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

To Appear, J. Computational and Graphical Statistics

CHONG GU and GRACE WAHBA*

November 11, 1992

Abstract

We study a multivariate smoothing spline estimate of a function of several variables, based on an ANOVA decomposition as sums of main effect 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 nominal confidence levels. Lastly, we 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.

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

*Chong Gu chong@pop.stat.purdue.edu 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 wahba@stat.wisc.edu 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.

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 we will study, either $\mathcal{T}^{(\alpha)} = [0, 1]$ or, $\mathcal{T}^{(\alpha)} = E^{k(\alpha)}$, Euclidean $k(\alpha)$ space, and then $\boldsymbol{t} = (t_1, \cdots, t_d)$ is in E^K space, where $K = \sum_\alpha k(\alpha)$. By setting $k(\alpha)$ 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, in addition, 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, 1991), the $\prod$-method (Breiman, 1991), Additive Models (Buja, Hastie and Tibshirani, 1989, Hastie and Tibshirani, 1990). Neural net research is partly concerned with multivariate function estimation in the sense that we use it here, see for example Moody and Utans (1991). Each method has unique problems and successes in providing accuracy statements which we will not discuss here.

In this paper, we will be providing accuracy statements within the framework of a general form of smoothing spline analysis of variance (SS-ANOVA) in reproducing kernel Hilbert spaces (RKHS). An overview of SS-ANOVA as it applies to polynomial splines and tensor products of polynomial splines can be found in Wahba(1990). More recently this framework has been generalized to show how to include thin plate splines in an SS-ANOVA model (Gu and Wahba, 1991a, 1993). The use of thin plate splines as part of the SS-ANOVA model allows the modeling of geographic and other variables as variables in main effects, interaction terms, and so forth.

In SS-ANOVA (and other ANOVA in function space approaches, see, e.g. Friedman(1991) and Stone(1985)), f has a representation of the form

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

where the expansion is made unique and (usually) truncated in some manner.

In the SS-ANOVA context, the estimate f_λ of f is obtained by finding f_λ in an appropriate RKHS to minimize an expression similar to

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

where $I_{\mathcal{M}}$ is the collection of indices for components to be included in the model, 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, satisfying an appropriate constraint for identifiability. 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). Numerical methods for fitting the smoothing spline models have been developed (Gu, Bates, Chen and Wahba, 1989; Gu and Wahba, 1991b), and publicly available code developed (RKPACK, Gu, 1989).

The goal of the present work is the establishment of component-wise Bayesian "confidence intervals" in the SS-ANOVA context, which generalize 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 $f_\alpha, f_{\alpha\beta}$, etc. to be included in the ANOVA decomposition. More importantly we obtain them in a manner which allows a stable and efficient calculation. In addition we demonstrate how they may be computed using RKPACK. We suggest their properties via a Monte Carlo study.

It is a major task of nonparametric regression to provide some sort of accuracy information 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. This property is an "across-the-function" property. "Across-the-function" means that when restricting the 95% confidence intervals to the n data points, around 95% of them will cover the values of the true curve there. A partly heuristic theoretical argument why this could be expected was given in Wahba(1983), 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). To us, it is important and useful that the weaker definition of "confidence interval" which is adopted in Wahba (1983) and assumed here leads to intervals which are easy to interpret psychologically. In simulations, when the intervals cover about 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. We note that these confidence intervals are not in general pointwise confidence intervals (there aren't many "free lunches" in nonparametric regression) — the coverage will tend to be less than nominal where the true curve has sharp peaks or kinks and more where the true curve is smooth. If the user interprets them appropriately across the function, he or she will have a reasonable feel for the *overall* accuracy of the estimate.

The results of the Monte Carlo study described here are suggestive that the componentwise confidence intervals roughly have the same "across-the function" coverage property *for each component*, in the examples we have chosen. The reader may judge from the plotted confidence intervals overlaying the true function the psychological information that is conveyed by the intervals.

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 propose using certain contours of constant posterior standard deviation to bound an area within which the estimated function is to be displayed.

In Section 2 we briefly review and slightly extend the SS-ANOVA framework given in Gu and Wahba(1991a, 1993). This will establish notation and demonstrate the key ingredients of a general SS-ANOVA. In Section 3 we give the component-wise posterior covariance functions. The proof is relegated to Appendix A. In Section 4 we review some known reproducing kernels which are useful in SS-ANOVA. In Section 5 we provide the details of how RKPACK may be used to carry out the calculations of the Bayesian "confidence intervals" , and in Section 6 we present the results of a small Monte-Carlo study on simulated data. In Section 7 we describe the application to some data on lake acidity as a function of geographical location and calcium concentration from the

Eastern Lake Survey (Douglas and Delampady (1990)). In our original submission we suggested what assumptions and lemmas might be necessary to extend the main theoretical results of Nychka (1988, 1990) concerning the properties of the (single component) Bayesian "confidence intervals" to the component-wise case considered here. This part has been deleted at the suggestion of the referees, but may be found in Gu and Wahba (1991c), Appendix B.

2 Analysis of Variance in RKHS

We will always assume that f is in some RKHS, that is, a Hilbert space of functions 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.

