

PENALIZED REGRESSION SPLINES

David Ruppert and Raymond J. Carroll *

June 26, 1997

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. Nonquadratic penalties are easily implemented and have interesting properties not shared by the usual quadratic penalties, for example ability to accommodate changepoints.

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

*David Ruppert is Professor, School of Operations Research & Industrial Engineering, Cornell University, Ithaca, New York 14853-3801 (E-mail: davidr@orie.cornell.edu). Ruppert’s research was supported by NSA Grant MDA 904-95-H-1025 and NSF Grant DMS-9306196. R.J. Carroll is Professor of Statistics, Nutrition and Toxicology, Texas A&M University, College Station, TX 77843-3143 (E-mail: carroll@stat.tamu.edu). Carroll’s research was supported by a grant from the National Cancer Institute (CA-57030) and was partially completed during visits to Sonderforschungsbereich 373 at the Humboldt Universität zu Berlin and the Division of Cancer Epidemiology and Genetics, National Cancer Institute. We thank an associate editor and four referees for their comments that have improved our presentation. We also thank Matt Briggs for his comments on an earlier version. Matt Wand kindly showed us his manuscript on a comparison of regression spline smoothing method.

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. 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 and Silverman's (1989) Turbo, Friedman's (1991) MARS algorithm, and Smith and Kohn's (1996) Bayesian knot selector based on Gibbs sampling.

In this paper, we combine features of smoothing splines and regression splines. Our models often have far fewer parameters than a smoothing spline, but unlike MARS and other approaches to regression splines, the location of the knots is not as crucial since the coefficients are shrunk. Moreover, selection of the smoothing parameter can be done through minimizing C_p , generalized cross-validation, or other methods that computationally 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 non-parametric 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, changepoint problems, and variance function estimation. One of the most promising features of our methodology is its flexibility in allowing a wide choice of penalties. The form of the penalty is more important than might be expected.

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

$$Y_i = m(X_i) + \epsilon_i, \quad (1)$$

and m is a smooth function giving the conditional mean of Y_i given X_i . We assume that the ϵ_i 's are mutually independent, mean zero random variables. To estimate m we can let $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p, \beta_{p+1}, \dots, \beta_{p+K})^T$ and use a regression spline model

$$m(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \sum_{k=1}^K \beta_{p+k} (x - \kappa_k)_+^p. \quad (2)$$

where $p \geq 1$ is an integer, $(u)_+^p = u^p I(u \geq 0)$, and $\kappa_1 < \dots < \kappa_K$ 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 $m(x; \boldsymbol{\beta})$. We could view this as a penalty on the $(p+1)$ th derivative of

$m(x; \boldsymbol{\beta})$ where that derivative is a generalized function. We recommend K between 5 and 40 and letting κ_k be the $k/(K+1)$ th sample quantile of the X_i 's—we call this choice of knots “equally-spaced sample quantiles.” Eilers and Marx (1996) have independently developed an estimation method similar to ours, and they have traced the original idea to O’Sullivan (1986, 1988). Eilers and Marx use equally-spaced knots and they use the B-spline basis whereas we use the power-function basis, but these differences seems inessential. What is new here compared to Eilers and Marx is that we introduce multiple smoothing parameters and nonquadratic penalties. As we show, both of these innovations can be quite useful in practice. Multiple smoothing parameters are essential for multiple predictor variables where main effects should be penalized differently than interactions; see section 4. Nonquadratic penalties are much more effective when estimating changepoints than quadratic penalties; see section 5.1.

We define $\hat{\boldsymbol{\beta}}(\alpha)$ to be the minimizer of

$$\sum_{i=1}^n \left\{ Y_i - m(x; \boldsymbol{\beta}) \right\}^2 + \alpha \sum_{k=1}^K \rho(\beta_{p+k}), \quad (3)$$

where ρ is a suitable nonnegative function, and α is a smoothing parameter. Because α 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 α is selected appropriately for each K and that $p \geq 2$. Selection of α will be discussed in the next section.

We start with the simplest case, where $\rho(x) = x^2$. We will see, however, that the choice of ρ is important and that nonquadratic penalties can have real advantages. Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ and \mathbf{X} be the “design matrix” for the regression spline so that the i th row of \mathbf{X} is

$$\mathbf{X}_i = (1, \quad X_i, \quad \dots \quad X_i^p, \quad (X_i - \kappa_1)_+^p, \quad \dots \quad (X_i - \kappa_K)_+^p). \quad (4)$$

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

$$\hat{\boldsymbol{\beta}}(\alpha) = (\mathbf{X}^T \mathbf{X} + \alpha \mathbf{D})^{-1} \mathbf{X}^T \mathbf{Y}. \quad (5)$$

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 α . The computational time for the matrices $\mathbf{X}^T \mathbf{X}$ and $\mathbf{X}^T \mathbf{Y}$ is linear in n , but these matrices need only be computed once. As Eilers and Marx (1996) mention, after these matrices are computed, only $K \times K$ matrices need to be manipulated.

This allows rapid selection of α by techniques such as minimizing C_p or generalized cross-validation when $\hat{\beta}(\alpha)$ is calculated over a grid of values of α .

In the next section, data-based selectors of α 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. Section 6 discuss logsplines for estimation of conditional variances for heteroscedastic data. Other uses of penalized regression spline, e.g., generalized regression and log spline density estimation, are discussed by Eilers and Marx (1996) and Ruppert and Carroll (1996).

2 SELECTION OF THE SMOOTHING PARAMETER

Using a suitable value of α is crucial to obtaining a satisfactory curve estimate. A simple method for selection of α is to minimize Mallows's C_p or the closely related generalized cross-validation (GCV) criterion. The use of C_p , GCV, cross-validation, and related criteria is controversial since they can be highly variable. However, we have found C_p and GCV to be satisfactory in the regression spline applications we have considered, although alternative methods of choosing the smoothing parameter should be investigated. Here we follow Hastie and Tibshirani (1990) closely, as do Eilers and Marx (1996). Let

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

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

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

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

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \{Y_i - m(X_i; \hat{\beta}(\alpha^*))\}^2}{n - \text{tr}\{2\mathbf{S}(\alpha^*) - \mathbf{S}^2(\alpha^*)\}}. \quad (6)$$

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

$$C_p(\alpha) = \text{ASR}(\alpha) + \frac{2\text{tr}(\mathbf{S}(\alpha))\hat{\sigma}^2}{n} \quad (7)$$

is the C_p statistic, and

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

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

We choose α by computing either $C_p(\alpha)$ or $\text{GCV}(\alpha)$ for a grid of α 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 $\mathbf{S}(\alpha)$ since if

$$\mathbf{C}(\alpha) = (\mathbf{X}^T \mathbf{X})(\mathbf{X}^T \mathbf{X} + \alpha \mathbf{D})^{-1},$$

then

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

As an example, we used the LIDAR data set discussed in Ruppert, Wand, Holst, and Hössjer (1997). The data were collected by emitting lasers at two frequencies into an emissions plume. The amounts of light reflected back can then be used to measure mercury concentration. There are 221 observations with x being range, that is, the distance from the light travels before reflected back to the source, and y being the log of the ratio of the received power on and off the resonance frequency of mercury.

The grid of thirty α values was log-spaced between 10^{-2} and 10^{10} , i.e., their base-10 logarithms were equally spaced between -2 and 10 , with the smallest grid value being used as α^* . There were six choices of (p, K) : $(2, 5)$, $(2, 10)$, $(2, 40)$, $(1, 5)$, $(1, 20)$, and $(3, 10)$. In each case the knots were equally-spaced sample quantiles and $\rho(x) = x^2$. Figure 1 shows the fits that minimize C_p for the six values of (p, K) . The raw data are plotted in the bottom right. The plots show that in this example, the value of K has little effect, provided there are at least 10 knots for $p = 1$ and at least 5 knots for $p = 2$ or 3. Computational time in MATLAB on a SPARC Ultra 1 is “interactive” (0.6 seconds) for $n = 221$, $K = 40$, $p = 2$, and 30 values of α .

