A SIMPLE ROUGHNESS PENALTY APPROACH TO REGRESSION SPLINE ESTIMATION

David Ruppert and Raymond J. Carroll *

May 30, 1996

Abstract

A regression spline is a piecewise polynomial function whose highest order nonzero derivative takes jumps at fixed “knots.” Usually regression splines are smoothed by deleting nonessential knots, or equivalently setting the jumps at those knots to zero. A method that is simpler to implement and has lower computational cost is to shrink the jumps at all knots towards zero by using a penalty function. The method is widely applicable, e.g., to multivariate regression, interaction models, and semiparametric estimators. We also consider a Bayesian approach with a new type of nonparametric prior.

Key words and phrases. Additive models, Bayesian models, Curve and surface fitting, Interaction models.

Short title. Smoothing Regression Splines.

*David Ruppert is Professor, School of Operations Research & Industrial Engineering, Cornell University, Ithaca, New York 14853 (E-mail: davidr@orie.cornell.edu). Ruppert’s research was supported by NSA Grant MDA 904-95-II-1025 and NSF Grant DMS-9306196. R.J. Carroll is Professor of Statistics, Nutrition and Toxicology, Texas A&M University, College Station, TX 77843-3801 (E-mail: carroll@stat.tamu.edu). Carroll’s research was supported by a grant from the National Cancer Institute (CA-57030).
1 INTRODUCTION

There are two general approaches to spline fitting, smoothing splines and regression splines. Smoothing splines require that many parameters be estimated, typically at least as many parameters as observations, and therefore special algorithms are needed to be computationally efficient; see, for example, Eubank (1988) or Green and Silverman (1994) for an introduction to these algorithms. While smoothing spline algorithms are readily available for univariate regression, they are not readily available or even nonexistent for more complex models. Regression splines can be fit by ordinary least squares once the knots have been selected, but knot selection requires sophisticated algorithms that can be computationally intensive; see, for example, Friedman’s (1991) MARS algorithm and Smith and Kohn’s (1995) Bayesian knot selector based on Gibbs sampling.

In this paper, we combine features of smoothing splines and regression splines. Our models have far fewer parameters than a smoothing spline, but unlike MARS and other approaches to regression splines, the location of the knots is not crucial since the coefficients are shrunk. Moreover, selection of the smoothing parameter can be done through minimizing $C_p$, generalized cross-validation, or other computational methods and are only moderately intensive.

Our primary intention is not to produce another univariate smoother, but rather to provide a flexible and easily implemented methodology for fitting complex nonparametric models. However, the basic ideas are more easily understood if presented first in the univariate case, which will be done in this section. In later sections, to show the power and flexibility of our approach, we will look at additive and interaction models, semiparametric models, changepoint problems, variance function estimation, generalized regression, and mixed models. Also, one of the most promising features of our methodology is its flexibility in allowing a wide choice of penalties and and, using Bayesian analysis, a broad choice of priors.

Suppose that we have data $(X_i, Y_i)$ where $X_i$ is univariate,

$$Y_i = m(X_i) + \epsilon_i,$$

and $m$ is a smooth function giving the conditional mean of $Y_i$ given $X_i$. To estimate $m$ we can let $\beta = (\beta_0, \ldots, \beta_p, \beta_{p1}, \ldots, \beta_{pK})^T$ and use a regression spline model

$$m(x; \beta) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^K \beta_{pk} (x - \kappa_k)^+,$$

where $p \geq 1$ is an integer, $(u)^+ = u^I(u \geq 0)$, and $\kappa_1 < \cdots < \kappa_K$ are fixed knots. The traditional method of “smoothing” the estimate is through knot selection. In this paper we use a different approach by allowing $K$ to be large and retaining all
knots, but using a roughness penalty on \( \{\beta_{p,k}\}_{k=1}^K \) which is the set of jumps in the \( p \)th
derivative of \( m(x; \beta) \). We could view this as a penalty on the \((p+1)\)th derivative of
\( m(x; \beta) \) where that derivative is a generalized function. We recommend \( K \) between
5 and 40 and letting \( \kappa_k \) be the \( k/(K+1) \)th sample quantile of the \( X_i \)'s—we call this
choice of knots “equally-spaced sample quantiles.”

We define \( \tilde{\beta}(\alpha) \) to be the minimizer of

\[
\sum_{i=1}^{n} \left\{ Y_i - m(x; \beta) \right\}^2 + \alpha \sum_{k=1}^{K} \rho(\beta_{p,k}),
\]

where \( \rho \) is a suitable nonnegative function, and \( \alpha \) is a smoothing parameter. Because
\( \alpha \) controls the amount of smoothing, the value of \( K \) is not crucial. As we will see,
for typical mean functions, \( K = 10 \) and \( K = 40 \), say, produce very similar estimates,
provided that \( \alpha \) is selected appropriately for each \( K \) and that \( p \geq 2 \). Selection of \( \alpha \)
will be discussed in the next section.

We start with the simplest case, where \( \rho(x) = x^2 \). Let \( Y = (Y_1, \ldots, Y_n)^T \) and \( X \)
be the “design matrix” for the regression spline so that the \( i \)th row of \( X \) is

\[
X_i = (1, X_i, \ldots, X_i^p, (X_i - \kappa_1)_+, \ldots, (X_i - \kappa_K)_+).
\]

Also, let \( D \) be a diagonal matrix whose first \((1+p)\) diagonal elements are 0 and whose
remaining diagonal elements are 1. Then for this \( \rho \) function, simple calculations show
that \( \tilde{\beta}(\alpha) \) is given by

\[
\tilde{\beta}(\alpha) = \left( X^T X + \alpha D \right)^{-1} X^T Y.
\]

This is a ridge regression estimator that shrinks the regression spline towards the
least-squares fit to a \( p \)th degree polynomial model (Hastie and Tibshirani, 1990,
Section 9.3.6).

Computing (5) is extremely quick, even for a relatively large number, say 30,
values of \( \alpha \). The computational time for the matrices \( X^T X \) and \( X^T Y \) is linear in \( n \),
but these matrices need only be computed once. After these matrices are computed,
only \( K \times K \) matrices need to be manipulated. This allows rapid selection of \( \alpha \) by
techniques such as \( C_p \) or generalized cross-validation where \( \tilde{\beta}(\alpha) \) must be calculated
over a grid of values of \( \alpha \).

In the next section, data-based selectors of \( \alpha \) are discussed. In Section 3, models
with several smoothing parameters are introduced and specific examples, e.g.,
multivariate, additive, and interaction models, are developed in Section 4. In Section 5,
nonquadratic penalties are discussed. A Bayesian approach to the roughness penalty
is found in Section 6. Later sections discuss logsplines for estimation of conditional
variances for heteroscedastic data, generalized regression where the response is in an
exponential family, and generalized mixed models.
2 SELECTION OF THE SMOOTHING PARAMETER

Using a suitable value of \( \alpha \) is crucial to obtaining a satisfactory curve estimate. A simple method for selection of \( \alpha \) is to minimize Mallows’s \( C_p \) or the closely related generalized cross-validation (GCV) criterion. Here we follow Hastie and Tibshirani (1990) closely. Let

\[
\text{ASR}(\alpha) = n^{-1} \sum_{i=1}^{n} \left[ Y_i - m(X_i; \beta(\alpha)) \right]^2
\]

be the average squared residuals using \( \alpha \). Again, assume that \( \rho(x) = x^2 \). Let

\[
S(\alpha) = X \left( X^T X + \alpha D \right)^{-1} X^T
\]

be the “smoother” or “hat” matrix. Let \( \alpha^* \) be a small value of \( \alpha \) implying little smoothing. Then

\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \{Y_i - m(X_i; \alpha^*)\}^2}{n - \text{tr}\{2S(\alpha^*) - S^2(\alpha^*)\}} \tag{6}
\]

is a nearly unbiased estimator of the variance of the \( e_i \)’s (Buja et al., 1989, Hastie and Tibshirani, 1990). Finally,

\[
C_p(\alpha) = \text{ASR}(\alpha) + \frac{2\text{tr}(S(\alpha))\hat{\sigma}^2}{n} \tag{7}
\]

is the \( C_p \) statistic, and

\[
\text{GCV}(\alpha) = \frac{\text{ASR}(\alpha)}{ \left[ 1 - \frac{\text{tr}(S(\alpha))}{n} \right]^2} \tag{8}
\]

is the generalized cross validation statistic. Following Hastie and Tibshirani (1990), \( df(\alpha) = \text{tr}\{S(\alpha)\} \) will be called the “effective degrees of freedom” of the fit.

We choose \( \alpha \) by computing either \( C_p(\alpha) \) or \( \text{GCV}(\alpha) \) for a grid of \( \alpha \) values and choosing the minimizer of that criterion. To compute the traces in (6), (7), and (8), one does not need to compute the \( n \times n \) matrix \( S(\alpha) \) since if

\[
C(\alpha) = (X^TX)(X^TX + \alpha D)^{-1},
\]

then

\[
df(\alpha) = \text{tr}\{S(\alpha)\} = \text{tr}\{C(\alpha)\} \quad \text{and} \quad \text{tr}\{S^2(\alpha)\} = \text{tr}\{C^2(\alpha)\}.
\]

As an example, we simulated a data set where \( n = 100 \), the \( X_i \)’s were equally spaced on \([0, 1] \), \( \sigma = .2 \), and \( m(x) = x \sin(19x) \). The grid of thirty \( \alpha \) values was log-spaced between \( 10^{-6} \) and \( 10^3 \), i.e., their base-10 logarithms were equally spaced.
between $-6$ and $3$, with the smallest grid value being used as $\alpha^*$. There were seven choices of $(p, K)$: $(2, 10)$, $(2, 20)$, $(2, 40)$, $(1, 10)$, $(1, 20)$, $(3, 10)$, and $(3, 20)$. In each case the knots were equally-spaced sample quantiles and $\rho(x) = x^2$. Figure 1 shows the fits (dotted curves) that minimize $C_p$ for the seven values of $(p, K)$. The true curve is plotted as a solid curve, and the raw data are plotted in the bottom right. The first three plots also have a dashed curve that is explained in Section 10 and can ignored for now. The plots show that in this example, the value of $K$ has little effect, provided that $p = 2$ or $3$ and $K$ least $10$. In fact, for $p = 2$ even $K = 5$ is roughly similar to the other choices. When one uses piecewise linear splines ($p = 1$) more knots are needed. It appears that $p = 2$ is very slightly superior to $p = 3$ in this example. Computational time in MATLAB on a SPARC 20 is “interactive” (1 to 2 seconds) for $n = 100$, $K = 40$, $p = 2$, and 30 values of $\alpha$.

