DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 898

December 20, 1992

Smoothing Splines and Analysis of Variance in Function Spaces

by

Chong Gu

and

Grace Wahba

Smoothing Splines and Analysis of Variance in Function Spaces

CHONG GU AND GRACE WAHBA*

December 1992

Abstract

This article presents an exposition of some recent developments in the smoothing spline approach to multivariate nonparametric regression. The essence of the methodology is highlighted via the detailed descriptions of a few mathematically simplest members of the spline family. Data analytical tools are discussed, and their use in data analysis is illustrated via simulated and real data examples. Following the systematic developments, a few interesting observations on certain aspects of the methodology are collected, including a comparative study of nonparametric analysis versus parametric analysis and an explanation of a certain curious behavior of generalized cross-validation.

KEY WORDS: ANOVA decomposition; Generalized cross-validation; Modeling; Penalty smoothing; Reproducing kernel Hilbert space; Smoothing spline.

1 Introduction

Regression analysis, analysis of variance (ANOVA), and analysis of covariance are among the most commonly used statistical methods in applications. The common structure of the problems is

$$y_i = f(t_i) + \epsilon_i, \qquad i = 1, \cdots, n$$

*Chong Gu is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, Indiana 47907. Grace Wahba is Bascom Professor, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706.

where y_i are observed responses, $\boldsymbol{t}_i$ are predictors or covariates, and ϵ_i are zero-mean common variance uncorrelated noise. Here we only consider the fixed effect models for ANOVA and analysis of covariance. Our primary interest is to estimate the systematic part f of the response.

In classical parametric analysis, $f(\boldsymbol{t})$ is assumed to be of certain parametric form $f(\boldsymbol{t}, \boldsymbol{\beta})$ where the only unknowns are the values of the parameter $\boldsymbol{\beta}$ to be estimated from the data. The dimension of the model space is the dimension of $\boldsymbol{\beta}$, presumably much smaller than n. When $f(\boldsymbol{t}, \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$, i.e., $f(\boldsymbol{t}, \boldsymbol{\beta}) = \boldsymbol{x}^T\boldsymbol{\beta}$ where $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{t})$ is a vector of known functions of $\boldsymbol{t}$, f is just a standard linear model. Dozens of standard textbooks are available on linear models, see, e.g., Draper and Smith (1981). When $f(\boldsymbol{t}, \boldsymbol{\beta})$ is nonlinear in $\boldsymbol{\beta}$, nonlinear regression methods are available; see, e.g., Bates and Watts (1988). The parametric form $f(\boldsymbol{t}) = f(\boldsymbol{t}, \boldsymbol{\beta})$ is a rigid constraint on f and should in principle be derived from the subject area knowledge of the problem. Sometimes, however, a parametric form might be imposed simply for the lack of alternatives. In such circumstances, the analysis is subject to potential model bias, in the sense that possibly no member of the specified parametric family is close to the underlying "true" systematic part.

To avoid possibly serious model bias in a parametric analysis, an alternative approach is to allow f to vary in a high (possibly infinite) dimensional function space, which leads to various nonparametric or semiparametric methods. Since the data are noisy, however, one needs to impose certain soft constraints on f to regulate its behavior and to effectively achieve noise reduction in the estimate. The most natural soft constraint, which is adopted by most if not all of the nonparametric methods, is that f is "smooth". Consequently, nonparametric/semiparametric modeling is also called *smoothing*. All smoothing methods are equivalent, to various extents, to locally averaging the data — local to control the bias and average to reduce the noise. Among the classical smoothing methods are the kernel method, the nearest neighbor method, and penalty smoothing (smoothing splines). Because of the curse of dimensionality (Huber 1985), many successful univariate smoothing methods (e.g., kernel method) face serious operational difficulties when extended to high dimensional space. Consequently, almost all practical multivariate smoothing methods impose appropriate constraints and/or have convenient schemes to control the model complexity. Some of the methods available are projection pursuit regression (Friedman and Stuetzle 1981; Huber 1985), additive models (Hastie and Tibshirani 1986, 1990; Buja *et al.* 1989), regression splines (Stone 1985), multivariate adaptive regression splines (MARS) (Friedman 1991), the $\prod$-method (Breiman

1991), and various multivariate smoothing splines (Wahba 1990).

In this article, we pursue an exposition of the smoothing spline approach to nonparametric regression for readers familiar with standard statistical theory and exposed to a few basic concepts in functional analysis. We try to highlight the essence of the methodology as well as to cover some recent developments in the multivariate setup, with an emphasis on the identification of the underlying model which seems largely overlooked in the nonparametric estimation literature. We shall describe the available modeling tools and illustrate their use and effectiveness via simulated and real data examples. We shall also compare nonparametric analysis with parametric analysis to demonstrate the pros and cons of the methodology. At the end we shall offer an explanation of a certain curious negative correlation behavior of generalized cross-validation, by discussing the proper indexing of the underlying models.

The rest of the article is organized as follows. Section 2 introduces the basic ideas and explains the essential ingredients of the methodology via examining a few simple examples. Section 3 discusses ANOVA decomposition on product domains and describes the construction of tensor product smoothing splines by relatively simple examples. The materials in Sections 2 and 3 could be treated more generally, but we choose not to do so due to the expository nature of this article. Section 4 collects a few data analytical tools for a nonparametric analysis via the models described in Sections 2 and 3. Section 5 illustrates the methodology by data examples. Section 6 demonstrates the plus and minus sides of the methodology compared to parametric modeling. Section 7 discusses the aforementioned negative correlation behavior of generalized cross-validation.

2 Smoothing Splines

2.1 Penalty smoothing

Smoothing spline is an instance of penalty smoothing. A few examples follow.

Example 2.1 *Cubic Spline.* A classical example of penalty smoothing is the famous cubic spline. Consider $y_i = f(t_i) + \epsilon_i$, $i = 1, \cdots, n$, where $t_i \in [0, 1]$ and $\epsilon_i \sim N(0, \sigma^2)$. Since one has only finite number of data to estimate the entire function f, it is necessary to impose certain soft constraint

such as smoothness on f. A good estimate of f can be obtained as the minimizer of

$$\frac{1}{n}\sum_{j=1}^{n}(y_j - f(t_j))^2 + \lambda \int_0^1 (\ddot{f})^2, \tag{2.1}$$

where the first term measures the goodness-of-fit, the second term penalizes the roughness of the estimate, and the smoothing parameter λ controls the tradeoff between the two conflicting goals. The minimization of (2.1) is implicitly over functions with square integrable second derivatives. The minimizer of (2.1) defines a cubic spline. As $\lambda \to 0$, the minimizer approaches the minimum curvature interpolator. As $\lambda \to \infty$, the minimizer approaches the simple linear regression line. Note that the linear polynomials form the null space of the roughness penalty $\int_0^1 (\ddot{f})^2$. □

Example 2.2 *Shrinkage Estimator.* A simpler example of penalty smoothing is related to the classical shrinkage estimators. Consider $y_i = f(t_i) + \epsilon_i$, where $t_i \in \{1, \cdots, K\}$ is a discrete covariate and ϵ_i are *i.i.d.* normal. f is now a vector $\boldsymbol{f} \in R^K$. The standard setup for shrinkage estimators is a special case of this setup where one observes exactly one sample at each of the K points. Following the standard empirical Bayes construction, one may assume a prior $\boldsymbol{f} \sim N(0, \tau^2 I)$, and the Bayes estimator under such a prior is a shrinkage estimator shrinking towards 0. It is easy to check that such an estimator is just the minimizer of

$$\frac{1}{n}\sum_{j=1}^{n}(y_j - f(t_j))^2 + \frac{\sigma^2}{n\tau^2}\sum_{t=1}^{K} f^2(t), \tag{2.2}$$

where $\sum_{t=1}^{K} f^2(t)$ is the roughness penalty and $\sigma^2/n\tau^2$ is the smoothing parameter. A smooth vector in this case is simply one with small Euclidean norm. Note that this roughness penalty has a nil null space. □

Example 2.3 *Shrinkage Estimator in One-Way ANOVA.* Elaborating a bit further on Example 2.2, one may write $\boldsymbol{f} = \mu \mathbf{1} + \boldsymbol{\alpha}$, $\mathbf{1}^T \boldsymbol{\alpha} = 0$, as in a one-way ANOVA with the standard side-condition. The prior $\boldsymbol{f} \sim N(0, \tau^2 I)$ could be decomposed accordingly as $\mu \sim N(0, \tau^2/K)$ and $\boldsymbol{\alpha} \sim N(0, \tau^2\{I - \mathbf{1}\mathbf{1}^T/K\})$. Note that the τ^2 in the decoupled priors could vary separately. Letting $\tau_\mu^2 \to \infty$ generates an uniform improper prior for the constant μ. The resulting Bayes estimator is a shrinkage estimator shrinking towards the constant, which can equivalently be defined as the

minimizer of

$$\frac{1}{n}\sum_{j=1}^{n}(y_j - f(t_j))^2 + \frac{\sigma^2}{n\tau_\alpha^2}\sum_{t=1}^{K}(f(t) - \bar{f})^2, \tag{2.3}$$

where the roughness penalty $\sum_{t=1}^{K}(f(t) - \bar{f})^2$ has $\{\mathbf{1}\}$ as its null space. A smooth vector in this case is one with small variance. □

2.2 Smoothing Splines and Reproducing Kernel Hilbert Spaces

In a statistical analysis, one needs data as well as models. Data carry noise but are "unbiased", while models help to reduce noise but are responsible for "biases". Models assumed by many nonparametric methods such as the kernel method are extremely vague and implicit, which might be responsible for the difficulties in their extensions to high dimensional spaces. Penalty smoothing, in contrary, represents a convenient, explicit and generic approach to model specification in a nonparametric analysis. An illustrative example follows. By the standard Fourier series expansion, any continuous function f on $[0, 1]$ can be written in the form of

$$f = \beta_0 + \beta_1 t + \sum_{\nu=1}^{\infty}(a_\nu \cos 2\pi\nu t + b_\nu \sin 2\pi\nu t), \tag{2.4}$$

but (2.4) can not serve as a statistical model because it involves too many unknown parameters. Nevertheless, families of "smooth function" models can be obtained by properly constraining the parameters in (2.4). One possible constraint is to require $\int_0^1 \ddot{f}^2 = \sum_\nu (a_\nu^2 + b_\nu^2)(2\pi\nu)^4 \leq \rho$ for some $\rho \geq 0$. A least-squares fit of such a model usually falls on the boundary $\int_0^1 \ddot{f}^2 = \rho$, and by Lagrange's method the fit is the minimizer of (2.1) over functions of the form (2.4) and with $\int_0^1 \ddot{f}^2 < \infty$, for a certain λ depending on ρ and y_i. This results in a slightly restricted version of Example 2.1. $\rho = 0$ and ∞ correspond to $\lambda = \infty$ and 0. As ρ increases, more and more high frequencies are allowed to enter the play, and the Lagrange multiplier λ indirectly but conveniently codes a "continuous" spectrum of explicit models for one to choose in data analysis.

On a generic domain $\mathcal{T}$, defining an appropriate roughness functional $J(f)$, statistical models can be specified via $J(f) \leq \rho$. Under such a model, the least-squares fit of f based on observations

$y_i = f(t_i) + \epsilon_i$, $t_i \in \mathcal{T}$, could be calculated as the minimizer of

$$\frac{1}{n}\sum_{j=1}^{n}(y_j - f(t_j))^2 + \lambda J(f) \tag{2.5}$$

for a certain λ over a collection $\mathcal{H}$ of "smooth" functions. The roughness functional $J(f)$ is usually taken as a quadratic form, and implicitly $J(f) < \infty$ in $\mathcal{H}$. It is necessary that $J_\perp = \mathcal{H} \cap \{J(f) = 0\}$ be of finite dimension to prevent interpolation. $J(f)$ forms a natural quadratic seminorm on $\mathcal{H}$, and with the supplement of a quadratic norm in $J_\perp$, makes $\mathcal{H}$ a Hilbert space. As ρ increases, $\{J(f) \leq \rho\}$ shall allow closer and closer fit of $f(t_i)$ to the data y_i, but such a relaxation should be gradual. A gradual relaxation requires that evaluation be continuous in the model space $\mathcal{H}$ (semi) normed by $J(f)$, which automatically assures the continuity of the functional (2.5) in $\mathcal{H}$. Putting things together, the minimizer of (2.5) in $\mathcal{H}$, a Hilbert space with $J(f)$ as the square seminorm in which evaluation is continuous, defines a smoothing spline.

