_{j=1}^{q} \theta_j R_j \).
The RK for subspaces $N_j$, $P_{m-1}$ and $S_m$ are $R_N(s,t) = 1$, $R_{P_{m-1}}(s,t) = \sum_{v=1}^{m-1} k_v(s)k_v(t)$ and $R_{S_m}(s,t) = k_m(s)k_m(t) + (-1)^{m-1}k_{2m}([s-t])$ respectively, where $k_v(\cdot) = B_v(\cdot)/v!$ and $B_v(\cdot)$ is the $v$-th Bernoulli polynomials defined by $B_0(t) \equiv 1$, $B'_v(t) = vB_{v-1}(t)$ and $\int_0^1 B_v(t)dt = 0$. $[t]$ denotes the fractional part of $t$. Since the RK of the tensor product space is the product of the RK's of the component spaces (Aronszajn (1950)), the computations of $R_j$'s are straightforward. For example, $R$ corresponding to the space $P_{m-1} \oplus S_m$ is $R_{P_{m-1}}(x_1, x_2)R_{S_m}(x_3, x_4, x_5)$, where $x_{k,l}$ denotes the $k$th coordinate of the $l$th "design point".

The solution to (6) has the form (Wahba 1990, O'Sullivan et al. 1986)

$$\eta(x) = \sum_{v=1}^{M} \phi_v(x) + \sum_{i=1}^{n} c_i(\sum_{j=1}^{q} \theta_j R_j(x_i, x)) = \phi^Td + \xi^Tc.$$ 

In this form, $\{\phi_v\}_{v=1}^{M}$ is a set of basis functions of $H^0$, $\phi^T(x) = (\phi_1(x), \cdots, \phi_M(x))$, $\xi^T(x) = (R_\Theta(x_1, x), \cdots, R_\Theta(x_n, x))$. $c$ and $d$ are vectors of coefficients to be computed as the minimizers of

$$I_\lambda(c, d) = -\sum_{i=1}^{n} l_i(\phi^T(x_i)d + \xi^T(x_i)c) + \frac{n}{2} \lambda c^T Q_\Theta c,$$

(7)

where $Q_\Theta$ is a $n \times n$ matrix with $Q_\Theta(i, j) = < R_\Theta(x_i, x), R_\Theta(x_j, x) > = R_\Theta(x_i, x_j)$. (7) can not be solved directly. But if all $l_i(\eta_i)$'s are strictly concave, we can use Newton-Raphson procedure to compute $c$ and $d$ for fixed $\lambda$ and $\Theta$. Denote $u_i = -dl_i/d\eta_i$, $u^T = (u_1, \cdots, u_n)$, $w_i = -d^2l_i/d\eta_i^2$, $W = \text{diag}(w_1, \cdots, w_n)$, $S = (\phi(x_1), \cdots, \phi(x_n))^T$. Then $\partial I_\lambda/\partial c = Q_\Theta u + n\lambda Q_\Theta c$, $\partial I_\lambda/\partial d = S^T u$, $\partial^2 I_\lambda/\partial c\partial c^T = Q_\Theta W Q_\Theta + n\lambda Q_\Theta$, $\partial^2 I_\lambda/\partial d\partial d^T = Q_\Theta W S$, $\partial^2 I_\lambda/\partial d\partial c^T = S^T W S$. The Newton-Raphson iteration satisfies the linear system

$$\begin{pmatrix}
Q_\Theta W Q_\Theta + n\lambda Q_\Theta & Q_\Theta W S \\
S^T W Q_\Theta & S^T W S
\end{pmatrix}
\begin{pmatrix}
c - c_- \\
d - d_-
\end{pmatrix}
= 
\begin{pmatrix}
-Q_\Theta u_+ - n\lambda Q_\Theta c_- \\
-S^T u_-
\end{pmatrix},$$

(8)

where the subscript minus indicates quantities evaluated at the previous step. Similar to Gu (1990), $\eta = Sd + Q_\Theta c$ is always unique as long as $S$ is of full column rank. If $Q_\Theta$ is
nonsingular, (8) is equivalent to the system

\[(W_\lambda Q_\Theta + n\lambda I)c + W_\lambda Sd = (W_\lambda \eta - u),\]
\[S^Tc = 0.\]  \hfill (9)