For practical applications, we recommend $p = 2$ when $m$ is smooth, i.e., at least having a continuous first derivative. If $m$ has a number of oscillations, then $K \geq 10$ is recommended, though in many applications $m$ will be monotonic or unimodal and $K = 5$ will often be quite adequate. Piecewise linear splines are inferior for smooth functions, but are useful if $m$ has a “kink” where $m'$ is discontinuous; see Section 5. When using $C_p$ for choosing $\alpha$, we recommend that $\alpha_*$ should clearly undersmooth the data. We evaluate $C_p$ on a log-spaced grid of $\alpha$ values and use the smallest grid value as $\alpha_*$. If $C_p$ is minimized over the grid by $\alpha_*$, then we lower the minimum value on the grid and start again.

### 3 MULTIPLE SMOOTHING PARAMETERS

The basic model (2) can be generalized in many ways, for example to additive and interaction models. We now let $m$ in (1) be a function of a possibly multivariate predictor $(X_{i1}, \ldots, X_{ij})^T$. To fit more complex models effectively, one may need to partition $\beta$ into blocks, e.g., representing main effects and interactions, and apply a different roughness penalty to each block. Thus, we will work with models of the form

$$Y = \sum_{m=1}^{M} X(m) \beta(m) + \epsilon = \underline{X}_{\beta} + \epsilon,$$

where $\underline{X} = \{X(1) \cdots X(M)\}$, $\underline{\beta} = \{\beta(1)^T \cdots \beta(M)^T\}^T$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. For example, (2) is of form (9) with $M = 2$, $X(1)$ having $i$th row $(1, X_{i1}, X_{i2}, \ldots, X_{ip})$ and $X(2)$ having $i$th row $\{(X_i - \kappa_1)_+, \ldots, (X_i - \kappa_K)_+\}$.

We propose to estimate $\underline{\beta}$ by $\hat{\beta}(\alpha)$ which is the minimizer over $\underline{\beta}$ of

$$\|Y - \underline{X}\beta\|_F^2 + \sum_{m=1}^{M} \alpha_m \sum_{j=1}^{dm} \rho(\beta(m)_j),$$

where $\alpha_m$, $\rho$, and $\beta(m)_j$ are defined in Section 5. This is an example of using multiple smoothing parameters to flexibly modeling the relationship between the response variable and the covariates. Each smoothing parameter $\alpha_m$ is associated with a specific block $m$ of the predictor, allowing for different levels of smoothness across different components of the model.
where $\| \cdot \|_2$ is the Euclidean (or $L_2$) norm, $\alpha_m \geq 0$ for $m = 1, \ldots, M$, and $\beta(m) = \{ \beta(m)_1, \ldots, \beta(m)_{d_m} \}^T$ so that $d_m$ is the dimension of $\beta(m)$. Equation (3) is a special case of (10) with $M = 2$, $\alpha_1 = 0$, and $\alpha_2 = \alpha$. Let $\alpha = (\alpha_1, \ldots, \alpha_M)^T$, and let $D(\alpha)$ be block diagonal with blocks $\alpha_1 I_d, \ldots, \alpha_M I_{d_M}$. Here $I_d$ is the $d \times d$ identity matrix. In the simple case where $\rho(x) = x^2$, we have that

$$\hat{\beta}(\alpha) = \left\{ X^T X + D(\alpha) \right\}^{-1} X^T Y. \quad (11)$$

Next, let

$$S(\alpha) = X \left\{ X^T X + D(\alpha) \right\}^{-1} X^T. \quad (12)$$

Then $C_p(\alpha)$ and GCV($\alpha$) are defined as in (7) and (8) with $S(\alpha)$ there replaced here by $S(\alpha)$ and with the scalar $\alpha^*$ replaced by a vector $\alpha^*$ which has small values of all its components.

4 MULTIVARIATE, ADDITIVE AND INTERACTION MODELS

Recall that $(X_{1}, \ldots, X_{J})^T$ is the vector of predictor variables. Suppose now that $J > 1$. As we will see, a full multivariate model for $m$ can be constructed using tensor-product regression splines. However, when $J$ is large the number of tensor-product basis functions is enormous, a problem often called the “curse of dimensionality.” To overcome this difficulty, we can use an appropriate subset of the tensor-product spline basis giving, for example, an additive model or a low-order interaction model. The idea is analogous to setting interactions, or at least higher order interactions, to 0 when fitting a factorial model.

For $j = 1, \ldots, J$, let $\{ \kappa_{kj} \}_{k=1}^K$ be a set of knots for the $j$th predictor. In practice, $K$ could vary with $j$, but for ease of notation $K$ will be independent of $j$ in this exposition. The basis functions for regression splines in this predictor are

$$B(j) = \{ 1 \} \cup B_P(j; p) \cup B_{PP}(j; p, \kappa_{1j}, \ldots, \kappa_{Kj}), \quad (13)$$

where 1 is the function identically equal to 1,

$$B_P(j; p) = \{ x_j, \ldots, x_j^p \}$$

is the set of polynomial basis functions, and

$$B_{PP}(j; p, \kappa_{1j}, \ldots, \kappa_{Kj}) = \{ (x_j - \kappa_{1j})^p, \ldots, (x_j - \kappa_{Kj})^p \}$$

is the set of piecewise polynomial basis functions. The subscripts “$P$” and “$PP$” denote “polynomial” and “piecewise polynomial.” We will often denote $B_P(j; p)$ by
\( B_P(j) \) and \( B_{PP}(j; p, \kappa_1, \ldots, \kappa_K) \) by \( B_P(j) \) to save space. The tensor-product regression spline basis is \( B(1, \ldots, J) \equiv_{def} B(1) \otimes \cdots \otimes B(J) \), i.e., the set of all products \( b(1) \cdots b(J) \) where \( b(j) \in B(j) \). We use the notation “\( a \equiv_{def} b \)” to mean that \( a \) equals \( b \) by definition of \( a \). The dimension of this basis, \( (1 + p + K)^J \), grows geometrically in \( J \) illustrating the curse of dimensionality.

### 4.1 Bivariate models

Consider the case \( J = 2 \). By (13)

\[
B(1, 2) \equiv_{def} B(1) \otimes B(2) = B_P(1, 2) \cup B_{PP}(1, 2)
\]

where

\[
B_P(1, 2) \equiv_{def} [\{1\} \cup B_P(1)] \otimes [\{1\} \cup B_P(2)]
\]

and

\[
B_{PP}(1, 2) \equiv_{def} [B_{PP}(1) \otimes B(2)] \cup [B(1) \otimes B_{PP}(2)].
\]

Thus, \( B_P(1, 2) \) is a basis of a space of polynomials in \( x_1 \) and \( x_2 \) and \( B_{PP}(1, 2) \) a basis for a space of piecewise polynomial functions of \( x_1 \) and \( x_2 \). Therefore, we let \( M = 2 \) and let \( X(1) \) and \( X(2) \) be generated\(^1\) by \( B_P(1, 2) \) and \( B_{PP}(1, 2) \), respectively, and then let \( \alpha_1 = 0 \) so only the coefficients of the piecewise polynomials are shrunk towards 0.

#### 4.1.1 Biomonitoring of mercury

Opsomer et al. (1995) analyze a data set obtained by biomonitoring of airborne mercury about the Warren Country Resource Recovery Facility (WCRRF), a solid-waste incinerator in New Jersey. Pots of sphagnum moss where placed in 16 sampling locations near the WCRRF and exposed to ambient conditions for a two-week period. At six locations there were replicate pots, for a total of 22 observations. The sampling locations and the location of the WCRRF are shown in Figure 2 as open circles and as an asterisk, respectively. The moss in each pot was collected and assayed for mercury both before and after drying. We will work with the dried moss data. Opsomer et al. (1995) fit bivariate local linear regression to estimate mercury concentration as a function of spatial location. Ruppert (1996) used local quadratic regression.

Figure 2 is a fit by a bivariate tensor-product spline using \( p = 2 \) and \( K = 4 \). It is clear that the estimated mercury concentration peaks near the WCRRF. The estimated concentration is highly variable near the edges due to data sparsity there. This high boundary variability was mitigated in the local polynomial fits of Opsomer

\(^1\)We say that \( X \) is generated by a certain basis if each column of \( X \) is generated by taking a basis function and evaluating it at the observed predictor vectors \( X_1, \ldots, X_n \).
et al. (1995) and Ruppert (1996) by using locally varying bandwidths. Local variation of the smoothing parameter is less easily implemented when fitting regression splines, though using knots at equally-spaced marginal quantiles of the predictors as we have done here is a step in that direction. A regression spline fit to a different model with stable boundary behavior is presented in Section 4.3.1.

### 4.2 Additive models

A function $m$ of $\mathbf{x} = (x_1, \ldots, x_J)^T$ is said to be additive if $m(\mathbf{x}) = \sum_{j=1}^{J} m_j(x_j)$ for univariate functions $m_j, j = 1, \ldots, J$. An additive model restricts $m$ in (1) to be an additive function. Additive models can be fit using the basis

$$B(1) \cup \cdots \cup B(J) = \{1\} \cup \left[ \bigcup_{j=1}^{J} \{B_P(j) \cup B_{PP}(j)\} \right] = \mathcal{P}(1, \ldots, J) \cup \mathcal{P}(1, \ldots, J),$$

where $\mathcal{P}(1, \ldots, J) \equiv \{1\} \cup B_P(1) \cup \cdots \cup B_P(J)$ is a basis for additive polynomial models and $\mathcal{P}(1, \ldots, J) \equiv \{B_P(1) \cup \cdots \cup B_{PP}(J)\}$ is a basis for additive piecewise polynomial models.

