3.05

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

TECHNICAL REPORT NO. 881

June 10, 1991

Smoothing Spline ANOVA with Component-Wise Bayesian " Confidence Intervals"

by

Chong Gu

and

Grace Wahba

Smoothing Spline ANOVA with Component-Wise Bayesian "Confidence Intervals"

CHONG GU and GRACE WAHBA*

June 1991

Abstract

We study a multivariate smoothing spline estimate of a function of several variables, based on an ANOVA decomposition as sums of main effects functions (of one variable), two-factor interaction functions (of two variables), etc. We derive the Bayesian "confidence intervals" for the components of this decomposition and demonstrate that, even with multiple smoothing parameters, they can be efficiently computed using the publicly available code RKPACK, which was originally designed just to compute the estimates. We carry out a small Monte Carlo study to see how closely the actual properties of these component-wise confidence intervals match their rated confidence levels. We also analyze some lake acidity data as a function of calcium concentration, latitude, and longitude, using both polynomial and thin plate spline main effects in the same model. Lastly we suggest what might be necessary to generalize the known frequentist properties of these confidence intervals in the undecomposed case to the ANOVA components.

KEY WORDS: Smoothing spline ANOVA; Bayesian "confidence intervals"; RKPACK; Multivariate function estimation.

---

*Chong Gu is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907. His research was supported by the National Science Foundation under Grant DMS-9101730. Grace Wahba is John Bascom Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706. Her research was supported by the National Science Foundation under Grant DMS-9002566 and by the Air Force Office of Scientific Research under Grant AFOSR-90-0103

1 Introduction

We consider the model

$$y_i = f(t_1(i), \cdots, t_d(i)) + \epsilon_i, \quad i = 1, \cdots, n \tag{1.1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2), \sigma^2$ unknown, and t_α, the αth "variable" is in $\mathcal{T}^{(\alpha)}$, where $\mathcal{T}^{(\alpha)}$ is some measurable space. In the examples given here $\mathcal{T}^{(\alpha)} = E^{d(\alpha)}$, Euclidean $d(\alpha)$ space, and then $t = (t_1, \cdots, t_d)$ is in E^d space, where $d = \sum_\alpha d(\alpha)$. By setting $d(\alpha)$ to be 2 or 3, we will be able to include geographic, atmospheric or oceanic variables, along with other concomitant variables, in a natural way. We wish to estimate f, given the data $y = (y_1, \cdots, y_n)'$ in such a way as to avoid the "curse of dimensionality", and, additionally, to provide useful information concerning the accuracy of such estimates.

Nonparametric function estimation is a major research area at the present time and we just mention representative examples of modern techniques for multivariate function estimation in several dimensions: ACE (Breiman and Friedman, 1985), MARS (Friedman, 1991), CART (Breiman, Friedman, Olshen and Stone, 1984), Projection Pursuit (Huber, 1985), Regression Splines (Stone, 1985), and the $\prod$-method (Breiman, 1991). Each method has unique problems and successes in providing accuracy statements which we will not discuss here.

In this paper, we will be working within the framework of a general form of analysis of variance in reproducing kernel Hilbert spaces (RKHS) as applied in particular to additive and interaction smoothing splines, (SS-ANOVA). Thin plate splines are specifically included, and it is their use that allows the modeling of geographic and other variables as mentioned above. See Wahba (1990) for an overview of additive and interaction polynomial smoothing splines. More recently, it has been shown how to include thin plate splines in an SS-ANOVA model (Gu and Wahba, 1990, 1991a). The main goal of the present work is the establishment of component-wise Bayesian "confidence intervals" in the SS-ANOVA context, generalizing the univariate Bayesian "confidence intervals" of Wahba (1983), and further studied by Nychka (1988, 1990), Cox (1989) and Hall and Titterington (1987), and recently extended to the non-Gaussian case by Gu (1992). In this paper, we derive these intervals for each component in the ANOVA decomposition, and, more importantly, obtain them in a form which allows a stable and efficient calculation, via the publicly available code RKPACK (Gu, 1989). We carry out a small Monte Carlo study to suggest how the confidence intervals might

work in practice. One achievement of this paper is a demonstration of the (workstation) feasibility of this type of calculation, another is the demonstration of the visual efficacy of the result. Finally we make some conjectures concerning the theoretical properties, and remarks on their relation to certain well-known error bounds, like the hypercircle inequality in function spaces.

We assume that $f \in \mathcal{H}$, a reproducing kernel Hilbert space (RKHS), that is, a Hilbert space in which all the point evaluations are bounded. See Aronszajn (1950), Weinert (1982), Mate (1989), and Wahba (1990). The last two give an expository description of facts about RKHS that are used here. We note that RKHS are the most general Hilbert spaces that are useful if one is interested in estimating the value of a function at a point.

Much recent work has focussed on the so-called additive, (or main-effect-only) models of the form

$$f(t) = C + \sum_{\alpha=1}^{d} f_\alpha(t_\alpha),$$

where $t_\alpha \in E^1$, and sufficient conditions, say

$$\int f_\alpha(t_\alpha) d\mu_\alpha(t_\alpha) = 0,$$

are imposed to insure that the model is identifiable. See Stone (1985), Buja, Hastie and Tibshirani (1989), Hastie and Tibshirani (1990) and references cited there. In this paper μ_α is a probability measure on $\mathcal{T}^{(\alpha)}$, satisfying some conditions. The present authors and others have been examining generalizations of the additive models to additive and interaction models, that is, models of the form

$$f(t) = C + \sum_{\alpha} f_\alpha(t_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \sum_{\alpha<\beta<\gamma} f_{\alpha\beta\gamma}(t_\alpha, t_\beta, t_\gamma) + \cdots \qquad (1.2)$$

and so forth. See Barry (1983, 1986), Chen (1989), Stone (1990). In previous work relevant to the present paper, a mathematical framework has been developed for fitting these models by penalized likelihood and in particular smoothing spline methods (Wahba, 1986; Chen, Gu and Wahba, 1989; Wahba, 1990; Gu, 1990). Numerical methods for fitting these models have been developed (Gu, Bates, Chen and Wahba, 1989; Gu and Wahba, 1991b), and publicly available code developed (Gu, 1989).

We first suppose that a model $\mathcal{M}$ has been selected, in our case $\mathcal{M}$ is the set of subspaces corresponding to the terms in the right hand side of (1.2) that will be retained in the model. The

estimate f_λ of f is then obtained by finding f_λ in $\mathcal{M}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{t}(i)))^2 + \lambda\left[\sum_{\alpha\in I_{\mathcal{M}}}\theta_\alpha^{-1}J_\alpha(f_\alpha) + \sum_{\alpha,\beta\in I_{\mathcal{M}}}\theta_{\alpha\beta}^{-1}J_{\alpha\beta}(f_{\alpha\beta}) + \cdots\right] \tag{1.3}$$

where $I_{\mathcal{M}}$ is the collection of indices for the terms included in $\mathcal{M}$, and the $J_\alpha, J_{\alpha\beta}$ and so forth are quadratic "smoothness" penalty functionals. λ is the main smoothing parameter, and the θ's are subsidiary smoothing parameters.

It is a major task of nonparametric regression to provide some sort of accuracy statements concerning the resulting estimate. Wahba (1983) described Bayesian "confidence intervals" for the (one component) smoothing spline model by deriving the posterior covariance for f given the Bayes model which is associated with spline smoothing, and showed by a Monte Carlo study that these confidence intervals appeared to have a certain frequentist property for f in certain function spaces. The property is, considering the n 95% confidence intervals at the n data points, about 95% of them will cover the values of the true curve there. A partly heuristic theoretical argument why this could be expected was given there, and later Nychka (1988, 1990), Hall and Titterington (1987), and Cox (1989) provided theorems concerning when and why they should work. Other definitions of confidence regions are of interest, in particular, a set of intervals that are required to cover 100% of the points with probability .95. Such intervals can be expected to be wider that the intervals considered in Wahba (1983). See for example Li (1989), Hall and Titterington (1988). We remark that the weaker definition of "confidence interval" which is adopted in Wahba (1983) leads to intervals which are easy to interpret psychologically. In Monte Carlo simulations, when they cover 95% of the values of the true curve at the data points, the intervals more or less "graze" the truth, and the width of the intervals is visually interpretable by an unsophisticated user as an accuracy indicator.

In this paper we generalize these confidence intervals to obtain the posterior covariance functions for the components $f_\alpha, f_{\alpha\beta}$, etc. of the model $\mathcal{M}$. Then we show how the generic algorithms in RKPACK can be used directly to compute the component-wise confidence intervals, with only trivial modifications. Once this is done, we carry out a Monte Carlo study to suggest whether or not the component-wise confidence intervals can be expected to inherit some of the favorable Monte Carlo and theoretical results available for the single variable case. The results are, with some caveats, suggestive that the answer is "yes". In the cases we have tried, the visual images

quite consistently reflected an interpretable "reality".

As a byproduct, we obtain another useful graphical tool: In estimating functions of two (or more) variables by nonparametric methods, the data are frequently arranged irregularly. This is particularly true for geographic data. While it is tempting to plot the estimate in, say, a rectangle, once one is sufficiently far from the data the nonparametric estimates become meaningless. We can obtain contours of constant posterior variance in two (or more) variables and use one of these contours to bound an area within which the estimated function will be displayed as a contour plot.

In Section 2 we describe a general form of ANOVA in function spaces. This represents a modest generalization on the construction proposed in Gu and Wahba (1990, 1991a), we include a summary here to make the exposition self-contained. In Section 3 we give the component-wise posterior covariance functions. A comparison of the result with the representation of a smoothing spline given in Kimeldorf and Wahba (1971) shows that, if an efficient tool for computing f_λ is available (as in RKPACK), then only trivial additions are required to compute the component-wise confidence intervals. In Section 4 we review old and generate some new methods for obtaining reproducing kernels which are used in obtaining explicit formulas for the minimizer of (1.3) with spline and spline-like penalty functionals. In Section 5 we provide the details of how RKPACK may be used to carry out the calculations, and in Section 6 we present the results of a small Monte-Carlo study. In Section 7 we review some of the theoretical properties of the single variable confidence intervals, and suggest what might have to be done to extend the main theorems of Nychka (1988, 1990) to the component-wise case.

2 Analysis of Variance in RKHS

Let $\mathcal{H}$ be an RKHS of real-valued functions of $\boldsymbol{t} = (t_1, \cdots, t_d)$. Here $t_\alpha \in \mathcal{T}^{(\alpha)}$, where $\mathcal{T}^{(\alpha)}$ is some index set. We need the following further properties:

1. $1 \in \mathcal{H}$, where 1 is the constant function of $\boldsymbol{t}$.

2. For each $\alpha = 1, \cdots, d$, we can construct a probability measure $d\mu_\alpha$ on $\mathcal{T}^{(\alpha)}$, and an averaging operator $\mathcal{E}_\alpha$, such that
$$(\mathcal{E}_\alpha f)(\boldsymbol{t}) = \int_{\mathcal{T}^{(\alpha)}} f(t_1, \cdots, t_d) d\mu_\alpha(t_\alpha)$$
is well defined and $\mathcal{E}_\alpha f \in \mathcal{H}$.

Consider

$$I = \prod_{\alpha}(\mathcal{E}_\alpha + (I - \mathcal{E}_\alpha)) = \prod_{\alpha}\mathcal{E}_\alpha + \sum_{\alpha}(I - \mathcal{E}_\alpha)\prod_{\beta\neq\alpha}\mathcal{E}_\beta + \sum_{\alpha<\beta}(I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta)\prod_{\gamma\neq\alpha,\beta}\mathcal{E}_\gamma + \cdots + \prod_{\alpha}(I - \mathcal{E}_\alpha). \tag{2.1}$$

This decomposition of the identity generates a unique (ANOVA-like) decomposition of f into

$$f = C + \sum_{\alpha} f_\alpha + \sum_{\alpha<\beta} f_{\alpha\beta} + \cdots + f_{1\cdots d}$$

where $C = \prod_\alpha \mathcal{E}_\alpha f, f_\alpha = (I - \mathcal{E}_\alpha)\prod_{\beta\neq\alpha}\mathcal{E}_\beta f, f_{\alpha\beta} = (I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta)\prod_{\gamma\neq\alpha,\beta}\mathcal{E}_\gamma f$, etc, are the mean, main effects, two factor interactions, etc. The closure of the range of each operator of the form $\prod_{\alpha_1,\cdots,\alpha_k}\mathcal{E}_\alpha\prod_{\alpha_{k+1},\cdots,\alpha_d}(I - \mathcal{E}_\beta)$ is clearly a subspace of $\mathcal{H}$, however, these subspaces are not necessarily orthogonal with respect to the inner product in $\mathcal{H}$.