A Hilbert space in which evaluation is continuous is called a reproducing kernel Hilbert space (RKHS). As a consequence of the Riesz representation theorem, there exists a reproducing kernel (RK) $R(\cdot, \cdot)$, a positive definite bivariate function on $\mathcal{T}$, such that $R(t, \cdot) = R(\cdot, t) \in \mathcal{H}$, $\forall t \in \mathcal{T}$, and $\langle R(t, \cdot), f(\cdot)\rangle = f(t)$ (the reproducing property), $\forall f \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathcal{H}$. The norm and the RK in an RKHS determine each other uniquely, but like other duals in mathematical structures, the interpretability, and the availability of an explicit form for one part is often at the expense of the same for the other part. It will be seen that the RK plays a central role in the construction of models on complex domains and in the computation of smoothing splines. A mathematical theory of RKHS was developed by Aronszajn (1950), which very much resembles the linear algebra theory. For the purpose of spline smoothing, only a few elementary properties are needed, which we shall quote below.

- *Construction of RKHS.* Given any positive definite function $R(\cdot, \cdot)$ on a domain $\mathcal{T}$, one can construct an RKHS $\mathcal{H} = \text{span}\{R(t, \cdot), \forall t \in \mathcal{T}\}$ with an inner product satisfying $\langle R(s, \cdot), R(t, \cdot)\rangle = R(s, t)$, which has $R(\cdot, \cdot)$ as its RK. The inner product may or may not have an explicit form.

- *Tensor Sum RKHS.* If $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ with an RK R, then $\mathcal{H}_0$ and $\mathcal{H}_1$ are RKHS's with RK's R_0 and R_1, and $R = R_0 + R_1$. This property generalizes naturally to multiterm decompositions.

A tensor product property shall be deferred to Section 3 after the discussion of product domains.

We now cast the examples of Section 2.1 in the light of the generic formulation of (2.5) to exemplify the framework. For Example 2.1, $J(f) = \int_0^1 \ddot{f}^2$ is a square seminorm in $\mathcal{H} = \{f : \int \ddot{f}^2 < \infty\}$. There are many ways of supplementing $J(f)$ to deduce a norm in $\mathcal{H}$. Two rather standard configurations follow. The first one takes $\|f\|^2 = f^2(0) + \dot{f}^2(0) + J(f)$ with the RK $R(s,t) = [1+st]+[\int_0^1 (s-u)_+(t-u)_+ du]$, where $(\cdot)_+$ is the positive part of $(\cdot)$. This configuration yields a tensor sum decomposition $\mathcal{H} = J_\perp \oplus \mathcal{H}_J$ where $J_\perp = \pi_1$, the linear polynomials, with the square norm $f^2(0)+\dot{f}^2(0)$, and $\mathcal{H}_J = \mathcal{H} \ominus J_\perp = \{f : f \in \mathcal{H}, f(0) = \dot{f}(0) = 0\}$ with the square norm $J(f)$, and the corresponding RK's of $J_\perp$ and $\mathcal{H}_J$ are the terms in brackets in the expression of R. The second one takes $\|f\|^2 = (\int_0^1 f)^2+(\int_0^1 \dot{f})^2+J(f)$ with the RK $R(s,t) = [1+k_1(s)k_1(t)]+[k_2(s)k_2(t)-k_4(|s-t|)]$, where $k_1 = (\cdot - .5)$, $k_2 = (k_1^2 - 1/12)/2$, and $k_4 = (k_1^4 - k_1^2/2 + 7/240)/24$; see, e.g., Craven and Wahba (1979). This configuration has $J_\perp = \pi_1$ with the square norm $(\int_0^1 f)^2 + (\int_0^1 \dot{f})^2$ and $\mathcal{H}_J = \mathcal{H} \ominus J_\perp = \{f : f \in \mathcal{H}, \int f = \int \dot{f} = 0\}$ with the square norm $J(f)$, and the decomposed RK's are as bracketed. Note that the norm in $J_\perp$ plays no role in the definition of smoothing spline, so these different configurations all lead to the same final result. Different (marginal) configurations do matter, however, in the construction of tensor product splines, to be discussed in Section 3.2.

A finite dimensional Hilbert space is a special case of RKHS, and familiar objects in linear algebra may serve as prototypes for perceiving objects in a general RKHS. A function on $\{1, \cdots, K\}$ is a K vector and an RK a $K \times K$ positive definite matrix, and an evaluation simply extracts a coordinate from a vector. For Example 2.2, $\mathcal{H} = R^K$ with the norm $J(f) = \boldsymbol{f}^T\boldsymbol{f}$, i.e., the standard Euclidean space, and the RK is simply the identity matrix I. For Example 2.3, $\mathcal{H} = R^K = \{\mathbf{1}\} \oplus \{\mathbf{1}\}^\perp$, $J(f) = \boldsymbol{f}^T(I - \mathbf{1}\mathbf{1}^T/K)\boldsymbol{f}$ is a norm in $\{\mathbf{1}\}^\perp$, and $I = [\mathbf{1}\mathbf{1}^T/K] + [I - \mathbf{1}\mathbf{1}^T/K]$ is the RK decomposition. In general, any nonnegative-definite matrix J may define a roughness penalty $J(f) = \boldsymbol{f}^T J \boldsymbol{f}$ with the complement of its column space as the null space $J_\perp$. For example, for an ordinal discrete covariate, $J(f) = \sum_{t=1}^{K-1}(f(t+1)-f(t))^2$ might be a more natural penalty than the one defined in Example 2.3. A norm in R^K can then be defined as $\|f\|^2 = \boldsymbol{f}^T(L+J)\boldsymbol{f}$ where $LJ = 0$ and $L + J$ positive-definite. It is easy to verify that the RK is simply $(L + J)^{-1} = [L^+] + [J^+]$, where the superscript $+$ indicates the Moore-Penrose inverse and the brackets indicate the RK decomposition. Again the choice of L does not affect the final result.

Finally we remark that a smoothing spline as defined in (2.5) is a Bayes estimator under a mean zero Gaussian process prior on $\mathcal{T}$. The prior process has two independent components, one

is diffuse on $J_\perp$, and the other has a covariance function proportional to R_J, the RK in $\mathcal{H} \ominus J_\perp$. In the discrete case $R_J = J^+$. Example 2.3 might be the simplest yet complete illustration of this classical duality result due to Kimeldorf and Wahba (1970) and Wahba (1978).

In very loose terms, we may summarize this section as follows. A convenient and generic approach to model specification in a nonparametric analysis is via $J(f) \leq \rho$ with an appropriately defined quadratic roughness functional $J(f)$, where appropriateness means that evaluation should be continuous with respect to $J(f)$. The Lagrange method converts constrained least-squares to penalized least-squares, and the continuity of evaluation induces an RK in the model space. The objects in an RKHS may be perceived via familiar objects in linear algebra, and the quadratic roughness penalty acts as a Gaussian prior.

3 ANOVA in Function Spaces

3.1 Function decomposition on product domains

An important aspect of statistical modeling, which distinguishes it from mere function approximation, is the interpretability of the results. Among the most interpretable notions in classical modeling are the notions of main effects and interactions in ANOVA. We describe below a simple generic operation to generalize these notions to a generic setup.

In a standard two-way ANOVA on $\{1, \cdots, K_1\} \times \{1, \cdots, K_2\}$, $f(t_1, t_2) = \mu + \alpha_{t_1} + \beta_{t_2} + \gamma_{t_1,t_2}$, where the main effects α_{t_1}, β_{t_2}, and the interaction γ_{t_1,t_2} have to satisfy certain side conditions to make the decomposition unique. Two sets of commonly used side conditions are

$$\sum_{t_1} \alpha_{t_1} = \sum_{t_2} \beta_{t_2} = \sum_{t_1} \gamma_{t_1,t_2} = \sum_{t_2} \gamma_{t_1,t_2} = 0 \tag{3.1}$$

and

$$\alpha_1 = \beta_1 = \gamma_{1,t_2} = \gamma_{t_1,1} = 0, \tag{3.2}$$

where (3.1) are the standard ones. In both cases one can write

$$\begin{aligned} f &= (E_1 + I - E_1)(E_2 + I - E_2)f \\ &= E_1E_2f + (I - E_1)E_2f + E_1(I - E_2)f + (I - E_1)(I - E_2)f \end{aligned}$$

$$= \mu + \alpha_{t_1} + \beta_{t_2} + \gamma_{t_1,t_2}, \tag{3.3}$$

where E_i are marginalization (or averaging) operators acting on $\{1, \cdots, K_i\}$. For (3.1) $Ef = \bar{f}$, for (3.2) $Ef = f(1)$, where in an abuse of notation we omitted the constant vector $\mathbf{1}$ in the right hand side of the equations.

Consider functions $f(t_1, \cdots, t_\Gamma)$ on a generic product domain $\prod_{\gamma=1}^{\Gamma} \mathcal{T}_\gamma$. Define E_γ to be a marginalization operator acting on the argument t_γ, which "averages" out t_γ from the active argument list of the function and satisfies $E_\gamma^2 = E_\gamma$. An ANOVA decomposition can be defined as

$$\begin{aligned} f &= [\prod_{\gamma=1}^{\Gamma} (I - E_\gamma + E_\gamma)]f \\ &= \sum_{A \subseteq \{1,\cdots,\Gamma\}} [\prod_{\gamma \in A} (I - E_\gamma) \prod_{\gamma \in A^c} E_\gamma]f \\ &= \sum_{A \subseteq \{1,\cdots,\Gamma\}} f_A \end{aligned} \tag{3.4}$$

where A is the active argument list in a component. $f_\emptyset = [\prod_{\gamma=1}^{\Gamma} E_\gamma]f$ is the constant term, $f_\gamma = f_{\{\gamma\}} = [(I - E_\gamma) \prod_{\alpha \neq \gamma} E_\alpha]f$ is the t_γ main effect, $f_{\gamma,\delta} = f_{\{\gamma,\delta\}} = [(I - E_\gamma)(I - E_\delta) \prod_{\alpha \neq \gamma,\delta} E_\alpha]f$ is the t_γ-t_δ interaction, and so on. The terms of such a decomposition satisfy the side conditions $E_\gamma f_A = 0$, $\forall A \ni \gamma$. The choice of E_γ, or the side conditions on each axis, is open to specialization.

The ANOVA decomposition of functions on a product domain not only makes the functions more interpretable, it also automatically provides a means of model simplification by selectively trimming off certain terms in the decomposition. Interactions of three or more variables are usually trimmed as in the classical ANOVA, for they are less perceivable and are more "expensive" to estimate. Such simplifications are almost necessary for a nonparametric multivariate fit since the data are scarce. The flexibility in the choice of E_γ can also be employed to facilitate the incorporation of certain constraints; for example, to enforce $f(1, t_2) = 0$ in a two-way ANOVA, one could simply take $E_1 f = f(1, t_2)$ and trim off the constant and the t_2 main effect from the model.

3.2 Tensor product splines

The explicit model specification in penalty smoothing makes it easier to incorporate structures on product domains. Specifically, ANOVA decompositions of multivariate functions, possibly with selective term trimming, can be conveniently constructed via tensor product splines, a specialization of (2.5) with $\mathcal{H}$ a tensor product RKHS to be discussed below. Given positive definite bivariate functions $R_1(s_1, t_1)$ and $R_2(s_2, t_2)$ on domains $\mathcal{T}_1$ and $\mathcal{T}_2$, it can be shown that $R((s_1, s_2), (t_1, t_2)) = R_1(s_1, t_1)R_2(s_2, t_2)$ is positive definite on $\mathcal{T}_1 \times \mathcal{T}_2$; see Aronszajn (1950). A simple tensor product RKHS on $\mathcal{T}_1 \times \mathcal{T}_2$ results from standard construction.

- *Tensor Product RKHS.* Given RKHS's $\mathcal{H}^\gamma$ on $\mathcal{T}_\gamma$ with RK's R_γ, $\gamma = 1, 2$, one can construct an RKHS $\mathcal{H}^1 \otimes \mathcal{H}^2$ on $\mathcal{T}_1 \times \mathcal{T}_2$ with $R = R_1 R_2$ as its RK. This property generalizes naturally to multiterm products.

To construct a composite tensor product RKHS on a product domain with an ANOVA decomposition built in, the first step is to cut appropriately configured marginal RKHS's $\mathcal{H}^\gamma$ using the tensor sum rule to two mutually orthogonal RKHS's $\mathcal{H}_0^\gamma$ and $\mathcal{H}_1^\gamma$ which separate $E_\gamma f$ and $(I - E_\gamma)f$; one then assembles simple tensor product RKHS's as modules from $\mathcal{H}_0^\gamma$ or $\mathcal{H}_1^\gamma$ using the tensor product rule, with each module representing a term in an ANOVA decomposition; the final step is to paste the modules together using the tensor sum rule, with the modules representing trimmed terms left out. We shall illustrate the construction with a few bivariate examples in the remainder of the section. General theory and more complicated examples can be found in, e.g., Wahba (1986) and Gu and Wahba (1991a, 1991b, 1993a).