When fitting an additive model, we can let $M = 2$ and let $X(1)$ be generated by $\mathcal{P}(1, \ldots, J)$ while $X(2)$ is generated by $\mathcal{P}(1, \ldots, J)$. Another possibility is to apply a different amount of shrinkage to each predictor variable, so that $M = J + 1$, $X(1)$ is as before, while $X(j + 1)$ is generated by $B_{PP}(j), j = 1, \ldots, J$. In either case, $\alpha_1 = 0$ so that polynomial coefficients are not shrunk.

Once the coefficients of the additive model basis functions have been estimated by penalized least squares, the component functions of $m(\mathbf{x}) = m_1(x_1) + \cdots + m_J(x_J)$ can be estimated by

$$\tilde{m}_j(x_j) = \sum_{i=1}^{p} \hat{\beta}_{ij} x_j^i + \sum_{k=1}^{K} \hat{\beta}_{kj} (x_j - \kappa_{kj})^k,$$

where $\hat{\beta}_{ij}$ is the estimated coefficient of $x_j^i$, etc. We recommend subtracting a “centering constant,” $C(j)$, from $\tilde{m}_j$ so that either

$$\sum_{i=1}^{n} \tilde{m}_j(X_{ij}) = 0,$$

where $\{X_{ij}, \ldots, X_{nj}\}$ are the observed values of the $j$th predictor or

$$\sum_{i=1}^{N} \tilde{m}_j(x_{ij}) = 0,$$

where $\{x_1, \ldots, x_{N_j}\}$ is an equally-spaced grid of points over some finite interval, say $[\min_i\{X_{ij}\}, \max_i\{X_{ij}\}]$. The intercept, $\hat{\beta}_0$, i.e., the estimated coefficient of $1$, would
then be replaced by $\hat{\beta}_0 + \sum_{j=1}^{J} C(j)$. The centering makes the $m_j$’s comparable to those from the backfitting algorithm of Hastie and Tibshirani (1990), where a constraint such as (15) or (16) is needed for identifiability.²

4.3 Interaction models

Let $B_{P,PP}(j) = B_P(j) \cup B_{PP}(j)$ be the set of polynomial and piecewise polynomial basis functions in the $j$th predictor variable. Then $B(1, \ldots, J)$ has the decomposition into main effects, two-way interactions, etc.:

$$B(1, \ldots, J) = \bigotimes_{j=1}^{J} [\{1\} \cup B_{P,PP}(j)]$$

$$= \{1\} \cup \left[ \bigcup_{j=1}^{J} B_{P,PP}(j) \right]$$

$$\cup \left[ \bigcup_{j_1=1}^{J} \bigcup_{j_2=j_1+1}^{J} B_{P,PP}(j_1) \otimes B_{P,PP}(j_2) \right] \cup \ldots$$

$$\cup \left[ \bigcup_{j_1=1}^{J} \cdots \bigcup_{j_J=j_{J-1}+1}^{J} B_{P,PP}(j_1) \otimes \cdots \otimes B_{P,PP}(j_J) \right].$$

Each of the main effects and interactions can be further decomposed into polynomial and piecewise polynomial components. There are many options for the assignment of penalties to the components of $B(1, \ldots, J)$ by decomposition of the design matrix $X$ into components $X(1), \ldots, X(M)$ with common penalties within components.

Here is a concrete recommendation. First, we see interaction models as alternatives for additive models, so we will restrict attention to two-way interactions by deleting three-way and higher interactions. Let

$$M_P = \bigcup_{j=1}^{J} B_P(j)$$

be the polynomial main effects basis functions, let

$$M_{PP} = \bigcup_{j=1}^{J} B_{PP}(j)$$

be the piecewise polynomial main effects, let

$$I_P = \bigcup_{j_1=1}^{J-1} \bigcup_{j_2=j_1+1}^{J} B_P(j_1) \otimes B_P(j_2)$$

²Identifiability constraints are not needed when fitting regression splines either by ordinary or penalized least squares, since a regression spline is a full-rank linear model.
be the polynomial two-way interaction basis functions, and let
\[
\mathcal{I}_{PP} = \bigcup_{j_1=1}^{J} \bigcup_{j_2=j_1+1}^{J} \left\{ \mathcal{B}_{PP}(j_1) \otimes \mathcal{B}_{PP}(j_2) \right\} \cup \left\{ \mathcal{B}_{PP}(j_1) \otimes \mathcal{B}_P(j_2) \right\}
\]
be the piecewise polynomial two-way interaction basis functions.

Our approach is to use \( M = 3 \) blocks of basis functions with \( X(1) \), \( X(2) \), and \( X(3) \) generated by \( \{1\} \cup \mathcal{M}_P, \mathcal{M}_{PP} \), and \( \mathcal{I}_P \cup \mathcal{I}_{PP} \), respectively. We let \( \alpha_1 = 0 \) so polynomial main effects are unpenalized. Furthermore, we suggest using less knots for interactions than for main effects, so \( \mathcal{M}_{PP} \) uses \( K_M \) knots in each predictor variable and \( \mathcal{I}_P \) and \( \mathcal{I}_{PP} \) use \( K_I \) knots, where typically \( K_I \) is smaller than \( K_M \). The idea is that we expect interactions to be smaller and less complex than main effects. Also, interactions cannot be estimated as precisely as main effects. Notice as well that since the polynomial interaction terms are in \( X(3) \), they are shrunk towards 0. This is in contrast to the polynomial main effects that are unpenalized.

We see the smoothing of interaction regression splines as a simpler alternative to interaction smoothing splines discussed by Wahba (1986, 1990), Chen (1991), and Gu and Wahba (1993).

4.3.1 Biomonitoring revisited

Besides regressing mercury concentration on spatial position \((x_1, x_2)\), Ruppert (1996) also tried the simpler model where mercury concentration is regressed on the variable \( d \) defined as the distance from the incinerator, a univariate function of \((x_1, x_2)\). The fit to distance alone explains a large part of the variation in mercury concentration, suggesting that the dependence of mercury concentration on \((x, y)\) is predominantly a function of \( d \).

However, mercury concentration is unlikely to be a function solely of \( d \), since the background concentration of mercury will not be exactly constant and because dispersal from the incinerator will not be constant in all directions. These considerations suggest the model
\[
Hg = m_1(d) + m_{23}(x_1, x_2) + \epsilon,
\]
where \( Hg \) is mercury concentration and \( \epsilon \) is random error. Since \( d \) is a function of \((x_1, x_2)\) the decomposition of \( E(Hg) \) into \( m_1(d) \) and \( m_{23}(x_1, x_2) \) is not unique, but this was not a problem since we only wished to model the sum \( m_1(d) + m_{23}(x_1, x_2) \). Thus, we built a model with main effects for \( d, x_1, \) and \( x_2 \) and a \( x_1 \) by \( x_2 \) interaction. The polynomial main effects for \( d \) generated \( X(1) \), the piecewise polynomial main effects for \( d \) generated \( X(2) \), and all main effects and interaction effects for \((x_1, x_2)\) formed \( X(3) \). We took \( \alpha_1 = 0 \) and \( \alpha_2 \) and \( \alpha_3 \) were chosed by minimizing \( C_p \). Thus, all effects of \( x_1 \) and \( x_2 \), even the polynomial main effects, had their coefficient shrunk
towards 0. In fact, this shrinkage was pronounced since $\alpha_3$ was five times larger than $\alpha_2$. ($\alpha_2$ was 1,000 and $\alpha_3$ was 5,000—these values minimized $C_p$ over a rectangular grid where each of $\alpha_1$ and $\alpha_2$ took 20 logarithmically-spaced values between $10^{-5}$ and $10^3$.) The predominance of $d$ is also reflected in our choice of numbers of knots. For the main effects of $d$, 15 knots were used, but for the main effects and interactions of $(x_1, x_2)$ only 5 knots were used.

The estimate $\hat{m}_1(d) + \hat{m}_23(x_1, x_2)$ is plotted in Figure 3. One can see that $\hat{m}_1(d)$ predominates since the contours are nearly circles centered somewhat southeast of the WCRRF. (The contours appear as ellipses since the horizontal and vertical scales differ.) Also, since $\hat{m}_1$ is univariate, the estimate plotted in Figure 3 is much more stable in sparse regions near the boundary than the estimate in Figure 2. We feel that the fit given by $\hat{m}_1(d) + \hat{m}_23(x_1, x_2)$ successfully handles the problem here of a very small sample size and is a striking improvement over an unstructured bivariate fit such as in Figure 2.

5 NONQUADRATIC PENALTY FUNCTIONS

Using $\rho(x) = x^2$ is convenient mathematically since this choice of objective function makes (10) a quadratic function of $\beta$ and lead to the linear estimator (11). As we have seen, the quadratic penalty function generally works well in practice and can be used as the default penalty. However, like all defaults, the quadratic penalty does not always lead to the best possible estimates. This section discusses alternative penalty functions such as the absolute value and Huber penalty function, the latter coming from the theory of robust estimation.