We will now provide a construction of $\mathcal{H}$ in which these subspaces are all orthogonal. Let $\mathcal{H}^{(\alpha)}$ be an RKHS of functions on $\mathcal{T}^{(\alpha)}$ with $\int_{\mathcal{T}^{(\alpha)}} f_\alpha(t_\alpha)d\mu_\alpha = 0$, $f \in \mathcal{H}^{(\alpha)}$, and let $\{1^{(\alpha)}\} = \{1\}$ be the one dimensional space of constant functions on $\mathcal{T}^{(\alpha)}$. Consider the space $\{1\} \oplus \mathcal{H}^{(\alpha)}$. Then f in this space will have a unique decomposition $f = P_c f + (f - P_c f)$, with $P_c f = \int f d\mu_\alpha \in \{1\}$ and $(f - P_c f) \in \mathcal{H}^{(\alpha)}$, we endow this space with the square norm $\|f\|^2 = (P_c f)^2 + \|f - P_c f\|^2_{\mathcal{H}^{(\alpha)}}$. Now, let

$$\mathcal{H} = \prod_{\alpha}[\{1\} \oplus \mathcal{H}^{(\alpha)}],$$

which can be expanded as

$$\mathcal{H} = \{1\} \oplus \sum_{\alpha}\mathcal{H}^{(\alpha)} \oplus [\sum_{\beta<\alpha}\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots$$

Here $f_\alpha \in \mathcal{H}^{(\alpha)}$ is called a main effect, $f_{\alpha\beta} \in \mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}$ is a two factor interaction, and so forth. The subspaces are orthogonal in the tensor product norm induced by the original inner products. Aronszajn (1950) can be consulted for details about tensor products of RKHS. Examples will be given later. Now $f \in \mathcal{H}$ has a unique orthogonal decomposition

$$f = C + \sum f_\alpha + \sum_{\alpha<\beta} f_{\alpha\beta} + \cdots$$

with $C = \int f \prod_\alpha d\mu_\alpha, f_\alpha \in \mathcal{H}^{(\alpha)}$, $f_{\alpha\beta} \in \mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}$ and so forth. For other interesting views of analysis of variance, see Antoniadis (1984) and Speed (1987).

We want one further decomposition, to allow for the imposition of spline and related penalty functionals. Let $\mathcal{H}^{(\alpha)}$ have an orthogonal decomposition $\mathcal{H}_\pi^{(\alpha)} \oplus \mathcal{H}_s^{(\alpha)}$, where $\mathcal{H}_\pi^{(\alpha)}$ is finite dimensional (the "parametric" part; usually, but not always, polynomials), and $\mathcal{H}_s^{(\alpha)}$, (the "smooth" part) is the orthocomplement of $\mathcal{H}_\pi^{(\alpha)}$ in $\mathcal{H}_\alpha$. We will later let $J_\alpha(f_\alpha) = \|P_s^{(\alpha)} f_\alpha\|^2_{\mathcal{H}^{(\alpha)}}$, where $P_s^{(\alpha)}$ is the orthogonal projection operator in $\mathcal{H}^{(\alpha)}$ onto $\mathcal{H}_s^{(\alpha)}$. Thus the null space of J_α in $\mathcal{H}^{(\alpha)}$ is $\mathcal{H}_\pi^{(\alpha)}$. $\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}$ will be a direct sum of four orthogonal subspaces:

$$\begin{aligned} \mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)} &= \mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)} && (2.2)\\ &+ \mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)} && (2.3)\\ &+ \mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)} && (2.4)\\ &+ \mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}. && (2.5) \end{aligned}$$

By convention the elements of the finite dimensional space $\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}$ are not penalized. We will in Section 4 let the penalties in the other subspaces be their square norms.

At this point we have (orthogonally) decomposed $\mathcal{H}$ into sums of products of unpenalized finite dimensional subspaces, plus main effects subspaces, plus two factor interaction spaces of the form parametric $\otimes$ smooth (π, s) of the form (2.3), smooth $\otimes$ parametric (s, π) of the form (2.4), and smooth $\otimes$ smooth (s, s) of the form (2.5), and so on for the three and higher factor subspaces.

Now we suppose that we have selected the model $\mathcal{M}$, that is, we have decided which subspaces will be included. Next, collect all of the included unpenalized subspaces into a subspace, call it $\mathcal{H}^0$, of dimension M, and relabel the other subspaces as $\mathcal{H}^\beta, \beta = 1, 2, \cdots, p$. Thus, $\mathcal{H}^\beta$ may stand for a subspace $\mathcal{H}_s^{(\alpha)}$, or one of the subspaces of the form (2.3), (2.4), (2.5), or a higher order subspace. Our model estimation problem becomes: find $f \in \mathcal{M} = \mathcal{H}^0 \oplus \sum_\beta \mathcal{H}^\beta$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{t}(i)))^2 + \lambda \sum_\beta \theta_\beta^{-1} \|P^\beta f\|^2, \tag{2.6}$$

where P^β is the orthogonal projector in $\mathcal{M}$ onto $\mathcal{H}^\beta$. Given a basis for $\mathcal{H}^0$, and reproducing kernels $R_\beta(\boldsymbol{s}, \boldsymbol{t})$ for $\mathcal{H}^\beta$, an explicit formula for the minimizer f_λ of (2.6) is well known; see, e.g., Chapter 10 of Wahba (1990). The code RKPACK (Gu, 1989) may be used to compute the GCV estimates of λ and the θ's. This code is available by writing netlib@ornl.gov with the words "`send index`", the robot mailserver will then respond with instructions. This code will be used later in the Monte Carlo experiments presented here.

3 Bayesian "Confidence Intervals" for Components

In this Section we derive general formulas for Bayesian "confidence intervals" for the components of f estimated by minimizing (2.6). The computation of the relevant quantities shall be discussed in Section 5.

We first review some relevant facts. Let $R_\beta(\boldsymbol{s},\boldsymbol{t})$ be the reproducing kernel for $\mathcal{H}^\beta$ and let $\phi_1,\cdots,\phi_M$ span $\mathcal{H}^0$. Let $X_\xi(\boldsymbol{t}), \boldsymbol{t}\in\mathcal{T}=\prod\mathcal{T}^{(\alpha)}$ be a stochastic process defined by

$$X_\xi(\boldsymbol{t}) = \sum_{\nu=1}^{M}\tau_\nu\phi_\nu(\boldsymbol{t}) + b^{1/2}\sum_{\beta=1}^{p}\sqrt{\theta_\beta}Z_\beta(\boldsymbol{t}),$$

where $\tau=(\tau_1,\cdots,\tau_M)'\sim\mathcal{N}(0,\xi I)$, the Z_β are independent, zero mean Gaussian stochastic processes, independent of the τ_ν, with $EZ_\beta(\boldsymbol{s})Z_\beta(\boldsymbol{t})=R_\beta(\boldsymbol{s},\boldsymbol{t})$. We have $Z(\boldsymbol{t})=\sum_\beta\sqrt{\theta_\beta}Z_\beta(\boldsymbol{t})$ satisfies $EZ(\boldsymbol{s})Z(\boldsymbol{t})=R(\boldsymbol{s},\boldsymbol{t})$ where $R(\boldsymbol{s},\boldsymbol{t})\equiv\sum_\beta\theta_\beta R_\beta(\boldsymbol{s},\boldsymbol{t})$.

Now, let

$$Y_i = X_\xi(\boldsymbol{t}(i))+\epsilon_i,\quad i=1,\cdots,n,$$

where $\epsilon=(\epsilon_1,\cdots,\epsilon_n)'\sim\mathcal{N}(0,\sigma^2 I)$. Let

$$f_\lambda(\boldsymbol{t}) = \lim_{\xi\to 0}E\{X_\xi(\boldsymbol{t})|Y_i=y_i, i=1,\cdots,n\}$$

and set $b=\sigma^2/n\lambda$. It is well known (Kimeldorf and Wahba, 1971), that

$$f_\lambda(\boldsymbol{t}) = \sum_{\nu=1}^{M}d_\nu\phi_\nu(\boldsymbol{t}) + \sum_{i=1}^{n}c_i R(\boldsymbol{t},\boldsymbol{t}(i)) \tag{3.1}$$

where $d=(d_1,\cdots,d_M)'$ and $c=(c_1,\cdots,c_n)'$ are given by

$$d = (S'M^{-1}S)^{-1}S'M^{-1}y \tag{3.2}$$

$$c = (M^{-1}-M^{-1}S(S'M^{-1}S)^{-1}S'M^{-1})y \tag{3.3}$$

where S is the $n\times M$ matrix with $i\nu$th entry $\phi_\nu(\boldsymbol{t}(i))$ and $M=\Sigma+n\lambda I$, where Σ is the $n\times n$ matrix with ijth entry $R(\boldsymbol{t}(i),\boldsymbol{t}(j))$. It is always being assumed that S is of full column rank. Furthermore, for any $\lambda>0$, f_λ is the minimizer of (2.6). See also Wahba (1978, 1990). The projections of f_λ on the various subspaces are the posterior means of the corresponding components and can be read off of (3.1). For example, let $g_{0,\nu}(\boldsymbol{t})=\tau_\nu\phi_\nu(\boldsymbol{t})$ and $g_\beta(\boldsymbol{t})=b^{1/2}\sqrt{\theta_\beta}Z_\beta(\boldsymbol{t})$, then we have

$$\begin{aligned} Eg_{0,\nu}(\boldsymbol{t})|y &= d_\nu\phi_\nu(\boldsymbol{t}) \\ Eg_\beta(\boldsymbol{t})|y &= \sum_{i=1}^{n}c_i\theta_\beta R_\beta(\boldsymbol{t},\boldsymbol{t}(i)). \end{aligned}$$

The posterior covariances of $g_{0,\nu}$ and g_β are summarized in the following theorem.

Theorem 3.1

$$\begin{aligned} Cov(g_{0,\nu}(\boldsymbol{s}), g_{0,\mu}(\boldsymbol{t}))/b &= \phi_\nu(\boldsymbol{s})\phi_\mu(\boldsymbol{t})e_\nu'(S'M^{-1}S)^{-1}e_\mu \\ Cov(g_\beta(\boldsymbol{s}), g_{0,\nu}(\boldsymbol{t}))/b &= d_{\nu,\beta}(\boldsymbol{s})\phi_\nu(\boldsymbol{t}) \\ Cov(g_\beta(\boldsymbol{s}), g_\beta(\boldsymbol{t}))/b &= \theta_\beta R_\beta(\boldsymbol{s},\boldsymbol{t}) - \sum_{i=1}^{n} c_{i,\beta}(\boldsymbol{s})\theta_\beta R_\beta(\boldsymbol{t},\boldsymbol{t}(i)) \\ Cov(g_\gamma(\boldsymbol{s}), g_\beta(\boldsymbol{t}))/b &= -\sum_{i=1}^{n} c_{i,\gamma}(\boldsymbol{s})\theta_\beta R_\beta(\boldsymbol{t},\boldsymbol{t}(i)) \end{aligned}$$

where e_ν is the νth unit vector, and $(d_{1,\beta}(\boldsymbol{s}),\cdots,d_{M,\beta}(\boldsymbol{s})) = d_\beta(\boldsymbol{s})'$ and $(c_{1,\beta}(\boldsymbol{s}),\cdots,c_{n,\beta}(\boldsymbol{s})) = c_\beta(\boldsymbol{s})'$ are given by

$$d_\beta(\boldsymbol{s}) = (S'M^{-1}S)^{-1}S'M^{-1}\begin{pmatrix} \theta_\beta R_\beta(\boldsymbol{s},\boldsymbol{t}(1)) \\ \vdots \\ \theta_\beta R_\beta(\boldsymbol{s},\boldsymbol{t}(n)) \end{pmatrix} \tag{3.4}$$

$$c_\beta(\boldsymbol{s}) = [M^{-1} - M^{-1}S(S'M^{-1}S)^{-1}S'M^{-1}]\begin{pmatrix} \theta_\beta R_\beta(\boldsymbol{s},\boldsymbol{t}(1)) \\ \vdots \\ \theta_\beta R_\beta(\boldsymbol{s},\boldsymbol{t}(n)) \end{pmatrix} \tag{3.5}$$

The proof is given in the appendix. It is clear that the calculation of the posterior covariances boils down to the calculation of $(S'M^{-1}S)^{-1}$, c_β and d_β, which we will pursue in Section 5.

4 Spline Penalty Functionals and Reproducing Kernels for SS-ANOVA Models

For $\mathcal{T}^{(\alpha)} = E^1$, the real line, we will let the main effect penalty functional be

$$J_m^1(f) = \int_{-\infty}^{\infty}(f^{(m)}(x))^2dx.$$

The null space of this penalty functional is the m-dimensional span of the polynomials of total degree less than m. If $d(\alpha) = k > 1$, then we let $\mathcal{T}^{(\alpha)} = E^k$, where E^k is Euclidean k-space. Then for the the main effect penalty functional we will use

$$J_m^k(f) = \sum_{\gamma_1+\cdots+\gamma_k=m}\frac{m!}{\gamma_1!\cdots\gamma_k!}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\frac{\partial^m f}{\partial x_1^{\gamma_1}\cdots\partial x_k^{\gamma_k}}\right)^2 dx_1\cdots dx_k, \tag{4.1}$$

which is the thin plate penalty functional. The null space of this penalty functional is the $\binom{m+k-1}{k}$ polynomials of total degree less than m in k variables. For technical reasons it is necessary that $2m-k>0$. In our examples we will let $\mathcal{T}^{(\alpha)}$ be E^1 or E^2, and $m=2$. The results generalize immediately to arbitrary $m,k,2m>k$ (provided there is enough data), but the notation becomes messy. See Wahba (1990) for details. In particular

$$J_2^2(f)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{\partial^2 f}{\partial x_1^2}\right)^2+2\left(\frac{\partial^2 f}{\partial x_1\partial x_2}\right)^2+\left(\frac{\partial^2 f}{\partial x_2^2}\right)^2 dx_1dx_2.$$

We will let the space $\{1\}\oplus\mathcal{H}^{(\alpha)}$ be the thin plate space $\mathcal{X}=\mathcal{X}_m^k$ described by Meinguet (1979), for our purposes we only need to know that $\mathcal{X}_m^k$ contains the $\binom{m+k-1}{k}$ polynomials of total degree less than m in k variables, and functions for which $J_m^k(\cdot)$ is well defined and finite. The $k=1$ case leads to polynomial splines, which have been discussed in the SS-ANOVA context from a slightly different point of view in Chen *et al.* (1989) and Gu *et al.* (1989). For our purposes it will be easiest to consider $k=1$ as a special case of the general k case we consider here.