In the bottom left panel, we plot the -1000 times the derivative of the curve estimate using $p = 2$ and $k = 40$. Scientific interest is in $m'(x)$ which is equal to the concentration of mercury at x times a known negative constant. A nice feature of a spline, e.g., compared to a local polynomial fit, is that a spline can be differentiated analytically. The peak in $-\hat{m}'$ occurs at an emissions plume at range approximately 550 to 650. The plume originates in an industrial smokestack, and the objective is to estimate the amount of mercury being emitted. Ideally, \hat{m}' should use a smoothing parameter chosen to minimize the average squared error of \hat{m}' itself, not \hat{m} —this would be an interesting area for future research. However, the minimum C_p value of α does give a visually appealing estimate of m' . Moreover, Ruppert, Wand, Holst, and Hössjer (1997) estimated both m and m' by local polynomial regression using the EBBS bandwidths of Ruppert (1997b). They found that the optimal bandwidths for m and m' were similar for this data set, which suggests that \hat{m} and \hat{m}' need about the same amount of smoothing.

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. We have found that piecewise linear splines are inferior for smooth functions with substantial curvature, e.g. a sinusoidal wave, but they are useful if m has a “kink” where m' is discontinuous; see Section 5. When using C_p for choosing α , we recommend that α^* should clearly undersmooth the data. We evaluate C_p on a log-spaced grid of α values and use the smallest grid value as α^* . If C_p is minimized over the grid by α^* , 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 $(X_{i1}, \dots, X_{iJ})^T$. To fit more complex models effectively, one may need to partition $\underline{\beta}$ 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

$$\mathbf{Y} = \sum_{m=1}^M \mathbf{X}(m) \underline{\beta}(m) + \epsilon = \underline{\mathbf{X}} \underline{\beta} + \epsilon, \quad (9)$$

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

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

$$\|\mathbf{Y} - \underline{\mathbf{X}} \underline{\beta}\|_2^2 + \sum_{m=1}^M \alpha_m \sum_{j=1}^{d_m} \rho\{\underline{\beta}(m)_j\}, \quad (10)$$

where $\|\cdot\|_2$ is the Euclidean (or L_2) norm, $\alpha_m \geq 0$ for $m = 1, \dots, M$, and $\underline{\beta}(m) = \{\beta(m)_1, \dots, \beta(m)_{d_m}\}^T$ so that d_m is the dimension of $\underline{\beta}(m)$. Equation (3) is a special case of (10) with $M = 2$, $\alpha_1 = 0$, and $\alpha_2 = \alpha$. Let $\alpha = (\alpha_1, \dots, \alpha_M)^T$, and let $\mathbf{D}(\alpha)$ be block diagonal with blocks $\alpha_1 I_{d_1}, \dots, \alpha_M I_{d_M}$. Here I_d is the $d \times d$ identity matrix. In the simple case where $\rho(x) = x^2$, we have that

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

Next, let

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

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

4 MULTIVARIATE, ADDITIVE AND INTERACTION MODELS

Recall that $(X_{i1}, \dots, X_{iJ})^T$ 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 $j = 1, \dots, J$, 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

$$\mathcal{B}(j) = \{\mathbf{1}\} \cup \mathcal{B}_P(j; p) \cup \mathcal{B}_{PP}(j; p, \kappa_{1j}, \dots, \kappa_{Kj}), \quad (13)$$

where $\mathbf{1}$ is the function identically equal to 1,

$$\mathcal{B}_P(j; p) = \{x_j, \dots, x_j^p\}$$

is the set of polynomial basis functions, and

$$\mathcal{B}_{PP}(j; p, \kappa_{1j}, \dots, \kappa_{Kj}) = \{(x_j - \kappa_{1j})_+^p, \dots, (x_j - \kappa_{Kj})_+^p\}$$

is the set of piecewise polynomial basis functions. The subscripts “ P ” and “ PP ” denote “polynomial” and “piecewise polynomial.” We will often denote $\mathcal{B}_P(j; p)$ by $\mathcal{B}_P(j)$ and $\mathcal{B}_{PP}(j; p, \kappa_{1j}, \dots, \kappa_{Kj})$ by $\mathcal{B}_{PP}(j)$ to save space. The tensor-product regression spline basis is $\mathcal{B}(1, \dots, J) \equiv_{\text{def}} \mathcal{B}(1) \otimes \dots \otimes \mathcal{B}(J)$, i.e., the set of all products $b(1) \dots b(J)$ where $b(j) \in \mathcal{B}(j)$. We use the notation “ $a \equiv_{\text{def}} b$ ” 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), i.e., $\mathcal{B}(j) = \{\mathbf{1}\} \cup \mathcal{B}_P(j) \cup \mathcal{B}_{PP}(j)$,

$$\mathcal{B}(1, 2) \equiv_{\text{def}} \mathcal{B}(1) \otimes \mathcal{B}(2) = \mathcal{B}_P(1, 2) \cup \mathcal{B}_{PP}(1, 2)$$

where

$$\mathcal{B}_P(1, 2) \equiv_{def} [\{\mathbf{1}\} \cup \mathcal{B}_P(1)] \otimes [\{\mathbf{1}\} \cup \mathcal{B}_P(2)]$$

and

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

Thus, $\mathcal{B}_P(1, 2)$ is a basis of a space of polynomials in x_1 and x_2 and $\mathcal{B}_{PP}(1, 2)$ a basis for a space of piecewise polynomial functions of x_1 and x_2 . Therefore, we let $M = 2$ and let $\mathbf{X}(1)$ and $\mathbf{X}(2)$ be generated by $\mathcal{B}_P(1, 2)$ and $\mathcal{B}_{PP}(1, 2)$, respectively, and then let $\alpha_1 = 0$ so only the coefficients of the piecewise polynomials are shrunk towards 0. (We say that \mathbf{X} is generated by a certain basis if each column of \mathbf{X} is constructed by taking a basis function and evaluating it at the observed predictor vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$.)

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 were 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) give the data. These authors fit bivariate local linear regression to estimate mercury concentration as a function of spatial location. Ruppert (1997a) used local quadratic regression. Ruppert found that the residuals have no apparent correlation, which suggests that a regression model with independent errors is appropriate.

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 et al. (1995) and Ruppert (1997a) 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.

One referee objected that regression splines are not invariant to rotation of the coordinates as a spatial model should be. In principle this might be a concern, but as a practical matter in this example the bivariate regression spline fit is similar to that of rotation invariant local polynomial regression.

An early attempt to estimate mercury concentration by kriging experienced difficulty due to the nonstationary variance; the data are more variable near the peak in mercury concentration and log or power transformation did not remove this problem. A result of this nonstationarity is that the sample variogram is nonmonotonic. Nonparametric regression does not suffer from this problem.

4.2 Additive models

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

$$\mathcal{B}(1) \cup \dots \cup \mathcal{B}(J) = \{\mathbf{1}\} \cup \left[\bigcup_{j=1}^J \{\mathcal{B}_P(j) \cup \mathcal{B}_{PP}(j)\} \right] = \mathcal{P}(1, \dots, J) \cup \mathcal{PP}(1, \dots, J), \quad (14)$$

where $\mathcal{P}(1, \dots, J) \equiv_{\text{def}} \{\mathbf{1}\} \cup \mathcal{B}_P(1) \cup \dots \cup \mathcal{B}_P(J)$ is a basis for additive polynomial models and $\mathcal{PP}(1, \dots, J) \equiv_{\text{def}} \mathcal{B}_{PP}(1) \cup \dots \cup \mathcal{B}_{PP}(J)$ is a basis for additive piecewise polynomial models.