To first understand the quadratic penalty function better, we discuss its Bayesian interpretation (Lindley and Smith, 1972). Suppose that $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, \sigma^2)$. Suppose as well that the prior distribution is that $\beta(1), \ldots, \beta(M)$ are independent with $\beta(m)$ having the multivariate normal distribution $N(0, \sigma^2_m I_{m})$. For now, assume that $\sigma^2, \sigma^2_1, \ldots, \sigma^2_M$ are known. Then the posterior log density of $\beta$ given $Y$ is, up to an additive function of $Y$ and $(\sigma^2, \sigma^2_1, \ldots, \sigma^2_M)$, given by

$$-\frac{1}{2} \left\{ \frac{1}{\sigma^2} \| Y - X\beta \|_2^2 + \sum_{m=1}^{M} \frac{1}{\sigma^2_m} \| \beta(m) \|_2^2 \right\}.$$ 

Thus, the maximum a posterior (MAP) estimator, i.e., the mode of the posterior density, minimizes (10) with $\alpha_m = (\sigma/\sigma_m)^2$, $m = 1, \ldots, M$. Of course, these variances usually will not be known in practice. An alternative to choosing $\alpha_1, \ldots, \alpha_M$ by $C_p$ or GCV would be to estimate $\sigma^2, \ldots, \sigma^2_M$ by hierarchical Bayesian or empirical Bayesian methods; see Section 6. The subvector $\beta(m)$ is unpenalized when $\alpha_m = 0$ which means one takes $\sigma^2_m$ to be $\infty$. 

10
For concreteness, consider the univariate regression spline model given by (2), so that $\beta(1) = (\beta_0, \ldots, \beta_p)^T$ are the polynomial coefficients and $\beta(2) = (\beta_{p1}, \ldots, \beta_{pK})^T$ are the jumps in the $p$th derivative. The Bayesian model leading to the quadratic penalty says that the jumps in $m^{(p)}$ at the knots are independent $N(0, \sigma^2)$ random variables. Also, one uses $\sigma^2 = \infty$ so that the polynomial coefficients are unpenalized since there is no prior information about them. The iid normal model for the jumps in $m^{(p)}$ will fit many functions reasonably well, but it is not entirely satisfactory for a function that has “change points” where the function’s behavior changes suddenly and dramatically. An example is the motorcycle impact data (Silverman, 1985) discussed latter in this section. At changepoints we would expect jumps that are much larger than the other jumps.

To accommodate large jumps, one can use a Bayesian model where the prior on the jumps is non-Gaussian. A simple choice is the double exponential or Laplace density, $(2\sigma)^{-1} \exp(-|x|/\sigma)$. The heavy tails of this prior allow for the occasional large jumps one expects at change points. The MAP estimator minimizes

$$
\|Y - X\beta\|_2^2 + \alpha\|\beta(2)\|_1, \tag{17}
$$

where $\|x\|_1 = \sum_{i=1}^{\dim(x)} |x_i|$ is the $L_1$ norm, so $\rho(x) = |x|$. The use of the $L_1$ norm in parametric regression was recently proposed by Tibshirani (1996) and called the lasso (least absolute shrinkage selection operator). Tibshirani mentions ongoing work with T. Hastie where the lasso is applied to the MARS (multivariate adaptive regression splines) algorithm of Friedman (1991); that work may be somewhat related to this paper.

To appreciate the potential advantages of the $L_1$ norm, suppose that $m(x) = |x|$ and that the data analyst knows that $m'$ is discontinuous but does not know the number or location of the discontinuities. Then the analyst might use a piecewise linear spline ($p = 1$) with a large set of knots with the intention that at least one knot will be near the location of each discontinuity of $m'$. Suppose, in fact, that several knots are near the single discontinuity at 0. The $L_2$ penalty will tend to choose a function with many small positive jumps in $\tilde{m}'$ around 0 with the sum of the jumps near +2. Such an estimate will have much smaller $L_2$ penalty than an estimate with a single jump at the knot closest to 0. On the other hand, the $L_1$ assigns the same penalty to one large positive jump as it does to many smaller positive jumps with the same total. Since a single large jump will tend to fit the data better, it will be selected by the $L_1$ penalty.

A compromise between the $L_1$ and the $L_2$ penalties uses the Huber (1964) \textquotedblleft rho function\textquotedblright

$$
\rho_H(x) = \begin{cases} 
\frac{x^2}{2} & |x| \leq k_H \\
0 & \text{otherwise}
\end{cases}
$$

11
where $k_H$ is a positive tuning constant. The Huber penalty allows $\tilde{m}'$ to take a few large jumps since large jumps receive the absolute value penalty. Generally, $k_H$ should be scaled to the size of $\beta(m)$. This scaling can be achieved by using $k_H$ between .5 and 1.5 (say) and standardizing $\beta(m)$ by its MAD (median absolute deviation). More specifically, let $\text{MED}\{\beta(m)\} = \text{median}\{\beta(m)_1, \ldots, \beta(m)_{d_m}\}$ and $\text{MAD}\{\beta(m)\} = \text{median}\{|\beta(m)_1 - \text{MED}\{\beta(m)\}|, \ldots, |\beta(m)_{d_m} - \text{MED}\{\beta(m)\}|\}$. Then the penalty in (10) is

$$
\sum_{m=1}^{M} \alpha_m \sum_{j=1}^{d_m} \rho \left( \frac{\beta(m)_j}{\text{MAD}\{\beta(m)\}} \right).
$$

(18)

To minimize (10) with a nonquadratic penalty, one can adapt the method of iterated reweighted least squares used in robust estimation. Let $\hat{\beta}^{(1)}$ be the estimate using the $L_2$ penalty. Then $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}, \ldots$ are calculated recursively by

$$
\hat{\beta}^{(N+1)} = \left[ X^T X + \frac{1}{2} D(\alpha) \text{diag} \left\{ \rho' \left( \frac{\hat{\beta}^{(N)}(m)_j/MAD_m}{\beta(m)_j/MAD_m} \right) \right\} \right]^{-1} X^T Y.
$$

(19)

Here $\hat{\beta}^{(N)} = (\hat{\beta}^{(N)}_1, \ldots, \hat{\beta}^{(N)}_d)^T$ where $d = d_1 + \cdots + d_M$ and $\text{MAD}_m = \text{MAD}\{\hat{\beta}^{(N)}(m)\}$. The derivation of (19) is found in the appendix. If one uses $\rho(x) = x^2$, then $\rho'(x)/x \equiv 2$ and (19) does not change from its starting estimate. We call (19) iteratively reweighted ridge regression (IRRR). We have found that IRRR is extremely easy to implement and that it produces satisfactory estimates for both the Huber and $L_1$ penalties, though for the $L_1$ penalty we do not know how closely IRRR finds the actual minimum of the objective function. For the $L_1$ penalty Tibshirani (1996) recommends a computational method based on quadratic programming.

From (19) we see that at convergence, the fitted values are $S(\alpha)Y$ where

$$
S(\alpha) = X \left[ X^T X + D(\alpha) \text{diag} \left\{ \rho' \left( \frac{\hat{\beta}(m)_j/MAD_m}{\beta(m)_j/MAD_m} \right) \right\} \right]^{-1} X^T.
$$

We use this $S(\alpha)$ as the smoother matrix to define $C_p(\alpha)$, GCV($\alpha$), $df(\alpha)$, and $\hat{\sigma}^2$ as in Section 2.

### 5.1 Motorcycle data

Silverman (1985) analyzes a data set from a simulated motorcycle crash where $Y =$ acceleration and $X =$ time in $ms$. At the time of impact, approximately $14ms$, the first derivative of $m$ appears to jump from 0 to a negative value. The presence of such
a jump is quite reasonable physically and should be accommodated by the estimation method. The exact time of impact is unknown but can be estimated from the data.

We fit model (2) to these data using piecewise quadratic splines \( p = 2 \) with \( L_1 \) and \( L_2 \) penalties and piecewise linear splines \( p = 1 \) using the \( L_1, L_2 \), and Huber penalties. There were 20 knots at equally-spaced sample quantiles for \( p = 2 \) and 30 knots for \( p = 1 \). The fitted functions are given if Figure 4. In each case, \( \alpha \) was first chosen by minimizing \( C_p \), but the plots in Figure 4 use \( \alpha \) slightly smaller than the minimizer of \( C_p \) to get a fit that is slightly more appealing visually near the point of impact. Away from the point of impact, the fits in Figure 4 appear the same as the minimum \( C_p \) fits. This choice of \( \alpha \) illustrates the common practice of using a data-based smoothing parameter as a starting point for choosing the smoothing parameter by visual inspection.

When \( p = 2 \), the fit is rather smooth and does not appear to depend on the choice of penalty function. The fit at the point of impact is not entirely satisfactory as there is an artificial “bump” immediately before the impact and the kink evident in the data is rounded off in the fit. The piecewise linear fits \( (p = 1) \) seem to adapt better to the data, especially when the absolute value or Huber penalty is used. The piecewise linear fits with nonquadratic penalties also have kinks at the trough and at the peak in acceleration after impact, though at these locations kinks appear to fit the data no better—though also no worse—than the smoother piecewise quadratic estimates.

Silverman (1985) noticed the substantial heteroscedasticity in these data. Besides the standard smoothing spline fit, he also performed a weighted analysis where the squared deviations of \( Y \) from the smoothing spline were weighted by the reciprocal of the estimated variance. He estimated the variance function, i.e., the conditional variance of acceleration given time, by a moving average of the squared residuals from a preliminary unweighted spline estimate. In Figure 5 we present a similar weighted analysis. The dispersion function there was estimated by fitting a logspline model to the absolute residuals; see Section 7.1 for details.

The absolute residuals and the estimated variance function are shown in the lower right panel. All fits seem improved by the weighted analysis, especially the ones with the most room for improvement. The \( p = 1 \) estimator with the \( L_1 \) penalty still appears to be the best fit near the discontinuity; this estimator is the least improved but only because it already gave a good fit without weighting.

6 A BAYESIAN APPROACH

The \( L_1 \) and \( L_2 \) penalties discussed in Section 5 are special cases of the \( L_q \) penalty

\[
\|x\|_q = \sum_{i=1}^{d_{\text{dim}(x)}} |x_i|^q, \quad q > 0.
\]
Despite the notation, $\| \cdot \|_q$ is not a norm for $q < 1$ but it is nonetheless a reasonable choice for a penalty function. In this section, we develop a hierarchical Bayesian model where the jumps in the $p$th derivative of $m$ are treated as random. This Bayesian approach leads to data-based selection of $q$ and of the smoothing parameters.