Let \(\tilde{Q}_\Theta = W^{1/2}_\Theta Q_\Theta W^{1/2}_\Theta\), \(\tilde{c} = W^{-1/2}_\lambda c\), \(\tilde{S} = W^{1/2}_\lambda S\), \(\tilde{d} = d\), and \(\tilde{y} = W^{-1/2}_\lambda (W_\lambda \eta - u);\) (9) becomes

\[(\tilde{Q}_\Theta + n\lambda I)\tilde{c} + \tilde{S}\tilde{d} = \tilde{y},\]
\[\tilde{S}^T\tilde{c} = 0.\]  \hfill (10)

It is interesting to note that the solution of (9) gives the minimizer of the following problem:

\[
\min \frac{1}{n} \sum_{i=1}^{n} w_i (\tilde{\eta}_i - \eta(x_i))^2 + \sum_{j=1}^{g} \lambda_j \|P_j \eta\|^2,
\]  \hfill (11)

where \(\tilde{\eta}_i = \eta_i - u_i / w_i\). So the procedure iteratively reformulates the problem to model \(\eta_i\)'s on pseudo-data \(\tilde{\eta}_i\)'s by the weighted least squares smoothing interaction spline.

## 3 Algorithm

So far \(\lambda\) and \(\Theta\) are fixed. Choosing appropriate smoothing parameters \(\lambda\) and \(\Theta\) is crucial for effectively estimating the true function from data by fitting spline models. Two of the commonly recognized data-driven methods for choosing smoothing parameters are the GCV and unbiased risk methods (see Wahba (1990)). The GCV score is

\[
V(\lambda, \Theta) = \frac{1/n \| (I - A(\lambda, \Theta)) W^{1/2} \tilde{\eta} \|^2}{[(1/n) Tr(I - A(\lambda, \Theta))]^2},
\]

where \(A(\lambda, \Theta)\) satisfies \((w_1^{1/2} \eta(x_1), \ldots, w_n^{1/2} \eta(x_n))^T = A(\lambda, \Theta)(w_1^{1/2} \tilde{\eta}_1, \ldots, w_n^{1/2} \tilde{\eta}_n)^T\), and \(\eta\) is computed from the solution of (10).
Writing
\[ \hat{U}(\lambda, \Theta) = \frac{1}{n} \| (I - A(\lambda, \Theta)) W^{1/2} \tilde{\eta} \|^2 + 2 \frac{\tilde{\sigma}^2}{n} tr A(\lambda, \Theta), \]
where \( \tilde{\sigma}^2 = 1/n \sum_{i=1}^{n} \tilde{u}_i^2 / w_i \) is an estimate of dispersion (Gu (1992)). Note that \( \hat{U}(\lambda, \Theta) \) is the unbiased risk estimate for Gaussian data with an estimated variance \( \tilde{\sigma}^2 \). If the dispersion is known to be 1, such as in the case of Bernoulli data and Poisson data, a better estimate is
\[ U(\lambda, \Theta) = \frac{1}{n} \| (I - A(\lambda, \Theta)) W^{1/2} \tilde{y} \|^2 + 2 \frac{\tilde{y}}{n} tr A(\lambda, \Theta). \]

However, even for these distributions we may still allow for overdispersion by estimating \( \theta \). We will choose \( \lambda \) and \( \Theta \) based on minimizing \( V(\lambda, \Theta) \) or \( U(\lambda, \Theta) \) (\( \hat{U}(\lambda, \Theta) \) for the case of over or under dispersion) for each Newton-Raphson iteration and will call them options \( V \) and \( U \) (\( \hat{U} \)) respectively. A genetic code RKPACK (Gu (1989), Gu and Wahba (1991)) is available to solve (10) and estimate \( \lambda \) and \( \Theta \) via GCV or the unbiased risk method at the same time. This suggests the following algorithm:

**Algorithm** Assuming the inputs of matrices \( S \), \( Q_j \)'s, the response vector \( y \) and the starting values \( \eta_0 \), perform as follows.

1. Compute \( u \) and \( W \). Compute transformations \( \tilde{S}, \tilde{Q}_j = W^{1/2} Q_j W^{1/2} \) and \( \tilde{\eta} \).