When fitting an additive model, we can let $M = 2$ and let $\mathbf{X}(1)$ be generated by $\mathcal{P}(1, \dots, J)$ while $\mathbf{X}(2)$ is generated by $\mathcal{PP}(1, \dots, J)$. Another possibility is to apply a different amount of shrinkage to each predictor variable, so that $M = J + 1$, $\mathbf{X}(1)$ is as before, while $\mathbf{X}(j + 1)$ is generated by $\mathcal{B}_{PP}(j), j = 1, \dots, J$. 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 $m(\mathbf{x}) = m_1(x_1) + \dots + m_J(x_J)$ can be estimated by

$$\widehat{m}_j(x_j) = \sum_{l=1}^p \widehat{\beta}_{lj} x_j^l + \sum_{k=1}^K \widehat{\beta}_{kj} (x_j - \kappa_{kj})_+^p$$

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

$$\sum_{i=1}^n \widehat{m}_j(X_{ij}) = 0, \quad (15)$$

where $\{X_{1j}, \dots, X_{nj}\}$ are the observed values of the j th predictor or

$$\sum_{i=1}^N \widehat{m}_j(x_{ij}) = 0, \quad (16)$$

where $\{x_{1j}, \dots, x_{Nj}\}$ is an equally-spaced grid of points over some finite interval, say $[\min_i \{X_{ij}\}, \max_i \{X_{ij}\}]$. The intercept, $\widehat{\beta}_0$, i.e., the estimated coefficient of $\mathbf{1}$, would

then be replaced by $\hat{\beta}_0 + \sum_{j=1}^J C(j)$. The centering makes the \widehat{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. (Identifiability 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.)

4.3 Interaction models

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

$$\begin{aligned} \mathcal{B}(1, \dots, J) &= \bigotimes_{j=1}^J [\{\mathbf{1}\} \cup \mathcal{B}_{P,PP}(j)] \\ &= \{\mathbf{1}\} \cup \left[\bigcup_{j=1}^J \mathcal{B}_{P,PP}(j) \right] \\ &\cup \left[\bigcup_{j_1=1}^J \bigcup_{j_2=j_1+1}^J \mathcal{B}_{P,PP}(j_1) \otimes \mathcal{B}_{P,PP}(j_2) \right] \cup \dots \\ &\cup \left[\bigcup_{j_1=1}^J \dots \bigcup_{j_J=j_{(J-1)}+1}^J \mathcal{B}_{P,PP}(j_1) \otimes \dots \otimes \mathcal{B}_{P,PP}(j_J) \right]. \end{aligned}$$

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 $\mathcal{B}(1, \dots, J)$ by decomposition of the design matrix \mathbf{X} into components $\mathbf{X}(1), \dots, \mathbf{X}(M)$ 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

$$\mathcal{M}_P = \bigcup_{j=1}^J \mathcal{B}_P(j)$$

be the polynomial main effects basis functions, let

$$\mathcal{M}_{PP} = \bigcup_{j=1}^J \mathcal{B}_{PP}(j)$$

be the piecewise polynomial main effects, let

$$\mathcal{I}_P = \bigcup_{j_1=1}^{J-1} \bigcup_{j_2=j_1+1}^J \mathcal{B}_P(j_1) \otimes \mathcal{B}_P(j_2)$$

be the polynomial two-way interaction basis functions, and let

$$\mathcal{I}_{PP} = \bigcup_{j_1=1}^{J-1} \bigcup_{j_2=j_1+1}^J \left[\left\{ \mathcal{B}_{P,PP}(j_1) \otimes \mathcal{B}_{PP}(j_2) \right\} \cup \left\{ \mathcal{B}_{PP}(j_1) \otimes \mathcal{B}_P(j_2) \right\} \right]$$

be the piecewise polynomial two-way interaction basis functions.

Our approach is to use $M = 3$ blocks of basis functions with $\mathbf{X}(1)$, $\mathbf{X}(2)$, and $\mathbf{X}(3)$ generated by $\{\mathbf{1}\} \cup \mathcal{M}_P$, \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 $\mathbf{X}(3)$, 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 an alternative to interaction smoothing splines discussed by Wahba (1986, 1990), Chen (1991), and Gu and Wahba (1993) and to nonpenalized regression spline models of Stone (1994).

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 due to wind effects and the hilly terrain. These considerations suggest the model

$$Hg = m_1(d) + m_{23}(x_1, x_2) + \epsilon,$$

where Hg is mercury concentration and ϵ 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 $\mathbf{X}(1)$, the piecewise polynomial main effects for d generated $\mathbf{X}(2)$, and *all* main effects and interaction effects for (x_1, x_2) formed $\mathbf{X}(3)$. We took $\alpha_1 = 0$ and α_2 and α_3 were chosen 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 α_3 was five times larger than α_2 . (α_2 was 1,000 and α_3 was 5,000—these values minimized C_p over a rectangular grid where each of α_1 and α_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 $\widehat{m}_1(d) + \widehat{m}_{23}(x_1, x_2)$ is plotted in Figure 3. One can see that $\widehat{m}_1(d)$ 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 \widehat{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 $\widehat{m}_1(d) + \widehat{m}_{23}(x_1, x_2)$ 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. This example illustrates the ability of regression splines with multiple smoothing parameters to provide estimators custom-designed for specific applications.

5 NONQUADRATIC PENALTY FUNCTIONS

Using $\rho(x) = x^2$ is convenient mathematically and computationally since this choice of objective function makes (10) a quadratic function of $\underline{\beta}$ and leads to the linear estimator (11) that can be computed non-iteratively. As we have seen, the quadratic penalty function often 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 $\epsilon_1, \dots, \epsilon_n$ are iid $N(0, \sigma^2)$. Suppose as well that the prior distribution is that $\beta(1), \dots, \beta(M)$ are independent with $\beta(m)$ having the multivariate normal distribution $N(0, \sigma_m^2 I_{d_m})$. For now, assume that $\sigma^2, \sigma_1^2, \dots, \sigma_M^2$ are known. Then the posterior log density of β given \mathbf{Y} is, up to an additive function of \mathbf{Y} and $(\sigma^2, \sigma_1^2, \dots, \sigma_M^2)$, given by

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

Thus, the maximum a posteriori (MAP) estimator, i.e., the mode of the posterior density, minimizes (10) with $\rho(x) = x^2$ and $\alpha_m = (\sigma/\sigma_m)^2$, $m = 1, \dots, M$. Of course,

these variances usually will not be known in practice. An alternative to choosing $\alpha_1, \dots, \alpha_M$ by C_p or GCV would be to estimate $\sigma^2, \dots, \sigma_M^2$ by hierarchical Bayesian or empirical Bayesian methods. The subvector $\boldsymbol{\beta}(m)$ is unpenalized when $\alpha_m = 0$ which means one takes σ_m^2 to be ∞ .

For concreteness, consider the univariate regression spline model given by (2), so that $\boldsymbol{\beta}(1) = (\beta_0, \dots, \beta_p)^T$ are the polynomial coefficients and $\boldsymbol{\beta}(2) = (\beta_{p+1}, \dots, \beta_{p+K})^T$ 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 $N(0, \sigma_2^2)$ random variables. Also, one uses $\sigma_1^2 = \infty$ 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

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

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 \widehat{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) “rho function”

$$\begin{aligned}\rho_H(x) &= \frac{x^2}{2} & |x| \leq k_H \\ &= k_H|x| - \frac{k_H^2}{2} & |x| > k_H,\end{aligned}$$

where k_H is a positive tuning constant. The Huber penalty allows \widehat{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 $\boldsymbol{\beta}(m)$. This scaling can be achieved by using k_H between .5 and 1.5 (say) and standardizing $\boldsymbol{\beta}(m)$ by its MAD (median absolute deviation). More specifically, let $\text{MED}\{\boldsymbol{\beta}(m)\} = \text{median}\{\boldsymbol{\beta}(m)_1, \dots, \boldsymbol{\beta}(m)_{d_m}\}$ and $\text{MAD}\{\boldsymbol{\beta}(m)\} = \text{median}\{|\boldsymbol{\beta}(m)_1 - \text{MED}\{\boldsymbol{\beta}(m)\}|, \dots, |\boldsymbol{\beta}(m)_{d_m} - \text{MED}\{\boldsymbol{\beta}(m)\}|\}$. Then the penalty in (10) is

$$\sum_{m=1}^M \alpha_m \sum_{j=1}^{d_m} \rho_H \left[\frac{\boldsymbol{\beta}(m)_j}{\text{MAD}\{\boldsymbol{\beta}(m)\}} \right]. \quad (18)$$

The L_1 and L_2 penalties are special cases of the L_q penalty

$$\|x\|_q^q = \sum_{i=1}^{\dim(x)} |x_i|^q, \quad q > 0.$$

Despite the notation, $\|\cdot\|_q$ is not a norm for $q < 1$ but it is nonetheless a reasonable choice for a penalty function. The L_q penalties also have a Bayesian derivation with prior

$$f(u; q, \sigma) \equiv_{\text{def}} \frac{1}{\sigma} \exp \left\{ -\frac{1}{2} \left| \frac{u}{\sigma} \right|^q - \eta(q) \right\}, \quad -\infty < u < \infty, \quad (19)$$

where $\sigma > 0$, $q > 0$, and $\eta(q) = \log\{q^{-1}2^{1+q^{-1}}\Gamma(q^{-1})\}$. This family of probability density functions includes the scaled double exponential, $(1/4\sigma) \exp(-\frac{1}{2}|\frac{x}{\sigma}|)$, when $q = 1$ and the normal when $q = 2$. See Box and Tiao (1973, Section 3.2.1) for discussion of (19) as a class of priors in parametric estimation and for earlier references.