Let now $\mathcal{H}$ be some RKHS of real-valued functions of $\boldsymbol{t} = (t_1, \cdots, t_d) \in \mathcal{T} = \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$, where we may allow $t_\alpha \in \mathcal{T}^{(\alpha)}$, an arbitrary measureable index set, and, furthermore, suppose the one dimensional space of constant functions on $\mathcal{T}$ is a subspace of $\mathcal{H}$. Then there are many ways that an ANOVA-like decomposition of the form (1.2) can be defined for f in such a space. We now give a general construction. For each $\alpha = 1, \cdots, d$, construct a probability measure $d\mu_\alpha$ on $\mathcal{T}^{(\alpha)}$, with the property that the symbol $(\mathcal{E}_\alpha f)(\boldsymbol{t})$, defined by

$$(\mathcal{E}_\alpha f)(\boldsymbol{t}) = \int_{\mathcal{T}^{(\alpha)}} f(t_1, \cdots, t_d) d\mu_\alpha(t_\alpha)$$

is well defined and finite for every $f \in \mathcal{H}$ and $\boldsymbol{t} \in \mathcal{T}$ (although of course $(\mathcal{E}_\alpha f)(\boldsymbol{t})$ will not vary with t_α). We need the further assumption, that considering $(\mathcal{E}_\alpha f)(\cdot)$ as a function of $\boldsymbol{t}$, then it defines an element of $\mathcal{H}$. We will henceforth assume that this condition holds (we will construct a generic example shortly), then we can consider $\mathcal{E}_\alpha$ as an operator from $\mathcal{H}$ to $\mathcal{H}$. We will call such operators averaging operators. 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 then always generates a unique (ANOVA-like) decomposition of f of the form (1.2) 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. Note that the components will depend on the measures $d\mu_\alpha$ and these should be chosen in a specific application so that the fitted mean, main effects, etc. have reasonable interpretations.

This construction specializes to the ordinary two way layout by taking $d = 2$ and $\mathcal{T}^{(\alpha)} = \{1, 2, ..., K_\alpha\}$ for $\alpha = 1, 2$, $\mathcal{T} = \mathcal{T}^{(1)} \otimes \mathcal{T}^{(2)}$ and letting $\mathcal{E}_1 f(\boldsymbol{t}) = \frac{1}{K_1} \sum_{\gamma=1}^{K_1} f(\gamma, t_2)$, and similarly for $\mathcal{E}_2$. $f(\cdot)$ and $(\mathcal{E}_\alpha f)(\cdot)$ should be thought of as $K = K_1 \times K_2$ vectors here. Although other averaging operators are obviously possible, this pair seems to be in common use in the usual two-way layout without much particular justification. Note that if we adopt the ordinary Euclidean inner product for functions defined on K dimensional Euclidean space, then the ranges of the four operators $\mathcal{E}_1\mathcal{E}_2, \mathcal{E}_1(I - \mathcal{E}_2), (I - \mathcal{E}_1)\mathcal{E}_2$,and $(I - \mathcal{E}_1)(I - \mathcal{E}_2)$ consist of four orthogonal subspaces of Euclidean K-space whose direct sum is Euclidean K-space. In that case the components are easy to estimate and have an intuitive meaning for the user. Note that in the usual d-way layout, the functions of interest are only defined on the design points, but that with the ANOVA that we will study, the functions may have a much larger domain, and, although the domain is required to have a tensor product structure, we will see that the design may not.

In the general RKHS case, the range of each operator of the form $\prod_{\alpha_1,\cdots,\alpha_k} \mathcal{E}_\alpha \prod_{\alpha_{k+1},...,\alpha_d}(I - \mathcal{E}_\beta)$ is a subspace of $\mathcal{H}$, however, these subspaces are not necessarily orthogonal with respect to the inner product in $\mathcal{H}$. In this paper we will restrict ourselves to ANOVA decompositions in RKHS such that the ranges of these operators are orthogonal. This will result in components that are relatively easy to estimate and that may have an intuitive meaning for the user.

We will now show how to construct generic RKHS's satisfying the above conditions, so that the subspaces which are ranges of sums of products of the $\mathcal{E}_\alpha$ and $I - \mathcal{E}_\alpha$ are all orthogonal in the inner product of the space. Let $\mathcal{H}^{(\alpha)}$ be an RKHS of functions on $\mathcal{T}^{(\alpha)}$ with $\int_{\mathcal{T}^{(\alpha)}} f(t_\alpha) d\mu_\alpha = 0$, $f \in \mathcal{H}^{(\alpha)}$, and let $[1^{(\alpha)}]$ be the one dimensional space of constant functions on $\mathcal{T}^{(\alpha)}$. Consider the space $[1^{(\alpha)}] \oplus \mathcal{H}^{(\alpha)}$, where $\oplus$ is tensor (or direct) sum. Then any f in this space will have a unique decomposition $f = P_c f + (f - P_c f)$, with $P_c f = \int f d\mu_\alpha \in [1^{(\alpha)}]$ 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

$$\mathcal{H} = \otimes_{\alpha=1}^{d} [[1^{(\alpha)}] \oplus \mathcal{H}^{(\alpha)}], \tag{2.2}$$

where $\otimes_{\alpha=1}^{d}$ is the tensor product of the d Hilbert spaces in brackets. See Aronszajn(1950) for a detailed discussion of tensor sums and tensor products of RKHS and Wahba(1990, Section 10) for examples. Further examples will be given later.

The right hand side of (2.2) can be expanded as

$$\mathcal{H} = [1] \oplus \sum_{\alpha} [\mathcal{H}^{(\alpha)}] \oplus \sum_{\beta<\alpha} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots, \tag{2.3}$$

where we have written $[1]$ to denote $\otimes_{\alpha=1}^{d}[1^{(\alpha)}]$, the constant functions on $\mathcal{T}$ and, with some abuse of notation, we have supressed $[1^{(\alpha)}]$ whenever it multiplies a term of a different form. That is, we have written $\mathcal{H}^{(1)}$ instead of $\mathcal{H}^{(1)} \otimes_{\alpha=2}^{d} [1^{(\alpha)}]$, and so forth. Hopefully this makes clear that the terms in brackets in (2.3) are all subspaces of functions on $\mathcal{T}$, even though the functions in them do not all depend on all of the variables $t_1, ..., t_d$.