Example 3.1 *Shrinkage Estimators in Two-Way ANOVA.* Consider a pure discrete case on $\{1, \cdots, K_1\} \times \{1, \cdots, K_2\}$ and adopt the standard side conditions of (3.1). The marginal RKHS on $\{1, \cdots, K_\gamma\}$ is taken as the standard Euclidean space R^{K_γ} which can be decomposed as $\{\mathbf{1}\} \oplus \{\mathbf{1}\}^\perp$ with RK's $R_0 = (\mathbf{1}\mathbf{1}^T/K_\gamma)$ and $R_1 = (I - \mathbf{1}\mathbf{1}^T/K_\gamma)$. A function on $\{1, \cdots, K_1\} \times \{1, \cdots, K_2\}$ can be written as a K_1K_2 vector $\boldsymbol{f} = (f(1,1), \cdots, f(1,K_2), \cdots, f(K_1,1), \cdots, f(K_1,K_2))^T$ and an RK a $(K_1K_2) \times (K_1K_2)$ matrix. The product RK $R_{0,0} = (\mathbf{1}\mathbf{1}^T/K_1) \otimes (\mathbf{1}\mathbf{1}^T/K_2)$ generates the constant space with the norm $\boldsymbol{f}^T J_{0,0} \boldsymbol{f}$ where $J_{0,0} = R_{0,0}^+ = R_{0,0}$ ($R_{0,0}$ is idempotent), where $\otimes$ indicates the Kronecker product of matrices. Similarly, $R_{1,0} = (I - \mathbf{1}\mathbf{1}^T/K_1) \otimes (\mathbf{1}\mathbf{1}^T/K_2)$ generates the t_1 main effect with the norm $\boldsymbol{f}^T J_{1,0} \boldsymbol{f}$ where $J_{1,0} = R_{1,0}^+ = R_{1,0}$, $R_{0,1} = (\mathbf{1}\mathbf{1}^T/K_1) \otimes (I - \mathbf{1}\mathbf{1}^T/K_2)$ generates

the t_2 main effect with the norm $\boldsymbol{f}^T J_{0,1}\boldsymbol{f}$ where $J_{0,1} = R_{0,1}^+ = R_{0,1}$, and $R_{1,1} = (I - \mathbf{1}\mathbf{1}^T/K_1)\otimes(I - \mathbf{1}\mathbf{1}^T/K_2)$ generates the interaction with the norm $\boldsymbol{f}^T J_{1,1}\boldsymbol{f}$ where $J_{1,1} = R_{1,1}^+ = R_{1,1}$. A composite RKHS can be constructed via an RK $R_\theta = \theta_{0,0}R_{0,0} + \theta_{1,0}R_{1,0} + \theta_{0,1}R_{0,1} + \theta_{1,1}R_{1,1}$ with the norm $\|f\|_\theta^2 = \boldsymbol{f}^T(\theta_{0,0}^{-1}J_{0,0} + \theta_{1,0}^{-1}J_{1,0} + \theta_{0,1}^{-1}J_{0,1} + \theta_{1,1}^{-1}J_{1,1})\boldsymbol{f}$, where $\theta \in [0,\infty)$. Statistical models could be specified via $J_\theta(f) = \boldsymbol{f}^T(\sum_\beta \theta_\beta^{-1}J_\beta)\boldsymbol{f} \le \rho$ which leads to penalty smoothing with J_θ as the roughness. A $\theta = \infty$ in J_θ puts a term in the null space of the roughness penalty, a $\theta = 0$ trims a term, and a $\theta \in (0,\infty)$ shrinks a term. In the equivalent Bayes model, $f(t_1,t_2) = \mu + \alpha_{t_1} + \beta_{t_2} + \gamma_{t_1,t_2}$, where $\mu \sim N(0,\tau_\mu^2/K_1K_2)$, $\boldsymbol{\alpha} \sim N(0,\tau_\alpha^2(I - \mathbf{1}\mathbf{1}^T/K_1)/K_2)$, $\boldsymbol{\beta} \sim N(0,\tau_\beta^2(I - \mathbf{1}\mathbf{1}^T/K_2)/K_1)$, and $\boldsymbol{\gamma} \sim N(0,\tau_\gamma^2(I-\mathbf{1}\mathbf{1}^T/K_1)\otimes(I-\mathbf{1}\mathbf{1}^T/K_2))$, independent of each other. Note that the side conditions are built into the covariance matrices of the priors. A diffuse prior ($\tau^2 = \infty$) leaves a term free, a degenerate prior ($\tau^2 = 0$) trims a term, and a proper prior ($\tau^2 \in (0,\infty)$) shrinks a term. □

Example 3.2 *Linear Spline and Tensor Product.* On $[0,1]$, an RKHS simpler than that in Example 2.1 is $\mathcal{H} = \{f : \int_0^1 \dot{f}^2 < \infty\}$ with a seminorm $J(f) = \int_0^1 \dot{f}^2$ and $J_\perp = \{1\}$. The resulting smoothing spline is known as a linear spline. With a null space norm $(f(0))^2$, $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_0 = \{1\}$ and $\mathcal{H}_1 = \{f : f \in \mathcal{H}, f(0) = 0\}$, with RK's $R_0 = 1$ and $R_1 = \min(s,t)$; this configuration fits the marginalization operator $Ef = f(0)$. With a null space norm $(\int_0^1 f)^2$, $\mathcal{H}_0 = \{1\}$ has an RK $R_0 = 1$ and $\mathcal{H}_1 = \mathcal{H} \ominus \mathcal{H}_0 = \{f : f \in \mathcal{H}, \int_0^1 f = 0\}$ has an RK $R_1 = k_1(s)k_1(t) + k_2(|s-t|)$ where k_1 and k_2 are given in Section 2.2; this configuration fits the marginalization operator $Ef = \int_0^1 f$. Taking $E_1 f = f(0)$ and $E_2 f = \int_0^1 f$ on $[0,1]^2$, a tensor product RKHS with ANOVA modules can be constructed as follows. $R_{0,0} = 1$ generates the constant function space with a norm $J_{0,0}(f) = (\int_0^1 f(0,t_2)dt_2)^2$, $R_{1,0} = \min(s_1,t_1)$ generates the t_1 main effect space with a norm $J_{1,0}(f) = \int_0^1(\int_0^1 \dot{f}_{t_1}dt_2)^2dt_1$, $R_{0,1} = k_1(s_2)k_1(t_2) + k_2(|s_2 - t_2|)$ generates the t_2 main effect space with a norm $J_{0,1}(f) = \int_0^1(\dot{f}_{t_2}(0,t_2))^2dt_2$, and $R_{1,1} = \min(s_1,t_1)(k_1(s_2)k_1(t_2) + k_2(|s_2 - t_2|))$ generates the interaction space with a norm $J_{1,1}(f) = \int_0^1\int_0^1 \ddot{f}_{t_1,t_2}^2 dt_1dt_2$. Pasting things together, an RK $R_\theta = \theta_{0,0}R_{0,0} + \theta_{1,0}R_{1,0} + \theta_{0,1}R_{0,1} + \theta_{1,1}R_{1,1}$ generates an RKHS with a norm $\|f\|_\theta^2 = \theta_{0,0}^{-1}J_{0,0}(f) + \theta_{1,0}^{-1}J_{1,0}(f) + \theta_{0,1}^{-1}J_{0,1}(f) + \theta_{1,1}^{-1}J_{1,1}(f)$, where $\theta \in [0,\infty)$. Statistical models could be specified via $J_\theta(f) = \sum_\beta \theta_\beta^{-1}J_\beta(f) \le \rho$. Usually $\theta_{0,0}$ is set to infinity in J_θ to put the constant in $J_\perp$. Other θ's can not be set to infinity or otherwise interpolation results. Setting a θ to 0 eliminates a term. □

Table 3.1: Norms in simple tensor product RKHS's in Example 3.3

R_β	J_β
$R_{c,c}$	$(\int_0^1 \int_0^1 f dt_1 dt_2)^2$
$R_{\pi,c}$	$(\int_0^1 \int_0^1 \dot{f}_{t_1} dt_1 dt_2)^2$
$R_{\pi,\pi}$	$(\int_0^1 \int_0^1 \ddot{f}_{t_1 t_2} dt_1 dt_2)^2$
$R_{s,c}$	$\int_0^1 (\int_0^1 \ddot{f}_{t_1^2 t_2} dt_2)^2 dt_1$
$R_{s,\pi}$	$\int_0^1 (\int_0^1 f^{(3)}_{t_1^2 t_2} dt_2)^2 dt_1$
$R_{s,s}$	$\int_0^1 \int_0^1 (f^{(4)}_{t_1^2 t_2^2})^2 dt_1 dt_2$

Example 3.3 *Tensor Product Cubic Spline.* The construction of RKHS on $[0,1]^2$ using $\mathcal{H} = \{f : \int_0^1 \ddot{f}^2 < \infty\}$ as marginals has a slight complication as we will see shortly. Take $E_1 f = E_2 f = \int_0^1 f$. With a norm $(\int_0^1 f)^2 + (\int_0^1 \dot{f})^2 + \int_0^1 \ddot{f}^2$, $\mathcal{H} = \mathcal{H}_c \oplus \mathcal{H}_\pi \oplus \mathcal{H}_s$, where $\mathcal{H}_c = \{1\}$ span the constant with an RK $R_c = 1$, $\mathcal{H}_\pi = \{(\cdot - .5)\}$ spans the "polynomial" with an RK $R_\pi = k_1(s)k_1(t)$, and $\mathcal{H}_s = \{f : f \in \mathcal{H}, \int_0^1 f = \int_0^1 \dot{f} = 0\}$ collects the "smooth" (meaning rough!) part with an RK $R_s = k_2(s)k_2(t) - k_4(|s-t|)$. $Ef \in \mathcal{H}_c$ and $(I-E)f \in \mathcal{H}_\pi \oplus \mathcal{H}_s$. Since $\mathcal{H}_\pi$ and $\mathcal{H}_s$ don't fit together naturally, we shall separate them in constructing the simple tensor product RKHS's, which entails the complication. There are altogether nine simple tensor product RKHS modules. The one with an RK $R_{c,c} = 1$ generates the constant term, the ones with RK's $R_{\pi,c} = R_\pi(s_1,t_1)$ and $R_{s,c} = R_s(s_1,t_1)$ generate the t_1 main effect, the ones with RK's $R_{c,\pi} = R_\pi(s_2,t_2)$ and $R_{c,s} = R_s(s_2,t_2)$ generate the t_2 main effect, and the ones with RK's $R_{\pi,\pi} = R_\pi(s_1,t_1)R_\pi(s_2,t_2)$, $R_{\pi,s} = R_\pi(s_1,t_1)R_s(s_2,t_2)$, $R_{s,\pi} = R_s(s_1,t_1)R_\pi(s_2,t_2)$ and $R_{s,s} = R_s(s_1,t_1)R_s(s_2,t_2)$ generate the interaction. An RK $R_\theta = \sum_\beta \theta_\beta R_\beta$ generates an RKHS with a norm $\|f\|_\theta^2 = \sum_\beta \theta_\beta^{-1} J_\beta(f)$, where the J_β, listed in Table 3.1, are norms in modules generated by R_β. Statistical models can be specified via $J_\theta(f) = \sum_\beta \theta_\beta^{-1} J_\beta(f) \le \rho$, where a $\theta = \infty$ in J_θ puts a term into $J_\perp$ and a $\theta = 0$ eliminates a term from the model. Usually $\theta_{c,c}$, $\theta_{\pi,c}$, $\theta_{c,\pi}$ and $\theta_{\pi,\pi}$ are set to infinity in J_θ. Other θ's can not be set to infinity or interpolation results. □

Example 3.4 *Analysis of Covariance.* Consider a product domain $\{1,\cdots,K\}\times[0,1]$. Let $E_1 f = \bar{f}$ and $E_2 f = \int_0^1 f$. We shall use the marginals of Examples 3.1 and 3.3 to construct tensor product RKHS's. There are six simple tensor product RKHS modules. The one with an RK $R_{0,c} = 1/K$ generates the constant term, the one with an RK $R_{1,c} = R_1(s_1,t_1)$ generates the t_1 main effect, the