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

$$\widehat{\boldsymbol{\beta}}^{(N+1)} = \left[\mathbf{X}^T \mathbf{X} + \frac{1}{2} \mathbf{D}(\boldsymbol{\alpha}) \text{diag} \left\{ \frac{\rho' \{ \widehat{\boldsymbol{\beta}}^{(N)}(m)_j / \text{MAD}_m \}}{\{ \widehat{\boldsymbol{\beta}}^{(N)}(m)_j / (\text{MAD}_m) \}} \right\} \right]^{-1} \mathbf{X}^T \mathbf{Y}. \quad (20)$$

Here $\widehat{\boldsymbol{\beta}}^{(N)} = (\widehat{\beta}_1^{(N)}, \dots, \widehat{\beta}_d^{(N)})^T$ where $d = d_1 + \dots + d_M$ and $\text{MAD}_m = \text{MAD}\{\widehat{\boldsymbol{\beta}}^{(N)}(m)\}$. The derivation of (20) is found in the appendix. If one uses $\rho(x) = x^2$, then

$\rho'(x)/x \equiv 2$ and (20) does not change from its starting estimate. We call (20) iteratively reweighted ridge regression (IRRR). We have found that IRRR is easy to implement and that it produces very satisfactory estimates for both the Huber and L_q penalties, though we have not investigated how closely IRRR finds the actual minimum of the objective function since our goal is only to produce satisfactory estimates. For the L_1 penalty Tibshirani (1996) recommends a computational method based on quadratic programming to find the actual minimum.

From (20) we see that at convergence, the fitted values are $\mathbf{S}(\boldsymbol{\alpha})\mathbf{Y}$ where

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

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

5.1 A jump function

Functions with discontinuities are common in two-dimensional image analysis. To get insight into the problems involved in estimating such functions, several authors have looked at one-dimensional jump functions; see Chu, Glad, Godtliebsen, and Marron (1997) and Donoho, Johnstone, Kerkycharian, and Picard (1995).

Figure 4 uses piecewise constant regression spline fits to such a function. The 200 simulated data points and the true function are shown in the upper left panel. The other three panels show the true function (dashed) and estimates (solid) using the L_q penalty with $q = 2, 1$, and $.05$. (We used $q = .05$ to try a value of q close to 0.) In each case, there were 60 knots and α was chosen by minimizing C_p . The average squared error, i.e., $n^{-1} \sum (m(x_i) - \hat{m}(x_i))^2$, was $.342$, $.293$, and $.278$, for $q = 2, 1$, and $.05$, respectively. The plots show clearly that the L_2 penalty estimator cannot estimate the changepoints. The L_1 penalty estimator is a major improvement and the $L_{.05}$ penalty estimator is essentially able to capture the jump function structure of m , even though that structure, amid the data scatter, is difficult for the human eye to detect.

5.2 LIDAR data with L_q penalties

To see how L_q penalties work with smooth data, we used the LIDAR data with 20 knots. When $p = 2$, there is little difference between the $q = .05, 1$, and 2 fits, and, in fact, virtually no difference between $q = 1$ and $q = 2$. When $p = 1$, the choice of q is important, with the fitted curve becoming less smooth and showing more kinks as q decrease; see Figure 5.

Although a reasonable choice of q can perhaps be made visually, data-based penalty selection is an important area for further research. We are currently developing Bayesian methods for estimating the best choice of q .

5.3 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 14*ms*, 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.

These data have been analyzed by many authors, including Eilers and Marx (1996). Unlike previous authors, we focus upon estimation of m near the time of impact.

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 $L_{1/2}$, 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 in a neighborhood of the time of impact are shown in Figure 6. In each case, α was first chosen by minimizing C_p .

When the quadratic penalty is used, the fit at the point of impact is not entirely satisfactory, especially for $p = 2$ where 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 ($p = 1$) also round off the kink when the quadratic penalty is used. The nonquadratic penalties and $p = 1$ best accommodate the apparent changepoint in m' .

Silverman (1985) noticed the substantial heteroscedasticity in these data, and he 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 by a moving average of the squared residuals from a preliminary unweighted spline estimate. In work not presented here, we performed a similar weighted analysis by fitting a logspline model to the absolute residuals as discussed in Section 6.

As a referee has mentioned, the data are a time series. One could model the errors as a correlated process. The first two autocorrelations of the residuals are $-.16$ and $-.18$, so there is some evidence of weak dependence. Given the structure of the mean and variance functions, one would probably retain the regression spline model or some other nonparametric model for m , rather than modeling the data as a realization of a stationary process.

6 ESTIMATING VARIANCE FUNCTIONS BY LOGSPINES

When $\text{Var}(Y_i|X_{i1}, \dots, X_{iJ})$ is a nonconstant function, there are several good reasons for modeling this variance function. First, a weighted estimator of $E(Y_i|X_{i1}, \dots, X_{iJ})$ 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 with

$$\text{Var}(Y_i|X_{i1}, \dots, X_{iJ}) = \exp(2\boldsymbol{\theta}^T \mathbf{X}_i), \quad (21)$$

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

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

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

$$\sum_{i=1}^n \left\{ \frac{e_i^2 - \exp(2\boldsymbol{\theta}^T \mathbf{X}_i)}{\exp(2\hat{\boldsymbol{\theta}}_{\text{prel}}^T \mathbf{X}_i)} \right\}^2 + \sum_{m=1}^{M_2} \alpha_m \|\boldsymbol{\theta}(m)\|_2^2$$

over $\boldsymbol{\theta}$. Least-squares estimation based on squared residuals is highly sensitive to outliers. A sensible alternative is to use least-squares based on absolute residuals; see Carroll and Ruppert (1988). Let $\epsilon_i = Y_i - E(Y_i|X_{i1}, \dots, X_{iJ})$. We will use the model:

$$E|\epsilon_i| = \exp(\boldsymbol{\theta}^T \mathbf{X}_i). \quad (22)$$

Model (22) will hold with the same value of $\boldsymbol{\theta}$ as in (21) except for a change in intercept if $\epsilon_i = \exp(\boldsymbol{\theta}^T \mathbf{X}_i)u_i$ where u_1, \dots, u_n are iid from any distribution with a finite variance. Using model (22), we estimate $\boldsymbol{\theta}$ by minimizing

$$\sum_{i=1}^n \left\{ \frac{|e_i| - \exp(\boldsymbol{\theta}^T \mathbf{X}_i)}{\exp(\hat{\boldsymbol{\theta}}_{\text{prel}}^T \mathbf{X}_i)} \right\}^2 + \sum_{m=1}^{M_2} \alpha_m \|\boldsymbol{\theta}(m)\|_2^2.$$

We can unify estimation based on absolute and squared residuals by minimizing the objective function

$$\sum_{i=1}^n \left\{ \frac{|e_i|^\nu - \exp(\nu \underline{\boldsymbol{\theta}}^T \mathbf{X}_i)}{\exp(\nu \hat{\underline{\boldsymbol{\theta}}}_{prel}^T \mathbf{X}_i)} \right\}^2 + \sum_{m=1}^{M_2} \alpha_m \|\boldsymbol{\theta}(m)\|_2^2, \quad (23)$$

where typically ν equals 1 or 2, though other values are possible. Differentiating (23) with respect to $\underline{\boldsymbol{\theta}}$, we get the estimating equation

$$0 = - \sum_{i=1}^n \left\{ \frac{|e_i|^\nu - \exp(\nu \underline{\boldsymbol{\theta}}^T \mathbf{X}_i)}{\exp(2\nu \hat{\underline{\boldsymbol{\theta}}}_{prel}^T \mathbf{X}_i)} \right\} \exp(\nu \underline{\boldsymbol{\theta}}^T \mathbf{X}_i) (\nu \mathbf{X}_i) + \mathbf{D}_2(\boldsymbol{\alpha}) \underline{\boldsymbol{\theta}}, \quad (24)$$

where $\mathbf{D}_2(\boldsymbol{\alpha})$ is block diagonal with blocks $\alpha_1 I_{d_1}, \dots, \alpha_{M_2} I_{d_{M_2}}$.