We want to decompose $\mathcal{H}^{(\alpha)}$ into $\mathcal{H}_\pi^{(\alpha)}\oplus\mathcal{H}_s^{(\alpha)}$, where $\mathcal{H}_\pi^{(\alpha)}$ is the $\binom{m+k-1}{k}-1$ dimensional space spanned by the polynomials of total degree less than m which satisfy $\int f d\mu_\alpha=0$. We will do this in such a way that $J_m^k(f)$ is the squared norm on $\mathcal{H}_s^{(\alpha)}$, and, finally, we will obtain reproducing kernels (RK's) for $\mathcal{H}_\pi^{(\alpha)}$ and $\mathcal{H}_s^{(\alpha)}$. Given these RK's we will immediately have RK's for spaces of the form (2.2) - (2.5), and higher order spaces, by multiplying and adding the appropriate RK's.

To carry out this program, let $\phi_1,\cdots,\phi_M, M=M(\alpha)=\binom{m+k-1}{k}$ span the polynomials of total degree less than m on E^k, and choose them so that $\phi_1=1$ and so that they are orthonormal under the inner product

$$<\phi_\mu,\phi_\nu>=\int\phi_\mu\phi_\nu d\mu_\alpha.$$

We are assuming here sufficient conditions on $d\mu_\alpha$ so that the Gram matrix of these polynomials has all finite entries and is of full rank. One must exclude certain cases, for example in two dimensions $d\mu_\alpha$ can not have all its mass restricted to a line. If the polynomials are lined up in say, lexicographic order, then the ϕ_ν may be chosen via the QR decomposition to have (say) their total degree non-decreasing with ν.

Now, let P_π be the projection operator in $\mathcal{H}^{(\alpha)}$ defined by

$$P_\pi f=\sum_{\nu=2}^{M}\phi_\nu\int f(x)\phi_\nu(x)d\mu_\alpha(x).$$

We note that if μ_α is a discrete measure, then P_π is well defined for any continuous function. In our examples we will use μ_α discrete, and to avoid technical details we will continue our development here under this assumption. However, we believe that the argument below can be carried out more generally. Now, any continuous function on E^k , and in particular any $f \in \mathcal{X}_m^k$ will have the unique decomposition

$$f = P_c f + P_\pi f + P_s f$$

where $P_c f = \int f(x) d\mu_\alpha(x) = \int f(x)\phi_1(x) d\mu_\alpha(x)$ and $P_s f = (I - P_c - P_\pi)f$. More importantly, it can be shown that this is an orthogonal decomposition of $\mathcal{X}_m^k$ endowed with the squared norm

$$\|f\|^2 = (P_c f)^2 + \sum_{\nu=2}^{M} \left(\int f(x)\phi_\nu(x) d\mu_\alpha(x) \right)^2 + J_m^k(f).$$

See Gu and Wahba (1990, 1991a) for details.

Now, we can let

$$\mathcal{H}^{(\alpha)} = \mathcal{H}_\pi^{(\alpha)} \oplus \mathcal{H}_s^{(\alpha)},$$

where $\mathcal{H}_\pi^{(\alpha)}$ is the span of the polynomials of total degree less than m which average to zero under μ_α, and $\mathcal{H}_s^{(\alpha)} = P_s(\mathcal{H}^{(\alpha)})$. The reproducing kernels R_π and R_s for these spaces can be found in Gu and Wahba (1990, 1991a) , and have a relatively simple form: For $\mathcal{H}_\pi^{(\alpha)}$:

$$R_\pi(x, x') = \sum_{\nu=2}^{M} \phi_\nu(x)\phi_\nu(x').$$

To write the RK for $\mathcal{H}_s^{(\alpha)}$, we need a few more definitions. Let $\tilde{P}_\pi f = (P_c + P_\pi)f$, and let $E_m^k(\tau)$ be the semi-kernel (variogram) associated with the thin plate splines, given by

$$\begin{aligned} E_m^k(\tau) &= c_{mk}|\tau|^{2m-k}, \quad k \quad \text{not an even integer} \\ &= c_{mk}|\tau|^{2m-k}\log|\tau|, \quad k \quad \text{an even integer} \end{aligned}$$

where

$$\begin{aligned} c_{mk} &= (-1)^{d/2+m+1} \,/\, \{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!\} \\ c_{mk} &= \Gamma(d/2-m) \,/\, \{2^{2m}\pi^{d/2}(m-1)!\} \end{aligned}$$

Let $E(x, x') = E_m^k(|x - x'|)$, where $|x - x'|$ is the Euclidean distance between x and x' in E^k, and let $\tilde{P}_{\pi(x)}$ be $\tilde{P}_\pi$ applied to what follows considered as a function of x. Then, it is shown in Gu

and Wahba (1990), that the RK for $\mathcal{H}_s^{(\alpha)}$ is given by

$$R_s(x,x') = (I - \tilde{P}_{\pi(x)})(I - \tilde{P}_{\pi(x')})E(x,x').$$

We remark that this result in the one dimensional case goes back to deBoor and Lynch (1966), see also Wahba and Wendelberger (1980).

5 The Use of RKPACK

Generic algorithms for computing smoothing splines have been developed by Gu *et al.* (1989) and Gu and Wahba (1991b), with the smoothing parameters θ_i's and λ either being selected via the generalized cross-validation (GCV) method of Craven and Wahba (1979) or being estimated by the ML-II (or generalized maximum likelihood - GML) method under the Bayes model. Portable code is available in RKPACK (Gu, 1989). We illustrate in this section that the quantities in Theorem 3.1 can be calculated via immediate adaptation of the generic algorithms.

We first outline the relevant steps in the generic algorithm (Gu and Wahba, 1991b). Let the QR decomposition of S be $S = FR = (F_1, F_2)\begin{pmatrix} R_1 \\ 0 \end{pmatrix}$ and let $z = F_2'y$. Let Σ_β be the $n \times n$ matrix with ijth entry $R_\beta(\boldsymbol{t}(i), \boldsymbol{t}(j))$, and let $\tilde{\Sigma}_\beta = F_2'\Sigma_\beta F_2$. Let $\tilde{\Sigma} = \sum_{\beta=1}^{p} \theta_\beta \tilde{\Sigma}_\beta$. The GCV score $V(\lambda,\theta)$ and the GML score $M(\lambda,\theta)$ which are minimized to obtain λ and θ are given by

$$V(\lambda,\theta) = \frac{z'(\tilde{\Sigma} + n\lambda I)^{-2}z}{(\text{trace}(\tilde{\Sigma} + n\lambda I)^{-1})^2},$$

$$M(\lambda,\theta) = \frac{z'(\tilde{\Sigma} + n\lambda I)^{-1}z}{(\det(\tilde{\Sigma} + n\lambda I)^{-1})^{1/(n-M)}},$$

see Wahba (1990). After calculating z and the $\tilde{\Sigma}_\beta$ the GCV or GML score is minimized with respect to θ_β's and λ iteratively. In this process each iteration consists of a θ-step followed by a λ-step, where the θ-step updates θ_β's to find a better orientation of λ/θ_β's and the λ-step conducts a line search along the updated orientation. The minimizing smoothing parameters are then used in calculating the fits. The initialization takes $O(n^2)$ flops, each θ-step takes $(2/3)(p-1)n^3 + O(n^2)$ flops, and each λ-step takes $(2/3)n^3 + O(n^2)$ flops. In the λ-step (Gu *et al.*, 1989), $\tilde{\Sigma}$ is decomposed as $\tilde{\Sigma} = UTU'$, where U is orthogonal and T is tridiagonal (Householder tridiagonalization), to facilitate the fast evaluation of the GCV or GML scores at different values of λ. Recalling that

$M = \Sigma + n\lambda I$ it can be shown that $M^{-1} - M^{-1}S(S'M^{-1}S)^{-1}S'M^{-1} = F_2U(T + n\lambda I)'U'F_2'$ and $(S'M^{-1}S)^{-1}S'M^{-1} = R_1^{-1}(F_1' - (F_1'\Sigma F_2)U(T + n\lambda I)^{-1}U'F_2')$, where $\Sigma = \sum_{i=1}^{p}\theta_\beta\Sigma_\beta$. So at the converged θ_β's and λ the algorithm returns

$$\begin{aligned} c &= F_2U(T + n\lambda I)'U'F_2'y \\ d &= R_1^{-1}(F_1'y - (F_1'\Sigma F_2)U(T + n\lambda I)^{-1}U'F_2'y), \end{aligned} \tag{5.1}$$

which are used to compute d and c of (3.2) and (3.3). Now it is clear that to obtain $d_\beta(\boldsymbol{s})$ of (3.4) and $c_\beta(\boldsymbol{s})$ of (3.5) one only needs to replace y by $(\theta_\beta R_\beta(\boldsymbol{s}, \boldsymbol{t}(1)), \ldots, \theta_\beta R_\beta(\boldsymbol{s}, \boldsymbol{t}(n)))'$ in (5.1). F and U are usually stored in factored form, the applications of F', F, and $(T + n\lambda I)^{-1}$ on vectors are of linear order, and the applications of U' and U on vectors are of quadratic order, so for a single $\boldsymbol{s}$ these quantities require $O(n^2)$ flops extra calculation. For $S'M^{-1}S$ we have

$$\begin{aligned} (S'M^{-1}S)^{-1} &= R_1^{-1}[(F_1'\Sigma F_1) - (F_1'\Sigma F_2)(F_2'\Sigma F_2 + n\lambda I)^{-1}(F_2'\Sigma F_1)](R_1^{-1})' \\ &= R_1^{-1}[(F_1'\Sigma F_1) - (F_1'\Sigma F_2)U(T + n\lambda I)^{-1}U'(F_2'\Sigma F_1)](R_1^{-1})', \end{aligned}$$

which can be calculated in $O(n^2)$ flops.

Finally, we need an estimate for b. In this paper we calculate it from an estimate of σ^2 via $b = \sigma^2/n\lambda$. The recommended variance estimate associated with the GCV selection of smoothing parameters is $\hat{\sigma}^2 = n\lambda z'(\tilde{\Sigma} + n\lambda I)^{-2}z/\text{trace}(\tilde{\Sigma} + n\lambda I)^{-1}$, and the GML estimate is $\hat{\sigma}^2 = n\lambda z'(\tilde{\Sigma} + n\lambda I)^{-1}z/(n - M)$. Wahba (1983, 1990), Gu (1989), and Gu and Wahba (1991b) have more details.

6 Numerical Experiments

We illustrate some applications of the component-wise Bayesian "confidence intervals" using examples in this section.

We first describe an experiment with $\boldsymbol{t} \in [0, 1]^3$ generated according to a pseudo-random uniform density and the test function $f(\boldsymbol{t}) = C + f_2(t_2) + f_3(t_3) + f_{1,2}(t_1, t_2)$. The decomposition is defined such that

$$\int f_2(t_2)d\mu_2(t_2) = \int f_3(t_3)d\mu_3(t_3) = \int f_{1,2}(t_1, t_2)d\mu_1(t_1) = \int f_{1,2}(t_1, t_2)d\mu_2(t_2) = 0.$$

We have here taken the $d\mu_\alpha$ as the marginal empirical design distribution, that is, if the design points are $\boldsymbol{t}(1), \cdots, \boldsymbol{t}(n)$, where $\boldsymbol{t}(i) = (t_1(i), \cdots, t_d(i))$, then μ_α has a mass of $1/n$ at $t_\alpha(i)$, $i = 1, \cdots, n$.

We first generated $\tilde{f}(\boldsymbol{t}) = 5\cos(2\pi(t_1 - t_2)) + \exp(3t_2) + (10^6 t_3^{11}(1-t_3)^6 + 10^4 t_3^4(1-t_3)^{10})$ on the randomly generated design points $\boldsymbol{t}(i)$, then subtracted the t_1 main effect of $\tilde{f}$ to obtain f, which then has a nil t_1 main effect. Here $d(\alpha) = 1$ for $\alpha = 1, 2, 3$, and $m(1) = m(2) = m(3) = 2$. We set $f_{1,3} = f_{2,3} = 0$ by setting the associated θ's to 0 in the model, and we estimate f_1, f_2, f_3, and $f_{1,2}$. The smoothing parameters for these terms were selected by minimizing the GCV function, as described in Section 5.