Table 3.2: Norms in simple tensor product RKHS's in Example 3.4

R_β	J_β
$R_{0,c}$	$(\sum_{t_1=1}^K \int_0^1 f dt_2)^2/K$
$R_{1,c}$	$\sum_{t_1=1}^K (\int_0^1 (f - \sum_{t_1=1}^K f/K) dt_2)^2$
$R_{0,\pi}$	$(\sum_{t_1=1}^K \int_0^1 \dot{f}_{t_2} dt_2)^2/K$
$R_{0,s}$	$\int_0^1 (\sum_{t_1=1}^K \dot{f}_{t_2^2})^2 dt_2/K$
$R_{1,\pi}$	$\sum_{t_1=1}^K (\int_0^1 (\dot{f}_{t_2} - \sum_{t_1=1}^K \dot{f}_{t_2}/K) dt_2)^2$
$R_{1,s}$	$\int_0^1 \sum_{t_1=1}^K (\ddot{f}_{t_2^2} - \sum_{t_1=1}^K \ddot{f}_{t_2^2}/K)^2 dt_2$

ones with RK's $R_{0,\pi} = R_\pi(s_2, t_2)/K$ and $R_{0,s} = R_s(s_2, t_2)/K$ generate the t_2 main effect, and the ones with RK's $R_{1,\pi} = R_1(s_1, t_1)R_\pi(s_2, t_2)$ and $R_{1,s} = R_1(s_1, t_1)R_s(s_2, t_2)$ generate the interaction. An RK $R_\theta = \sum_\beta \theta_\beta R_\beta$ generates an RKHS with a norm $\|f\|_\theta^2 = \sum_\beta \theta_\beta^{-1} J_\beta(f)$, where the J_β, listed in Table 3.2, are norms in modules generated by R_β. Statistical models can be specified via $J_\theta(f) = \sum_\beta \theta_\beta^{-1} J_\beta(f) \leq \rho$, where a $\theta = \infty$ in J_θ puts a term into $J_\perp$ and a $\theta = 0$ eliminates a term from the model. The constant is usually unpenalized, i.e., $\theta_{0,c} = \infty$. $R_{1,c}$, $R_{0,\pi}$ and $R_{1,\pi}$ are all of finite dimension, and can be put into $J_\perp$ without interpolating the data. Setting $\theta_{1,\pi} = \theta_{1,s} = 0$ enforces a main-effect-only model, which amounts to parallel cubic splines. Forcing $\theta_{0,s} = \theta_{1,s}$ and setting all other θ's to infinity yields separate cubic splines at different t_1 values but with a common smoothing parameter. Finally we note that to obtain separate cubic splines at different t_1 values with separate smoothing parameters, one needs to decompose the RK in R^K into K terms $I = \sum_{t_1=1}^K J_{t_1}$ where J_i has 1 at (i, i) and 0 everywhere else, and construct tensor product modules from J_i. □

It is clear that the models specified via $J_\theta(f) \leq \rho$ in a multimodule RKHS depend on ρ as well as the parameters θ_β. In general, the R_β generate functions of different nature and the scalings are arbitrary, so it is necessary that the θ_β appearing in J_θ, together with ρ or equivalently the Lagrange multiplier λ, be selected adaptively in data analysis. RK's are the prime object in the construction of tensor product RKHS's and explicit forms are necessary. Norms only serve to help the perception of models so explicit forms are dispensable.

4 Modeling Tools

Sections 2 and 3 concern the (conceptual) construction of nonparametric models via the smoothing spline approach. To make the approach applicable in data analysis, further tools are needed. In this section, we briefly describe a few modeling tools for model fitting, model checking, and precision assessment. These tools are considerably different from their counterparts in a parametric statistical analysis, as we will see shortly, and that is not surprising because the basic principle of a nonparametric analysis is sufficiently different from that of a parametric analysis. We remark that the development of modeling tools, especially for model checking and precision assessment, is by and large immature at the present time, and the reader will see that there are many open problems.

4.1 Calculation of cross-validated fit

This subsection is about model fitting. Consider the following generic problem.

$$\min \frac{1}{n}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \sum_{\beta=1}^{p} \theta_\beta^{-1} J_\beta(f), \tag{4.1}$$

where $f = \sum_{\beta=0}^{p} f_\beta \in \mathcal{H} = \oplus_{\beta=0}^{p} \mathcal{H}_\beta$, $f_\beta \in \mathcal{H}_\beta$, and $J_\beta(f) = J_\beta(f_\beta)$ is a square norm on $\mathcal{H}_\beta$ with the associated RK R_β. The θ_β appearing in the penalty $J_\theta(f) = \sum_{\beta=1}^{p} \theta_\beta^{-1} J_\beta(f)$ are all in $(0, \infty)$ and $J_\perp = \mathcal{H}_0$. It is easily seen that all our examples in Sections 2 and 3 are specializations of (4.1), with $p = 1$ for the examples of Section 2. The solution of (4.1) has an expression

$$f = \sum_{\nu=1}^{M} \phi_\nu(\cdot) d_\nu + \sum_{i=1}^{n} (\sum_{\beta=1}^{p} \theta_\beta R_\beta(t_i, \cdot)) c_i = \boldsymbol{\phi}^T(\cdot)\boldsymbol{d} + \boldsymbol{\xi}^T(\cdot)\boldsymbol{c}, \tag{4.2}$$

where $\{\phi_\nu\}_{\nu=1}^{M}$ span $\mathcal{H}_0$, $\boldsymbol{\xi}^T(\cdot) = (\xi_1(\cdot), \cdots, \xi_n(\cdot))$, $\xi_i(\cdot) = \sum_{\beta=1}^{p} \theta_\beta R_\beta(t_i, \cdot)$, and $\boldsymbol{c}$ and $\boldsymbol{d}$ are the minimizers of

$$(\boldsymbol{y} - S\boldsymbol{d} - Q\boldsymbol{c})^T(\boldsymbol{y} - S\boldsymbol{d} - Q\boldsymbol{c}) + n\lambda \boldsymbol{c}^T Q \boldsymbol{c}, \tag{4.3}$$

where S is $n \times M$ with (i, ν)th entry $\phi_\nu(t_i)$, $Q = \sum_{\beta=1}^{p} \theta_\beta Q_\beta$, and Q_β is $n \times n$ with (i, j)th entry $R_\beta(t_i, t_j)$; see, e.g., Kimeldorf and Wahba (1971) and Gu and Wahba (1991a). It can be shown that (4.2) is unique as the solution of (4.1) provided that S is of full column rank, whereas (4.3)

could have multiple numerical solutions of $\boldsymbol{c}$. All we need, however, is one solution of (4.3), which can be obtained by solving the well-behaving surrogate linear system

$$\begin{aligned} (Q + n\lambda I)\boldsymbol{c} + S\boldsymbol{d} &= \boldsymbol{y} \\ S^T\boldsymbol{c} &= 0. \end{aligned} \tag{4.4}$$

The choice of smoothing parameters λ and θ_β in (4.1) determines the behavior of a smoothing spline estimate. A good choice is via the generalized cross-validation of Craven and Wahba (1979), which aims to minimize the mean square error of the resulting estimate; relevant asymptotic analysis and empirical results may be found in Wahba (1990) and references cited therein. Generic algorithms for solving (4.4) with cross-validated smoothing parameters appear in Gu *et al.* (1989) and Gu and Wahba (1991a), where further details can be found. The algorithms are implemented in a collection of Ratfor subroutines under the name RKPACK (Gu 1989) available from Statlib and Netlib. To use the software, the user has to construct the S and Q_β matrices and input them together with the response vector $\boldsymbol{y}$ into one of the drivers, and the driver will return the cross-validated fit in terms of $n\lambda$, θ_β, $\boldsymbol{c}$, and $\boldsymbol{d}$. The drivers also return a variance estimate $\hat{\sigma}^2$ recommended by Wahba (1983).

4.2 Cosine diagnostics

This subsection is about model checking. Similar to the fact that the rigid constraint in a parametric analysis makes lack of fit the main concern there, the flexibility in a nonparametric analysis makes overinterpretation the prime target of the current development. More precisely, we consider an *interpretable* decomposition of the fit $f = \sum_{\beta=0}^{p} f_\beta$, such as the ANOVA decomposition of Section 3, and check for the identifiability and the nontriviality of the terms in such a decomposition. By convention f_0 is taken as the constant function. Note that this decomposition is in general different from the computation-oriented decomposition in (4.1). Also note that such checks are not necessary if the sole purpose of the analysis is for prediction.

Assume that the decomposition $f = \sum_{\beta=0}^{p} f_\beta$ is well-defined on the domain $\mathcal{T}$. When a fit is calculated from the data, however, information comes from the design points t_i, and the credibility of the decomposition depends on how well it is supported on the design points. Evaluating the fit

at t_i, one gets a retrospective linear model

$$y = \tilde{f}_0 + \cdots + \tilde{f}_p + \tilde{e}, \tag{4.5}$$

where $\tilde{f}_\beta$ are f_β evaluated at t_i and $\tilde{e}$ is the residual vector. Removing the constant by projecting (4.5) onto $\{\mathbf{1}\}^\perp$, one gets

$$z = f_1 + \cdots + f_p + e. \tag{4.6}$$

The collinearity indices κ_β's of $(f_1, \cdots, f_p)$ (Stewart 1987), which can be calculated from the cosines between the f_β's, measure the identifiability of the terms in the decomposition $\sum_{\beta=1}^p f_\beta$, and in turn the identifiability of the terms in the decomposed fit $f = \sum_{\beta=0}^p f_\beta$. The f_β's are supposed to predict the "response" z so a near orthogonal angle between a f_β and z indicates a noise term. Signal terms should be reasonably orthogonal to the residuals hence a large cosine between a f_β and e makes a term suspect. $\cos(z, e)$ and $R^2 = \|z - e\|^2/\|z\|^2$ are informative ad hoc measures for the signal to noise ratio in the data. A *very* small norm of a f_β compared to that of z disqualifies the cosines as reliable measures, but it itself indicates a negligible term. We will treat the cosine diagnostics as absolute measures for cross-validated fits. Our limited experience suggests that a term with $\cos(z, f) < .25$ may be discarded and a term with $\cos(z, f) > .4$ and with a reasonable magnitude is not likely all noise. More discussion can be found in Gu (1992). These measures are intuitively reasonable and have been used successfully in examples. It would be nice to have further understanding of their operating properties.

4.3 Bayesian confidence intervals

This subsection is about precision assessment. As noted at the end of Section 2, a smoothing spline is an empirical Bayes estimator under a Gaussian prior. More precisely, it can be verified that the solution of (4.1) is just the posterior mean of a model

$$y_i = \sum_{\nu=1}^M \psi_\nu(t_i) + \sum_{\beta=1}^p g_\beta(t_i) + \epsilon_i, \tag{4.7}$$

where g_β are independent mean zero Gaussian processes on $\mathcal{T}$ with covariance functions $\mathrm{Cov}(g_\beta(s), g_\beta(s')) = b\theta_\beta R_\beta(s, s')$ where $b = \sigma^2/n\lambda$, $\psi_\nu = d_\nu \phi_\nu$ where d_ν have uniform improper prior on $(-\infty, \infty)$, and

$\epsilon_i \sim N(0, \sigma^2)$. Let S and Q_β be as defined in (4.3) and $M = \sum_{\beta=1}^{p} \theta_\beta Q_\beta + n\lambda I$. The posterior distributions are summarized in the following theorem.

Theorem 4.1 *Fix $n\lambda$, θ_β, and σ^2 in (4.7).*

$$\begin{aligned}
\mathrm{E}(\psi_\nu(s)|\boldsymbol{y}) &= \phi_\nu(s)\boldsymbol{e}_\nu^T(S^TM^{-1}S)^{-1}S^TM^{-1}\boldsymbol{y} && (4.8)\\
\mathrm{E}(g_\beta(s)|\boldsymbol{y}) &= \theta_\beta R_\beta(s,\boldsymbol{t}^T)(M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1})\boldsymbol{y} && (4.9)\\
\mathrm{Cov}(\psi_\nu(s),\psi_\mu(s')|\boldsymbol{y})/b &= \phi_\nu(s)\phi_\mu(s')\boldsymbol{e}_\nu^T(S^TM^{-1}S)^{-1}\boldsymbol{e}_\mu && (4.10)\\
\mathrm{Cov}(\psi_\nu(s),g_\beta(s')|\boldsymbol{y})/b &= -\phi_\nu(s)\boldsymbol{e}_\nu^T(S^TM^{-1}S)^{-1}S^TM^{-1}\theta_\beta R_\beta(\boldsymbol{t},s') && (4.11)\\
\mathrm{Cov}(g_\beta(s),g_\gamma(s')|\boldsymbol{y})/b &= -\theta_\beta R_\beta(s,\boldsymbol{t}^T)(M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1})\theta_\gamma R_\gamma(\boldsymbol{t},s') \\
&\quad +\delta_{\beta,\gamma}\theta_\beta R_\beta(s,s') && (4.12)
\end{aligned}$$

where $\boldsymbol{t}$ is the vector of the design points, $\boldsymbol{e}_\nu$ is the νth unit vector, and $\delta_{\beta,\gamma}$ is the Kronecker delta.