Equation (24) is the basis of an iterative estimation scheme where given $\hat{\underline{\boldsymbol{\theta}}}_{old}$ we solve for $\hat{\underline{\boldsymbol{\theta}}}_{new}$ in

$$\begin{aligned} 0 = & - \sum_{i=1}^n \left[\left\{ \frac{|e_i|^\nu - \exp(\nu \hat{\underline{\boldsymbol{\theta}}}_{old}^T \mathbf{X}_i) (1 + \nu (\hat{\underline{\boldsymbol{\theta}}}_{new} - \hat{\underline{\boldsymbol{\theta}}}_{old})^T \mathbf{X}_i)}{\exp(\nu \hat{\underline{\boldsymbol{\theta}}}_{old}^T \mathbf{X}_i)} \right\} \nu \mathbf{X}_i \right] \\ & + \mathbf{D}_2(\boldsymbol{\alpha}) \hat{\underline{\boldsymbol{\theta}}}_{old} + \mathbf{D}_2(\boldsymbol{\alpha}) (\hat{\underline{\boldsymbol{\theta}}}_{new} - \hat{\underline{\boldsymbol{\theta}}}_{old}). \end{aligned}$$

The solution is

$$\begin{aligned} \hat{\underline{\boldsymbol{\theta}}}_{new} = & \hat{\underline{\boldsymbol{\theta}}}_{old} + \left\{ \sum_{i=1}^n \nu^2 \mathbf{X}_i \mathbf{X}_i^T + \mathbf{D}_2(\boldsymbol{\alpha}) \right\}^{-1} \\ & \left[\sum_{i=1}^n \left\{ \frac{|e_i|^\nu - \exp(\nu \hat{\underline{\boldsymbol{\theta}}}_{old}^T \mathbf{X}_i)}{\exp(\nu \hat{\underline{\boldsymbol{\theta}}}_{old}^T \mathbf{X}_i)} (\nu \mathbf{X}_i) \right\} - \mathbf{D}_2(\boldsymbol{\alpha}) \hat{\underline{\boldsymbol{\theta}}}_{old} \right]. \end{aligned} \quad (25)$$

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 residuals to get weights.

Within each step of 2. in the main iteration there is a subloop where one iterates (25). 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; see Carroll, Wu, and Ruppert (1987). As in their results, we expect that two or three iterations will be acceptable for penalized estimation.

The GCV criterion is given by (8) with

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

and $\mathbf{S}(\boldsymbol{\alpha}) = \underline{\mathbf{X}} \{ \underline{\mathbf{X}}^T \underline{\mathbf{X}} + \mathbf{D}_2(\boldsymbol{\alpha}) \}^{-1} \underline{\mathbf{X}}^T$. Here $\underline{\mathbf{X}}$ is the matrix with i th row given by $\nu \mathbf{X}_i$.

6.1 LIDAR Data — Modeling the Variance Function

Figure 7 is an estimate of $E(|\epsilon_i| \mid X_i)$ for the LIDAR data. We used two iterations of the algorithm just described with $\nu = 1$. In the subloop, step (25) was iterated 15 times, or until the relative change in $\hat{\theta}$ was less than 10^{-6} , whichever came first.

This is an example where a log or power transformation will not stabilize the variance, since the variance does not depend on the mean but rather on the variable “range.” One can see in Figures 1 and 7 that as range increases from 400 to 550, the mean response changes little if at all while the estimated $E(|\epsilon_i| \mid X_i)$ more than doubles. Another difficulty with power or log transformations is that the response is mostly negative, though there are some positive values, so that a shift parameter must be chosen in addition to the transformation.

7 Spatial variability and local smoothing parameters

Consider again the univariate case given by model (1). The estimator $m(x; \hat{\beta}(\alpha))$ has its smoothness controlled by a single smoothing parameter α . If m has rapid oscillations in some regions of x 's but is relatively smooth in other regions, then a single smoothing parameter may not achieve a good overall fit. Rather, to get a satisfactory smooth, α must vary in accordance with the spatial variability in m . The problems caused by having a single smoothing parameter are shown by Wand (1997).

Here is a simple approach to spatially varying α . Fix a subset of the knots including the smallest and largest knots, say, $\{\kappa_{k(1)}, \dots, \kappa_{k(M)}\}$ where $k(1) = 1$ and $k(M) = K$ and M is much smaller than K . The penalty at the subknot $\kappa_{k(j)}$ is controlled by an independently varying smoothing parameter $\alpha_{k(j)}$. The penalties at knots other than the subknots are determined by cubic interpolation of the penalties at the subknots, the interpolation being done on the log-penalty scale to ensure positivity of the penalties. Thus, we have a penalty α_k at each κ_k but with only $\alpha_{k(j)}$, $j = 1, \dots, M$, to be selected. Given these penalty weights, $\hat{\beta}(\alpha_{k(1)}, \dots, \alpha_{k(M)})$ is defined by (5) with \mathbf{D} replaced by $\text{diag}(\alpha_1, \dots, \alpha_K)$.

The smoothing parameters $\alpha_{k(1)}, \dots, \alpha_{k(M)}$ can be determined by minimizing C_p . A search over an M -dimensional grid is not recommended because of computational cost. Rather, we recommend that one start with $\alpha_{k(1)}, \dots, \alpha_{k(M)}$ each equal to the best global value of α chosen by minimizing C_p , call it $\hat{\alpha}_{gl}$. Then the $\alpha_{k(j)}$ are varied, one-at-a-time with the others fixed, over a one-dimensional grid centered at $\hat{\alpha}_{gl}$ and then set equal to the value minimizing C_p on this grid. After minimizing over each $\alpha_{k(j)}$ this procedure could be repeated, but we have not investigated this possibility since the estimator without repetition has worked quite well. Although minimizing

C_p over the $\alpha_{k(j)}$'s one at a time in this manner does not guarantee finding the global minimum of C_p over $\alpha_{k(1)}, \dots, \alpha_{k(M)}$, our experience shows that this procedure is effective in selecting the right amount of local smoothing.

7.1 A simulation example

We performed a small Monte Carlo experiment using the “spatial variability” scenario in Wand (1997). The x 's were equally spaced on $[0,1]$, n was 400, and the ϵ_i 's were independent $N(0, (0.2)^2)$. The regression function, whose spatial variability was controlled by a parameter j , was

$$m(x; j) = \sqrt{x(1-x)} \sin \left\{ \frac{2\pi(1 + 2^{(9-4j)/5})}{x + 2^{(9-4j)/5}} \right\}.$$

We used both $j = 3$ which gave low spatial variability and $j = 6$ which gives rather severe spatial variability; see panels (a) and (b) of Figure 8. We used 40 and 80 knots. When we used 40 knots, then $\{\kappa_{k(j)} : j = 1, 10, 20, 30, 40\}$ were the subknots used for the local penalty. For 80 knots, $\{\kappa_{k(j)} : j = 1, 20, 40, 60, 80\}$ were the subknots. In all cases, quadratic splines were used. For each of the four combinations of j and K , we simulated 250 data sets and applied the global and local penalty function estimators to each. Boxplots of

$$\log_{10}(\text{RMSE}) = \log_{10} \left(n^{-1} \sum_{i=1}^n (\hat{m}(x_i) - m(x_i))^2 \right).$$

are shown in Figure 8. These are comparable to the boxplots in Figure 5 of Wand (1997); his middle left panel is $j = 3$ while his bottom right panel is $j = 6$.