Thus, the model fitted was

$$f(\boldsymbol{t}) = C + f_1(t_1) + f_2(t_2) + f_3(t_3) + f_{1,2}(t_1, t_2). \tag{6.1}$$

y_i's were generated according to (1.1). The estimates of the five components in (6.1) are obtained by projecting f_λ (given by (3.1)), onto $\{1\}$, $\mathcal{H}^{(\alpha)}, \alpha = 1, 2, 3$ and $\mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)}$ respectively. Six smoothing parameters were estimated, one each for the 3 smooth components of f_1, f_2, and f_3, and three for the three penalized components of the form (2.3), (2.4), and (2.5) of $f_{1,2}$. Letting $\hat{g}(\boldsymbol{t})$ stand for any one of the four (nonconstant) estimated components the 95% Bayesian "confidence interval" at $\boldsymbol{t}$ is then given by $\hat{g}(\boldsymbol{t}) \pm 1.96 s_g(\boldsymbol{t})$, where $s_g^2(\boldsymbol{t})$ is the posterior variance for $\hat{g}(\boldsymbol{t})$ obtained from Theorem 3.1 by collecting the relevant terms, including cross-terms, from the penalized and unpenalized components. Since in this example $\mathcal{H}^{(\alpha)}, \alpha = 1, 2, 3$ and $\mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)}$ each have a one-dimensional unpenalized subspace, there will be $\binom{2}{1} + \binom{2}{2} = 3$ terms to be collected for each of the main effects and $\binom{4}{1} + \binom{4}{2} = 10$ terms for the interaction term. However, the three penalized components in the interaction component, which had been kept separate for smoothing parameter selection, can be lumped together for the purpose of computing the posterior variance, so that there will again be three terms to collect for the posterior variance.

Six experiments were run, with two levels of n (100, 200), crossed with three levels of σ (1, 3, 10). 100 replicates were generated for each experiment, and data for the 50%, 75%, 90% and 95% confidence intervals were collected. In each case, the number of data points at which the confidence interval covered the true value of f, f_1, f_2, f_3, and $f_{1,2}$ was recorded. The averages, over all 100 replicates, are given in Table 6.1 for $n = 200$, and in Tables 6.2 and Table 6.3 for $n = 100$. Considering the $n = 200$ case, with the exception of f_1, which we will discuss separately, the $\sigma = 1$ and $\sigma = 3$ percentages are extremely close to their nominal values. The $\sigma = 10$ case is somewhat less close. We note that the test function f ranges between -8.64 and 30.56 on the data points, and so in this example σ is about 1/4 of the range of f. For $n = 100$, with $\sigma = 1$, six of the

Table 6.1: Average coverage percentage of a simulation study with $n = 200$ and 100 replicates.

Nominal Coverage	Average Coverage Percentage				
	f	f_1	f_2	f_3	$f_{1,2}$
			$\sigma = 1$		
95%	94.85	83.70	93.24	96.14	95.20
90%	89.83	73.41	88.84	91.39	90.40
75%	75.09	56.17	74.66	76.49	75.79
50%	50.29	36.00	51.31	51.24	50.71
			$\sigma = 3$		
95%	94.14	82.44	90.32	95.92	94.69
90%	89.06	71.60	85.29	90.97	89.75
75%	74.50	53.42	72.05	75.53	75.51
50%	49.50	36.10	49.88	49.76	50.26
			$\sigma = 10$		
95%	87.36	73.00	82.20	91.49	77.63
90%	80.86	66.06	77.34	85.59	70.68
75%	64.76	51.94	63.80	69.81	55.00
50%	41.76	39.96	41.65	46.28	34.96

Table 6.2: Average coverage percentage of a simulation study with $n = 100$ and 100 replicates: Smoothing replicates

Nominal Coverage	Average Coverage Percentage				
	f	f_1	f_2	f_3	$f_{1,2}$
			$\sigma = 1$ (94 replicates)		
95%	93.03	89.17	91.15	96.40	93.56
90%	87.06	80.67	85.94	91.37	88.49
75%	72.18	65.94	71.62	77.59	73.49
50%	47.32	37.69	48.95	52.11	49.55
			$\sigma = 3$ (98 replicates)		
95%	91.00	86.46	82.74	94.61	87.55
90%	84.85	80.18	75.16	89.78	81.18
75%	68.98	62.40	58.64	76.73	64.91
50%	44.44	42.68	37.13	50.51	41.54
			$\sigma = 10$ (100 replicates)		
95%	86.82	91.85	57.22	90.53	66.39
90%	79.99	87.77	51.18	84.70	60.95
75%	63.75	72.55	39.57	70.29	47.98
50%	41.41	45.21	24.78	47.48	30.57

Table 6.3: Average coverage percentage of a simulation study with $n = 100$ and 100 replicates: Interpolating replicates.

Nominal Coverage	Average Coverage Percentage				
	f	f_1	f_2	f_3	$f_{1,2}$
	$\sigma = 1$ (6 replicates)				
95%	0.00	83.33	75.33	65.33	72.17
90%	0.00	50.00	65.50	57.33	64.83
75%	0.00	50.00	44.00	43.50	47.67
50%	0.00	16.67	27.33	27.17	27.67
	$\sigma = 3$ (2 replicates)				
95%	0.00	100.0	47.50	81.00	62.50
90%	0.00	50.00	44.50	69.00	51.00
75%	0.00	50.00	32.50	44.50	39.50
50%	0.00	50.00	18.50	23.50	27.50

100 replicates were near-interpolants, and for $\sigma = 3$, two of the replicates were near-interpolants. These eight cases are listed separately in Table 6.3. Once they are removed, the the remainder of the average of the average coverages are given in Table 6.2 and are about the same as those in Table 6.1. It has been reported previously, see Wahba (1983, 1990), that, for small sample sizes, there is a small but non-zero probability that the GCV estimate of λ will be much too small. These cases can usually be spotted in practice as they are characterized by an estimate of σ^2 that is too small by several orders of magnitude. The problem goes away as the sample size becomes larger. See, for example, the hypotheses of Theorem 1.1 of Nychka (1990).

Figures 6.1 and 6.2 describe a sample of two extremes of the 100 replicates of this experiment, with $n = 200$ and $\sigma = 3$. Four examples were first chosen for inspection by counting the number of 95% confidence intervals (out of 200) for f that failed to cover the true value. Then we selected the replicates with the 5th, 25th, 75th, and 95th largest number of points out. They had, respectively, 23, 15, 7 and 3 points out. (Recall that the nominal number in this case is 10). Ties were broken by taking the earliest replicate among the tied values. The imagery from the 5th and 25th largest cases and the 75th and 95th largest cases was essentially the same, so that only the 5th and 95th largest cases are shown. We believe they "bracket" the behavior of the confidence intervals in this population of 100 replicates. Figure 6.1 plots the main effects and the interaction $f_{1,2}$ in the 5th largest case. The solid curves in (a)-(c) are the true main effects components f_1, f_2, and f_3 and

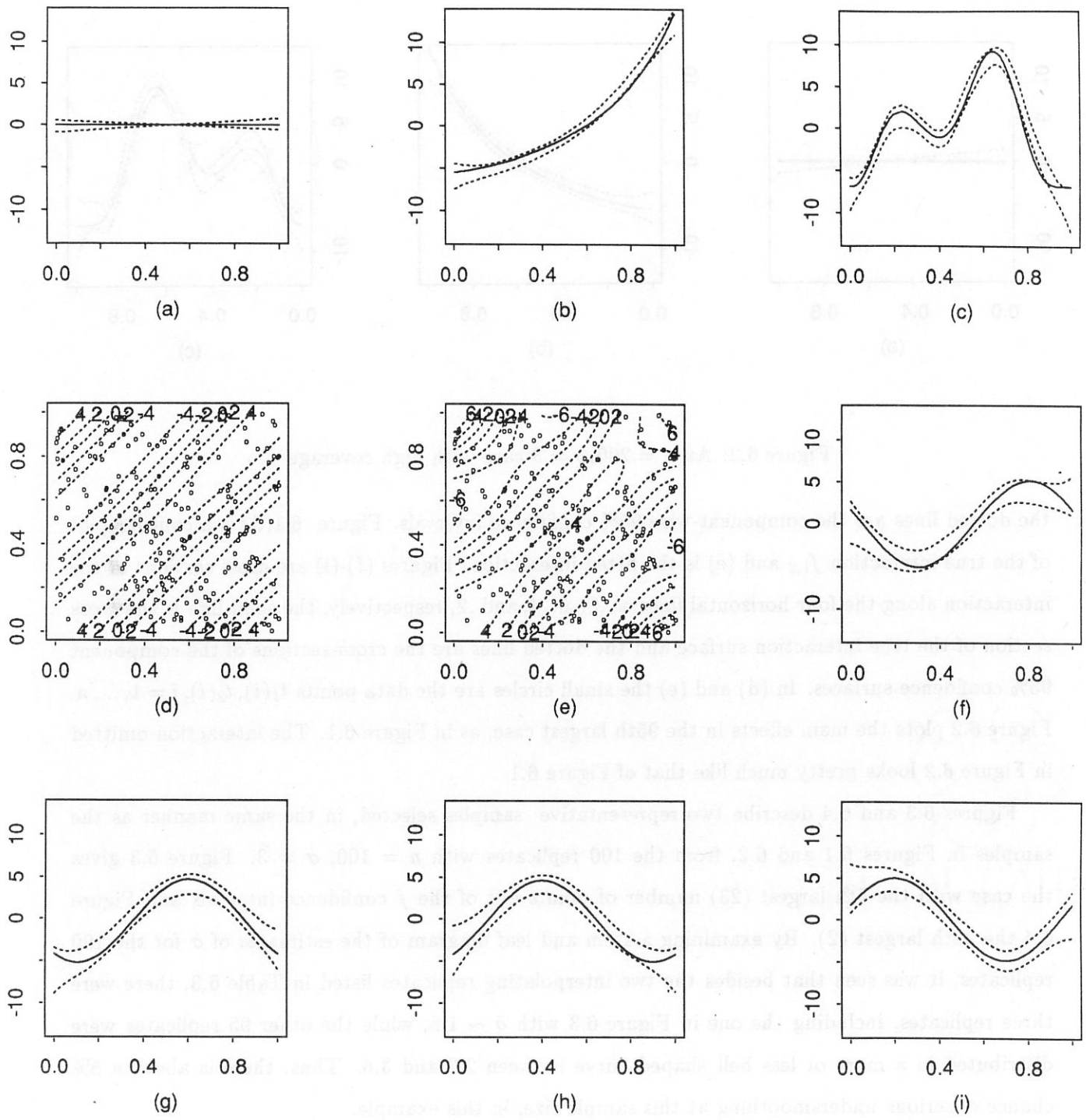

Figure 6.1: An $n = 200, \sigma = 3$ case with low coverage.

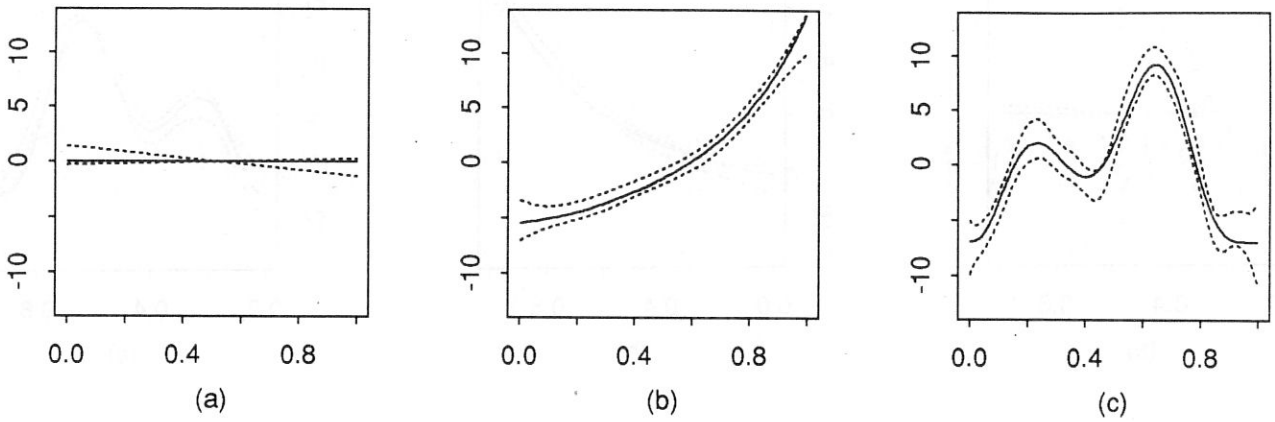

Figure 6.2: An $n = 200, \sigma = 3$ case with high coverage

the dotted lines are the component-wise 95% confidence intervals. Figure 6.1(d) is a contour plot of the true interaction $f_{1,2}$ and (e) is the fitted interaction. Figures (f)-(i) are cross sections of the interaction along the four horizontal lines at .8, .6, .4, and .2, respectively, the solid line is the cross section of the true interaction surface and the dotted lines are the cross-sections of the component 95% confidence surfaces. In (d) and (e) the small circles are the data points $t_1(i), t_2(i), i = 1, ..., n$. Figure 6.2 plots the main effects in the 95th largest case, as in Figure 6.1. The interaction omitted in Figure 6.2 looks pretty much like that of Figure 6.1.