A proof of the theorem can be found in Gu and Wahba (1993b). Based on (4.8) – (4.12), posteriors of all linear combinations of ψ_ν and g_β, specifically those of the terms in an ANOVA decomposition on a product domain, can be readily derived. The calculations of these quantities can be conveniently conducted using the RKPACK facilities; see Gu and Wahba (1993b). One may plug in the cross-validation estimates for the smoothing parameters appearing in the formulas and use $b = \hat{\sigma}^2/n\lambda$, where the $\hat{\sigma}^2$ is the variance estimate recommended by Wahba (1983). Based on the posterior analysis, component-wise Bayesian confidence intervals can be easily constructed for any linear combinations of $\psi_\nu(s)$ and $g_\beta(s)$, including terms in an ANOVA decomposition and f itself. These confidence intervals were first studied in Wahba (1983). See also Wecker and Ansley (1983). Under certain conditions, these intervals have a "correct" asymptotic *average* coverage, in the sense that the coverage of the true component by the 1.96σ intervals averaged over the design points t_i centers around 95%. Further details can be found in Wahba (1983), Nychka (1988, 1990), and Gu and Wahba (1993b).

5 Examples

We will analyze three data sets in this section using the techniques presented in the previous sections.

Table 5.1: Diagnostics for Pure Noise

	f_1	f_2	$f_{1,2}$	e	z
κ	1.07	1.02	1.05	$R^2 = 0.044$	
$\cos(e, \cdot)$	0.00	0.00	0.02	1	0.98
$\cos(z, \cdot)$	0.07	0.01	0.20	0.98	1
$\|\| \cdot \|\|$	1.16	0.10	2.08	9.98	10.26

5.1 Pure noise

The first example is a trivial exercise. We generated $n = 100$ design points from $U(0,1)^2$ and attached 100 pseudo $N(0,1)$ deviates as y_i to these points. We used the tensor product cubic spline of Example 3.3 to fit the data. The fit was calculated with $\theta_{c,c} = \theta_{c,\pi} = \theta_{\pi,c} = \theta_{\pi,\pi} = \infty$ and with the other five smoothing parameters cross-validated. The nine fitted terms were then collapsed into one constant, two main effects, and one interaction terms. The diagnostics are summarized in Table 5.1. The conclusion is self-evident.

5.2 NOX data

The data were from an experiment in which a single-cylinder engine was run with ethanol. There were 88 measurements of compression ratio (C), equivalence ratio (E), and NO_x in the exhaust. The purpose of the analysis was to see how NO_x depends on E and C. Cleveland and Devlin (1988) have more details about the data and an analysis using the multivariate loess. Breiman (1991) analyzed the same data using the $\prod$ method. We followed Cleveland and Devlin (1988) by taking the cube root transformation of NO_x. Since C only varied on 5 distinct values, we could treat it both as a continuous covariate and as a discrete covariate, which we did in different analyses.

The covariate E was translated into $[0,1]$ by $t_1 = (E - .535)/.697$. First we treated C as continuous and translated it by $t_2 = (C - 7.5)/10.5 \in [0,1]$. A tensor product cubic spline fit was calculated the same way as in the pure noise example. The diagnostics are summarized in Table 5.2. f_2 and $f_{1,2}$ were basically orthogonal to z. Clearly, there wasn't enough evidence in the data to support the C main effect and the interaction.

Treating C as a nominal discrete covariate, we also calculated a tensor product spline model using the terms in Table 3.2 (with t_1 and t_2 switched) with $\theta_{c,0} = \theta_{c,1} = \theta_{\pi,0} = \theta_{\pi,1} = \infty$. The

Table 5.2: Diagnostics for NOX Model: Continuous C.

	f_1	f_2	$f_{1,2}$	e	z
κ	1.08	1.07	1.02	$R^2 = .971$	
$\cos(e, \cdot)$	0.04	0.00	0.07	1	0.18
$\cos(z, \cdot)$	0.96	-0.02	0.04	0.18	1
$\|\| \cdot \|\|$	10.80	2.43	1.70	1.31	10.57

Table 5.3: Diagnostics for NOX Model: Discrete C.

	f_1	f_2	$f_{1,2}$	e	z
κ	1.06	1.08	1.02	$R^2 = .974$	
$\cos(e, \cdot)$	0.04	-0.00	0.06	1	0.17
$\cos(z, \cdot)$	0.96	-0.02	0.12	0.17	1
$\|\| \cdot \|\|$	10.65	2.45	1.68	1.28	10.57

diagnostics are summarized in Table 5.3. The conclusion remains unchanged. To exercise extra caution to protect the interaction which was declared eminent by both Cleveland and Devlin (1988) and Breiman (1991) in their analyses, we further attached five separate smoothing parameters to the slices at the five different C values so the five curves are not shrunk towards each other, and calculated the cross-validated fit and evaluated the ANOVA decomposition with the side conditions $\int_0^1 f dt_1 = \sum_C f = 0$ at the design points. The diagnostics are summarized in Table 5.4. Despite the special protection, the C main effect and the interaction are still beyond our sights.

We finally calculated a cubic spline fit of $NO_x^{1/3}$ on E, which is plotted in Figure 5.1 together with the connected 1.96σ Bayesian confidence intervals.

Table 5.4: Diagnostics for NOX Model: Separate θ for Different C.

	f_1	f_2	$f_{1,2}$	e	z
κ	1.05	1.06	1.01	$R^2 = .979$	
$\cos(e, \cdot)$	0.06	0.00	0.12	1	0.17
$\cos(z, \cdot)$	0.96	-0.02	0.19	0.17	1
$\|\| \cdot \|\|$	10.55	2.31	1.84	0.91	10.57

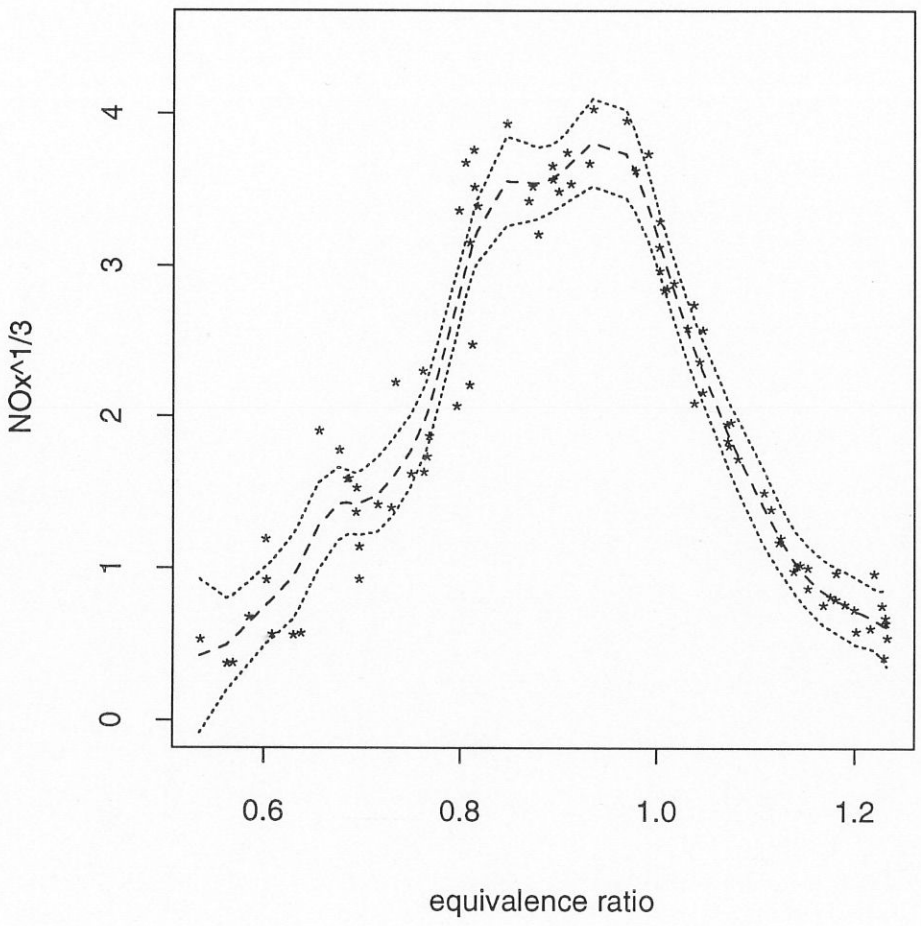

Figure 5.1: The NOX Model. The dashed line is the cross-validated cubic spline fit. The dotted lines are connected 1.96σ Bayesian confidence intervals. The data are superimposed as stars.

5.3 Ozone data

The data are 330 daily measurements of ozone concentration and eight other meteorological variables in the Los Angles basin in 1976. The purpose of the analysis is to build a predictive model of the ozone concentration on the other variables. The data were analyzed by Breiman and Friedman (1985) using ACE and by Buja *et al.* (1989) using additive regression models. A data description, a scatter plot matrix of the data, and a comparative study of various modeling techniques applied on the data can be found in Section 10.3 of Hastie and Tibshirani (1990). We used the variable code of Hastie and Tibshirani (1990) in our analysis (except that `humidity` is shortened as `hum`), and followed their suggestion in taking the log transform of the ozone concentration as the response. From the scatter plot matrix, the three variables `vh`, `temp`, and `ibt` are highly linearly correlated, and we picked `vh` and discarded the other two in our analysis. We also discarded the variable `wind` which showed no relation with any of the other variables. A square root transform is applied to the variable `vis` to make it more uniformly scattered on its range.

Our first attempt was to fit a model on the variables `vh`, `hum`, `ibh`, `dpg`, and `vis`. The translation $(\cdot - \min)/(\max - \min)$ was applied to all the variables to map the data into $[0,1]^5$. We first used tensor product linear spline with all five marginalization operators as $Ef = \int_0^1 f$. Linear splines give rougher looking fits but the main features of the fits are the same as those of cubic spline fits; see Gu and Wahba (1991a) for some simulation results. The reason for using linear splines in the screening stage was to save the number of smoothing parameters we had to deal with, noting that an interaction in a tensor product linear spline carries only one smoothing parameter while a two-factor interaction in a tensor product cubic spline can have as many as four. This is a computational advantage of linear splines over cubic splines in a multivariate setup since the cost of computing is proportional to the number of free smoothing parameters; see Gu and Wahba (1991a). We included the five main effects and the ten pairwise two-factor interactions of the five variables, altogether 16 terms (including the unpenalized constant). The cross-validated fit has a $R^2 = 0.741$. The 7 terms with small $\cos(z, f)$ and very small $\|f\|$ are listed in Table 5.5. Note that these include all pairwise interactions but those among the three variables `vh`, `ibh`, and `vis`. A refit was calculated with the terms in Table 5.5 deleted. The diagnostics are summarized in Table 5.6, where the last line records the maximum ratio (in absolute values) on the design points of the posterior mean over the posterior standard deviation of each term. It can be seen that f_{hum}, f_{ibh}, $f_{\text{vh,ibh}}$, and $f_{\text{vh,vis}}$ are

Table 5.5: Diagnostics for Ozone Data: Noise Interactions.

	$f_{\text{vh,hum}}$	$f_{\text{vh,dpg}}$	$f_{\text{hum,ibh}}$	$f_{\text{hum,dpg}}$	$f_{\text{hum,vis}}$	$f_{\text{ibh,dpg}}$	$f_{\text{dpg,vis}}$	e	z
$\cos(z, \cdot)$	0.59	0.57	0.21	0.38	0.20	0.16	0.18	0.53	1
$\|\| \cdot \|\|$	0.03	0.00	2.32	0.00	0.00	1.28	0.00	5.23	13.57

Table 5.6: Diagnostics for Ozone Data: Linear Spline Fit.

	f_{vh}	f_{hum}	f_{ibh}	f_{dpg}	f_{vis}	$f_{\text{vh,ibh}}$	$f_{\text{vh,vis}}$	$f_{\text{ibh,vis}}$	e	z
κ	1.78	2.92	4.80	2.73	2.38	2.33	1.86	2.30	$R^2 = .667$	
$\cos(e, \cdot)$	0.05	0.09	0.06	0.08	0.06	0.11	0.13	0.14	1	0.59
$\cos(z, \cdot)$	0.63	0.37	0.67	0.42	0.48	0.48	0.41	0.50	0.59	1
$\|\| \cdot \|\|$	6.35	0.62	1.27	3.70	2.73	1.02	0.57	2.14	6.55	13.57
$\max(f/\sigma_f)$	8.67	1.27	1.89	4.52	3.95	1.84	1.07	3.53		

very weak, both in that their norms are small and in that their 1.96σ Bayesian confidence intervals completely cover zero. Four of the five estimated main effects and their 1.96σ Bayesian confidence intervals are plotted in Figure 5.2.