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. We are continuing with this notational convention, that is, f_α is considered as an element of $\mathcal{H}$ even though it is a constant function of all the t_β's except for $\beta = \alpha$.

The subspaces in brackets in (2.2) are all orthogonal in the tensor product norm induced by the original inner products. Thus the decomposition of f of the form (1.2) with $C = (\prod_\alpha \mathcal{E}_\alpha)f, f_\alpha \in \mathcal{H}^{(\alpha)}$, $f_{\alpha\beta} \in \mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}$ will be an orthogonal decomposition. 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:

$$\begin{aligned}
\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)} &= \mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)} && (2.4)\\
&+ \mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)} && (2.5)\\
&+ \mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)} && (2.6)\\
&+ \mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}. && (2.7)
\end{aligned}$$

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 (2.5), smooth $\otimes$ parametric (s, π) of the form (2.6), and smooth $\otimes$ smooth (s, s) of the form (2.7), 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 $\mathcal{H}^\beta, \beta = 1, 2, \cdots, p$. For example, in the case $J_\alpha(f_\alpha) = \int (f_\alpha^{(m)}(u))^2 du$, $\mathcal{H}_\pi^{(\alpha)}$ is spanned by the polynomials of degree less than m in t_α which average to 0 under $\mathcal{E}_\alpha$, and $\mathcal{H}^0$ is sums and products of such polynomials. $\mathcal{H}^\beta$ may stand for a subspace $\mathcal{H}_s^{(\alpha)}$, or one of the subspaces of the form (2.5), (2.6), (2.7), 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

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{t}(i)))^2 + \lambda \sum_\beta \theta_\beta^{-1} \|P^\beta f\|^2, \tag{2.8}$$

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 (2.8) 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.

We end this section with a few remarks concerning the choice of the probability measure μ_α. In the case $\mathcal{T}^{(\alpha)}$ consists of K_α points it is natural to use the uniform measure on the points, in any case this is the common practice in (parametric) ANOVA. In the case $\mathcal{T}^{(\alpha)}$ a finite interval, a natural choice, which would lead to interpretable results, would be to let μ_α be (a multiple of) Lebesgue measure. In the case that the uniform measure on $\mathcal{T}^{(\alpha)}$ cannot be scaled to be a probability measure (i.e., if $\mathcal{T}^{(\alpha)} = E^{k(\alpha)}$), another choice must be made. A uniform measure over a finite region of interest or a measure reflecting the observational density could be used. In the examples in this paper we will use Lebesgue measure when $\mathcal{T}^{(\alpha)}$ is [0,1] and uniform measure on the (marginal) design points when $\mathcal{T}^{(\alpha)} = E^{k(\alpha)}$.

3 Bayesian Posterior Covariances for Components

In this Section we provide general formulas for the Bayesian posterior covariances for the components of f estimated by minimizing (2.8). The component-wise Bayesian "confidence intervals' are then computed from the relevant posterior standard deviations, generalizing the (single-component)

Bayesian "confidence intervals" given in Wahba (1983). The computation of the relevant quantities will 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 $X_\xi(\boldsymbol{t}),\boldsymbol{t}\in\mathcal{T}=\otimes_\alpha\mathcal{T}^{(\alpha)}$ 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 $\epsilon=(\epsilon_1,\cdots,\epsilon_n)'\sim\mathcal{N}(0,\sigma^2I)$. Let

$$f_\lambda(\boldsymbol{t})=\lim_{\xi\to0}E\{X_\xi(\boldsymbol{t})|Y_i=y_i,i=1,\cdots,n\}$$

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

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

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 (2.8). 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

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

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

Theorem 3.1

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

where e_ν *is the* ν*th 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 Appendix A. 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

We remind the reader (see Aronszajn(1950)) that reproducing kernels (RK's) for tensor products of RKHS are just the products of the individual RK's. In symbols, if $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ are RKHS of functions defined on $\mathcal{T}^{(1)}$ and $\mathcal{T}^{(2)}$ respectively with RK's $R^{(1)}(t_1, t_1')$ and $R^{(2)}(t_2, t_2')$ then the RK R for $\mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)}$ is the function of (t_1, t_2) and (t_1', t_2') given by $R(t_1, t_2; t_1', t_2') = R^{(1)}(t_1, t_1')R^{(2)}(t_2, t_2')$. By iterating this process (and silently using the fact that the RK for $[1^{(\alpha}]$ with the norm defined implicitly just before (2.2) is the constant 1) it can be seen that all of the $R_\beta(\boldsymbol{s},\boldsymbol{t})$ that we need will be known once we know the reproducing kernels for the $\mathcal{H}_\pi^{(\alpha)}$ and $\mathcal{H}_s^{(\alpha)}$. In the simulations below we will use $\mathcal{H}^{(\alpha)}$ that correspond to polynomial and thin plate splines respectively. Examples of

reproducing kernels for these cases appear in the literature and we will just display the results that we will use in the Monte Carlo and data analysis studies below:

4.1 Univariate Polynomial Splines

For $\mathcal{T}^{(\alpha)} = [0,1]$ the polynomial spline penalty functional is $J_m^1(f) = \int_0^1 (f^{(m)}(t))^2 dt$. The null space of this penalty functional is the m-dimensional span of the polynomials of degree less than m. $\mathcal{H}_\pi^{(\alpha)}$ is of dimension $m-1$, and is the span of these polynomials satisfying the side condition that they integrate to 0 with respect to μ_α. The elements of $\mathcal{H}_s^{(\alpha)}$ also integrate to 0 with respect to μ_α and will satisfy $m-1$ side conditions to guarantee orthogonality with $\mathcal{H}_\pi^{(\alpha)}$. RK's are given in Gu, Bates, Chen and Wahba(1989) and Wahba(1990, Chapter 10), for μ_α Lebesgue measure and side conditions periodic boundary conditions. We will use the case of $m = 2$ in the Monte Carlo study in Section 6, and the RK's R_π and R_s for $\mathcal{H}_\pi^{(\alpha)}$ and $\mathcal{H}_s^{(\alpha)}$ in this case are reproduced here: Letting $k_\ell = B_\ell/\ell!$, where B_ℓ is the ℓth Bernoulli polynomial,

$$R_\pi(t,t') = k_1(t)k_1(t') \tag{4.1}$$

$$R_s(t,t') = k_2(t)k_2(t') - k_4([t-t']) \tag{4.2}$$

where $[\tau]$ is the fractional part of τ. For future reference we remark that in this case $\mathcal{H}_\pi^{(\alpha)}$ is spanned by k_1.

4.2 Thin Plate Splines

For $\mathcal{T}^{(\alpha)} = E^k$, $k = 1, 2, ..$ the thin plate penalty functional is

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

For technical reasons it is necessary that $2m - k > 0$. The null space of this penalty functional is the $M = \binom{m+k-1}{k}$ polynomials of total degree less than m in k variables, and $\mathcal{H}_\pi^{(\alpha)}$ is the $M-1$ dimensional space spanned by these polynomials constrained to integrate to 0 with respect to μ_α. This integration must be well defined, so we cannot take μ_α as Lebesgue measure over $\mathcal{T}^{(\alpha)}$ here. Reproducing kernels for this case have been given in Gu and Wahba(1993) for μ_α any probability measure nontrivially supported on a finite unisolvent set of points. (A unisolvent set of points is a set for which least squares regression on polynomials of total degree less than m in k variables is

unique.) The elements of $\mathcal{H}_s^{(\alpha)}$ integrate to 0 with respect to this measure and must satisfy $M-1$ additional moment conditions. In the data analysis study in Section 7 we will let $\mathcal{T}^{(\alpha)}$ be E^1 or E^2, and $m=2$. The reproducing kernel R_π for $\mathcal{H}_\pi^{(\alpha)}$ in Gu and Wahba(1993) was obtained by choosing M linearly independent polynomials $\phi_1, \cdots, \phi_M$, in the k variables, of total degree less than m so that $\phi_1 = 1$ and so that they are orthonormal under the inner product $\langle \phi_\mu, \phi_\nu \rangle = \int \phi_\mu \phi_\nu d\mu_\alpha$. Then the RK R_π for $\mathcal{H}_\pi^{(\alpha)}$ with this inner product is

$$R_\pi(t,t') = \sum_{\nu=2}^{M} \phi_\nu(t)\phi_\nu(t').$$

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

$$P_\pi f = \sum_{\nu=1}^{M} \phi_\nu \int f \phi_\nu d\mu_\alpha,$$

suitable moment conditions on elements in $\mathcal{H}_s^{(\alpha)}$ are defined by $P_\pi f = 0$. The RK $R_s(t,t')$ for $\mathcal{H}_s^{(\alpha)}$, was obtained as a function of the semi-kernel (variogram) E_m^k associated with thin plate splines, given by $E_m^k(\tau) \propto |\tau|^{2m-k}$, k not an even integer, and $E_m^k(\tau) \propto |\tau|^{2m-k} \log|\tau|$, k an even integer. Letting $E(t,t') = E_m^k(|t-t'|)$, where $|t-t'|$ is the Euclidean distance between t and t' in E^k, and letting $P_{\pi(t)}$ be P_π applied to what follows considered as a function of t, then, it is shown in Gu and Wahba (1993), that the RK for $\mathcal{H}_s^{(\alpha)}$ is given by

$$R_s(t,t') = (I - P_{\pi(t)})(I - P_{\pi(t')})E(t,t').$$

We remark that this result in the one dimensional case ($k=1$) goes back to deBoor and Lynch (1966), see also Wahba and Wendelberger (1980). The $k=1$ case results in polynomial splines for the main effect, although the side conditions on $\mathcal{H}_s^{(\alpha)}$ are different here than in Section 4.1.

5 Computation

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. These algorithms are implemented in RKPACK. 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

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

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

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

$$\begin{aligned} (S'M^{-1}S)^{-1} &= R_1^{-1}[(F_1'\Sigma F_1 + n\lambda I) - (F_1'\Sigma F_2)(F_2'\Sigma F_2 + n\lambda I)^{-1}(F_2'\Sigma F_1)](R_1^{-1})' \\ &= R_1^{-1}[(F_1'\Sigma F_1 + n\lambda I) - (F_1'\Sigma F_2)U(T + n\lambda I)^{-1}U'(F_2'\Sigma F_1)](R_1^{-1})', \end{aligned} \tag{5.2}$$

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 computational form for the variance estimate usually associated with the GCV selection of smoothing parameters is $\hat{\sigma}^2_{GCV} = n\lambda z'(\tilde{\Sigma} + n\lambda I)^{-2}z/\text{trace}(\tilde{\Sigma} + n\lambda I)^{-1}$, and the GML estimate is $\hat{\sigma}^2_{GML} = n\lambda z'(\tilde{\Sigma} + n\lambda I)^{-1}z/(n - M)$. A more familiar form for $\hat{\sigma}^2_{GCV}$ is $\hat{\sigma}^2_{GCV} = (\textit{residual sum of squares/degrees of freedom for noise})$, see Wahba (1983, 1990), Gu (1989), and Gu and Wahba (1991b), Hall and Titterington(1987) for further information about σ^2_{GCV}.