Figures 6.3 and 6.4 describe two representative samples selected, in the same manner as the samples in Figures 6.1 and 6.2, from the 100 replicates with $n = 100$, $\sigma = 3$. Figure 6.3 gives the case with the 5th largest (23) number of points out of the f confidence intervals and Figure 6.4 the 95th largest (2). By examining a stem and leaf diagram of the estimates of $\hat{\sigma}$ for the 100 replicates, it was seen that besides the two interpolating replicates listed in Table 6.3, there were three replicates, including the one in Figure 6.3 with $\hat{\sigma} \sim 1.8$, while the other 95 replicates were distributed in a more or less bell shaped curve between 2.3 and 3.6. Thus, there is about a 5% chance of serious undersmoothing at this sample size, in this example.

We remark here on the distinctive appearance of the f_1 confidence intervals in Figures 6.1, 6.2

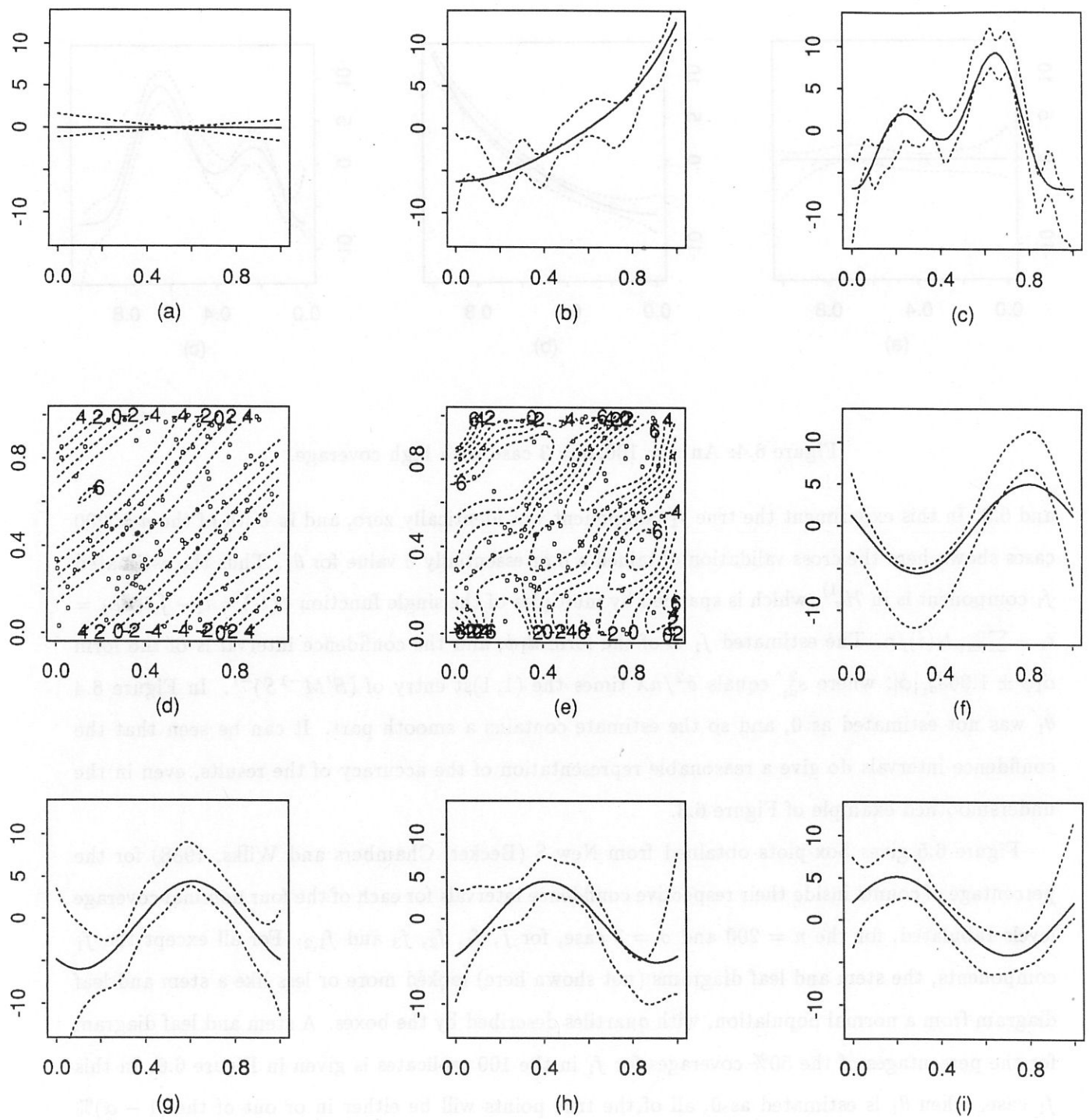

Figure 6.3: An $n = 100, \sigma = 3$ case with low coverage.

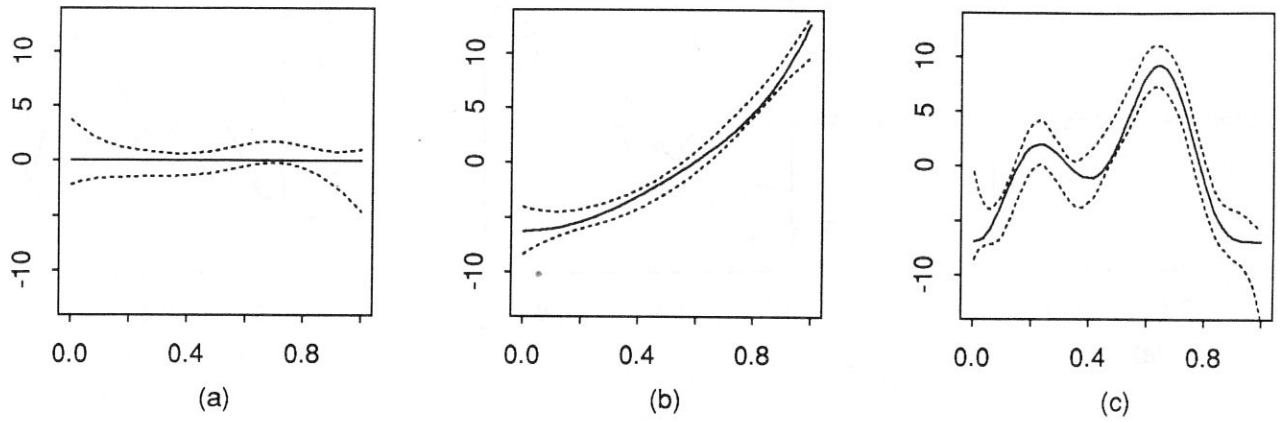

Figure 6.4: An $n = 100, \sigma = 3$ case with high coverage.

and 6.3. In this experiment the true f_1 component was identically zero, and in both of the $n = 200$ cases shown here the cross validation obtained a 0 or essentially 0 value for θ_1. Thus the estimated f_1 component is in $\mathcal{H}_\pi^{(1)}$, which is spanned by multiples of the single function $\phi(t_1) = t_1 - \int t_1 d\mu_1 = t_1 - \sum_{i=1}^n t_1(i)/n$. The estimated f_1 is of the form $d_1\phi$, and the confidence interval is of the form $d_1\phi \pm 1.96 s_{d_1}|\phi|$, where $s_{d_1}^2$ equals $\hat{\sigma}^2/n\hat{\lambda}$ times the $(1,1)$st entry of $(S'M^{-1}S)^{-1}$. In Figure 6.4 θ_1 was not estimated as 0, and so the estimate contains a smooth part. It can be seen that the confidence intervals do give a reasonable representation of the accuracy of the results, even in the undersmoothed example of Figure 6.3.

Figure 6.5 gives box-plots obtained from New S (Becker, Chambers and Wilks, 1988) for the percentage of counts inside their respective confidence intervals for each of the four nominal coverage levels tabulated, for the $n = 200$ and $\sigma = 3$ case, for f, f_1, f_2, f_3 and $f_{1,2}$. For all except the f_1 components, the stem and leaf diagrams (not shown here) looked more or less like a stem and leaf diagram from a normal population, with quartiles described by the boxes. A stem and leaf diagram for the percentages of the 50% coverages for f_1 in the 100 replicates is given in Figure 6.6. In this f_1 case, when θ_1 is estimated as 0, all of the true points will be either in or out of the $(1-\alpha)\%$ confidence intervals together, according as the interval $d_1 \pm z_{\alpha/2} s_{d_1}$ covers 0 or not. The cases where

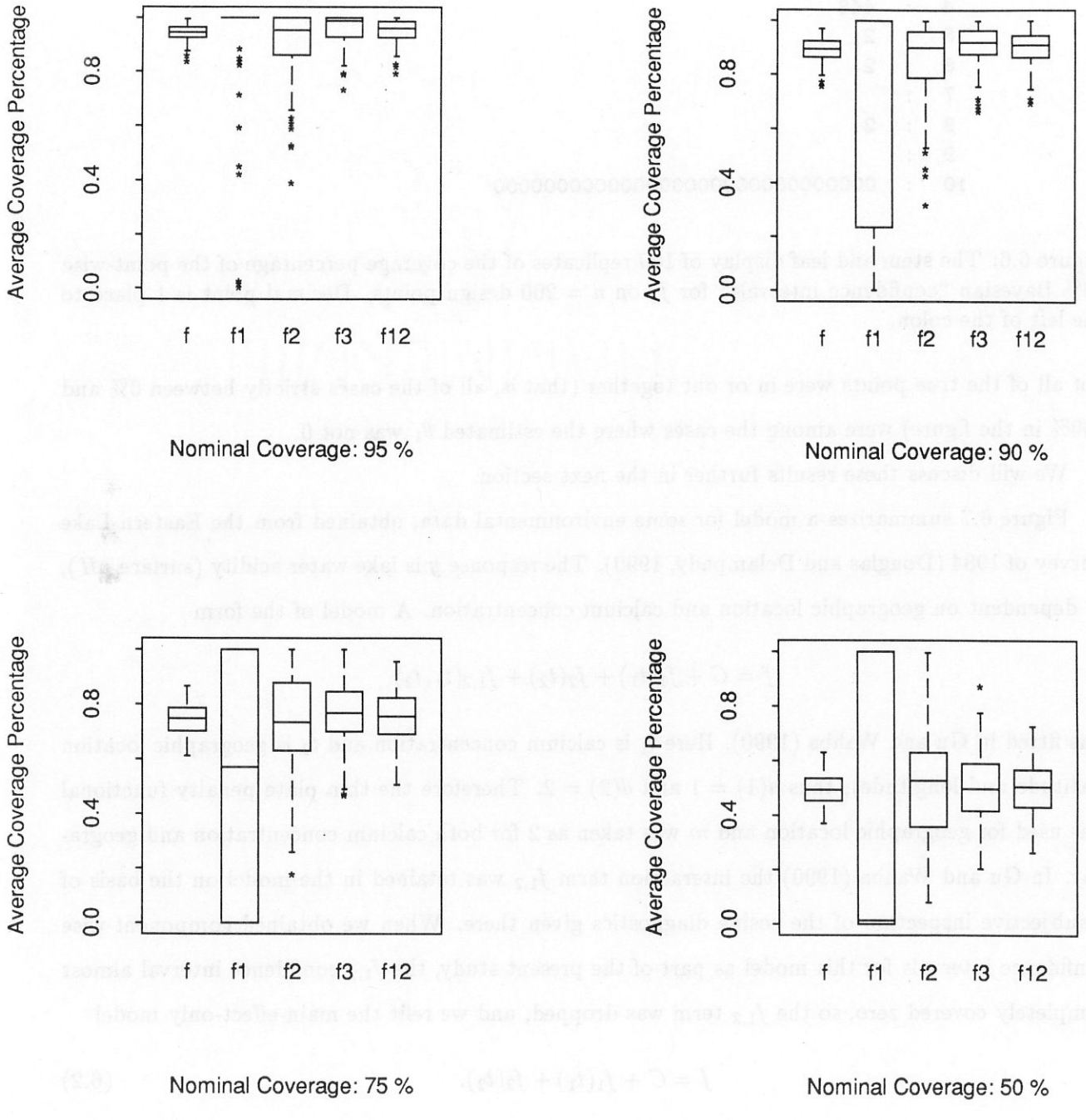

Figure 6.5: Summary of 100 replicates of the coverage percentage of the Bayesian "confidence intervals" on $n = 200$ design points.

```
 0  :  00000000000000000000000000000000000000000000001126
 1  :  2369
 2  :  445
 3  :  0028
 4  :  448
 5  :  2
 6  :  2
 7  :
 8  :  2
 9  :
10  :  00000000000000000000000000000
```

Figure 6.6: The stem and leaf display of 100 replicates of the coverage percentage of the point-wise 50% Bayesian "confidence intervals" for f_1 on $n = 200$ design points. Decimal point is 1 place to the left of the colon.

not all of the true points were in or out together (that is, all of the cases strictly between 0% and 100% in the figure) were among the cases where the estimated θ_1 was not 0.

We will discuss these results further in the next section.