A five term cubic spline refit was then calculated, including f_{vh}, f_{ibh}, f_{dpg}, f_{vis}, and $f_{\text{ibh,vis}}$, where f_{ibh} was included because that $\cos(z, f_{\text{ibh}})$ in Table 5.6 is big and that the interaction $f_{\text{ibh,vis}}$ was included. The diagnostics of the refit are summarized in Table 5.7. $f_{\text{ibh,vis}}$ became the next target of deletion. We finally fit a cubic spline main-effect-only model with f_{vh}, f_{ibh}, f_{dpg}, and f_{vis}. The diagnostics of the refit are summarized in Table 5.8. Everything looks normal. The terms in the final model are plotted in Figure 5.3.

Table 5.7: Diagnostics for Ozone Data: Cubic Spline Fit.

	f_{vh}	f_{ibh}	f_{dpg}	f_{vis}	$f_{\text{ibh,vis}}$	e	z
κ	1.47	1.83	1.15	1.35	1.37	$R^2 = .712$	
$\cos(e, \cdot)$	0.00	0.02	0.02	0.03	0.04	1	0.54
$\cos(z, \cdot)$	0.61	0.68	0.42	0.42	0.38	0.54	1
$\|\| \cdot \|\|$	5.90	3.21	4.79	2.79	2.22	6.97	13.57
$\max(f/\sigma_f)$	11.42	1.53	6.44	2.68	1.62		

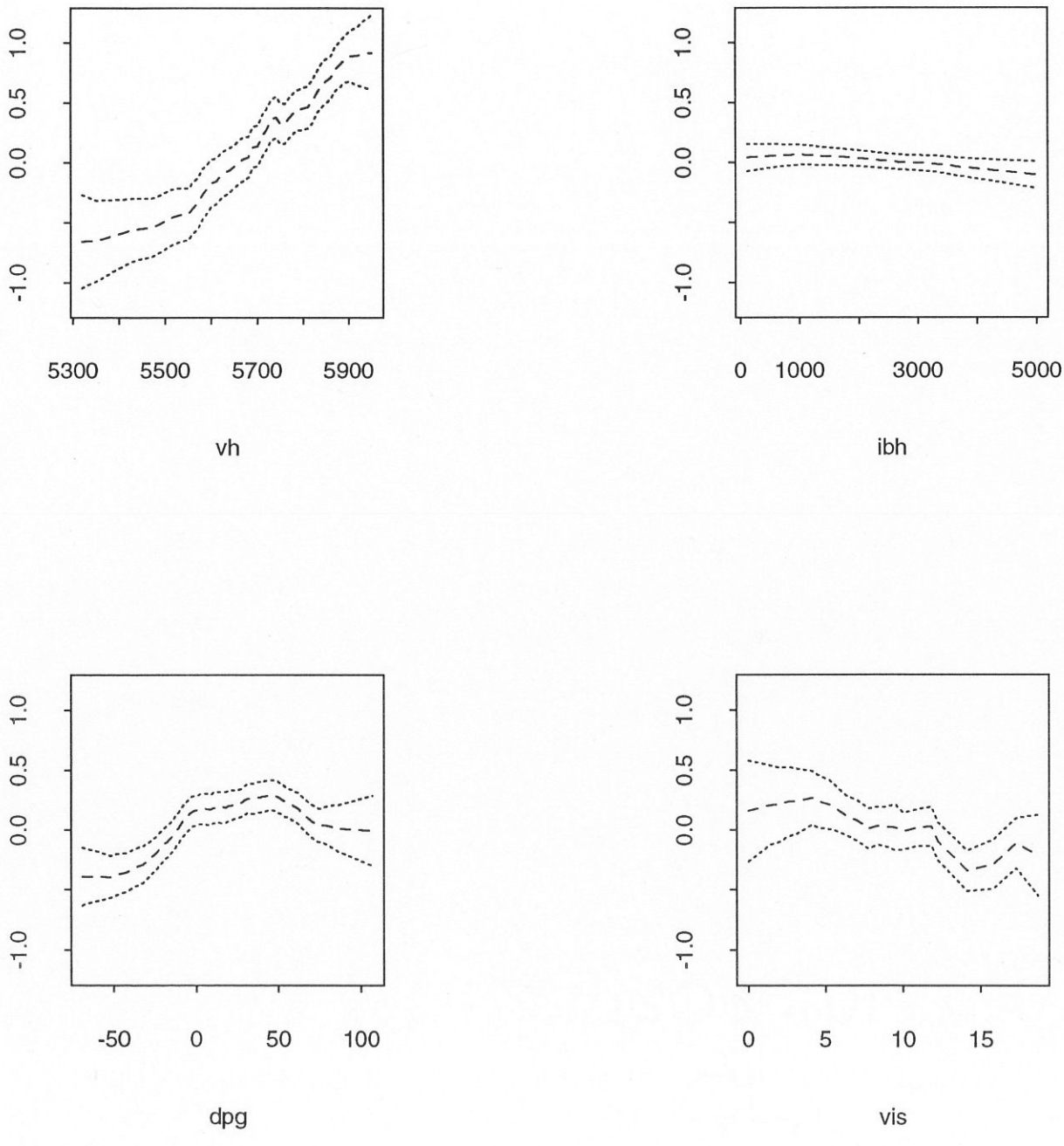

Figure 5.2: The Linear Spline Ozone Model: Main Effects. The dashed lines are the posterior means. The dotted lines are connected 1.96σ Bayesian confidence intervals.

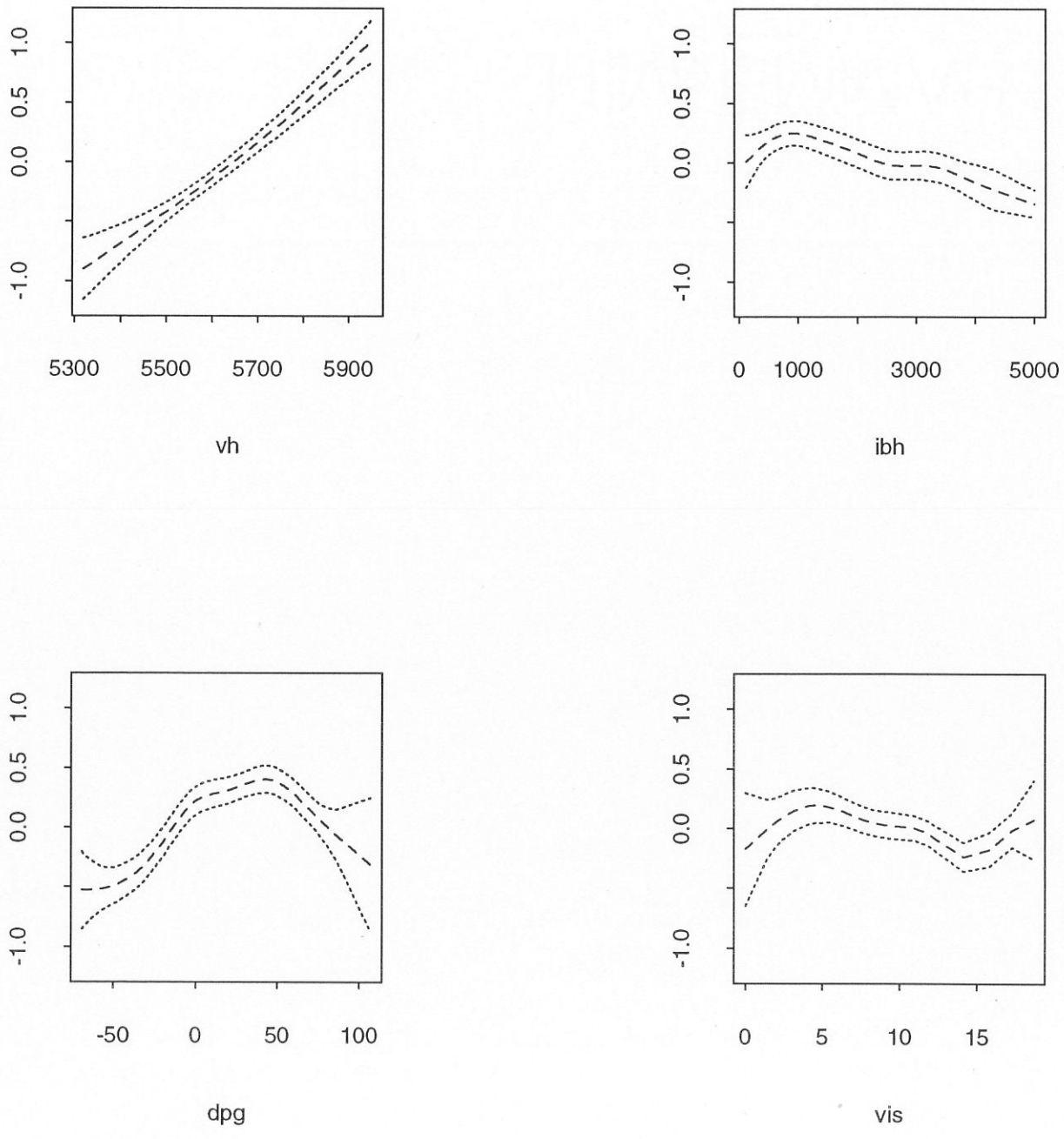

Figure 5.3: The Cubic Spline Ozone Model. The dashed lines are the posterior means. The dotted lines are connected 1.96σ Bayesian confidence intervals.

Table 5.8: Diagnostics for Ozone Data: Final Model.

	f_{vh}	f_{ibh}	f_{dpg}	f_{vis}	e	z
κ	1.46	1.58	1.11	1.17	$R^2 = .694$	
$\cos(e, \cdot)$	0.00	0.01	0.02	0.04	1	0.55
$\cos(z, \cdot)$	0.61	0.67	0.42	0.45	0.55	1
$\|\| \cdot \|\|$	6.04	4.22	4.96	2.02	7.27	13.57
$\max(f/\sigma_f)$	11.52	5.92	6.93	3.83		

6 Comparison of Parametric and Nonparametric Analyses

Data analysis does not produce information. The amount of information in the output of an analysis can not exceed the amount of information contained in its input, namely *the data* and *the assumptions.* In a nonparametric analysis one assumes less, so naturally the conclusions of a nonparametric analysis shall be weaker than those of a parametric analysis based on the same data. In this section, we present simulated examples to illustrate some implications of this simple fact.

Consider $f(t_1, t_2) = 1.5 + .5(e^{3t_1} - 1) + 3\sin(2\pi t_2 - \pi)$ on $[0, 1]^2$. We generated $n = 50$ design points $(t_{1,i}, t_{2,i})$ from $U(0, 1)^2$ and calculated $y_i = f(t_{1,i}, t_{2,i}) + \epsilon_i$, where ϵ_i were generated from $N(0, 1)$. We then conducted analyses using the ordinary linear regression technique, the parametric nonlinear regression technique, and the smoothing spline technique under decreasing amount of assumptions. Note that the function f is written as $f = f_\emptyset + f_1(t_1) + f_2(t_2)$, where $f_1(0) = \int_0^1 f_2 = 0$. We shall compare the confidence intervals of f_1 and f_2 from the three analyses. Note that the standard confidence intervals in a parametric analysis could be viewed as Bayesian confidence intervals under a uniform improper prior in the parametric space and they do carry a correct average coverage, so the intervals are comparable to each other under appropriate interpretations.

In the first analysis, we fitted a linear model

$$y = \beta_1 + \beta_2(e^{3t_1} - 1) + \beta_3 \sin(2\pi t_2 - \pi) + \epsilon.$$

The least squares fit gives $\hat{\beta}^T = (1.691, .468, 2.826)$ The estimated $f_1(t_1)$ and $f_2(t_2)$ are simply $\hat{\beta}_2(e^{3t_1} - 1)$ and $\hat{\beta}_3 \sin(2\pi t_2 - \pi)$ with standard deviations $s_{\hat{\beta}_2}|e^{3t_1} - 1|$ and $s_{\hat{\beta}_3}|\sin(2\pi t_2 - \pi)|$.