2. Call RKPACK with inputs \( \tilde{S}, \tilde{Q}_i \) and \( \tilde{\eta} \). This is to solve (10) and choose \( \lambda \) and \( \Theta \) according to GCV or unbiased risk method.

3. Compute the new \( \eta \). Stop if the algorithm converges under some criteria (for example, weighted mean square error \( \sum_{i=1}^{n} w_i ((\eta_i - \eta_i) / (1 + |\eta_i|))^2 / \sum_{i=1}^{n} w_i \), which will be used in the later sections); otherwise go to step 1.

Since changing \( \lambda \) and \( \Theta \) along the iteration means modifying the problem successively, convergence is not guaranteed. Nevertheless, the algorithm converges most of the time except for some extreme cases (for example, almost all 1's in binary data).
4 A Practical Example

In this section, we apply the algorithm of Section 3 to the analysis of a data set collected at the University of Chicago Hospitals. The data concern the survival of patients who have colorectal cancer between 1965 and 1981 (Michelassi, Vanucci, Montag, Goldberg, Chappell, Dytch, Bibbo and Block (1988), Michelassi, Vanucci, Chappell, Goldberg and Block (1990)). There are three different responses: stage, morphology and microinvasion. Each of these variables is a strong predictor of whether one will live, given that one has colorectal cancer, and is of biological interest in its own right. Higher stage and microinvasion are more dangerous, and exophytic morphology is less dangerous; stage reflects the degree to which the cancer has spread to adjacent organs, microinvasion reflects whether it has spread within the lymph nodes, and exophytic morphology reflects whether the tumor sticks out (and is easy to excise). We will create a binary variable $y$ for each response and fit them separately. For stage as response, we let $y = 0$ if stage is B, $y = 1$ if stage is C. For morphology as response, $y = 1$ if morphology is exophytic, $y = 0$ otherwise. For microinvasion as response, $y = 1$ if microinvasion is +, $y = 0$ otherwise.

Each of the responses is available in two data sets. The two data sets are: rectal tumors (148 subjects) and colonic tumors (392 subjects). One data with age 15 was deleted since it is far from the other data. The covariates in each case are year (we will denote it as $x_1$) and age (we will denote it as $x_2$), which are the year of and age at diagnosis. This duplication provides a primitive validation of our methods. The interest is to investigate how probability of tumor characteristics depends on year and age at diagnosis. Affine transformations were made on $x_1$ and $x_2$ into $[0, 1]$. For each case, we first fit the largest model. Taking $m = 2$ in section 2, our model space is

$$
\mathcal{M} = \mathcal{H}^0 \oplus \mathcal{H}^1 \oplus \mathcal{H}^2 \oplus \mathcal{H}^3 \oplus \mathcal{H}^4 \oplus \mathcal{H}^5
$$

where $\mathcal{H}^0 = \{1\} \oplus \mathcal{P}_1^1 \oplus \mathcal{P}_2^1 \oplus (\mathcal{P}_1^1 \otimes \mathcal{P}_1^1); \mathcal{H}^1 = \mathcal{S}_1^1; \mathcal{H}^2 = \mathcal{S}_2^2; \mathcal{H}^3 = \mathcal{S}_2^1 \otimes \mathcal{P}_1^1$ (the space with interactions of smooth on year and linear on age); $\mathcal{H}^4 = \mathcal{P}_1^1 \otimes \mathcal{S}_2^2$ (the space with interactions of linear on year and smooth on age); $\mathcal{H}^5 = \mathcal{S}_2^1 \otimes \mathcal{S}_2^2$ (the space with interactions of
smooth on year and smooth on age). So $M = 4$ and $\phi_1(x_1, x_2) = 1$, $\phi_2(x_1, x_2) = x_1 - 1/2$, $\phi_3(x_1, x_2) = x_2 - 1/2$, $\phi_4(x_1, x_2) = \phi_2(x_1, x_2)\phi_3(x_1, x_2)$, $R_{S_x}(s, t) = k_2(s)k_2(t) - k_4(s - t)$, $k_2(s) = s^2/2 - s/2 + 1/12$, $k_4(s) = s^4/24 - s^3/12 + s^2/24 - 1/720$. The RKs for $\mathcal{H}^1$, $\mathcal{H}^2$, $\mathcal{H}^3$, $\mathcal{H}^4$ and $\mathcal{H}^5$ are $R_{S_x}(x_{1i}, x_{1j})$, $R_{S_x}(x_{2i}, x_{2j})$, $R_{S_x}(x_{1i}, x_{1j})\phi_3(x_{1i}, x_{2j})\phi_3(x_{1j}, x_{2j})$, $\phi_2(x_{1i}, x_{2i})\phi_2(x_{1j}, x_{2j})R_{S_x}(x_{2i}, x_{2j})$ and $R_{S_x}(x_{1i}, x_{1j})R_{S_x}(x_{2i}, x_{2j})$ respectively. Unbiased risk method was used to select the smoothing parameter.