Figure 6.7 summarizes a model for some environmental data, obtained from the Eastern Lake Survey of 1984 (Douglas and Delampady, 1990). The response y is lake water acidity (surface pH), as dependent on geographic location and calcium concentration. A model of the form

$$f = C + f_1(t_1) + f_2(t_2) + f_{1,2}(t_1, t_2)$$

was fitted in Gu and Wahba (1990). Here t_1 is calcium concentration and t_2 is geographic location (latitude and longitude), thus $d(1) = 1$ and $d(2) = 2$. Therefore the thin plate penalty functional was used for geographic location and m was taken as 2 for both calcium concentration and geography. In Gu and Wahba (1990) the interaction term $f_{1,2}$ was retained in the model on the basis of a subjective inspection of the cosine diagnostics given there. When we obtained component-wise confidence intervals for this model as part of the present study, the $f_{1,2}$ confidence interval almost completely covered zero, so the $f_{1,2}$ term was dropped, and we refit the main-effect-only model

$$f = C + f_1(t_1) + f_2(t_2). \tag{6.2}$$

From the right frame of Figure 6.7 it can be seen that the lakes (112 in total) run roughly along the Blue Ridge mountains which run SW to NE in GA, TN, SC, NC and VA. The contours represent the

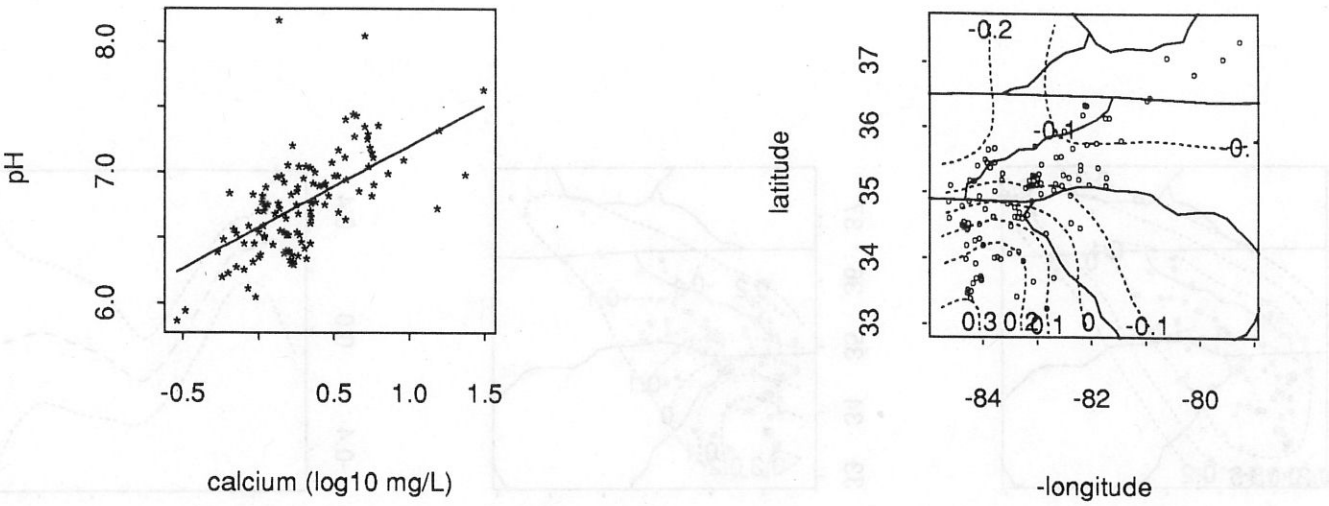

Figure 6.7: Main-effect-only model of Blue Ridge lake acidity.

fitted main effects for geography. Although there is no data in the NW and SE corners, an estimate of f_2 is available over all of E^2. However, it is clear that as one gets far enough away from the data this estimate carries no real information. Our first task here then is to obtain a reasonable graphical display of what is hopefully the meaningful part of f_2. To this end, we obtained a contour plot of the (estimated) posterior standard deviation of f_2, which is given in the (1,1)st frame of Figure 6.8. We arbitrarily selected the posterior standard deviation contour of .15 as our cutoff, based on the observation that it was approximately 3 times the minimum posterior standard deviation. In the (1,2)nd frame of Figure 6.8 we display the .15 posterior standard deviation contour, and a contour plot of f_2 in the region enclosed by this contour (that is, the region with a smaller posterior standard deviation). Visually, the results appear much more sensible than in Figure 6.7. The (1,3)rd frame of Figure 6.8 presents a cross sections of the 95% confidence interval taken along the diagonal from the lower left to the upper right corner. The minimum value of the estimated geographic component of the lake acidity occurs roughly where this diagonal intersects the 82 degrees longitude line - this is roughly the location of Mt. Mitchell, the highest point in NC, at the high point of the crest

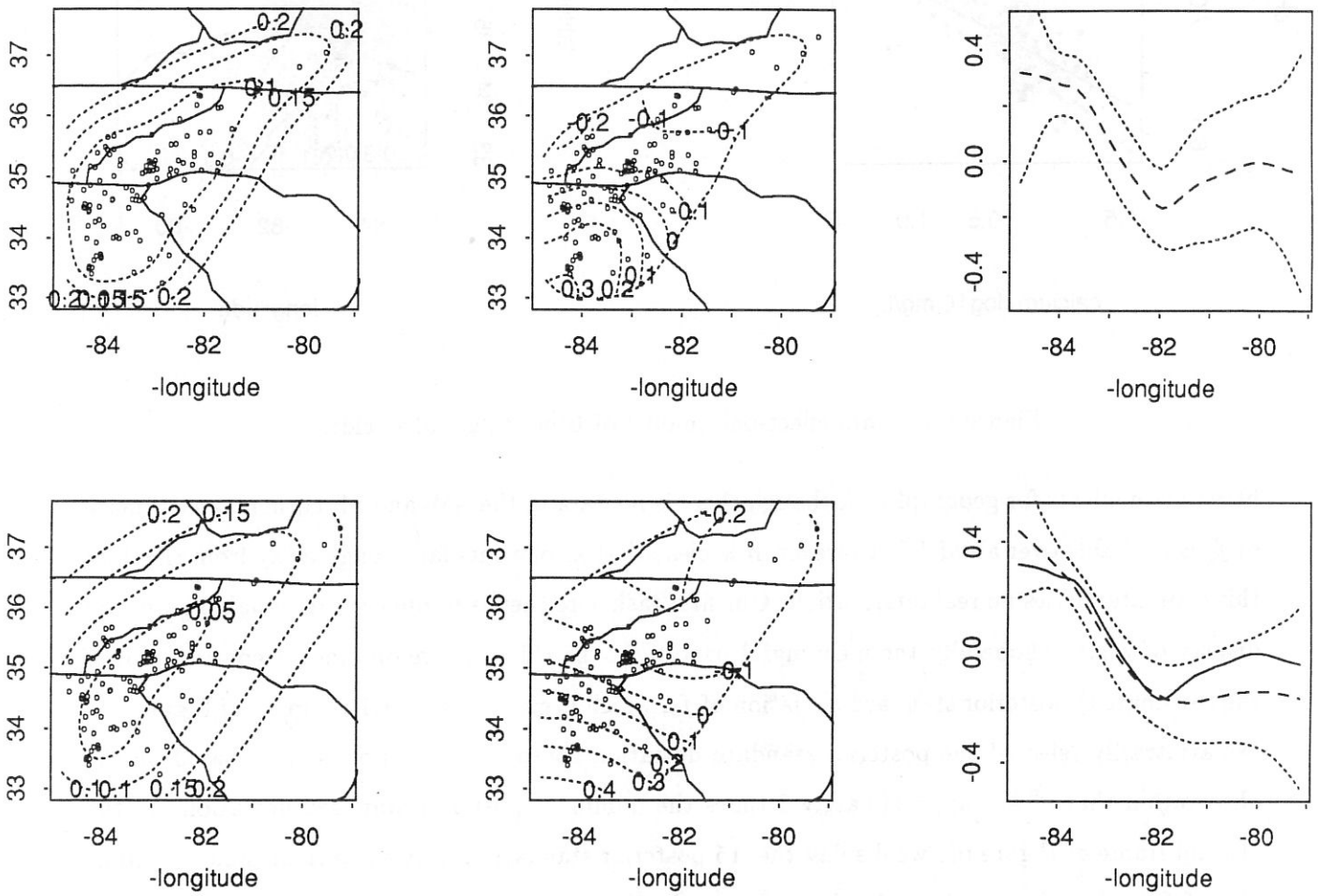

Figure 6.8: Geography effect of lake acidity model.

Table 6.4: Average coverage percentage of a simulation study of the Blue Ridge lake acidity model with 100 replicates.

Nominal Coverage	Average Coverage Percentage					
	f	f_{calc}	f_{geog}	f	f_{calc}	f_{geog}
	94 smoothing replicates			6 interpolating replicates		
95%	88.44	85.11	90.11	0.00	66.67	23.96
90%	83.60	81.79	85.40	0.00	50.15	18.90
75%	70.45	66.00	71.19	0.00	33.33	14.58
50%	48.08	40.42	46.64	0.00	0.00	8.48

of the Blue Ridge mountains. To see whether these intervals were reasonable, we simulated data from the model (6.2), using the model that we have just fitted as the truth, and generating *i.i.d.* normally distributed ϵ_i's with variance equal to the variance which had been estimated for the real data. The second row of Figure 6.8 shows the simulation results of a single replicate in parallel to the first row, with the exception that the true function is added in the (2,3)rd frame as the solid line. It can be seen that the CI's give a reasonable visual image of the accuracy of the estimate.

We then simulated 100 replicates from the model (6.2), with the same $n = 112$ data points, and, for f, $f_1 = f_{\text{calc}}$ and $f_2 = f_{\text{geog}}$ counted the percent of true data points inside their confidence intervals. In this experiment, there were 6 out of 100 interpolating cases. The percent of data points inside their respective confidence intervals is given in Table 6.4.

7 When and Why Should the CI's Work?

To discuss the question of this section we will consider the case of a purely parametric component and the case of a not purely parametric component separately. Note that in our synthetic example the true f_1 was purely parametric, in fact 0, and of the four plots in Figures 6.1(a) – 6.4(a), f_1 was fitted in its one-dimensional parametric subspace in three of them. The subspaces for f_2, f_3 and $f_{1,2}$ all contain a one dimensional unpenalized subspace. For ease of exposition consider first the case of a single variable and now let

$$f = f_0 + f_1$$

where $f_0 \in \mathcal{H}^0$, the M-dimensional subspace spanned by $\phi_1, \cdots, \phi_M$ and $f_1 \in \mathcal{H}^1$, a single penalized subspace, with one smoothing parameter. First consider the Bayesian "confidence interval" for the

purely parametric component f_0. Let $f_0 = \sum_{\nu=1}^{M} d_{0,\nu}\phi_\nu$, and let $\tilde{f}_1 = (f_1(\boldsymbol{t}(1)), \cdots, f_1(\boldsymbol{t}(n)))'$. Then $y = Sd_0 + \tilde{f}_1 + \epsilon$, where $d_0 = (d_{0,1}, \cdots, d_{0,M})'$. Letting d be the estimate of d_0, we have

$$d - d_0 = (S'M^{-1}S)^{-1}S'M^{-1}(Sd_0 + \tilde{f}_1 + \epsilon) - d_0.$$

Thus

$$d - d_0 \sim \mathcal{N}((S'M^{-1}S)^{-1}S'M^{-1}\tilde{f}_1, \sigma^2(S'M^{-1}S)^{-1}S'M^{-2}S(S'M^{-1}S)^{-1}) \tag{7.1}$$

whereas the Bayesian "confidence interval" for $f_0(\boldsymbol{t})$ is treating $d - d_0$ as though

$$d - d_0 \sim \mathcal{N}(0, \sigma^2(S'M^{-1}S)^{-1}/n\lambda). \tag{7.2}$$

Note that as $n\lambda \to \infty$ we have that the right hand sides of (7.1) and (7.2) tend to $\mathcal{N}((S'S)^{-1}S'\tilde{f}_1, \sigma^2(S'S)^{-1})$ and $\mathcal{N}(0, \sigma^2(S'S)^{-1})$ respectively, and if the orthogonality of $\mathcal{H}_\pi$ and $\mathcal{H}_s$ has been defined via P_π with $d\mu_1$ as the design measure then $S'\tilde{f}_1 = 0$, so that asymptotically in this case the Bayesian "confidence intervals" are doing the "right thing". Of course it is only in the $M = 1$ case that 100% of the points will all be in or out of their confidence intervals according as d_0 is in or out of its confidence interval. Also, it is only in the single variable case, with inner product as defined here, that we can always arrange to have $S'\tilde{f}_1 = 0$. Shiau (1985) and Nychka (1988) have noted that if the square bias is not large relative to the variance, then treating a random variable as $\mathcal{N}(0, (bias)^2 + variance)$ is not a bad approximation to $\mathcal{N}(bias, variance)$ for confidence intervals purposes. Here we have

$$E(d - d_0)(d - d_0)' = \frac{\sigma^2}{n\lambda}(S'M^{-1}S)^{-1}B(S'M^{-1}S)^{-1}$$

where

$$B = S'M^{-1}(\tilde{f}_1\tilde{f}_1{}'/b + n\lambda I)M^{-1}S$$

and in this case to the extent that $B \sim (S'M^{-1}S)$, we will be making the approximation that Shiau and Nychka have found is reasonable. Shiau (1985) did an extensive series of simulations in the one-dimensional case with $M = 1$, testing the validity of the Bayesian "confidence intervals" for the estimate d of d_0, in a series of examples. In 8 of 10 examples, the observed coverages were very close to the rated coverages, and in two they were not. She found that the 8 "successful" cases had relatively small (less than 1) square bias-variance ratios, while the unsuccessful cases had large ratios. Referring to Figure 6.6 it can be seen that there were 30 cases of all points in their CI, and

49 cases of all points out of their CI. If our approximations were good, then the probability of all-in would be .5 whereas we have observed $30/79 = .379$, which is 2.72 standard deviations below the nominal value of .5.