For $\sigma > 0$ and $q > 0$,

$$f(u; q, \sigma) \equiv \frac{1}{\sigma} \exp \left\{ -\frac{1}{2} \left| \frac{u}{\sigma} \right|^q - \eta(q) \right\}, \quad -\infty < u < \infty,$$

where $\eta(q) = \log \{ q^{-1/2} \Gamma(q^{-1}) \}$, is a probability density function which includes the scaled double exponential, $(1/4\sigma) \exp(-1/2|u|)$, when $q = 1$ and the normal when $q = 2$. See Box and Tiao (1973, Section 3.2.1) for discussion of (20) as a class of priors in parametric estimation and for earlier references.

We partition $\beta$ into $\beta(1), \ldots, \beta(M)$ as before. We use a prior on $\beta$ such that the components of $\beta$ are mutually independent and such that each component of $\beta(m)$ has density $f(u; q_m, \sigma_m)$, $m = 1, \ldots, M$. As before $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, \sigma^2)$. The log of the conditional density of $(Y, \beta)$ given $(\sigma, \sigma_1, \ldots, \sigma_M, q_1, \ldots, q_M)$ is, up to an additive constant,

$$\log p(Y, \beta; \sigma, \sigma_1, \ldots, \sigma_M, q_1, \ldots, q_M) =$$

$$-\frac{1}{2} \left\{ \frac{1}{\sigma^2} \| Y - X\beta \|_2^2 + \sum_{m=1}^M \left\| \frac{\beta(m)}{\sigma_m} \right\|_q^{q_m} \right\}$$

$$+ n \log \sigma^2 + \sum_{m=1}^M d_m \log \sigma_m^2 \} - \sum_{m=1}^M d_m \eta(q_m).$$

Now we put a diffuse reference prior on $\sigma^2, \sigma_1^2, \ldots, \sigma_M^2$ which is proportional to the inverse of their product. Also, we put a uniform prior on $q_1, \ldots, q_M$. The MAP estimator of $(\beta, \sigma^2, \sigma_1^2, \sigma_2^2, \ldots, \sigma_M^2)$ is defined as follows. First let $\alpha_m = \sigma^2 \sigma_m^{-q_m}$ so that $\log \sigma_m^2 = \frac{q_m}{2} \left[ \log \sigma^2 - \log \alpha_m \right]$, define $\alpha = (\alpha_1, \ldots, \alpha_M)^T$ as before, and define

$$SS(\beta, \alpha) = \| Y - X\beta \|_2^2 + \sum_{m=1}^M \alpha_m \| \beta(m) \|_{q_m}^{q_m}.$$ 

Then the MAP estimator maximizes

$$\frac{1}{2} \left\{ \frac{SS(\hat{\beta}, \alpha)}{\sigma^2} + \left\{ n + 1 + \sum_{m=1}^M \frac{2(d_m + 1)}{q_m} \right\} \log \sigma^2 \right\}$$

$$- \sum_{m=1}^M \frac{2(d_m + 1)}{q_m} \log \alpha_m \} - \sum_{m=1}^M d_m \eta(q_m).$$

(21)

Given $\alpha$, (21) is maximized over $\beta$ by $\hat{\beta}(\alpha)$ that is defined to be the argument minimizing $SS(\beta, \alpha)$ over $\beta$. Given $\hat{\beta}$ and $\alpha$, (21) is maximized over $\sigma^2$ by

$$\hat{\sigma}^2(\beta, \alpha) = \frac{SS(\beta, \alpha)}{n + 1 + \sum_{m=1}^M \frac{2(d_m + 1)}{q_m} \log \alpha_m}.$$ 

(22)
By substituting $\hat{\beta}(\alpha)$ in (22) and then substituting (22) into (21), we see that to find the MAP estimator of $\alpha$ and $(q_1, \ldots, q_M)$ we maximize

$$
-\frac{1}{2} \left\{ n + 1 + \sum_{m=1}^{M} \frac{2(d_m + 1)}{q_m} \right\} \left\{ 1 + \log \sigma^2(\hat{\beta}(\alpha), \alpha) \right\} + \sum_{m=1}^{M} \left\{ (d_m + 1) \frac{\log \alpha_m}{q_m} - \frac{d_m \eta(q_m)}{q_m} \right\}
$$

over $\alpha$ and $(q_1, \ldots, q_m)$, say by a grid search. When appropriate, we might restrict $q_m$ to a small grid, say \{1/2, 1, 1 1/2, 2\}, and we might require $q_1 = \cdots = q_m$ in order to reduce the computational cost of such a search. Another possibility would be to use a Newton-Raphson algorithm to maximize (23) over $\alpha$ and $(q_1, \ldots, q_m)$. Also, if we do not wish to penalize $\beta(1)$ then we take $\sigma^2 = \infty$, $\alpha_1 = 0$, $q_1$ is not used, and the summations over $m$ in (23) are from 2 to $M$.

### 6.1 Motorcycle Data

We applied this Bayesian method to piecewise-linear modeling of the motorcycle data with $K = 40$ knots, $M = 2$, $\alpha_1 = 0$ so $q_1$ is not defined, $q_2$ confined to the grid \{1/2, 1, 2\}, and $\alpha_2$ taking 30 log-spaced values between $10^{-1}$ and $10^3$. For each value of $(q_2, \alpha_2)$, the posterior density was maximized over $(\hat{\beta}, \sigma^2)$. The log-posterior maximized over $\hat{\beta}$ and $\sigma^2$ with $\alpha_2$ fixed is plotted against $\log(\alpha_2)$ in the middle column of Figure 6, a separate plot for each value of $q_2$. The fits with the MAP estimates of $(\hat{\beta}, \sigma^2, \alpha_2)$ are in the left column, again with a separate plot for each value of $q_2$. In the right column one finds the histograms of the 40 estimated jumps in $m'$.

From Figure 6 we can draw several conclusions:

- The log-posterior is much higher for $q = 1/2$ than for $q = 1$ or 2, suggesting that $q = 1/2$ provides a better fit, at least from the perspective of the Bayesian model.

- If we interpret all values in the middle bins of the histograms as essentially equal to zero, then $q = 1/2$ produces very few, 10, non-zero values of the estimated jumps in $m'$ while $q = 1$ produces more, 17, and $q = 2$ produces 31 non-zero jumps. Tibshirani (1996) advocates $q = 1$ rather that $q = 2$ to reduce the number of non-zero coefficients, which in the context of parametric regression may lead to more interpretable models. We see here that using $q < 1$ may further reduce the number of nonzero coefficients, though doing this is not necessarily desirable when fitting splines.

- $q = 1/2$ produces a fit with an obvious piecewise linear appearance, while $q = 2$ gives a far smoother fit since it has many small but non-zero jumps in $m'$. $q = 1$ is intermediate in smoothness between $q = 1/2$ and $q = 2$. 

15
• The value of $\alpha$ that we have selected as “maximizing” the posterior will be only a local maximum. However, when $q = 1$ or 2, the other local maxima occur at points well separated from the local maxima shown in Figure 6 and produce estimates (not shown here) that clearly do not fit the data. The log-posterior plot for $q = 1/2$ has several local maxima and suggests that one should exercise extreme care when using $q$ less than 1.

The reason that we must select a local maximum is that the global maximum is $+\infty$. Using (23) with $M = 1$ and $\alpha_1 = \alpha$, we can see that the posterior will approach a supremum of $+\infty$ as $\alpha \to \infty$ since

$$SS(\hat{\beta}(\alpha),\alpha) \leq SS(0,\alpha) = \|Y\|_2^2 \quad \forall \alpha$$

while the other term in $\alpha$, $(d + 1) \log(\alpha)$, diverges as $\alpha \to \infty$. The problem of an unbounded posterior can be avoided by using a prior on $\sigma_1^2$ that converges to 0 sufficiently fast as its argument approaches 0. The choice of such priors and estimation using them will be the subject of another paper.

### 7 ESTIMATING VARIANCE FUNCTIONS BY LOGSPLINES

When $\text{Var}(Y_i|X_{i1}, \ldots, X_{iJ})$ is a nonconstant function, then there are several good reasons for modeling this variance function. First, a weighted estimator of $E(Y_i|X_{i1}, \ldots, X_{iJ})$ using reciprocals of estimated variances as weights is usually more efficient than an unweighted estimator. Second, prediction and calibration intervals require an estimator of the variance function; see Carroll and Ruppert (1988).