In the second analysis, we fitted a nonlinear model

$$y = \beta_1 + \beta_2(e^{\beta_3 t_1} - 1) + \beta_4 \sin(2\pi t_2 - \beta_5) + \epsilon.$$

The least squares fit gives $\hat{\beta}^T = (1.771, .394, 3.170, 2.812, 3.142)$. To make inferences concerning a nonlinear model, a standard approach is to calculate the linear approximation of the model at the fit, which we did. The approximating linear model in this case is

$$\begin{aligned} y &= \gamma_1 + \gamma_2(e^{\hat{\beta}_3 t_1} - 1) + \gamma_3 e^{\hat{\beta}_3 t_1} t_1 + \gamma_4 \sin(2\pi t_2 - \hat{\beta}_5) + \gamma_5 \cos(2\pi t_2 - \hat{\beta}_5) + \epsilon \\ &= \gamma_1 + \gamma_2 x_1(t_1) + \gamma_3 x_2(t_1) + \gamma_4 x_3(t_2) + \gamma_5 x_4(t_2) + \epsilon, \end{aligned}$$

where $e^{\beta_3 t_1} t_1 = d(e^{\beta_3 t_1} - 1)/d\beta_3$ and $\cos(2\pi t_2 - \beta_5) = -d(\sin(2\pi t_2 - \beta_5))/d\beta_5$. As expected, the least squares fit gives $\hat{\gamma}^T = (1.771, .394, .000, 2.812, .000)$. Note that $x_1(0) = x_2(0) = \int_0^1 x_3 = \int_0^1 x_4 = 0$. The estimated $f_1(t_1)$ is $\hat{\beta}_2(e^{\hat{\beta}_3 t_1} - 1) = \hat{\gamma}_2 x_1(t_1) + \hat{\gamma}_3 x_2(t_1)$ with an approximate standard deviation $(s^2_{\hat{\gamma}_2} x_1^2(t_1) + 2 s_{\hat{\gamma}_2} s_{\hat{\gamma}_3} r(\hat{\gamma}_2, \hat{\gamma}_3) x_1(t_1) x_2(t_1) + s^2_{\hat{\gamma}_3} x_2^2(t_1))^{1/2}$. The estimated $f_2(t_2)$ is $\hat{\beta}_4 \sin(2\pi t_2 - \hat{\beta}_5) = \hat{\gamma}_4 x_3(t_2) + \hat{\gamma}_5 x_4(t_2)$ with an approximate standard deviation $(s^2_{\hat{\gamma}_4} x_3^2(t_2) + 2 s_{\hat{\gamma}_4} s_{\hat{\gamma}_5} r(\hat{\gamma}_4, \hat{\gamma}_5) x_3(t_2) x_4(t_2) + s^2_{\hat{\gamma}_5} x_4^2(t_2))^{1/2}$.

In the third analysis, we used the two different configurations of cubic splines in Section 2 on the two axes to comply with the two different side conditions $f_1(0) = 0$ and $\int_0^1 f_2 = 0$. The interaction is eliminated and the penalty on the remaining components is $J(f) = \theta_1^{-1} \int_0^1 \ddot{f}_1 dt_1 + \theta_2^{-1} \int_0^1 \ddot{f}_2 dt_2$. The null space basis is $\{1, t_1, t_2 - .5\}$, from which the matrix S was generated. $R_1(s_1, t_1) = \int_0^1 (s_1 - u)_+ (t_1 - u)_+ du = (3t_1 - s_1)s_1^2/6$ for $s_1 \le t_1$, and $R_2(s_2, t_2) = k_2(s_2)k_2(t_2) - k_4(|s_2 - t_2|)$, from which Q_1 and Q_2 were constructed. The fit has an expression

$$\begin{aligned} f(t_1, t_2) &= d_1 + d_2 t_1 + d_3(t_2 - .5) + \sum_{i=1}^{n} c_i(\theta_1 R_1(t_{1,i}, t_1) + \theta_1 R_2(t_{2,i}, t_2)) \\ &= [d_1] + [d_2 t_1 + \theta_1 \sum_{i=1}^{n} c_i R_1(t_{1,i}, t_1)] + [d_3(t_2 - .5) + \theta_2 \sum_{i=1}^{n} c_i R_2(t_{2,i}, t_2)], \end{aligned}$$

where the brackets indicate the decomposition $f = f_\emptyset + f_1 + f_2$. Cross-validated fit and the related posterior standard deviations were calculated using RKPACK facilities as described in Section 4.

The results of the analyses are summarized in Figure 6.1. The two columns of Figure 6.1

correspond to the results for f_1 and f_2 respectively. The first three rows of Figure 6.1 correspond to the linear model analysis, the nonlinear model analysis, and the smoothing spline analysis respectively, where the solid lines are the truth, the dashed lines are the fitted, and the dotted lines are the 1.96σ Bayesian confidence intervals. The last row of Figure 6.1 compares the standard deviations in the three analyses, with `o` indicating the ordinary linear model, `n` indicating the nonlinear model, and `s` indicating the smoothing spline. As expected, the fewer the assumptions, the wider the intervals. For f_1, it can be seen that the impact of $f_1(0) = 0$ fades out much faster in the spline case than in the parametric cases. For f_2, the smoothing spline is less sure about its estimation near the boundaries of the data region.

Now consider a function $f(t_1, t_2) = 1.5 + [.4(e^{3t_1} - 1) + 2(1 - e^{-2t_1})] + [2.75\sin(2\pi t_2 - \pi) - .5\sin(4\pi t_2 - \pi)] = f_\emptyset + f_1 + f_2$. Note that both f_1 and f_2 are just slight modifications of the previous ones. We generated new y_i by evaluating this function on the same 50 design points and adding the same 50 pseudo $N(0, 1)$ perturbations. The maximum pairwise difference between the two sets of y_i is 1.045. The same three analyses conducted above were repeated on the new data set. The results of the analyses are summarized in Figure 6.2 with further details omitted. Based on the inaccurate assumptions, the 1.96σ confidence intervals in the linear model analysis missed f_1 almost entirely and missed f_2 over more than half of the $[0, 1]$ interval. The nonlinear parametric analysis gave a better estimate for f_1 because of the extra flexibility. However, the nonlinear f_2 point estimates are almost the same as in the linear model since the phase flexibility didn't help, although the interval estimates are more honest because of the extra uncertainty in the assumptions. In contrast to the parametric analyses, the performance of the smoothing spline analysis stays the same, and is comparatively better than the parametric analyses on the new data set. The conclusion is clear. More assumptions yield stronger claims, which are honest (hence better) when the assumptions are accurate, but could be misleading when the assumptions are inaccurate.

7 Model Indexing and Cross-Validation

As a model selection tool, generalized cross-validation aims to minimize the mean square error of the resulting estimate. Naturally, one would expect the cross-validated model to follow the optimal model which delivers the smallest mean square error among the class of available models. In examining the merit of cross-validation or any other model selection tools from this perspective,

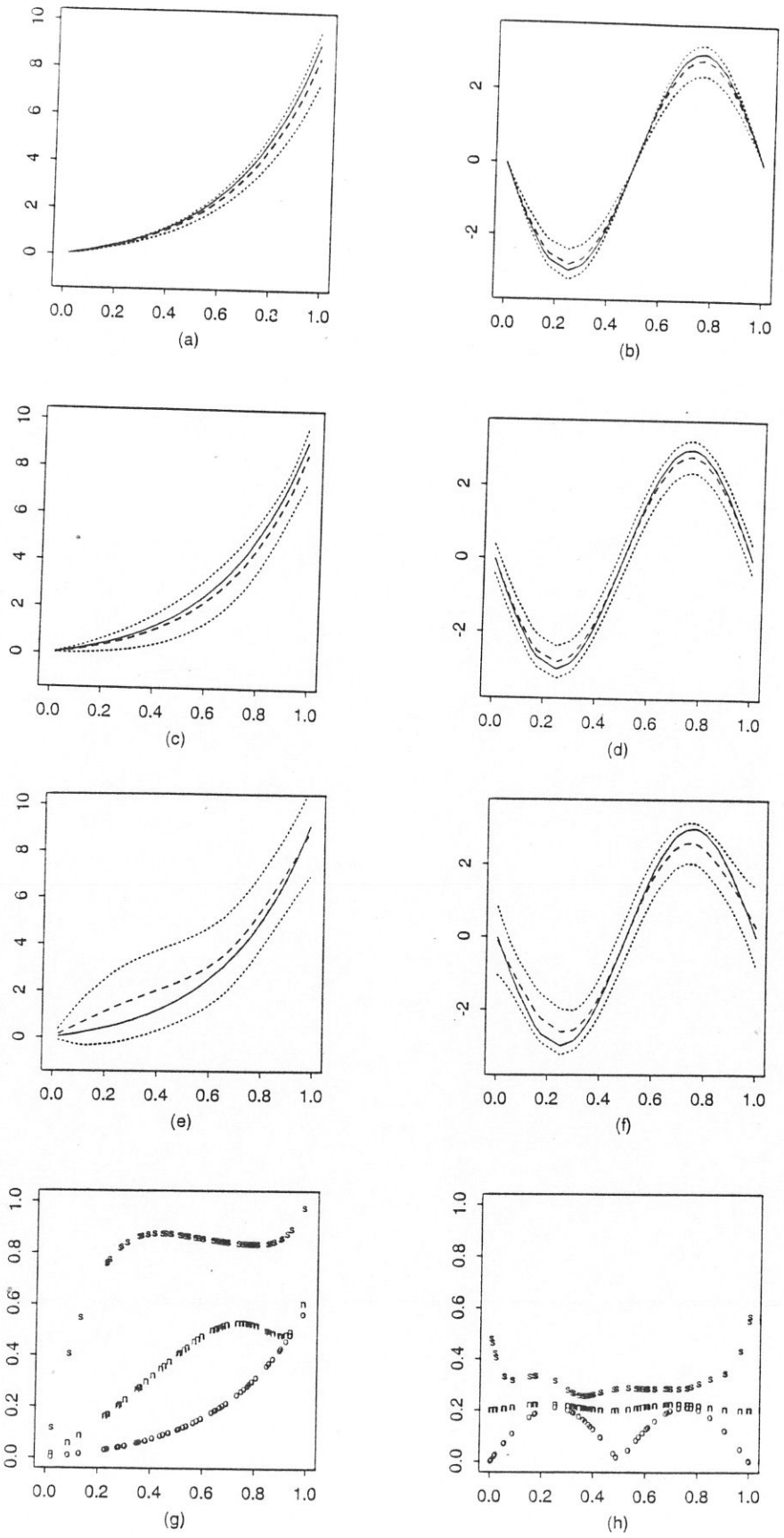

Figure 6.1: A Comparison of Nonparametric Analysis and Parametric Analyses with Correct Parametric Families. (a-b): Linear; (c-d): Nonlinear; (e-f): Nonparametric; (g-h): Comparison of Standard Deviations.

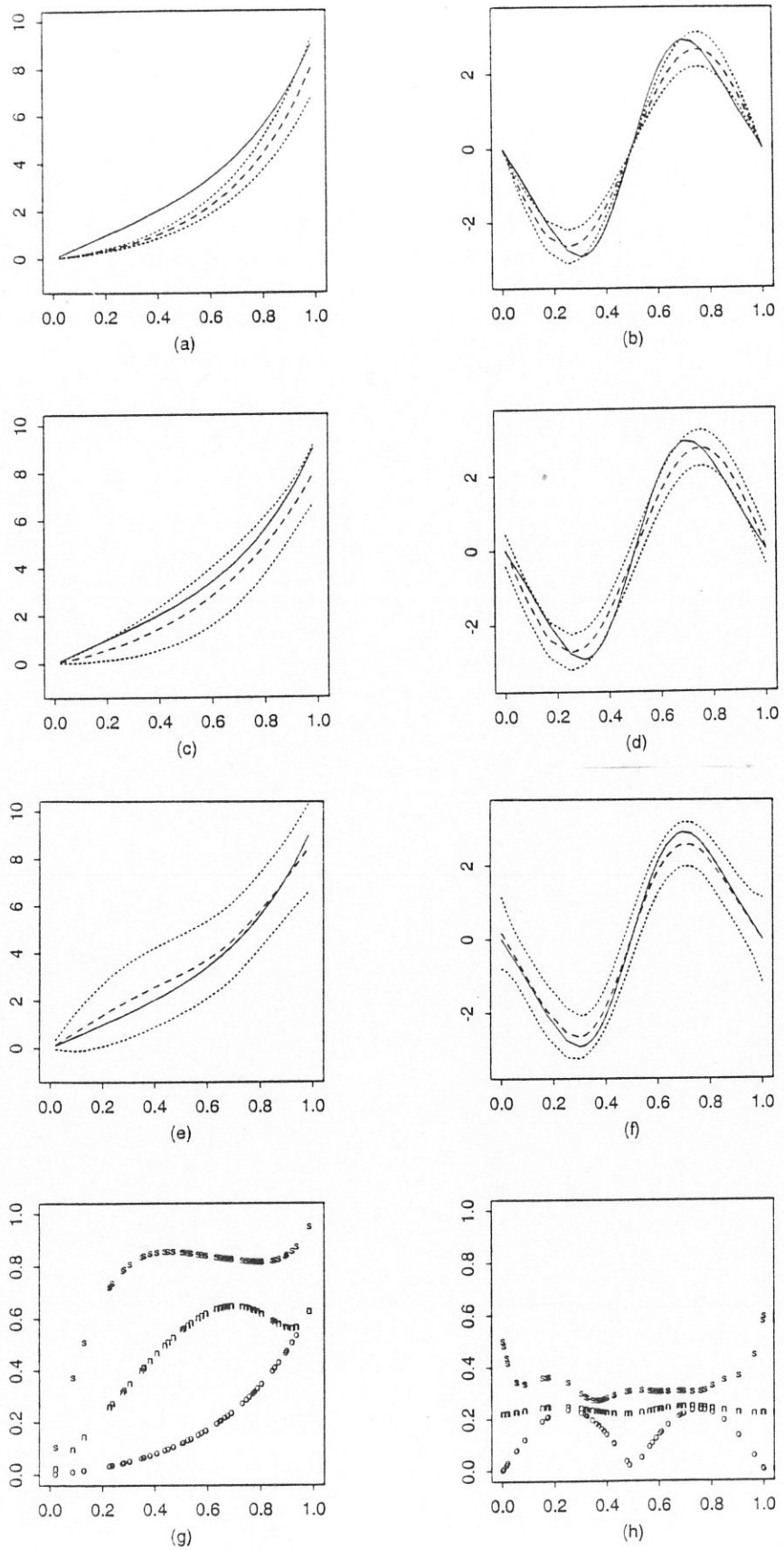

Figure 6.2: A Comparison of Nonparametric Analysis and Parametric Analyses with Incorrect Parametric Families. (a-b): Linear; (c-d): Nonlinear; (e-f): Nonparametric; (g-h): Comparison of Standard Deviations.

however, the proper indexing of the models is subtle. We shall illustrate this point via a set of simple simulations, and briefly discuss some implications of our finding.