For *stage* as response, the plots of main effects and interactions estimates are shown in Figures 1, 2 and 3. For both rectal tumor data and colonic tumor data, the contour plots indicate that the linear-smooth and smooth-smooth interactions are not important. A more appropriate model space is $\mathcal{M} = \mathcal{H}^0 \oplus \mathcal{H}^1 \oplus \mathcal{H}^2 \oplus \mathcal{H}^3$. Probability estimates from the full model are given in Figure 4. The influence of year on the effect age is simple: in early years of the study, the elderly are at higher risk of stage C tumors while in the late 70's the elderly are at lower risk. This pattern is probably explained by stronger testing programs among the elderly, which results in tumors being detected at an earlier stage.

For *morphology* as response, the plots of main effects and interactions estimates are shown in Figures 5, 6 and 7, 8. Figure 5 demonstrates the improvement interaction smoothing splines can make upon the GLIM model. For the rectal tumor data, the contour plot indicate that linear-linear ($\mathcal{P}^1 \otimes \mathcal{P}^2$) and smooth-linear interactions are important while others are not. The overall interaction is important at the ages over 80. The $\lambda_i$'s corresponding to the five penalized spaces (smooth part of year, smooth part of age, smooth-linear, linear-smooth and smooth-smooth) are 0.0653, 0.0226, 30509.72, 0.0009 and 1101.456 respectively. The deviances corresponding to the GLIM model, additive model and interaction model are 198.19, 194.86 and 188.03 respectively. All of these suggest that an additive model for rectal data is not adequate. A more appropriate model space is $\mathcal{M} = \mathcal{H}^0 \oplus \mathcal{H}^1 \oplus \mathcal{H}^2 \oplus \mathcal{H}^4$. Similarly, for the colonic data, the contour plots, $\lambda_i$'s and the deviances indicate that an appropriate model space is $\mathcal{M} = \mathcal{H}^0 \oplus \mathcal{H}^1 \oplus \mathcal{H}^2 \oplus \mathcal{H}^3 \oplus \mathcal{H}^4$.

The main effect fit using the full model (Figure 5) shows a consistent, nonlinear age effect across the two data sets. Clinicians have speculated on a high hazard (associated with exophytic morphology) for young cancer victims which tapers off at middle age. The plots justify that. Age effect changes with year of study again possibly due to changes in screening
behavior.

One notable feature of Figure 8 is the presence of "corner effects", that is steep gradients at the corners. This may be real, but especially if they go different directions for the two groups of tumors (for example, high age / early year), may be artifacts.

For \textit{microinvasion} as response, the plots of main effects, interactions and final probability estimations are shown in Figure 9, 10, 11, 12. For both rectal tumor data and colonic tumor, the contour plots indicate that the smooth*linear, linear*smooth and smooth*smooth interactions are not important. A more appropriate model space is $M = \mathcal{H}^0 \oplus \mathcal{H}^1 \oplus \mathcal{H}^2$.

Formal statistical tests are under study.