For modeling variance functions, a natural candidate is a logspline model where

$$\text{Var}(Y_i|X_{i1}, \ldots, X_{iJ}) = \exp(2\theta^T X_i),$$

(24)

where $X_i$ is the $i$th row of $X$ consisting of polynomial and piecewise polynomial terms in $X_{i1}, \ldots, X_{iJ}$ and $\theta$ is an unknown coefficient vector. Let $e_i$ be the residual of $Y_i$ from a preliminary fit to a spline model for $E(Y_i|X_{i1}, \ldots, X_{iJ})$. A simple estimate of $\theta$ is a penalized least squares estimator. Assume that for some $M_2 \geq 1$ the parameter vector $\theta$ has been partitioned into $\theta(1), \ldots, \theta(M_2)$ and for simplicity assume that we are using quadratic penalties. Then the penalized least squares estimate minimizes

$$\sum_{i=1}^n \left\{ e_i^2 - \exp(2\theta^T X_i) \right\}^2 + \sum_{m=1}^{M_2} \alpha_m \|\theta(m)\|_2^2$$

16
over $\theta$. Once we have a preliminary estimator $\hat{\theta}_{\text{prel}}$ of $\theta$ then we can reestimate $\theta$ by weighted least squares which minimizes

$$
\sum_{i=1}^{n} \left\{ \frac{\varepsilon_i^2 - \exp(\nu \hat{\theta}_{\text{prel}}^T X_i)}{\exp(\nu \hat{\theta}_{\text{prel}}^T X_i)} \right\}^2 + \sum_{m=1}^{M_2} \alpha_m \| \theta(m) \|^2_2
$$

over $\theta$. Least-squares estimation based on squared residuals is highly sensitive to outliers. An sensible alternative is to use least-squares based on absolute residuals; see Davidian and Carroll (1987) and Carroll and Ruppert (1988). Let $\varepsilon_i = Y_i - E(Y_i | X_1, \ldots, X_d)$. We will use the model:

$$
E|\varepsilon_i| = \exp(\theta^T X_i).
$$

Model (25) will hold with the same value of $\theta$ as in (24) except for a change in intercept if $\varepsilon_i = \exp(\theta^T X_i) u_i$ where $u_1, \ldots, u_n$ are iid. Using model (25), we estimate $\theta$ by minimizing

$$
\sum_{i=1}^{n} \left\{ \frac{|\varepsilon_i\nu - \exp(\nu \hat{\theta}_{\text{prel}}^T X_i)}{\exp(\nu \hat{\theta}_{\text{prel}}^T X_i)} \right\}^2 + \sum_{m=1}^{M_2} \alpha_m \| \theta(m) \|^2_2,
$$

where typically $\nu$ equals 1 or 2, though other values are possible. Differentiating (26) with respect to $\theta$, we get the estimating equation

$$
0 = -\sum_{i=1}^{n} \left\{ \frac{|\varepsilon_i\nu - \exp(\nu \hat{\theta}_{\text{prel}}^T X_i)}{\exp(\nu \hat{\theta}_{\text{prel}}^T X_i)} \right\} \exp(\nu \hat{\theta}_{\text{prel}}^T X_i) (\nu X_i) + D_2(\alpha) \theta.
$$

where $D_2(\alpha)$ is block diagonal with blocks $\alpha_1 I_{d_1}, \ldots, \alpha_{M_2} I_{d_{M_2}}$.

Equation (27) is the basis of an iterative estimation scheme where given $\hat{\theta}_{\text{old}}$ we solve for $\hat{\theta}_{\text{new}}$ in

$$
0 = -\sum_{i=1}^{n} \left\{ \frac{|\varepsilon_i\nu - \exp(\nu \hat{\theta}_{\text{old}}^T X_i)(1 + \nu(\nu \hat{\theta}_{\text{new}} - \hat{\theta}_{\text{old}})^T X_i)}{\exp(\nu \hat{\theta}_{\text{old}}^T X_i)} \right\} \nu X_i
$$

$$
+ D_2(\alpha) \hat{\theta}_{\text{old}} + D_2(\alpha)(\hat{\theta}_{\text{new}} - \hat{\theta}_{\text{old}}).
$$

The solution is

$$
\hat{\theta}_{\text{new}} = \hat{\theta}_{\text{old}} + \left\{ \sum_{i=1}^{n} \nu^2 X_i^T X_i + D_2(\alpha) \right\}^{-1} \left[ \sum_{i=1}^{n} \left\{ \frac{|\varepsilon_i\nu - \exp(\nu \hat{\theta}_{\text{old}}^T X_i)(\nu X_i)}{\exp(\nu \hat{\theta}_{\text{old}})} - D_2(\alpha) \hat{\theta}_{\text{old}} \right\} \right].
$$

(28)
Generally, one estimates the mean and variance functions iteratively by alternating between

1. weighted estimation of the mean by a spline fit to the $Y_i$’s to get residuals

2. estimation of the variance function by a logspline fit to absolute or squared residual to get weights.

Within each step of 2. in the main iteration there is a subloop where one iterates (28). The main iteration starts with an unweighted estimate of the mean. For parametric modeling, i.e., without a roughness penalty, the best number of iterations of the main loop is a complicated issue, but iterating to convergence is not recommended. Rather, two or three iterations seems best; see Carroll, Wu, and Ruppert (1987). We expect that two or three iterations will also be best for penalized estimation.

The GCV criterion is given by (8) with

$$\text{ASR}(\alpha) = \sum_{i=1}^{n} \left\{ \frac{|e_i|^\nu - \exp(\nu \hat{\theta}_i X_i)}{\exp(\nu \hat{\theta}_i X_i)} \right\}$$

and $S(\alpha) = X \{X^T X + D_2(\alpha)\}^{-1} X^T$. Here $X$ is the matrix with $i$th row given by $\nu X_i$.

7.1 Motorcycle Data — Modeling the Variance Function

The lower right plot in Figure 5 is an estimate of $E(\{|e_i| \mid X_i\})$ for the motorcycle data. We used two iterations of the algorithm just described with $\nu = 1$. In the subloop, step (28) was iterated 15 times, or until the relative change in $\hat{\theta}$ was less than $10^{-6}$, whichever came first.

8 GENERALIZED REGRESSION

In generalized linear regression models, the distribution of $Y$ is in an exponential family and the mean of $Y$ is $g(\eta)$ where $g$ is a known “link” function and $\eta$ is the “linear predictor” following a linear model; see, for example, McCullagh and Nelder (1989) and Hastie and Tibshirani (1990). As a nonparametric alternative to a generalized linear model, one can use a regression spline model in place of the linear predictor, e.g., as in Stone (1994). With this substitution, we have

$$E(Y|X_1, \ldots, X_J) = \mu(X^T \beta) = \mu\{ \sum_{m=1}^{M} X(m) \beta(m) \};$$

$$\text{var}(Y|X_1, \ldots, X_J) = \sigma^2 V(X^T \beta) = \sigma^2 V\{ \sum_{m=1}^{M} X(m) \beta(m) \},$$

18
where if $\mu(\cdot)$ is the mean function and $Z = (Z_1, \ldots, Z_n)^T$ is any vector, $\mu(Z) = \{\mu(Z_1), \ldots, \mu(Z_n)\}^T$, and similarly for the variance functions $V(\cdot)$ and $\mathbf{V}(\cdot)$ and other functions to follow. Also, $\mathbf{X}$ is the vector of spline basis functions evaluated at $X_1, \ldots, X_J$ and $\mu^{-1}$ is the link function. For functions $\theta(\cdot), b_1(\cdot)$ and $b_2(\cdot)$, the loglikelihood can be written as

$$\ell\{\mu(\mathbf{X}\mathbf{\beta}), \sigma^2\} = \frac{Y^T \theta\{\mu(\mathbf{X}\mathbf{\beta})\} - 1^T b_1\left[\theta\{\mu(\mathbf{X}\mathbf{\beta})\}\right]}{\sigma^2} + 1^T b_2(Y, \sigma^2),$$

where $1$ is a vector of ones. The penalized loglikelihood is

$$\ell\{\mu(\mathbf{X}\mathbf{\beta}), \sigma^2\} - \sum_{m=1}^{M} (\alpha_m/\sigma^2) \sum_{j=1}^{d_m} \rho(\beta(m)_j).$$

Let $\mu^{(1)}$ be the derivative of $\mu$, $\rho^{(1)}$ the derivative of $\rho$, $c(\cdot) = \mu^{(1)}(\cdot)/V(\cdot)$ and $d(\cdot) = c(\cdot)\mu^{(1)}(\cdot)$, in which case the penalized estimator solves the penalized estimating equation

$$0 = \mathbf{X}^T \text{diag}\{c(\mathbf{X}\mathbf{\beta})\} \{\mathbf{Y} - \mu(\mathbf{X}\mathbf{\beta})\} - \sum_{m=1}^{M} \alpha_m \sum_{j=1}^{d_m} \rho^{(1)}(\beta(m)_j).$$

(29)

Let $\psi(x) = \rho^{(1)}(x)/x$, and note that (29) can be recast as solving

$$0 = \mathbf{X} \text{diag}\{c(\mathbf{X}\mathbf{\beta})\} \{\mathbf{Y} - \mu(\mathbf{X}\mathbf{\beta})\} - \sum_{m=1}^{M} \alpha_m \sum_{j=1}^{d_m} \psi(\beta(m)_j)\beta(m)_j.$$

As in (19), we rescale $\rho$, so we make the definitions

$$\mathbf{W}(\mathbf{\beta}) = \text{diag}\{d(\mathbf{X}\mathbf{\beta})\};$$
$$\mathbf{Q}(\mathbf{\beta}) = \text{diag}\{\psi(\mathbf{\beta})/\text{MAD}\};$$
$$\mathbf{T}(\mathbf{\beta}) = \mathbf{X}\mathbf{\beta} + \left[\text{diag}\{\mu^{(1)}(\mathbf{X}\mathbf{\beta})\}\right]^{-1} \left\{\mathbf{Y} - \mu(\mathbf{X}\mathbf{\beta})\right\};$$
$$\mathbf{R}(\mathbf{\beta}) = \left\{\mathbf{X}^T \mathbf{W}(\mathbf{\beta}) \mathbf{X} + \mathbf{D}(\alpha) \mathbf{Q}(\mathbf{\beta})\right\}^{-1}.$$

Hence, $\hat{\mathbf{\beta}}$ is obtained by iterative updating:

$$\hat{\mathbf{\beta}}_{\text{new}} = \mathbf{R}(\mathbf{\beta}_{\text{old}}) \mathbf{X}^T \mathbf{W}(\mathbf{\beta}_{\text{old}}) \mathbf{T}(\mathbf{\beta}_{\text{old}}).$$

(30)

Letting $\hat{\mu}_{\text{sat}}$ be the maximizer in $\mu$ of the saturated likelihood and letting $\hat{\mathbf{\beta}}(\alpha)$ be the estimator of $\mathbf{\beta}$ for given $\alpha$, the GCV criterion (Hastie & Tibshirani, 1990, page 159) for choosing $\alpha$ is to minimize

$$\sigma^2 \left(\ell(\hat{\mu}_{\text{sat}}, \sigma^2) - \ell(\hat{\mu}(\hat{\mathbf{\beta}}(\alpha), \sigma^2)\right) / \left(1 - \text{tr} \left[\mathbf{R}(\hat{\mathbf{\beta}}(\alpha)) \mathbf{X}^T \mathbf{W}(\hat{\mathbf{\beta}}(\alpha)) \mathbf{X}\right]/n\right)^2.$$