To facilitate the application of the technique developed in this article, two self-documented utility routines have been added to RKPACK to carry out the calculations in (5.1) and (5.2). The routines are to be used in conjunction with the RKPACK drivers and the usage is illustrated in our simulation code. (See §6 for simulations.) The RKPACK package including the simulation code is available from `statlib@temper.stat.cmu.edu` and `netlib@research.att.com`.

6 Simulations

We present the results of a pilot simulation study in this section to illustrate the practical performance of the component-wise Bayesian "confidence intervals". Since the method applies to a generic class of nonparametric models we are not attempting a definitive study. Instead, we apply the technique to a single arbitrary but nontrivial test example and collect and describe the results. Our simulation code has been briefly commented and is available to the public so that interested readers may choose to augment our simulations by running the code on our examples or on test examples of their own choice.

Our test example is on $\mathcal{T} = \mathcal{T}^{(1)} \otimes \mathcal{T}^{(2)} \otimes \mathcal{T}^{(3)} = [0,1]^3$, using a model built up from the $\mathcal{H}^{(\alpha)}$ spaces in Section 4.1 with μ_α Lebesgue measure. We generated design points $\boldsymbol{t}(i)$ (once and for all) from the uniform distribution on $[0,1]^3$ and generated responses by $y = C + f_1(t_1) + f_2(t_2) + f_{12}(t_1,t_2) + \epsilon$ with $\epsilon \sim N(0,\sigma^2)$. Note that there was no dependence of the response on t_3. The components of the test function used in our simulations were $C = 5$, $f_1(t_1) = e^{3t_1} - (e^3 - 1)/3$, $f_2(t_2) = 10^6[t_2^{11}(1-t_2)^6 - Be(12,7)] + 10^4[t_2^3(1-t_2)^{10} - Be(4,11)]$, and $f_{12}(t_1,t_2) = 5\cos(2\pi(t_1 - t_2))$, where $Be(p,q)$ is the Beta function. These component functions satisfy the side conditions $0 = \mathcal{E}_1 f_1 = \mathcal{E}_2 f_2 = \mathcal{E}_1 f_{12} = \mathcal{E}_2 f_{12}$, where $\mathcal{E}_\alpha(\cdot) = \int_0^1 (\cdot) dt_\alpha$.

We chose to fit a model with three main effects and one two factor interaction:

$$f(\boldsymbol{t}) = C + f_1(t_1) + f_2(t_2) + f_{12}(t_1, t_2) + f_3(t_3). \tag{6.1}$$

The unpenalized space $\mathcal{H}^0$ is the five dimensional span of $\{1, k_1(t_1), k_1(t_2), k_1(t_3), k_1(t_1)k_1(t_2)\}$ and there are six penalized spaces each with a separate smoothing parameter, consisting of three spaces of the form $\mathcal{H}_s^{(\alpha)}$ for the three main effects and three for the t_1-t_2 interaction of the form (2.5), (2.6) and (2.7).

For the simulations, we have chosen to use a model that is (one component) bigger than the true model, to see whether the method will correctly suggest that the spurious component is not present. In the simulations below we have taken $n = 100$ and 200, and the six smoothing parameters that we have in the model are probably about as many smoothing parameters as one can expect to deal with with these small sample sizes. Thus, we deleted *a priori* the t_1-t_3 and t_2-t_3 interactions. The point of view we are taking here is that the SS-ANOVA should, strictly speaking, be thought of as a top-down approach, that is, the models being entertained should ideally contain the true model. Model selection is, of course very important, but beyond the scope of the present paper. We just note that 1) in the simulations below the confidence intervals do indicate that the f_3 term is not present, and 2) if the model is too small, then of course the estimate of σ^2 is likely to be inflated. See Gu and Wahba(1993) for a discssion of some model selection methods. Although we have not done a full fit including all the two-factor interactions, we conjecture that it can be done with a larger sample size, and that the results would correctly suggest that dependence on t_3 is absent.

Letting $\hat{g}(\boldsymbol{t})$ stand for any one of the four estimated components f_1, f_2, f_{12} or f_3, 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. We note that the three penalized spaces in the t_1-t_2 interaction term, which had been kept separate for smoothing parameter selection, are lumped together as a single penalized term, for display and for computing 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 intervals covered the true values of f, f_1, f_2, f_{12}, and f_3 were recorded. These numbers were then divided by the corresponding sample sizes to form the coverage percentages of the intervals on the

design points. We summarized these coverage percentages using box-plots and some of them are presented in Figures 6.1 and 6.2. Figure 6.1 collects the coverage percentages of the 95% intervals in all the six experiments, with the two rows corresponding to $n = 100$ (top) and $n = 200$ (bottom) and the three columns corresponding to $\sigma = 1$ (left), $\sigma = 3$ (center), and $\sigma = 10$ (right). Figure 6.2 augments the center bottom frame of Figure 6.1 with the coverage percentages of the 90%, 75%, and 50% intervals for the $n = 200$ and $\sigma = 3$ experiment. The sample means of the coverage percentages are marked as plusses in the boxplots and the nominal coverages are superimposed as dotted lines. Box plots corresponding to Figure 6.2 for the cases not shown had a similar appearance.