From the results in Figure 8 we may draw the following conclusion:

- Locally varying penalties are as effective as a global penalty when there is little spatial variability. This conclusion does not necessarily apply, of course, to situations where n is small.
- For severe spatial variability, the local penalty approach is far superior to a global penalty.
- There is little difference between using 40 and 80 knots, except for one important situation. If one uses a local penalty and $j = 6$, then 80 knots is significantly better than 40. The reason is that 80 knots allows the spline to track the rapid oscillations on the left, but only if a local penalty is used.

Also, comparing the results in Figure 8 to the results in Wand (1997) for $j = 6$, the local penalty approach is somewhat better than the Bayesian method of Smith and

Kohn (1996) and the stepwise selection method of Stone, Hansen, Kooperberg, and Truong (1997). However, Wand’s simulations used code provide by Smith that had 35 knots “hard-wired” into it (Wand, personal communication). With more knots, the Smith and Kohn method could very well be competitive with the local penalty method.

We have also looked at moderate spatial variability ($j = 4$ or 5). There the local penalty estimator is better than the global penalty estimator, and again the local penalty estimator is as good as the Bayesian and stepwise methods studied by Wand.

We also tried cubic regression splines with both global and local penalties. We found cubic splines somewhat inferior to quadratic splines. Overall, we recommend against using cubic splines in our penalty method of regression spline estimation, since we have not seen any situations where they outperform quadratic splines.

A recent paper by Luo and Wahba combines knot selection and smoothing by a penalty. Their results show that they their “hybrid splines” are spatially adaptive.

8 FURTHER DISCUSSION

8.1 Robust Estimation

In (3), the residuals are squared but an arbitrary rho-function is used for the roughness penalty. To remove sensitivity to outlying responses, one can follow the method of M -estimation in robust regression and replace the quadratic goodness-of-fit penalty on the residuals by a second rho-function penalty, e.g., using the Huber’s rho. Cunningham, Eubank, and Hsing (1991) have developed such “ M -type” estimators for smoothing splines but only with a quadratic roughness penalty.

8.2 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 $\mathbf{X}(1)$ 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. This is especially appropriate

if some of the predictor variables are categorical. One could also model main effects and possibly low order interactions nonparametrically and model some higher order interactions parametrically.

8.3 Pseudosplines

There are some connections between our work and Hastie's (1996) idea of a *pseudospline*. A linear smoother can be written as $\mathbf{S}\mathbf{Y}$ where \mathbf{S} is a $n \times n$ matrix. Typically, \mathbf{S} is of full rank and all eigenvalues of \mathbf{S} are positive but most of the eigenvalues of \mathbf{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 may be difficult to find the eigenvalue/eigenvector decomposition of \mathbf{S} if n is at all large and \mathbf{S} is of full rank. Hastie's idea is to approximate \mathbf{S} by a low rank matrix.

Following Hastie's notation, let the "design matrix" \mathbf{P} be generated by an orthogonal basis of dimension k and let $D_\theta = \text{diag}(\theta_1, \dots, \theta_k)$. Then Hastie defines a pseudospline as $\mathbf{P}\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ minimizes

$$Q_\lambda(\boldsymbol{\beta}, \mathbf{Y}) = \|\mathbf{Y} - \mathbf{P}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^T \mathbf{D}_\theta \boldsymbol{\beta}. \quad (26)$$

Hastie also mentions that the pseudospline equals $\mathbf{S}_\lambda(\mathbf{P}, \theta)\mathbf{Y}$ where

$$\mathbf{S}_\lambda(\mathbf{P}, \theta) = \mathbf{P}(\mathbf{P}^T \mathbf{P} + \lambda \mathbf{D}_\theta)^{-1} \mathbf{P}^T.$$

The smoother matrix $\mathbf{S}_\lambda(\mathbf{P}, \theta)$ 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 (26) with (10) we see that if we use the L_2 penalty then our penalized regression spline is a pseudospline.

The point is that the rank of our smoother matrix $\mathbf{S}(\boldsymbol{\alpha})$ in (12) equals the rank of \mathbf{X} which is often much less than n . Moreover, the eigenvalues and eigenvectors of $\mathbf{S}(\boldsymbol{\alpha})$ can be found from those of a matrix whose size is $\text{rank}(\mathbf{X})$,

$$\mathbf{C}(\boldsymbol{\alpha}) \equiv_{\text{def}} (\mathbf{X}^T \mathbf{X} + \mathbf{D}(\boldsymbol{\alpha}))^{-1} (\mathbf{X}^T \mathbf{X}).$$

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

8.4 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. The finite dimensional parameter space easily allows non-Gaussian priors which correspond to the nonquadratic penalties we have explored in this paper. 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. 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 $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ plus (18). The gradient of the first component is $-2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta)$. If in (18) we replace $\text{MAD}\{\beta(m)\}$ by MAD_m which is considered fixed and not depending on $\beta(m)$, then the gradient of (18) is

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

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

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

which leads to (20).

REFERENCES

- Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.
- Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear smoothers and additive models (with discussion)," *The Annals of Statistics*, 17, 453–555.
- Carroll, R.J., and Ruppert, D. (1988), *Transformation and Weighting in Regression*, London: Chapman and Hall.
- Carroll, R.J., Wu, C.J.F., and Ruppert, D. (1988), "The effect of estimating weights in generalized least squares" *J. American Statistical Association*, 83, 1045–1054.

- Chen, Z. (1991), "Interaction spline models and their convergence rates," *The Annals of Statistics*, 19, 1855–1868.
- Chu, C.K., Glad, I., Godtliebsen, F., and Marron, J.S. (1997), "Edge preserving smoothers for image processing," *Journal of the American Statistical Association*, to appear.
- Cunningham, J. K., Eubank, R. L., and Hsing, T. (1991), "M-type smoothing splines with auxiliary scale estimation," *Computational Statistics and Data Analysis*, 11, 43–51.
- Donoho, D.L., Johnstone, I.M., Kerkycharian, G., and Picard, D. (1995), "Wavelet shrinkage: asymptotia (with discussion)," *Journal of the Royal Statistical Society*, 2, 301–370.
- Eilers, P.H.C., and Marx, B.D. (1996), "Flexible smoothing with B-splines and penalties (with discussion)," *Statistical Science*, 11, 89–121.
- Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, New York and Basil: Marcel Dekker.
- Friedman, J.H. (1991), "Multivariate adaptive regression splines (with discussion)," *The Annals of Statistics*, 19, 1–141.
- Friedman, J.H., and Silverman, B.W. (1989), "Flexible parsimonious smoothing and additive modeling (with discussion)," *Technometric*, 31, 3–39.
- Good, I. J., and Gaskins, R. A. (1971), "Non-parametric roughness penalties for probability densities," *Biometrika*, 58, 255–277.
- Green, P. J. (1987), "Penalized likelihood for general semi-parametric regression models," *International Statistical Review*, 55, 245–259.
- Green, P. J., and Silverman, B. (1994), *Nonparametric Regression and Generalized Linear Models*, London: Chapman and Hall.
- Gu, C., and Wahba, G. (1994), "Semiparametric analysis of variance with tensor product thin plate splines," *Journal Royal Statistical Society, Series B*, 55, 353–368.
- Hastie, T. (1996), "Pseudosplines," *Journal Royal Statistical Society, Series B*, 58, 379–396.
- Hastie, T.J., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.
- Kooperberg, C., and Stone, C. J. (1992), "A study of logspline density estimation," *Computational Statistics and Data Analysis*, 12, 327–347.
- Lindley, D. V., and Smith, A. F. M. (1972), "Bayes estimates for the linear model (with discussion)," *Journal Royal Statistical Society, Series B*, 34, 1–41.
- Luo, Z., and Wahba, G. (1997). "Hybrid adaptive splines," *Journal of the American Statistical Association*, 92, 107–116.

- McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models, 2nd Edition*, London: Chapman and Hall.
- Nychka, D. (1988), "Bayesian confidence intervals for smoothing splines," *Journal American Statistical Association*, 83, 1134–1143.
- Opsomer, J.D., Agras, J., Carpi, A., and Rodrigues, G. (1995), "An application of locally weighted regression to airborne mercury deposition around an incinerator site," *Environmetrics*, 6, 205–219.
- O'Sullivan, F. (1986), "A statistical perspective on ill-posed inverse problems (with discussion)," *Statistical Science*, 1, 505–527.
- O'Sullivan, F. (1988), "Fast computation of fully automated log-density and log-hazard estimators," *SIAM Journal of Scientific and Statistical Computation*, 9, 363–379.
- Ruppert, D. (1997a), "Local polynomial regression and its applications in environmental statistics," to appear in *Statistics for the Environment*, Volume 3 (V. Barnett and F. Turkman, editors) Chichester: John Wiley.
- Ruppert, D. (1997b). "Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation," *Journal of the American Statistical Association*, 92, to appear.
- Ruppert, D., and Carroll, R.J. (1996), "A simple roughness penalty approach to regression spline estimation," Technical Report #1167, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY (available at <http://www.orie.cornell.edu/trlist/trlist.html>).
- Ruppert, D., Wand, M., Holst, U., and Hossjer, O. (1997). "Local polynomial variance function estimation," *Technometrics*, 39, to appear.
- Scott, D. W., Tapia, R. A., and Thompson, J. R. (1980), "Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria," *The Annals of Statistics*, 8, 820–832.
- Silverman, B. (1985), "Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion)," *Journal Royal Statistical Society, Series B*, 47, 1–52.
- Smith, M., and Kohn, R. (1996), "Nonparametric regression using Bayesian variable selection," *Journal of Econometrics*, 75, 317–344.
- Stone, C.J., (1994), "The use of polynomial splines and their tensor products in multivariate function estimation," *The Annals of Statistics*, 22, 118–170.
- Stone, C.J., Hansen, M., Kooperberg, C., and Truong, Y. K. (1997). "Polynomial splines and their tensor products in extended linear modeling," *The Annals of Statistics*, to appear.
- Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal Royal Statistical Society, Series B*, 58, 267–288.

- Wahba, G. (1983), “Bayesian ‘confidence intervals’ for the cross-validated smoothing spline,” *Journal Royal Statistical Society, Series B*, 45, 133–150.
- Wahba, G. (1986), “Partial and interaction splines for the semiparametric estimation of functions of several variables,” In *Computer Science and Statistics: Proc. 18th Symposium Interface* (ed. T. J. Boardman), pp. 75–80. Washington, DC: American Statistical Association.
- Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: Society for Industrial and Applied Mathematics.
- Wand, M.P. (1997), “A Comparison of Regression Splines Smoothing Procedures,” Manuscript.

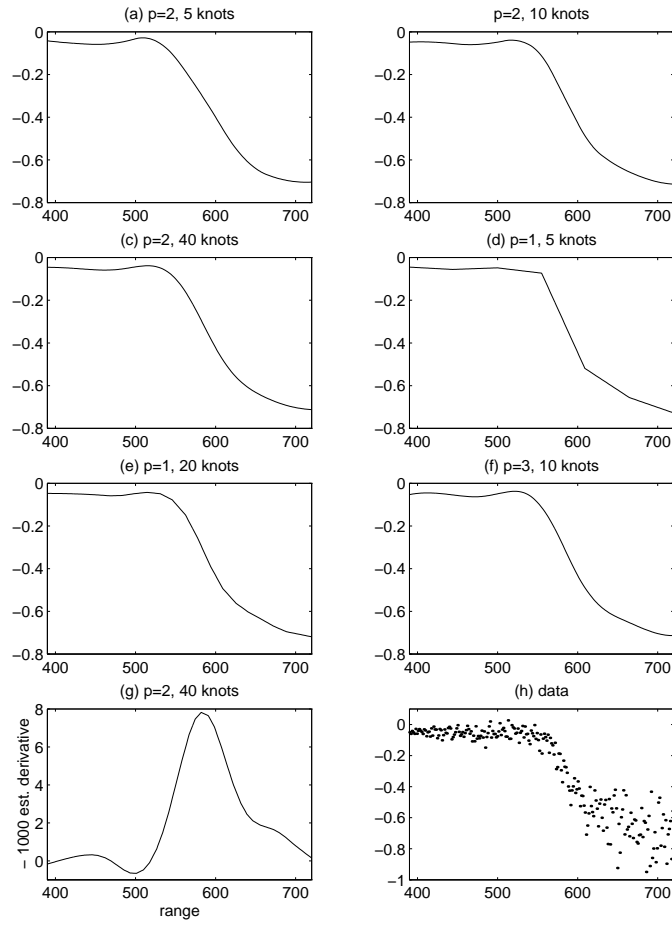


Figure 1: *LIDAR data. (a)–(f): estimates of m with various choices of p and K . (g) estimate of m' with $p = 2$ and $K = 40$. (h) data.*

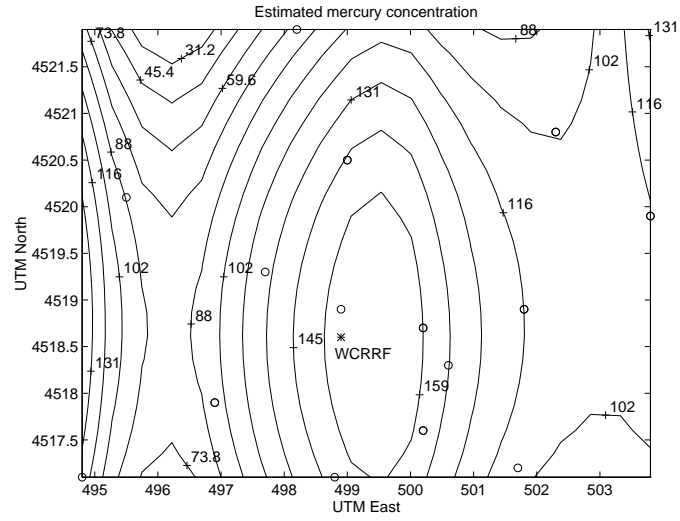


Figure 2: *Biomonitoring of airborne mercury. Bivariate tensor-product spline fit. Open circles are sampling locations, the asterisk is the location of the solid-waste incinerator, and the plus signs label the contours.*

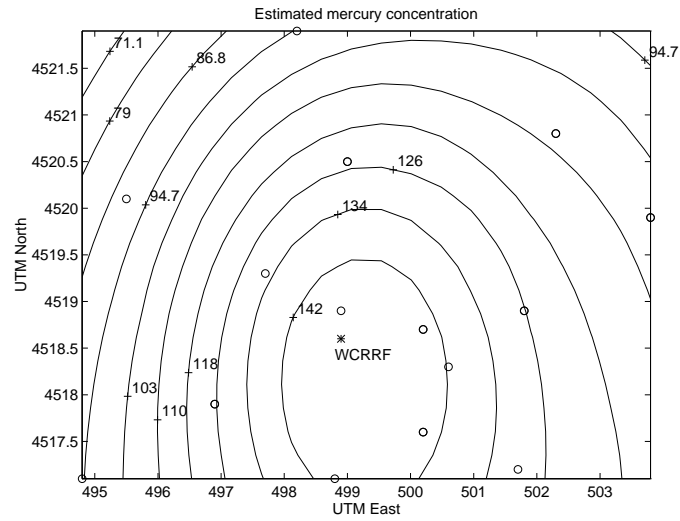


Figure 3: *Biomonitoring example. Regression spline fit with a main effect for distance from the WCRRF and main effects and interaction for spatial location.*