(31)
In the case of logistic regression, $\mu^{(1)} = V = \mu(1 - \mu)$, $c(x) = 1$, $\theta(x) = \mu^{-1}(x)$, $b_1(x) = -\log\{1 - \mu(x)\}$, $d(x) = \mu^{(1)}(x)$, (30) is

$$\hat{\beta}_{\text{new}} = \left[ X^T \text{diag}\{\mu^{(1)}(X\hat{\beta}_{\text{old}})\} X + D(\alpha)Q(\hat{\beta}_{\text{old}}) \right]^{-1} \left[ X^T \text{diag}\{\mu^{(1)}(X\hat{\beta}_{\text{old}})\} \hat{\beta}_{\text{old}} 
+ X^T \{ Y - \mu(X\hat{\beta}_{\text{old}}) \} \right] = \hat{\beta}_{\text{old}} + \left[ X^T \text{diag}\{\mu^{(1)}(X\hat{\beta}_{\text{old}})\} X + D(\alpha)Q(\hat{\beta}_{\text{old}}) \right]^{-1} \left[ X^T \{ Y - \mu(X\hat{\beta}_{\text{old}}) \} \right] - D(\alpha)Q(\hat{\beta}_{\text{old}}) \hat{\beta}_{\text{old}};$$

and the numerator of (31) equals

$$-Y^T X\hat{\beta}(\alpha) - 1^T \log \left[ 1 - \mu \left\{ X\hat{\beta}(\alpha) \right\} \right].$$

### 8.1 Framingham Study

As an example, we used the Framingham heart data (Kannel, et al., 1986) and regressed incidence of coronary heart disease against “age” = (actual age – 45)/20. We used the logistic link function and a piecewise quadratic spline with 5 knots at equally-spaced quantiles. Although we focused on the effect of age, to correct for other covariates, smoking status (smoker or nonsmoker), systolic blood pressure at exam 2, and cholesterol level at exam 2 were entered linearly into the logistic model. The $L_2$ penalty was used for the piecewise polynomial basis coefficients of age and the smoothing parameter $\alpha$ was chosen by GCV. In Figure 7 we display the logits of the fitted values (top) and the GCV function (bottom).

### 9 GENERALIZED MIXED MODELS

The same idea can be used to fit generalized linear mixed models when $m(x; \beta)$ enters as a fixed effect.

For example, consider the simple mixed logistic regression model

$$\text{pr}(Y_{ij} | X_{ij}, \zeta_j) = H \{ \zeta_j + m(X_{ij}; \beta) \},$$

for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$, where $n = \sum_j n_j$. In this model, $\zeta_j$ are independent mean zero normal random variables with common variance $\theta$. In this problem, there is interest both in the semiparametric regression function $m(x; \beta)$ as well as in the variance $\theta$.

One method that is relatively simple to implement in this context is the so-called $L2$ method of Breslow & Lin (1995). The method works as follows. For fixed $(\theta, \beta)$,
estimate \( \zeta_j \) for \( j = 1, \ldots, J \) to maximize
\[
-\frac{\zeta_j^2}{2\theta} + \sum_{i=1}^{n_j} (Y_{ij} \log \{H \{\zeta_j + m(X_{ij}, \beta)\}\} + (1 - Y_{ij}) \log \{1 - H \{\zeta_j + m(X_{ij}, \beta)\}\})
\]
\[= -\frac{\zeta_j^2}{2\theta} + \ell_j(\zeta_j, \beta, \theta) . \tag{32}\]

Denote the vector of these estimates by \( \hat{\zeta}(\beta, \theta) \). The L2 method is to maximize
\[
\mathcal{L}(\beta, \theta) = \sum_{j=1}^{J} \left\{ -(1/2) \log(1 - \theta \ell_j^{(2)}) + \ell_j - \tilde{\zeta}(\beta, \theta)^2/(2\theta) + \frac{\theta^2 \ell_j^{(4)}}{8(1 - \theta \ell_j^{(2)})^2} \right\} , \tag{33}\]
where
\[
\ell_j = \ell_j \{\hat{\zeta}(\beta, \theta), \beta, \theta\} ; \\
H_{ij} = H \{\hat{\zeta}(\beta, \theta) + m(X_{ij}; \beta)\} ; \\
H_{ij}^{(1)} = H_{ij}(1 - H_{ij}) ; \\
\ell_j^{(2)} = -\sum_{i=1}^{n_j} H_{ij}^{(1)} ; \\
\ell_j^{(4)} = \sum_{i=1}^{n_j} \left\{ 2(H_{ij}^{(1)})^2 + 4H_{ij} H_{ij}^{(1)} - 4H_{ij}^2 H_{ij}^{(1)} - H_{ij}^{(1)} \right\} .
\]

The penalized L2 method then maximizes a penalized version of (33), namely
\[
\mathcal{L}(\beta, \theta) - \alpha \beta^T \beta / 2. \tag{34}\]

Obtaining \( \hat{\zeta}(\beta, \theta) \) is easy using Newton-Raphson, since the first and second derivatives in \( \zeta_j \) of (32) are given, respectively, by
\[
-\zeta_j / \theta + \sum_{i=1}^{n_j} [Y_{ij} - H \{\zeta_j + m(X_{ij}; \beta)\}] ; \\
-1/\theta - \sum_{i=1}^{n_j} H^{(1)} \{\zeta_j + m(X_{ij}; \beta)\} .
\]

The maximization of (34) is done iteratively: fix \( \alpha \) at its current value, then maximize (34) as a function only of \( (\beta, \theta) \), then update \( \alpha \), etc.

We obtain starting values iteratively in three steps, as follows. For fixed \( (\theta, \alpha) \), estimate \( \beta \) by maximizing (34); this is an easy modification of the methods of Section 8, with known “offsets” \( (\zeta_1, \ldots, \zeta_J) \) added to the intercept term in the regression spline. Next, for fixed \( (\alpha, \beta) \), estimate \( \theta \) by maximizing (34) in \( \theta \) alone; we evaluate (34) on a grid of values from 0.01 to 2.00 to obtain this maximum. The third step is to update \( \alpha \) by maximizing (32).
We ran a small simulation using these methods with $J = 15$, $n_j = 20$, $\theta = 0.36$ and the true regression function sine$(X)$. The $X$’s were generated as follows. First, we set up $n = \sum_j n_j = 300$ equally spaced observations on the interval $[-4, 4]$. We then randomly allocated these observations to the $J = 15$ populations, with a different random allocation for each simulation. We used a degree 2 polynomial with 5 knots. We ran the simulation 25 times, obtaining a mean estimate of $\theta$ equal to 0.35. In Figure 8, we display the results of this simulation, not only the actual estimated functions, but also their mean, which it remarkably similar to the sine function.

10 QUADRATIC INTEGRAL PENALTIES

The roughness penalties that we have used so far differ from the penalty that is the basis for smoothing splines,

$$\int \{m^{(\nu)}(x)\}^2 \, dx$$

where $\nu = 2$ (Eubank, 1988; Wahba, 1990). A smoothing spline is the function $\overline{m}$ minimizing

$$\sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 + \alpha \int \{m^{(\nu)}(x)\}^2 \, dx$$

over all $m$ with two derivatives where again $\nu$ is equal to 2. It is possible to apply penalty (35) to regression splines by minimizing (36) over the set of all regression splines of a fixed degree with a fixed set of knots and with $\nu$ not necessarily equal to 2. In this section we show how this minimization can be achieved and compare the resulting estimates with the estimates in Section 2.

We will work with univariate regression splines given by (2). Model (2) is of the form $m(x; \beta) = B(x)^T \beta$ where $B(x) = (B_1(x), \ldots, B_{1+p+K}(x))^T$ is the vector of regression spline basis functions evaluated at $x$ so that $X_i = B(X_i)$ where $X_i$ is given by (4). Let $\kappa_0 \leq \min(X_i)$ and $\kappa_{K+1} \geq \max(X_i)$ be given. We assume that $p \geq \nu$ to avoid the trivial case where the penalty is 0. If the limits of integration in (35) are $\kappa_0$ to $\kappa_{K+1}$, then as in Green (1987)

$$\int \{m^{(\nu)}(x)\}^2 \, dx = \beta^T \Sigma \beta$$

where

$$\Sigma = \int_{\kappa_0}^{\kappa_{K+1}} B^{(\nu)}(x) \{B^{(\nu)}(x)\}^T \, dx.$$

Therefore, (36) is minimized over regression splines by $\overline{m}(x) = B(x)^T \hat{\beta}$ where

$$\hat{\beta} = \left( X^T X + \alpha \Sigma \right)^{-1} X^T Y;$$

see Green (1987). Letting $S(\alpha) = X \{X^T X + \alpha \Sigma\}^T X^T$, $df(\alpha)$, $\hat{\sigma}^2$, $C_p(\alpha)$, and GCV($\alpha$) are defined as in Section 2.
10.1 Quadratic splines with a quadratic penalty on the second derivative

As an example, suppose that \( \nu = 2 \) and \( p = 2 \). Then

\[
B^{(2)}(x) = (0, 0, 2, 2I(x > \kappa_1), \ldots, 2I(x > \kappa_K))^T,
\]

\( \Sigma_{ij} = 0 \) if \( i \leq 2 \) or \( j \leq 2 \), and \( \Sigma_{i+2, j+2} = 4(\kappa_{K+1} - \kappa_{\text{max}(i,j)-1}) \).

In Figure 1, the first three plots illustrate regression quadratic splines. As discussed in Section 2, the dotted curves are the spline estimates with a penalty on the sum of the coefficients on the piecewise polynomial basis function, or, equivalently, on the third derivative as a generalized function. The dashed curves are quadratic splines with penalty (35) with \( \nu = 2 \). In all cases, \( \alpha \) is chosen by minimizing \( C_p \). One can see that in this example, penalizing the third derivatives leads to nearly the same estimate as penalizing the second derivative.