Generate samples via $y_i = 3\sin(2\pi t_i - \pi) + \epsilon_i$, $t_i = (i - .5)/100$, $\epsilon_i \sim N(0, 1)$, $i = 1, \cdots, 100$, and calculate estimates using the cubic spline of Example 2.1. The choice of the test function is arbitrary and the signal to noise ratio is at best moderate. Knowing the truth, the optimal model could be obtained via a fine grid search in $\log_{10} n\lambda$. After simulating 100 replicates, the cross-validation $\log_{10} n\lambda$ and the optimal $\log_{10} n\lambda$ are plotted in the left frame of Figure 7.1. The cross-validation λ captured the magnitude of the optimal λ, but the apparent negative correlation appears bothersome. There is something wrong.

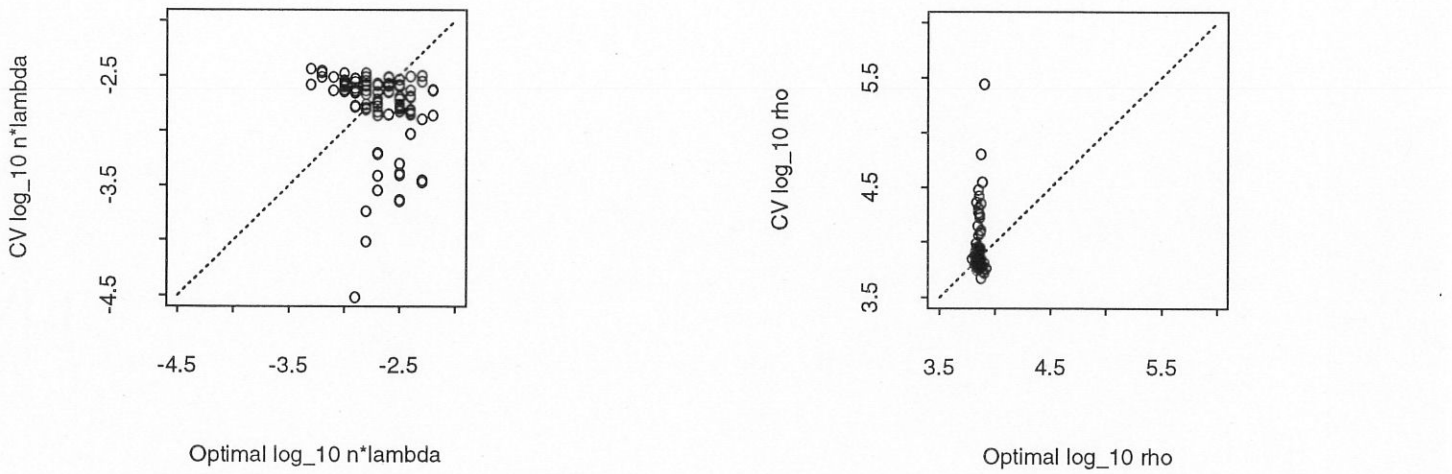

Figure 7.1: Behavior of Cross-Validation under Different Model Indexing. The left frame uses the λ indexing; the right frame uses the ρ indexing.

If the plot were not misleading us, then certainly cross-validation worked towards the wrong direction! But wait, what is the meaning of λ in this context? Recall Section 2, the underlying models are most directly specified via $J(f) \leq \rho$, which unfortunately gets ignored once we start to work with the convenient penalized least squares score (2.5). The subtle point rests right here: The ρ indexing of models is apparently data independent whereas the indexing by the Lagrange

multiplier λ is data dependent through the least squares. This tells us that *the same λ value in (2.5) generally implies different underlying models for different data*, so the points in the left frame of figure 7.1 are not comparable with each other. Now comes the right plot using the correct indexing: We calculated $\rho = J(\hat{f})$ which codes the selected models, and plotted the cross-validation $\log \rho$ and the optimal $\log \rho$ of the replicates in the right frame of Figure 7.1. Note that the penalty enters (2.5) in the form of $\lambda J(f)$, so the axes of the two plots are carefully scaled to the same unit. Clearly, the counter-intuitive negative correlation disappears. Furthermore, the range of the optimal models now appears much tighter, as it should be because the replicates are from a single data source.

The negative correlation demonstrated in the λ plot seems not a coincidence; it occurred in all similar simulations we conducted. The same phenomenon has also been reported many times in the literature for cross-validated kernel estimation applied to replicates generated from single data source, where the parameter λ is to be replaced by the kernel width; see, e.g., Hall and Johnstone (1992) for a recent account.

The implications of this subtle observation are straightforward but somewhat striking; what was deemed natural may now sound awkward. For example, it may not make sense to talk about a fixed "optimal" λ under repeated sampling from a single data source. Similarly, practical smoothing parameter selection under the λ indexing should only involve the observed data; this technically rules out the use of resampling techniques in the selection of λ. These implications do not extend to the ρ indexing, which however is very inconvenient to work with.

Of course our finding is in the spline smoothing context to which the implication is limited. Similar negative correlation in the kernel estimation literature between various forms of cross validation kernel widths and optimal kernel widths under repeated sampling from a single data source tempts us to submit that the kernel width indexing of kernel estimates may share more properties with the λ indexing of splines than with the ρ indexing of spines. A thorough understanding of related facts would require far more complicated analysis, however, and for now we shall leave it to the reader to judge how far the implications of our finding may extend.

8 Concluding Remarks

With the materials presented in Sections 2 through 6, we hope to bring to our readers' attention some of the recent developments in spline smoothing, their usefulness in data analysis, and the pros and cons of a nonparametric analysis compared to a parametric analysis. The main features of the smoothing spline technique are its generic explicit model specification and its generic continuous model coding via the smoothing parameters. The former makes it easier to incorporate structures on complex domains and the latter makes it possible to develop generic code for computation. The role of the underlying models in nonparametric estimation has been largely neglected in the literature, and Section 7 reminds us the possible pitfalls due to such a negligence. We omitted several important topics in our exposition, such as the thin plate splines and the smoothing of non Gaussian data; a comprehensive treatment can be found in Wahba (1990). Topics other than regression such as density estimation and hazard estimation are also omitted in our treatment.

In Sections 2 and 3, discrete domain smoothing splines are described in some detail for the first time, and are used as primary examples in our exposition. Although mathematically the simplest, these models are probably the least understood from a nonparametric perspective. The pure discrete models are potentially useful in handling large sparse contingency tables, and the mixed-covariate models provide a means for conducting nonparametric analysis of covariance. Further study is needed before routine use of these models can be recommended, however.

Acknowledgements

Chong Gu's research was supported by NSF under Grant DMS-9101730. Grace Wahba's research was supported by NSF under Grant DMS-9002566.

References

Aronszajn, N. (1950), "Theory of Reproducing Kernels," *Transaction of the American Mathematical Society*, 68, 337 – 404.

Bates, D. and Watts, D. (1988), *Nonlinear Regression Analysis and Its Applications.* Wiley.

Breiman, L. (1991), "The $\prod$ Method for Estimating Multivariate Functions from Noisy Data" (with discussion), *Technometrics*, 33, 125 – 160.

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models" (with discussion), *The Annals of Statistics*, 17, 453 – 555.

Cleveland, W. and Devlin, S. (1988), "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, 83, 596 – 610.

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

Draper, N. R. and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.) Wiley.

Friedman, J. (1991), "Multivariate Adaptive Regression Splines" (with discussion), *The Annals of Statistics*, 19, 1 – 141.

Friedman, J. and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817 – 823.

Gu, C. (1989), "RKPACK and Its Applications: Fitting Smoothing Spline Models," *Proceedings of Statistical Computing Section: American Statistical Association*, 42 – 51.

—— (1992), "Diagnostics for Nonparametric Regression Models with Additive Terms," *Journal of the American Statistical Association*, 87, 000 – 000.

Gu, C., Bates, D. M., Chen, Z., and Wahba, G. (1989), "The Computation of GCV Functions through Householder Tridiagonalization with Application to the Fitting of Interaction Spline Models," *SIAM Journal on Matrix Analysis and Applications*, 10, 457 – 480.

Gu, C. and Wahba, G. (1991a), "Minimizing GCV/GML Scores with Multiple Smoothing Parameters via the Newton Method," *SIAM Journal on Scientific and Statistical Computing*, 12, 383 – 398.

—— (1991b), Discussion of "Multivariate Adaptive Regression Splines" by J. Friedman, *The Annals of Statistics*, 19, 115 – 123.

—— (1991c), "Smoothing Spline ANOVA with Component-Wise Bayesian "Confidence Intervals"," *Journal of Computational and Graphical Statistics*, tentatively accepted.

—— (1993), "Semiparametric ANOVA with Tensor Product Thin Plate Splines," *Journal of the Royal Statistical Society*, Ser. B, 55, 000 – 000.

Hall, P. and Johnstone, I. (1992), "Empirical Functionals and Efficient Smoothing Parameter Selection" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 54, 475 – 530.

Hastie, T. and Tibshirani, R. (1986), "Generalized Additive Models," *Statistical Science*, 1, 297 – 318.

—— (1990), *Generalized Additive Models*. Chapman and Hall.

Huber, P. (1985), "Projection Pursuit" (with discussion), *The Annals of Statistics*, 13, 435 – 475.

Kimeldorf, G. and Wahba, G. (1970), "A Correspondence between Bayesian Estimation of Stochastic Processes and Smoothing by Splines," *The Annals of Mathematical Statistics*, 41, 495 – 502.

—— (1971), "Some Results on Tchebycheffian Spline Functions," *Journal of Mathematical Analysis and Applications*, 33, 82 – 95.

Nychka, D. (1988), "Bayesian Confidence Intervals for Smoothing Splines," *Journal of the American Statistical Association*, 83, 1134 – 1143.

—— (1990), "The Average Posterior Variance of a Smoothing Spline and a Consistent Estimate of the Average Squared Error," *The Annals of Statistics*, 18, 415 – 428.

Stewart, G. W. (1987), "Collinearity and Least Square Regression," *Statistical Science*, 2, 68 – 100.

Stone, C. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689 – 705.

Wahba, G. (1978), "Improper Priors, Spline Smoothing and the Problem of Guarding against Model Errors in Regression," *Journal of the Royal Statistical Society*, Ser. B, 40, 364 – 372.

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

—— (1986), "Partial and Interaction Splines for the Semiparametric Estimation of Functions of Several Variables," in *Computer Science and Statistics: Proceedings of the 18th Symposium on the interface*, ed. T.J. Boardman, American Statistical Association, pp. 75 – 80.

—— (1990), *Spline Models for Observational Data*, CBMS–NSF Regional Conference Series in Applied Mathematics, Vol. 59, SIAM.

Wecker, W. and Ansley, C. (1983), "The Signal Extraction Approach to Nonlinear Regression and Spline Smoothing," *Journal of the Americam Statistical Association*, 78, 81 – 89.