We now consider the Bayesian "confidence intervals" for f where f is not purely parametric. We briefly review the known results for these CI's for f in the single variable case, based mainly on Nychka (1988, 1990), see also Wahba (1983) and Hall and Titterington (1987). Below it is assumed that $f \in \mathcal{H}$ but not in $\mathcal{H}^0$. Then we will outline what would have to be proved to extend the known theoretical results to the component cases.

Let $A(\lambda)$ be the $n \times n$ influence matrix defined by

$$\begin{pmatrix} f_\lambda(t(1)) \\ \vdots \\ f_\lambda(t(n)) \end{pmatrix} = A(\lambda)y$$

and let $\tilde{f} = (f(t(1)), \cdots, f(t(n)))'$. Let

$$\begin{aligned} b(\lambda) &= (I - A(\lambda))\tilde{f} \\ v(\lambda) &= -A(\lambda)\epsilon, \end{aligned}$$

the vectors of the bias and variance components of the error $f - f_\lambda$ at the data points, respectively. Let the predictive mean square error $T(\lambda)$ be defined by

$$T(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(f(t(i)) - f_\lambda(t(i)))^2 = \frac{1}{n}\|b(\lambda) + \epsilon(\lambda)\|^2.$$

Since $f \notin \mathcal{H}^0, \|b(\lambda)\|^2 \neq 0$ in general. To understand the import of what follows, the reader needs to know that the posterior covariance matrix at the data points, with ijth entry $Q_\lambda(t(i), t(j))$ reduces to $\sigma^2 A(\lambda)$, see Wahba (1983). The next fact that is necessary, is that (it can be shown that) if λ^* minimizes $ET(\lambda)$ and $\hat{\lambda}$ is the GCV estimate of λ, then $\hat{\lambda}$ and λ^* are close enough so that they can be interchanged in what follows. We also note that if the constant function is not penalized, then $A(\lambda)(1, \cdots, 1)' = (1, \cdots, 1)'$ so that $\sum_{i=1}^{n} b_i(\lambda)/n = 0$.

Then, under certain assumptions (see Nychka (1990), Theorem 1.1 for a precise statement)

$$\lim_{n\to\infty} \frac{(\sigma^2/n)\text{trace}A(\lambda^*)}{ET(\lambda^*)} \to \kappa_1 \sim 1 \tag{7.3}$$

$$\lim_{n\to\infty} \frac{\|(b(\lambda*)\|^2/n}{\|v(\lambda^*)\|^2/n} \sim \kappa_2 < 1 \tag{7.4}$$

for large n.

Now, let $a_{ii}(\hat{\lambda})$ be the iith entry of $A(\hat{\lambda})$, and let $C(z_{\alpha/2}, \boldsymbol{t}(i))$ be the interval $f_{\hat{\lambda}}(\boldsymbol{t}(i)) \pm z_{\alpha/2}\hat{\sigma}\sqrt{a_{ii}(\hat{\lambda})}$, which is the confidence interval at $\boldsymbol{t}(i)$. Define the average coverage (AC) as

$$AC = \frac{1}{n}\sum_{i=1}^{n} I\{f(\boldsymbol{t}(i)) \in C(z_{\alpha/2}, \boldsymbol{t}(i))\} = \frac{1}{n}\sum I\{|b_i(\hat{\lambda}) + v_i(\hat{\lambda})|/\hat{\sigma}(\sqrt{a_{ii}(\hat{\lambda})}) \leq z_{\alpha/2}\}$$

where I is the indicator function of the event in brackets. Letting $\mathcal{U}$ be the random variable which takes on the value $b_i(\hat{\lambda}) + v_i(\hat{\lambda})$ with probability $1/n$, $i = 1, \cdots, n$, Nychka argues that if the b_i are not dominated by a decreasing number of increasingly large values, and (7.4) is true, then $\mathcal{U}$ behaves like the convolution of a zero mean discrete distribution and a Normal distribution and approximately

$$\mathcal{U} \sim \mathcal{N}(0, \frac{1}{n}E\|b(\hat{\lambda}) + v(\hat{\lambda})\|^2) \tag{7.5}$$

and so

$$EI\{\|\mathcal{U}\|/\sqrt{\frac{1}{n}E\|b(\hat{\lambda}) + v(\hat{\lambda})\|^2} \leq z_{\alpha/2}\} \sim 1 - \alpha. \tag{7.6}$$

Furthermore, if $a_{ii}(\hat{\lambda}) \sim \text{trace}A(\hat{\lambda})/n$ and (7.3) is true then it will follow from (7.6) and (7.3) that $E(AC) \sim (1 - \alpha)$.

We now suggest some of the steps that would be required to extend the known theory concerning the properties of these CI's to the component-wise case. First, write the representation of $Q_\lambda(\boldsymbol{s}, \boldsymbol{t})$ as given in Wahba (1983), which is

$$Q_\lambda(\boldsymbol{s}, \boldsymbol{t}) = Q_0(\boldsymbol{s}, \boldsymbol{t}) + \sigma^2\delta'(\boldsymbol{s})A(\lambda)\delta(\boldsymbol{t}) \tag{7.7}$$

where Q_0 is obtained from Q_λ by setting $\lambda = 0$, and $\delta(\boldsymbol{s}) = (\delta_1(\boldsymbol{s}), \cdots, \delta_n(\boldsymbol{s}))'$ is given by

$$\delta'(\boldsymbol{s}) = (\phi_1(\boldsymbol{s}), \cdots, \phi_M(\boldsymbol{s}))(S'\Sigma^{-1}S)^{-1}S'\Sigma^{-1} + (R(\boldsymbol{s}, \boldsymbol{t}(1)), \cdots, R(\boldsymbol{s}, \boldsymbol{t}(n)))(\Sigma^{-1} - \Sigma^{-1}S(S'\Sigma^{-1}S)^{-1}S'\Sigma^{-1})$$

Here the kth component δ_k of δ is that element in $\mathcal{H}$ minimizing the penalty functional (term in brackets in (1.3)) and interpolating to data which is 1 at $\boldsymbol{t}(k)$ and 0 at $\boldsymbol{t}(i)$ for $i \neq k$. Note that although $Q_0(\boldsymbol{s}, \boldsymbol{t}(i)) = 0$, all $\boldsymbol{s}$, $P^\beta_{(\boldsymbol{s})}P^\gamma_{(\boldsymbol{t})}Q_0(\boldsymbol{s}, \boldsymbol{t})_{\boldsymbol{t}=\boldsymbol{t}(i)}$ is not necessarily 0. Here the subscript $(\boldsymbol{s})$ means the operator P^β applied to what follows considered as a function of $\boldsymbol{s}$. However, we will assume that Q_0 is negligible in what follows. Q_0 is concerned with the unobservables and we will discuss it later. In what follows we are suppressing the θ's.

Now, we can write

$$f_\lambda(\boldsymbol{s}) = \delta'(\boldsymbol{s})A(\lambda)y = \delta'(\boldsymbol{s})A(\lambda)(\tilde{f} + \epsilon)$$

to obtain

$$P^\beta f_\lambda(\boldsymbol{s}) = P^\beta \delta'(\boldsymbol{s})A(\lambda)(\tilde{f}_0 + \tilde{f}_1 + \cdots + \tilde{f}_p + \epsilon)$$

where, in an obvious notation we have written out the components $\tilde{f} = \tilde{f}_0 + \tilde{f}_1 + \cdots + \tilde{f}_p$. Let G_β be the $n \times n$ matrix with ijth entry $(P^\beta \delta_j)(\boldsymbol{t}(i))$. Note that $\sum_\beta G_\beta = I$, where the projection onto $\mathcal{H}^0$ is included in the sum.

Letting the vector of biases of $P^\beta f_\lambda$ at the data points be defined as

$$b_\beta(\lambda) = \begin{pmatrix} P^\beta f(\boldsymbol{t}(1)) \\ \vdots \\ P^\beta f(\boldsymbol{t}(n)) \end{pmatrix} - \begin{pmatrix} P^\beta f_\lambda(\boldsymbol{t}(1)) \\ \vdots \\ P^\beta f_\lambda(\boldsymbol{t}(n)) \end{pmatrix},$$

we have the following expressions for the bias and variance of the βth component $P^\beta f_\lambda$:

$$\begin{aligned} b_\beta(\lambda) &= (I - G_\beta A(\lambda))\tilde{f}_\beta - G_\beta A(\lambda) \sum_{\gamma \neq \beta} \tilde{f}_\gamma \\ v_\beta(\lambda) &= -G_\beta A(\lambda)\epsilon \end{aligned}$$

Furthermore, the posterior covariance of $P^\beta f_\lambda$ is $\sigma^2 G_\beta A(\lambda) G'_\beta +$ (terms from Q_0). Let $T_\beta(\lambda) = \|b_\beta(\lambda) + v_\beta(\lambda)\|^2/n$. To extend the known results to the component-wise confidence intervals it is sufficient to show, for large n:

1. $P^\beta_{(\boldsymbol{s})} P^\beta_{(\boldsymbol{t})} Q_0(\boldsymbol{s}, \boldsymbol{t})_{\boldsymbol{s},\boldsymbol{t}=\boldsymbol{t}(i)}$ is negligible,

2. The distribution of $\mathcal{U}_\beta$, the random variable which takes on the value $b_{\beta,i}(\hat{\lambda}) + v_{\beta,i}(\hat{\lambda})$ with probability $1/n, i = 1, \cdots, n$ approximately satisfies

$$\mathcal{U}_\beta \sim \mathcal{N}(0, \frac{1}{n} E\|b_\beta(\hat{\lambda}) + v_\beta(\hat{\lambda})\|^2),$$

3. The diagonal entries of $G_\beta A(\hat{\lambda}) G'_\beta$ are not too far from their average value,

4. $\lim_{n\to\infty} \dfrac{(\sigma^2/n)\mathrm{trace}(G_\beta A(\lambda^*) G'_\beta)}{E T_\beta(\lambda^*)} \to \kappa_1 \sim 1,$

5. $\lim_{n\to\infty} \dfrac{\|b_\beta(\lambda*)\|^2/n}{\|v_\beta(\lambda^*)\|^2/n} \sim \kappa_2 < 1.$

Note that we now have to be concerned with "leakage" of one component entering into the bias of another component, thus, we would like $G_\beta A(\lambda)\tilde{f}_\gamma$ to be small for $\gamma \neq \beta$.

We remark on the requirement that the distribution of the components of $b_\beta(\hat{\lambda})$ not be asymptotically characterized by a decreasingly small number of increasingly large values. If this happens, then $\mathcal{U}$ may not look sufficiently Normal for $E(AC)$ to be near its rated value. If in the single variable, $d = 1$ case f has $2m$ square integrable derivatives, but has large absolute mth derivative near the extreme data points, then the bias error can be dominated by its few values at these extremes. See Nychka (1988) and references cited there for further discussion. We have not seen this phenomena to be a problem at the sample sizes that we have been using, however, it cannot be ruled out.

We note what is known about κ_1 and κ_2 in the single variable case. For any $f \in \mathcal{X}_m^d$, it can be shown that $\|b(\lambda)\|^2/n \leq \lambda J_m^d(f)$, see Wahba (1990). If further conditions are imposed on f (i.e., additional square integrable derivatives and boundary behavior), then we may have $\|b(\lambda)\|^2/n = c_p(f)\lambda^p(1 + o(1))$ as $\lambda \to 0$, for some c_p and $p \in (1, 2]$. Also, if the design points are "nicely distributed", then it is known that the eigenvalues of $A(\lambda)$ are such that $\text{trace}A(\lambda)/n \sim (c_{m,d}/n\lambda^{d/2m})(1 + o(1))$, for some $c_{m,d}$ as $n \to \infty$, $\lambda \to 0$, and $n\lambda^{d/2m} \to \infty$. (These conditions are satisfied by λ^*). In the single variable case with $d = d(1) = 1$ the argument in Wahba (1983) is suggestive that

$$\begin{aligned} \kappa_1 &= \frac{2m}{2m-1}\frac{2mp}{2mp+1} \\ \kappa_2 &= \frac{1}{2mp}. \end{aligned}$$

This has been proved by Nychka (1990) for the $p = 2$ case. Cox (1989) has a rather strong related theorem in something very similar to our $p = 2$ case, however, he has some related partly pessimistic theoretical results for the $p < 2$ case, whose practical import remains to be studied. It is a good conjecture that if $d > 1$ we can replace $2m$ in (7.8) and (7.8) by $2m/d$. See, for example Cox (1984) and Wahba (1979).

We think it is also a good conjecture that useful frequentist properties of these component-wise "confidence intervals" can be proved, in some generality, possibly subject to the caveats described above. We do note that although the components of the ANOVA models are orthogonal in function space, they are not in general orthogonal in data space. These confidence intervals are component-

wise, and do not take into account correlations between components. One has to exercise exactly the same caution that one would use in a multiple linear regression where the design matrix might be ill-conditioned. Multicollinearity here between components $P^\beta f$ and $P^\gamma f$ can be diagnosed by examining $\text{trace}(\tilde{\Sigma}_\beta \tilde{\Sigma}_\gamma)/(\text{trace}\,\tilde{\Sigma}_\beta^2 \text{trace}\,\tilde{\Sigma}_\gamma^2)^{1/2}$, $\text{trace}\,G_\beta A G_\gamma'/(\text{trace}\,G_\beta A G_\beta')^{1/2}(\text{trace}\,G_\gamma A G_\gamma')^{1/2}$ or by examining the cosine diagnostics of Gu (1990). If two components are highly collinear, the user may consider combining them into one component or deleting one of them, depending on the context.