However, there will be cases where penalizing the second derivative will lead to something different than penalizing the third derivative. Penalizing the third derivative puts no penalty on the coefficient of \( x^2 \) and will lead to overfitting in situations where a least-squares fit to a quadratic polynomial is already overfitting. This can happen if there are few observations and \( m \) is nearly linear.

11 FURTHER DISCUSSION

11.1 Density estimation

The idea of putting a roughness penalty on a large, but finite dimensional, parameter can be found in an interesting paper by Scott, Tapia, and Thompson (1980) that partially inspired the present paper. These authors looked at nonparametric estimation of a probability density function by the method of penalized maximum likelihood introduced by Good and Gaskins (1971). Rather than maximizing the penalized likelihood over an infinite dimensional space as in Good and Gaskins, Scott et al. maximize it over a space of piecewise constant, or piecewise linear, splines with fixed knots. The penalty is quadratic on the first derivative. Constrained maximization is used to ensure that the estimate is nonnegative and integrates to 1.

Kooperberg and Stone (1992) and Stone (1994) discuss density estimation by logspline models similar to those in Section 7. They smooth by knot selection. The logspline model ensures nonnegativity, and the estimate can be normalized to integrate to 1. As Stone (1994) discusses, density estimation in high dimensions is facilitated by using additive or interaction logspline models.

The penalty function approach presented here could be readily applied to logspline density estimation. One can either maximize a penalized log-likelihood over a space
of logsplines or convert the density estimation problem to a regression problem by binning the data to form a histogram. The bin centers are taken as the $x$ values and the bin counts, normalized so that the histogram is a pdf, are the $Y$'s. By the Poisson approximation to the binomial, these $Y$ values are nearly Poisson variants multiplied by known constants. Therefore, it is appropriate to fit them by generalized regression as in Section 8 using a log link function and the scaled Poisson likelihood. This approach combines features of the Scott, Tapia, and Thompson approach, e.g., penalized estimation, with features of the Kooperberg and Stone approach, e.g., a logspline model.

### 11.2 Semiparametric Models

Semiparametric models are particularly easy to fit when the nonparametric components are modeled as regression splines. Using the general structure given by (9), one incorporates all parametric components of the model into $X(1)$ which usually will also contain the polynomial basis functions of the nonparametric components. Then setting $\alpha_1 = 0$ ensures that the parametric components are not penalized. The parametric and nonparametric components can then be estimated simultaneously in the same way that we have discussed for purely nonparametric models.

For example, suppose that we wish to fit a quadratic regression spline but wish to allow a kink, i.e., a jump in the first derivative, at a known changepoint, call it $x_c$. Then we simply add $(x - x_c)_+$ to the polynomial basis functions.

In the case of multiple predictors, we can combine nonparametric models for some predictors with parametric models for other predictors, e.g., as in the Framingham study in Section 8.1. This is especially appropriate if some of the predictor variables are categorical, e.g., smoking status in the Framingham study is binary. Once could also model main effects and possibly low order interactions nonparametrically and model some higher order interactions parametrically.

### 11.3 Pseudosplines

There are some connections between our work and Hastie’s (1996) idea of a pseudospline. A linear smoother can be written as $SY$ where $S$ is a $n \times n$ matrix. Typically, $S$ is of full rank and all eigenvalues of $S$ are positive but most of the eigenvalues of $S$ are close to 0. The behavior of the smoothing matrix can be understood in terms of these significantly nonzero eigenvalues and their eigenvectors. However, it is difficult to find the eigenvalue/eigenvector decomposition of $S$ if $n$ is at all large and $S$ is of full rank. Hastie’s idea is to approximate $S$ by a low rank matrix.

Following Hastie’s notation, let the “design matrix” $P$ be generated by an orthogonal basis of dimension $k$ and let $D_\theta = \text{diag}(\theta_1, \ldots, \theta_k)$. Then Hastie defines a
pseudospline as \( P\hat{\beta} \) where \( \hat{\beta} \) minimizes
\[
Q_\lambda(\beta, Y) = \|Y - P\beta\|^2 + \lambda\beta^T D\beta.
\] (38)

Hastie also mentions that the pseudospline equals \( S_\lambda(P, \theta)Y \) where
\[
S_\lambda(P, \theta) = P(P^TP + \lambda D\theta)^{-1}P^T.
\]

The smoother matrix \( S_\lambda(P, \theta) \) is of rank \( k \) and typically \( k \) is far smaller than \( n \). Hastie concentrates on bases of orthogonal polynomials and on using pseudosplines to approximate other smoothers, e.g., smoothing splines and local polynomial fits.

Comparing (38) with (10) or (37), we see that if we use the \( L_2 \) penalty or (35), then our penalized regression spline is a pseudospline.

The point is that the rank of our smoother matrix \( S(\alpha) \) in (12) equals the rank of \( X \) which is typically much less than \( n \). Moreover, the eigenvalues and eigenvectors of \( S(\alpha) \) can be found from those of a matrix whose size is rank(\( X \)),
\[
C(\alpha) \equiv_{\text{ef}} (X^T X + D(\alpha))^{-1}(X^T X).
\]

More precisely, simple algebra shows that if \( u \) is an eigenvector of \( C(\alpha) \) with eigenvalue \( \nu \), then \( X(X^T X)^{-1}u \) is an eigenvector of \( S(\alpha) \) also with eigenvalue \( \nu \).

11.4 Bayesian Inference

A promising area for future research is Bayesian inference applied to regression splines, e.g., tests of submodels and confidence intervals and bands for univariate curves or components of additive and interaction models.

Regression splines seem particularly amenable to Bayesian analysis since the prior (and posterior) distributions are on finitely dimensional spaces, i.e., the space of coefficients of the basis functions. In contrast, smoothing splines require inference on infinite dimensional function spaces; see Wahba (1983) and Nychka (1988) for Bayesian confidence intervals for smoothing splines. As we have seen in Section 6, the finite dimensional parameter space easily allows non-Gaussian priors. Priors with nonconstant variance could be used to model situations where the mean function is flat in some regions but has high curvature in other regions. The simple Bayesian approach of Section 6 models the coefficients of the basis functions as independent according to the prior. This approach leads to satisfactory point estimation, but for inference could be expanded to allow a priori dependence, say according to an AR model. These and other issues in the choice of prior will be explored in another paper.
APPENDIX: DERIVATION OF THE ITERATIVELY
REWEIGHTED RIDGE REGRESSION ALGORITHM

The estimator \( \hat{\beta}(\alpha) \) minimizes \( \|Y - X\hat{\beta}\|_2^2 \) plus (18). The gradient of the first component is \(-2X^T(Y - X\beta)\). If in (18) we replace \( \text{MAD}\{\beta(m)\} \) by \( \text{MAD}_m \) which is considered fixed and not depending on \( \beta(m) \), then the gradient of (18) is

\[
\sum_{m=1}^{M} \frac{\alpha_m}{\text{MAD}_m^2} \sum_{j=1}^{d_m} \left[ \rho' \left( \frac{\hat{\beta}^{(N)}(m)_j / \text{MAD}_m}{\{\hat{\beta}^{(N)}(m)_j / \text{MAD}_m\}} \right) \right] \beta^{(N)}(m)_j.
\]

We can subsume \( 1/\text{MAD}_m^2 \) into \( \alpha_m \), since the \( \alpha_m \)'s will be varied to minimize GCV or \( C_p \). Then, the minimizer of \( \|Y - X\hat{\beta}\|_2^2 \) plus (18) solves

\[
2X^TX\hat{\beta} - 2X^TY + D(\alpha)\text{diag} \left[ \frac{\rho' \left( (\hat{\beta}^{(N)}(m)_j / \text{MAD}_m) \right) / \text{MAD}_m}{(\hat{\beta}^{(N)}(m)_j / \text{MAD}_m)} \right] \beta
\]

which leads to (19).

REFERENCES


Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *J. Royal


Figure 1: Simulated data with $m(x) = x \sin(19x)$, $\sigma = .2$, $n = 100$, equally spaced X’s, and normally distributed errors. True $m$ (solid) and $\hat{m}$ (dotted). The first three plots also give the estimates that use $\int \{\hat{m}^{(2)}(x)\}^2 dx$ as a roughness penalty (dashed).
Figure 2: Biomonitoring of airborne mercury. Bivariate tensor-product spline fit. Open circles are sampling locations, the asterisk is the location of the solid-waste incinerator, and the plus signs label the contours.
Figure 3: Biomonitoring example. Regression spline fit with a main effect for distance from the WCRRF and main effects and interaction for spatial location.
Figure 4: Unweighted regression spline fits to the motorcycle impact data.
Figure 5: Weighted regression spline fits to the motorcycle impact data. The lower right panel shows the absolute residuals and the estimate of the expected absolute residual given time.
Figure 6: Unweighted piecewise-linear regression spline fits using the hierarchical Bayesian approach. The penalty uses $q$ equal to $1/2$, $1$, and $2$ in the top, middle, and bottom rows, respectively. The right column gives the histogram of the estimated jumps in the first derivative at the 40 knots. The middle column is a plot of the log-posterior maximized over all parameters but $\alpha_2$ against $\log(\alpha_2)$. 

34
Figure 7: Framingham data. Logistic regression fit by a penalized quadratic spline with 5 knots. The response is the indicator of coronary heart disease and the predictor is “age” = (actual age – 45)/20. Systolic blood pressure, cholestrol, and smoking status enter the model linearly on the logit scale. Top: logit of the spline fit to “age” centered to have mean 0. Bottom: GCV function.
Figure 8: Simulation of logistic random effects model with $J = 15$, $n_j = 20$ and random effects being normally distributed with mean zero and variance $\theta = 0.36$. The upper plot consists of the true function $\sin(X)$ along with 25 realizations from the simulation. The lower plot consists again of $\sin(X)$ along with the mean of the 25 estimated functions.