5 Simulations

We will use the same independent variables year ($x_1$) and age ($x_2$) in Section 4 as our covariates. Binary data are generated based on a additive model

$$
\text{logit}[Pr(\text{survive}|x_1, x_2)] = \eta(x_1, x_2) = C + \beta_1(x_1) + \beta_2(x_2)
$$

where \( C \) is a constant, \( \beta_1(x_1), \beta_2(x_2) \) are smooth functions which integrate to zero over the range of year and age. We write $\beta_i(x_i) = \alpha_i x_i, \; i = 1, 2$ when these functions are linear. Three different $\beta_i(x_i)$'s are chosen here: \textit{linear effects}, \textit{weak nonlinear effects} and \textit{strong nonlinear effects}. For linear effects, both $\beta_1(x_1)$ and $\beta_2(x_2)$ are linear. The coefficients $\alpha_i$ were chosen to be "moderate": for $n=150$, $\alpha_1 = .0408$ ($\alpha_2 = .1061$) was sufficient to achieve 90\% power against the hypothesis $H_0 : \alpha_1 = 0 \; (H_0 : \alpha_2 = 0)$ at the .05 level. The constant $C = -3.2464$ was picked to yield an overall success rate of approximately 30\%, as in the real data. For weak nonlinear effects, we have

$$
s_1 + \beta_1(x) = \begin{cases} 
\alpha_1 x + h_1 \times (1 + \cos(x\pi/p_1))/2 & \; x < p_1 \\
\alpha_1 x & \; p_1 \leq x < p_2 \\
\alpha_1 x + h_1 \times (1 + \cos((2x - (p_2 + p_3))\pi/(p_3 - p_2)))/2 & \; p_2 \leq x < p_3 \\
\alpha_1 x & \; p_3 \leq x,
\end{cases}
$$
\[ s_2 + \beta_2(x) = \begin{cases} 
\alpha_2 x + h_2 \times (1 + \cos(x\pi/p_5))/2 & x < p_4 \\
\alpha_2 x & p_4 \leq x < p_5 \\
\alpha_2 x + h_2 \times (\cos((x - p_5)\pi/(\max(x_2) - p_5)) - 1) & p_5 \leq x,
\end{cases} \]

where \( p_1, p_2 \) and \( p_3 \) are the 20th, 40th and 60th percentiles of \( x_1 \); \( p_4 \) and \( p_5 \) are the 20th and 80th percentiles of \( x_2 \); \( \max(x_2) \) is the maximum value of \( x_2 \). \( s_i \)'s are the integrations of the right hand side, such that \( \beta_i(x_i) \)'s integrate to zero. Their heights \( h_i \) were calculated to be 1/2 of the total variation in linear effect; for example, \( h_2 = 0.0408 \times (93 - 24) \) for \( \beta_2(x_2) \) since age has a range of 24 to 93 years. The functions and their first derivatives are continuous. They consist of two spikes in the response for each variable, positioned at the lower end and in the middle for \( \beta_1(x_1) \) and at the two ends for \( \beta_2(x_2) \). Nonlinear effects at the extremes of age, and in the early calendar years (startup effects), with a spike for a few years (epidemiologic factors) are medically realistic. The spikes were added onto the underlying linear effect. The strong nonlinear effects consist of exactly the same functions as the weak linear effects; merely multiplying the \( h_i \)'s by 4. All main effects are plotted in Figure 13.

For each of the effects, three sample sizes are used: \( n=150, n=200 \) and \( n=300 \). So we have a 3 x 3 factorial structure and will denote each case as linear-200, weak-150, strong-300 etc. For sample size \( n \), we picked the first \( n \) cases' covariate values in the order listed in our file, which seems rather random. The data were generated using the Fortran routine uni of the Core Mathematics Library (Cmlib) from the National Bureau of Standards after the null call which passed the seed 30 to the routine. For each case, we ran a simulation with 100 repetitions. That is, for each case such as strong-200, 100 data sets with sample size 200 were generated and were fitted using the algorithm with the unbiased risk method for choosing smoothing parameters. All cases converged though one data set in the zero-200 case overflowed. Symmetrized Kullback-Leibler discrepancy (SKL)

\[
SKL = \frac{1}{n} \sum_{i=1}^{n} (p_i - \hat{p}_i) \times (\eta_i - \hat{\eta}_i)
\]

was used to measure the performance of the fit, where \( p = e^\eta/(1 + e^\eta) \) is the probability
of survival. A hat indicates the fits. Figure 14 shows that when the underlying model has linear effects only, the GLIM model is always at least as good as than the interaction spline for all the sample sizes. When the underlying model has weak nonlinear effects, GLIM model is generally but not always better for sample sizes 150 and 200. They are about the same for sample size 300. When the underlying model has strong nonlinear effects, the interaction spline is always better for all the sample sizes.