The confidence intervals can also be "calibrated" essentially as we have done with the lake data, by running a Monte Carlo study about the estimated $f_{\hat{\lambda}}$, although this approach must be used with caution. When extremely large data sets are available, as is happening in many present day medical, environmental, and meteorological contexts, the data may be divided in half and the model fitted on the first half, and checked or calibrated on the second half.

We note that $Q_0(s,t) = E(X(s) - \hat{X}(s))(X(t) - \hat{X}(t))$, where $X = \lim_{\xi\to\infty} X_\xi$ and $\hat{X}(t) = EX(t)|\{X(t(1),...,X(t(n))\}$. Thus $X(t) - \hat{X}(t)$ is a stochastic process which is independent of the data and hence we have no information about it from the data. Analogously, if we let $\hat{f}(t)$ be the minimal semi-norm interpolant to f at the data points, that is, $\hat{f}$ minimizes the penalty, call it $J(f)$, of (1.3) subject to $\hat{f}(t(i)) = f(t(i)), i = 1,\cdots,n$, then $f - \hat{f}$ will be orthogonal to f_λ in $\mathcal{H}$, and in fact, we have no information about $f - \hat{f}$ from the data. Thus, one should exercise caution in making inferences concerning $f - \hat{f}$. It is safer to think of f_λ as an estimate of $\hat{f}$. We remark that bounds on $|f(t) - \hat{f}(t)|$ can be given in terms of $J(f)$ and $Q_0(t,t)$ via the hypercircle inequality; see Wahba (1990, p.96). We omit the details.

A Proof of Theorem 3.1

We can see how to prove the various parts of the Theorem by a unified method if we first prove the following: $Q_\lambda(s,t) = E((f_\lambda(s) - f(s))(f_\lambda(t) - f(t))|Y = y)$ is given by:

$$Q_\lambda(s,t)/b = (\phi_1(s),\cdots,\phi_M(s))(S'M^{-1}S)^{-1}\begin{pmatrix}\phi_1(t)\\ \vdots \\ \phi_M(t)\end{pmatrix}$$

$$-(\phi_1(\boldsymbol{s}),\cdots,\phi_M(\boldsymbol{s}))(S'M^{-1}S)^{-1}S'M^{-1}\begin{pmatrix} R(\boldsymbol{t},\boldsymbol{t}(1)) \\ \vdots \\ R(\boldsymbol{t},\boldsymbol{t}(n)) \end{pmatrix}$$

$$-(\phi_1(\boldsymbol{t}),\cdots,\phi_M(\boldsymbol{t}))(S'M^{-1}S)^{-1}S'M^{-1}\begin{pmatrix} R(\boldsymbol{s},\boldsymbol{t}(1)) \\ \vdots \\ R(\boldsymbol{s},\boldsymbol{t}(n)) \end{pmatrix}$$

$$+R(\boldsymbol{s},\boldsymbol{t})-(R(\boldsymbol{s},\boldsymbol{t}(1)),\cdots,R(\boldsymbol{s},\boldsymbol{t}(n)))[M^{-1}-M^{-1}S(S'M^{-1}S)^{-1}S'M^{-1}]\begin{pmatrix} R(\boldsymbol{t},\boldsymbol{t}(1)) \\ \vdots \\ R(\boldsymbol{t},\boldsymbol{t}(n)) \end{pmatrix}.$$

After we prove this, which is equivalent to Theorem 2 of Wahba (1983), we show that by a simple substitution in the proof, each of the posterior covariances of the components is obtained by the same technique.

Let $y = f + \epsilon$, where f and ϵ are 0 mean Gaussian random (column) vectors with $Eff' = b\Sigma_{ff}, E\epsilon\epsilon' = \sigma^2 I, E\epsilon f' = 0$, and let g, h be zero mean Gaussian random vectors with $Egh' = b\Sigma_{gh}, Egf' = b\Sigma_{gf}$ and $Efh' = b\Sigma_{fh}$. Let $\sigma^2/b = n\lambda$. Then we have

$$Cov(g,h|y) = b(\Sigma_{gh} - \Sigma_{gf}(\Sigma_{ff} + n\lambda I)^{-1}\Sigma_{fh}). \tag{A.1}$$

Let $f(\boldsymbol{t}) = \sum_{\nu=1}^{M}\tau_\nu\phi_\nu(\boldsymbol{s}) + bZ(\boldsymbol{t})$, where $\tau = (\tau_1,\cdots,\tau_M)' \sim N(0,I), EZ(\boldsymbol{s})Z(\boldsymbol{t}) = R(\boldsymbol{s},\boldsymbol{t})$, and τ and $Z(\boldsymbol{t})$ are independent. Letting $\xi = \eta/b$, then

$$Ef(\boldsymbol{s})f(\boldsymbol{t}) = b[\eta\sum_{\nu=1}^{M}\phi_\nu(\boldsymbol{s})\phi_\nu(\boldsymbol{t}) + R(\boldsymbol{s},\boldsymbol{t})]$$

Now, let $f = (f(\boldsymbol{t}(1)),\cdots,f(\boldsymbol{t}(n)))$, $g = f(\boldsymbol{s})$ and $h = f(\boldsymbol{t})$ and let S, Σ, and M be as in the text. Let $\phi(\boldsymbol{s}) = (\phi_1(\boldsymbol{s}),\cdots,\phi_M(\boldsymbol{s}))'$, and let $R(\boldsymbol{s}) = (R(\boldsymbol{s},\boldsymbol{t}(1)),\cdots,R(\boldsymbol{s},\boldsymbol{t}(n)))'$. We have, upon substituting these into (A.1),

$$Q_\lambda(\boldsymbol{s},\boldsymbol{t})/b = \eta\phi'(\boldsymbol{s})\phi(\boldsymbol{t}) + R(\boldsymbol{s},\boldsymbol{t}) - (\eta\phi(\boldsymbol{s})'S' + R(\boldsymbol{s}))(\eta SS' + M)^{-1}(\eta S\phi(\boldsymbol{t}) + R(\boldsymbol{t})). \tag{A.2}$$

Upon collection terms the right hand side of (A.2) becomes

$$\phi'(\boldsymbol{s})[\eta I - \eta S'(\eta SS' + n\lambda I)^{-1}\eta S]\phi(\boldsymbol{t})$$
$$- \quad \eta\phi'(\boldsymbol{s})S'(\eta SS' + M)^{-1}R(\boldsymbol{t})$$

$$
\begin{aligned}
&- \quad R(\boldsymbol{s})'(\eta SS' + M)^{-1}\eta S\phi(\boldsymbol{t}) \\
&+ \quad R(\boldsymbol{s},\boldsymbol{t}) - R(\boldsymbol{s})'(\eta SS' + M)^{-1}R(\boldsymbol{t}). \qquad \text{(A.3)}
\end{aligned}
$$

Now, the following formulas are known (Wahba, 1983, Eq. (2.14), and 1978, Eqs. (2.8) and (2.7) respectively):

$$
\begin{aligned}
\lim_{\eta\to\infty} \eta I - \eta S'(\eta SS' + M)^{-1}S\eta &= (S'M^{-1}S)^{-1} \\
\lim_{\eta\to\infty} \eta S'(\eta SS' + M)^{-1} &= (S'M^{-1}S)^{-1}S'M^{-1} \\
\lim_{\eta\to\infty} (\eta SS' + M)^{-1} &= M^{-1} - M^{-1}S(S'M^{-1}S)^{-1}S'M^{-1}. \qquad \text{(A.4)}
\end{aligned}
$$

Substitution of (A.4) into (A.3) gives the result. In order to get the posterior covariances of the components of f_λ, as given in the theorem, we can now see that by letting g and h in the above proof be any of $\tau_\nu\phi_\nu(\boldsymbol{s})$, $b^{1/2}\sqrt{\theta_\beta}Z_\beta(\boldsymbol{s})$, $\tau_\mu\phi_\mu(\boldsymbol{t})$, and $b^{1/2}\sqrt{\theta_\beta}Z_\gamma(\boldsymbol{t})$, instead of $f(\boldsymbol{s})$ and $f(\boldsymbol{t})$, we will obtain the posterior covariances of the theorem. Similarly, the posterior covariance of components which are the sum of several components may be obtained by letting g and h be the relevant sums.

References

Antoniadis, A. (1984). Analysis of variance on function spaces. *Math. Operationsforsch. u. Statist.*, **15** 59 – 71.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, **68** 337 – 404.

Becker, R., Chambers, J. and Wilks, A. (1988). *The New S Language.* Wadsworth.

Breiman, L. (1991). The $\prod$-method for estimating multivariate functions from noisy data. *Technometrics*, **33** 125 – 160.

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.*, **80** 580 – 619.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth.

Buja, A., Hastie, T., and Tibshirani, R. (1989), Linear smoothers and additive models. *Ann. Statist.*, **17** 453 – 555.

Chen, Z. (1989). Interaction spline models and their convergence rates. Technical Report 854, Dept. of Statistics, University of Wisconsin, Madison, to appear *Ann. Statist.*

Chen, Z., Gu, C. and Wahba, G. (1989). Discussion of "Linear smoothers and additive models" by Buja, Hastie and Tibshirani. *Ann. Statist.*, **17** 515 – 521.

Cox, D. (1984). Multivariate smoothing spline functions. *SIAM J. Numer. Anal.*, **21** 789 – 813.

—— (1989). Coverage probability of Bayesian confidence intervals for smoothing splines. Technical Report 24, Dept. of Statistics, University of Illinois, Champaign.

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. *Numer. Math.*, **31** 377 – 403.

Douglas, A. and Delampady, M. (1990). Eastern lake survey - phase i: Documentation for the data base and the derived data sets. SIMS Technical Report 160, Dept. of Statistics, University of British Columbia, Vancouver.

Gu, C. (1989). RKPACK and its applications: Fitting smoothing spline models. In *Proc. Statist. Comput. Section*, pp. 42 – 51. American Statistical Association.

—— (1990). Diagnostics for nonparametric additive models. Technical Report 92, Dept. of Statistics, University of British Columbia.

—— (1992). Penalized likelihood regression: A Bayesian analysis. *Statistica Sinica*, **2**, 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 J. Matrix Anal. Appl.*, **10** 457 – 480.

Gu, C. and Wahba, G. (1990). Semiparametric ANOVA with tensor product thin plate splines. Technical Report 90-61, Dept. of Statistics, Purdue University, West Lafayette.

—— (1991a). Discussion of "Multivariate adaptive regression splines" by J. Friedman. *Ann. Statist.*, **19** 115 – 123.

—— (1991b). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.*, **12** 383 – 398.

Hall, P. and Titterington, D. M. (1987). Common structure of techniques for choosing smoothing parameters in regression problems. *J. Roy. Statist. Soc. Ser. B*, **49** 184 – 198.

—— (1988). On confidence bands in nonparametric density estimation and regression. *J. Mult. Anal.*, **27** 228 – 254.

Hastie, T. and Tibshirani, R. (1990). *Generalizes Additive Models.* Chapman and Hall.

Huber, P. (1985). Projection pursuit. *Ann. Statist.*, **13** 435 – 525.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, **33** 82 – 95.

Li, K.-C. (1989). Honest confidence intervals for nonparametric regression. *Ann. Statist.*, **17** 1001 – 1008.

Mate, L. (1989). *Hilbert Space Methods in Science and Engineering.* Hilger.

Meinguet, J. (1979). Multivariate interpolation at arbitrary points made simple. *J. Appl. Math. Phys. (ZAMP)*, **30** 292 – 304.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Amer. Statist. Assoc.*, **83** 1134 – 1143.

—— (1990). The average posterior variance of a smoothing spline and a consistent estimate of the average squared error. *Ann. Statist.*, **18** 415 – 428.

Shiau, J.-J. (1985). Smoothing spline estimation of functions with discontinuities. Technical Report 768, Dept. of Statistics, University of Wisconsin, Madison.

Speed, T. (1987). What is an analysis of variance? *Ann. Statist.*, **15** 885 – 941.

Stone, C. (1985). Additive regression and other nonparametric models. *Ann. Statist.*, **13** 689 – 705.

—— (1990). l_2 rate of convergence for interaction spline regression. Technical Report 268, Dept. of Statistics, University of California, Berkeley.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B*, **40** 364 – 372.

—— (1979). Convergence rates of "thin plate" smoothing splines when the data are noisy. In T. Gasser and M. Rosenblatt, editors, *Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics, No.757*, pp. 232 – 246.

—— (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B*, **45** 133 – 150.

—— (1986). Partial and interaction splines for the semiparametric estimation of functions of several variables. In T. Boardman, editor, *Computer Science and Statistics: Proceedings of the 18th Symposium*, pp. 75 – 80. American Statistical Association, Washington, DC.

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

Wahba, G. and Wendelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review*, **108** 1122 – 1145.

Weinert, H., editor. (1982). *Reproducing kernel Hilbert spaces: Application in signal processing.* Hutchinson Ross.