In the $n = 100$ experiments here, the GCV criterion chose to (nearly) interpolate the data in four of the one hundred $\sigma = 1$ replicates and in one of the $\sigma = 3$ replicates. These cases can be readily detected by estimates of σ^2 which are many orders of magnitude too small, and by their plots, which are highly wiggly. There were no such near-interpolating cases in the $\sigma = 10$ replicates nor in any of the three hundred $n = 200$ replicates. This phenomenon, a small fraction of unacceptable results in small sample sizes which disappears in larger sample sizes, has been noted elsewhere. See, for example Wahba(1983). The 5 near-interpolating cases have been omitted from Figure 6.1. We note that as the difference between the fit and the data comes close to machine 0, the calculation of posterior variances by adding terms in Theorem 3.1 may become unstable.

We visually inspected many of the plotted intervals and (with the above five exceptions) they all convey a similar visual impression. Therefore we will just display "typical" $n = 100$ and $n = 200$ 95% cases for the $\sigma = 3$ runs. These cases were actually the first replicates in the two runs. Here the data for the $n = 100$ case form a subset of the $n = 200$ case data. In Figure 6.3, the main effects are plotted in the top row and three slices of the interaction are plotted in the bottom row, where the solid lines are the test functions, the dashed lines the $n = 200$ intervals, and dotted lines the $n = 100$ intervals.

Note that although we have plotted the intervals as continuous bands for a clear visual interpretation, these curves should not be considered as defining classical simultaneous confidence bands. In order to make statements about their frequentist properties for fixed f's (as discussed in Wahba(1983), Nychka(1988, 1990) and Gu and Wahba(1991c)), we have to evaluate the coverage at the n data points and then average across the data points. Coverage percentages aside, the 95% intervals appear to have the right magnitude as judged by the fact that they just about graze

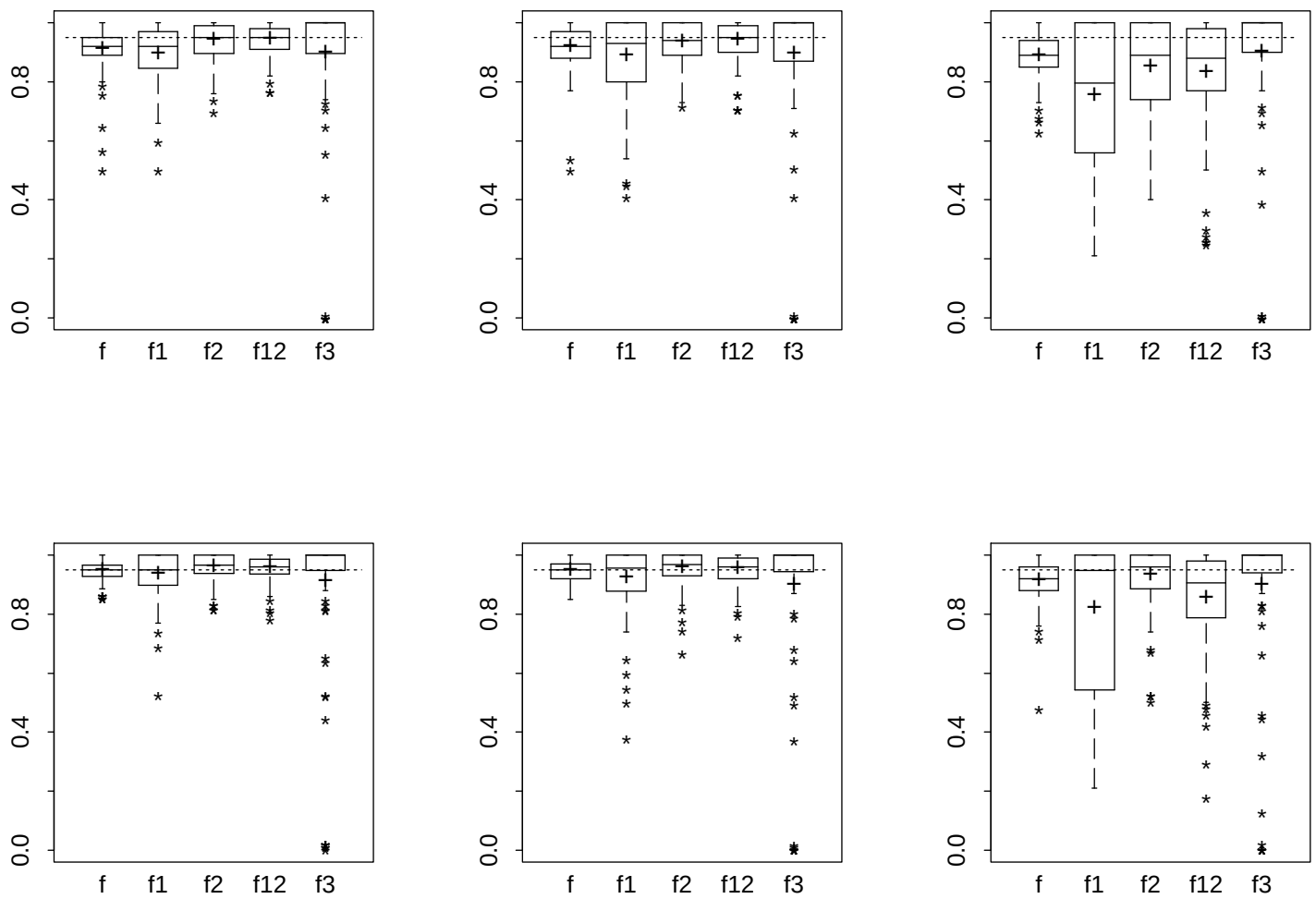

Figure 6.1: Coverage percentages of 95% intervals in the experiments. Top row: $n = 100$; bottom row: $n = 200$; left column: $\sigma = 1$; center column: $\sigma = 3$; right column: $\sigma = 10$. Plusses: sample means; dotted lines: nominal coverage.