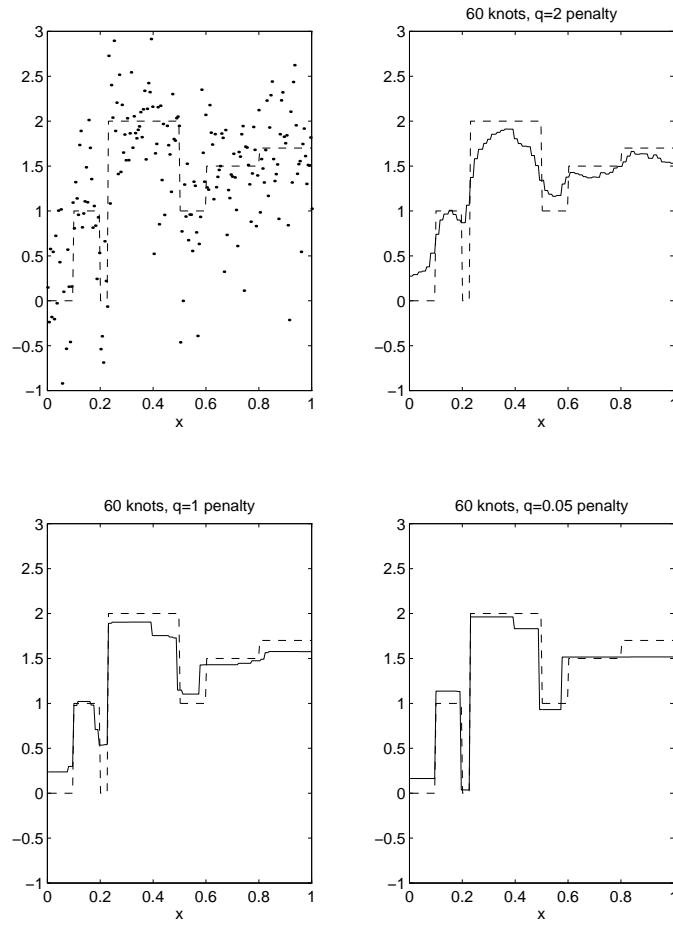


Figure 4: *Piecewise constant ($p = 0$) regression spline fits to a jump function.*

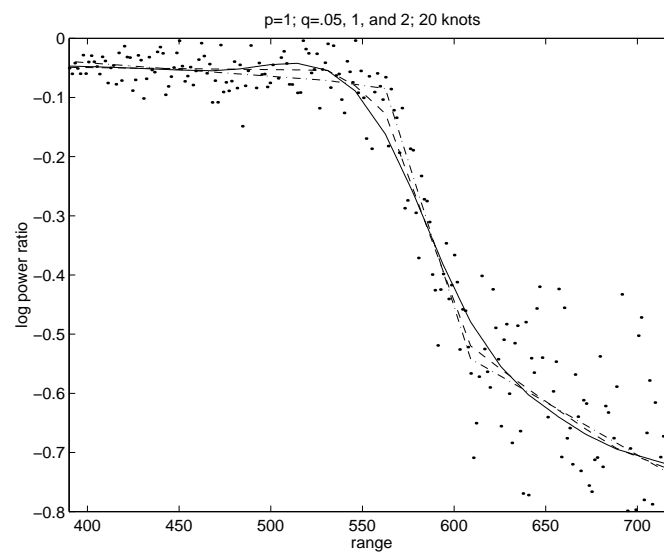


Figure 5: *Linear ($p = 1$) regression spline fits to the LIDAR data with $q = 2$ (solid), $q = .5$ (dashed), and $q = .05$ (dotted and dashed).*

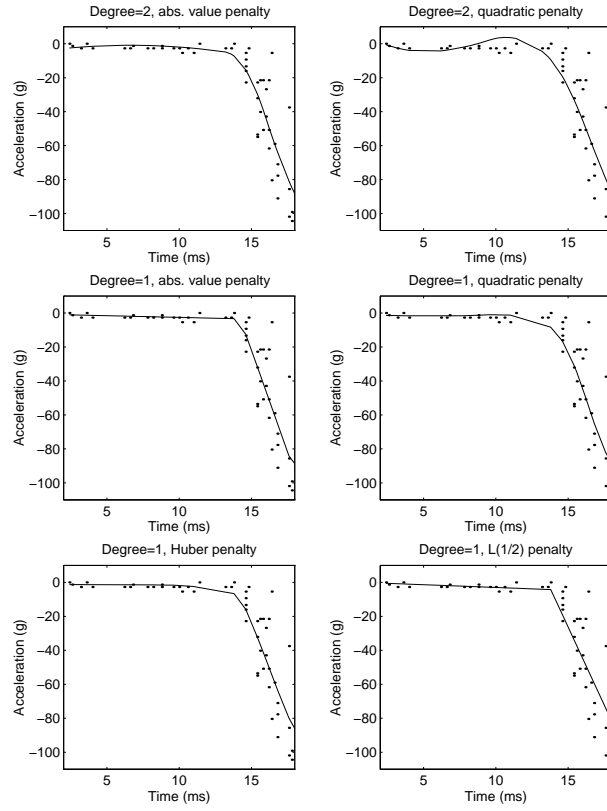


Figure 6: *Regression spline fits to the motorcycle impact data. Detailed views about point of impact.*

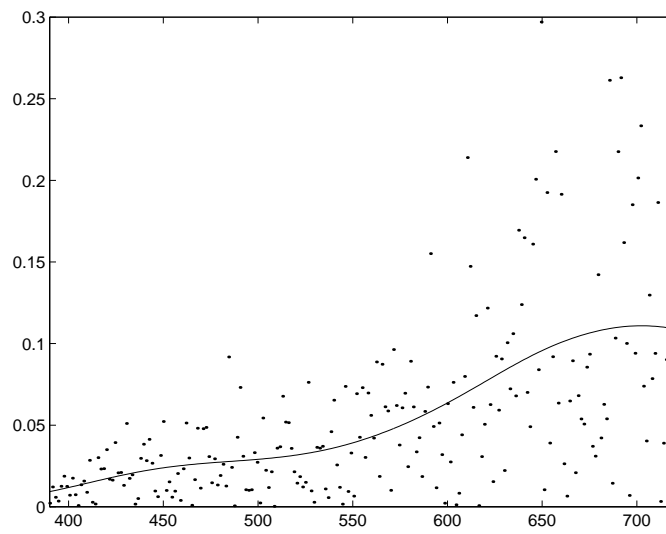


Figure 7: *Regression spline estimate of conditional dispersion of the LIDAR data. Absolute residuals (dots) and the estimate of the expected absolute residual given range (solid curve).*

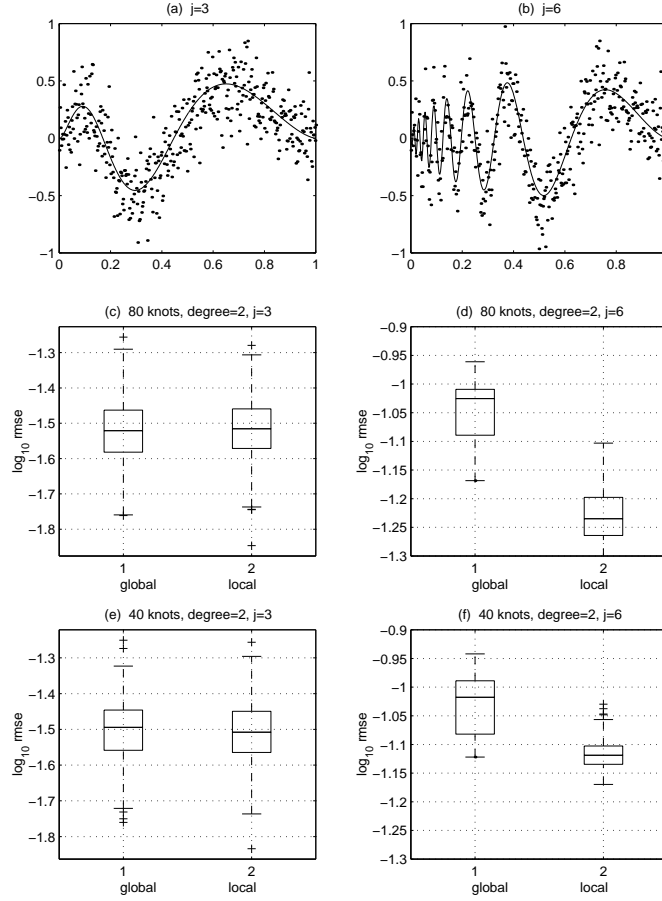


Figure 8: *Comparison of global and local penalty parameters under low ($j = 3$) and severe ($j = 6$) spatial variability in the oscillations of the regression function. (a) The regression function (solid) and one sample (dots) when $j = 3$. (b) Same as (a) but $j = 6$. (c) Boxplots of $\log_{10}(\text{RMSE})$ for 250 simulated samples using global and local penalty parameters. 80 knots, quadratic splines, and $j = 3$. (d) Same as (c) but $j = 6$. (e) Same as (c) but 40 knots. (f) Same as (e) by $j = 6$.*