Figure 15, 16, 17 and 18 display the main effects and final probability estimations for linear effects, weak nonlinear effects and strong nonlinear effects with sample size 300. They captured the shape of the underlying model.

To compare the option U and option V, we ran the same simulations with option V. Figure 19 shows that the U option always got a smaller SKL except in a few cases, though the differences become smaller and smaller when sample size increases. So the U option is recommended if there is no overdispersion.

To assess the performance of the algorithm with option U, the simplex method was used to search the real minimum point as \( \lambda_1 \) and \( \lambda_2 \) vary. That is, for 3 fixed pairs of \( \lambda_1 \) and \( \lambda_2 \), we used the RKPACK to calculate the estimation of \( \eta \). Then SKL's for each pair of \( \lambda_1 \) and \( \lambda_2 \) are calculated since we know the true function. A new pair of \( \lambda_1 \) and \( \lambda_2 \) was found by the simplex method. This procedure was iterated till converge. The converged point had the minimum SKL value. Figure 20 shows that the algorithm did a decent job.

Table 1 gives the summary of average CPU time (in minute) and average number of iterations. All simulations were conducted on a DECstation 3100 using FORTRAN.

<table>
<thead>
<tr>
<th>CASE</th>
<th>CPU</th>
<th>ITER</th>
<th>CASE</th>
<th>CPU</th>
<th>ITER</th>
<th>CASE</th>
<th>CPU</th>
<th>ITER</th>
</tr>
</thead>
<tbody>
<tr>
<td>zero-150</td>
<td>2.75</td>
<td>9.69</td>
<td>weak-150</td>
<td>2.61</td>
<td>9.90</td>
<td>strong-150</td>
<td>4.63</td>
<td>16.17</td>
</tr>
<tr>
<td>zero-200</td>
<td>5.67</td>
<td>8.69</td>
<td>weak-200</td>
<td>5.39</td>
<td>8.88</td>
<td>strong-200</td>
<td>9.89</td>
<td>16.37</td>
</tr>
<tr>
<td>zero-300</td>
<td>14.59</td>
<td>9.12</td>
<td>weak-300</td>
<td>13.02</td>
<td>9.30</td>
<td>strong-300</td>
<td>22.28</td>
<td>14.73</td>
</tr>
</tbody>
</table>

Table 1: Summary of Simulations. CPU times are minutes. ITER is the average number of iteration.
6 Future Research

Many questions remain open since the interaction spline is a new area. In Section 4, we selected a model heuristically. Formal model selection procedure based on hypothesis tests or some criteria like unbiased risk need to be studied. Test sample, two layer cross-validation and bootstrap methods may be used.

Theoretical results can provide insight into and justification for the method. Large sample properties like consistency, convergence rate, strong or weak GCV theorem are desirable.

Some special models for polytomous data are also of interest. Let \( \gamma_j = Pr(Y \leq j) \), \( j = 0, \ldots, k \), a proportional-odds model is

\[
\log \{ \gamma_j/(1 - \gamma_j) \} = \theta_j - \eta(x), \quad j = 1, \ldots, k - 1,
\]

with \( \theta_1 \leq \theta_2 \leq \cdots \theta_{k-1} \). We can also consider a more general model

\[
\log \{ \gamma_j/(1 - \gamma_j) \} = \eta_j(x), \quad j = 1, \ldots, k - 1,
\]

with \( \eta_1(x) \leq \eta_2(x) \leq \cdots \leq \eta_{k-1}(x) \). A statistical test between these two models will be interesting.

7 Conclusion

We have shown, using practical examples and simulations, that interaction smoothing splines are useful extensions of the generalized linear models and that the extensions apply to models with two or more covariates. In our simulations, the unbiased risk method is usually superior to generalized cross-validation method as a method of objectively choosing the smoothing parameters for binary data.