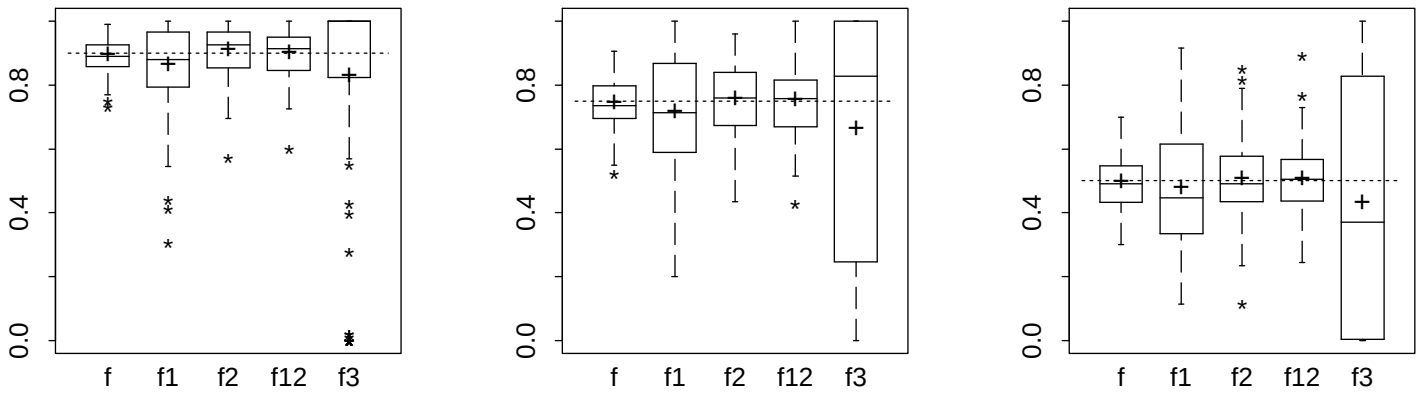

Figure 6.2: Coverage percentages of $n = 200$ and $\sigma = 3$ experiment. Left: nominal 90%; center: nominal 75%; right: nominal 50%. Plusses: sample means; dotted lines: nominal coverages.

the true curves over a few percent of the domain. For the $n = 200$ case shown, considering the 95%, 90%, 75% and 50% confidence intervals for the entire function (not shown), they covered, respectively 97%, 89.5%, 72.5% and 43% of the values of the true function at the data points, the corresponding numbers for the $n = 100$ case were 98%, 95%, 78% and 48%.

A point worth noting is the apparently different behavior of the intervals for f_3 compared to the other components, as can be seen in the box plots of Figures 6.1 and 6.2 and in the f_3 curves of Figure 6.3. Since in the test function $f_3 = 0$, the GCV criterion often effectively removed the penalized space for f_3 so that the fitted f_3 was dominated by the parametric term, which is a multiple of $k_1(t) = t - \frac{1}{2}$. In this case s_{f_3} is also a multiple of k_1. Therefore, if the penalized space component is removed completely then confidence intervals evaluated at the data points would cover the test function $f_3 = 0$ at all or none of the design points depending on whether the confidence interval for the parametric coefficient covered 0 or not. This has happened in the example shown in Figure 6.3. This can explain why the means of the f_3 box plots are roughly at their nominal values but the spread tends to be larger than for the other components. The all-or-none coverage pattern has been diluted somewhat by the fact that in some of the replicates a small "smooth" component has been included in the f_3 fit.