Spline estimation fails in some few instances due to lack of convergence, and also exhibits some bias due to edge effects (for example, in the results shown for age in 15). However, in situations which are not adequately described by linear or polynomial models, interaction smoothing splines provide an attractive alternative.
References


Figure 1: Estimations for the main effects for the two data sets using stage as responses. The solid lines are the main effects estimates calculated with all interaction terms. The dashed lines are the main effects estimates from a GLIM model. Segments are the logit of proportions within each decile adjusted so they add to zero.
Figure 2: Contour plots of interactions in spaces $\mathcal{H}^0$, $\mathcal{H}^3$, $\mathcal{H}^4$ and $\mathcal{H}^5$ respectively. The response is stage and the data set is rectal tumors. Notice that the scale of the two plots on the bottom are small.
Figure 3: Contour plots of interactions in spaces $\mathcal{H}^0$, $\mathcal{H}^3$, $\mathcal{H}^4$ and $\mathcal{H}^5$ respectively. The response is stage and the data set is colonic tumors. Notice that the scale of the two plots on the bottom are small.
Figure 4: Estimations of the probability for the two data sets using stage as response. Above is the fit for rectal data, below is the fit for colonic data. Both used the full model.
Figure 5: Estimations for the main effects for the two data sets using morphology as response. The solid lines are the main effects estimates calculated with all interaction terms. The dashed lines are the main effects estimates from a GLIM model. Segments are the logit of proportions within each decile adjusted so they add to zero.
Figure 6: Contour plots of interactions in spaces $\mathcal{H}^0$, $\mathcal{H}^3$, $\mathcal{H}^4$ and $\mathcal{H}^5$ respectively. The response is *morphology* and the data set is rectal tumors. Notice the two plots on the right are on $10^{-8}$ scale.
Figure 7: Contour plots of interactions in spaces $\mathcal{H}^0$, $\mathcal{H}^3$, $\mathcal{H}^4$ and $\mathcal{H}^5$ respectively. The response is *morphology* and the data set is colonic tumors. Notice that the last plot is on $10^{-6}$ scale.
Figure 8: Estimations of the probability for the two data sets using *morphology* as response. Above is the fit for rectal data, below is the fit for colonic data. Both used the full model.
Figure 9: Estimations for the main effects for the two data sets using microinvasion as response. The solid lines are the main effects estimates calculated with all interaction terms. The dashed lines are the main effects estimates from a GLIM model. Segments are the logit of proportions within each decile adjusted so they add to zero.
Figure 10: Contour plots of interactions in spaces $\mathcal{H}^0$, $\mathcal{H}^3$, $\mathcal{H}^4$ and $\mathcal{H}^5$ respectively. The response is *microinvasion* and the data set is rectal tumors. Notice that the scale of the last three plots are very small.
Figure 11: Contour plots of interactions in spaces $\mathcal{H}^0$, $\mathcal{H}^3$, $\mathcal{H}^4$ and $\mathcal{H}^5$ respectively. The response is microinvasion and the data set is colonic tumors. Notice that the scale of the last three plots are very small.
Figure 12: Estimations of the probability for the two data sets using microinvasion as response. Above is the fit for rectal data, below is the fit for colonic data. Both used the full model.
Figure 13: Main effects plots. The first row is the linear effects, second row is the weak nonlinear effects, third row is the strong nonlinear effects.
Figure 14: Comparison the SKL between GLIM and interaction spline under different underlying models and different sample sizes.
Figure 15: Estimations for the main effects. Solid lines are real main effects. Five dashed lines in each graph are main effect fits: 1, 2, 3, 4, 5 corresponding to the 5th, 25th, 50th, 75th, 95th best fit ordered by SKL. Underlying models of the first, second and third row are linear, weak and strong effects. Sample sizes are all equals to 200.
Figure 16: 3d plots of the probability estimations for linear effect with sample size 300. The plots are the true probability function, 5th, 25th, 50th, 75th and 95th best fit ordered by SKL.
Figure 17: 3d plots of the probability estimations for weak effect with sample size 300. The plots are the true probability function, 5th, 25th, 50th, 75th and 95th best fit ordered by SKL.
Figure 18: 3d plots of the probability estimations for strong effect with sample size 300. The plots are the true probability function, 5th, 25th, 50th, 75th and 95th best fit ordered by SKL.

33a.
Figure 19: Comparison of option U and option V using SKL.
Figure 20: Relative effectiveness of the option U. Best SKLs are calculated by simplex search on a plane.
Figure 19: Comparison of option U and option V using SKL.
Figure 20: Relative effectiveness of the option U. Best SKLs are calculated by simplex search on a plane.