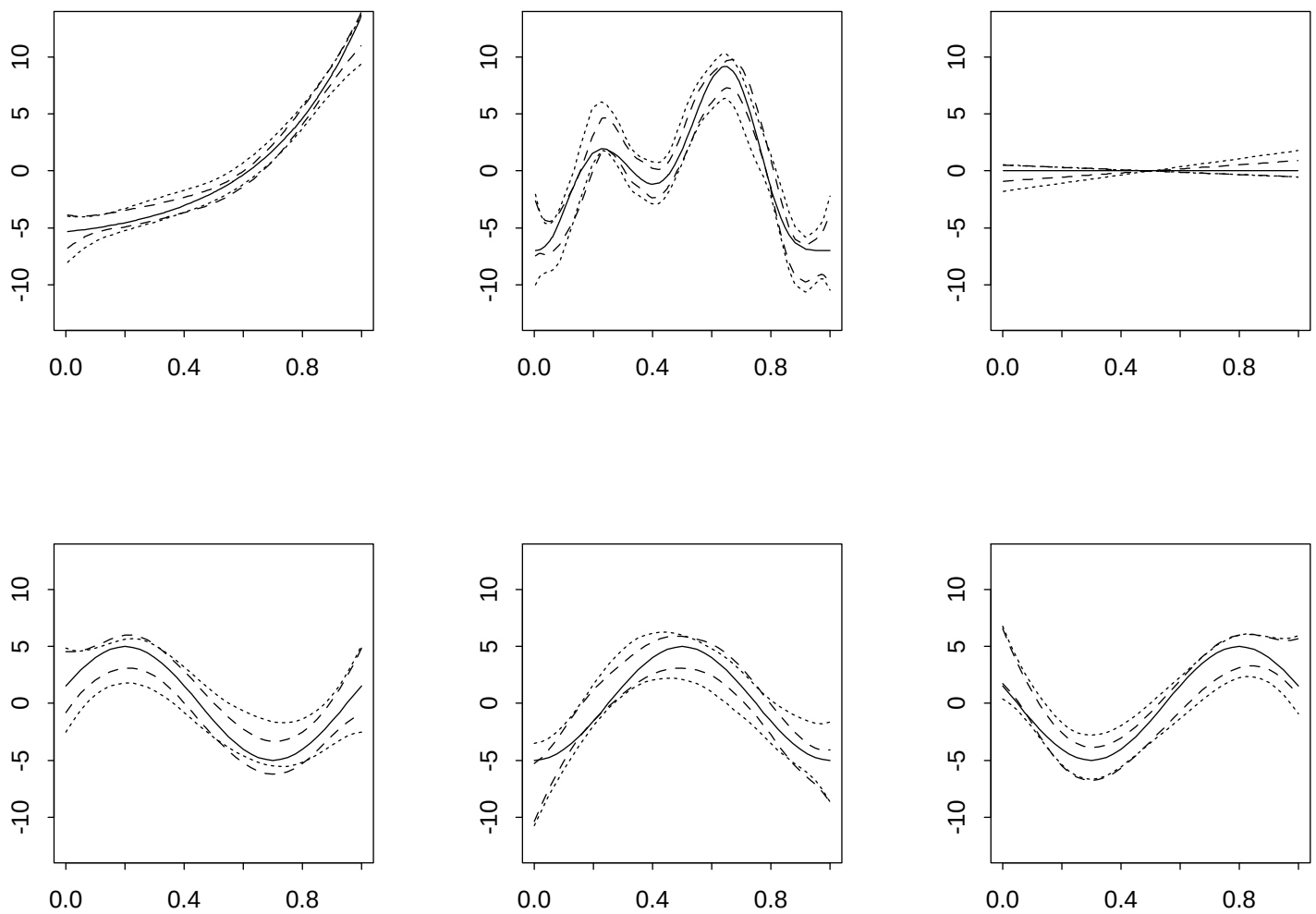

Figure 6.3: Display of the 95% intervals in a "typical" $n = 100$ and $\sigma = 3$ case and its superset $n = 200$ case. Top row: main effects $f_1(t_1)$, $f_2(t_2)$, and $f_3(t_3)$; bottom row: sliced interaction $f_{12}(t_1, .2)$, $f_{12}(t_1, .5)$, and $f_{1,2}(t_1, .8)$. Solid lines: test function; dashed lines: $n = 200$ intervals; dotted lines: $n = 100$ intervals.

7 Application: Lake Acidity Study

We further illustrate possible applications of the Bayesian "confidence intervals" on a real data problem in this section. From the data edited by Douglas and Delampady (1990) based on the Eastern Lakes Survey of 1984 by EPA, we extracted observations on 112 lakes in the southern Blue Ridge mountains area. The response y is lake water acidity (surface pH), as dependent on geographic location and calcium concentration. A model of the form $f = C + f_1(t_1) + f_2(t_2) + f_{12}(t_1, t_2)$ was fitted in Gu and Wahba (1993). Here t_1 is calcium concentration and $t_2 = (x_1, x_2)$ is geographic location. We let $\mathcal{T}^{(1)} = E^1$ and $\mathcal{T}^{(2)} = E^2$ and μ_1 and μ_2 be uniform measures over the (marginal) design points. We let $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ be as in Section 4.2 with $m = 2$. The norm in $\mathcal{H}^{(2)}$ is "rotation-invariant" and hence is suitable for modeling geographic effect; more technical details can be found in Gu and Wahba (1993). In Gu and Wahba (1993) the interaction term f_{12} was retained in the model. When we obtained component-wise Bayesian "confidence intervals" for this model as part of the present study, however, the 95% intervals for f_{12} almost completely covered zero on the design points. Thus we dropped the f_{12} term and refit the main-effect-only model $f = C + f_1(t_1) + f_2(t_2)$. As usual, $\mathcal{E}_1 f_1 = \mathcal{E}_2 f_2 = 0$, here $\mathcal{E}_\alpha f_\alpha = \frac{1}{n}\sum_{i=1}^n f_\alpha(t_\alpha(i))$.

The left and right frames of Figure 7.1 give the fitted main effects for pH and geography respectively. The constant term has been added to the f_1 main effect so that the fitted f_1 can be visually compared to the data marked by *'s. The fitted f_1 main effect is essentially a straight line. Contours for the estimated f_2 main effect are the solid lines in the right frame. The locations of the 112 lakes in the present study are marked as circles in the right frame of Figure 7.1, where the dotted lines indicate the state borders. It can be seen that the lakes run roughly along the Blue Ridge mountains which run SW to NE in GA, TN, SC, NC and VA. Although there is no data in the NW and SE corners, the estimate of f_2 is defined everywhere. It is clear that as one gets far enough away from the data the estimate carries little information. Our first task here then is to obtain a reasonable graphical display of what is hopefully the meaningful part of the estimated f_2. To this end, we plotted as the dashed lines in the left frame of Figure 7.2 the contours of the posterior standard deviation calculated according to Theorem 3.1. We plotted the contours of the estimated f_2 as solid lines in the center frame of Figure 7.2 but only within a region with the posterior standard deviation smaller than .15 (pH). The choice of .15 is about 3 times the minimum posterior standard deviation in the left frame of Figure 7.2 but otherwise arbitrary. The right frame

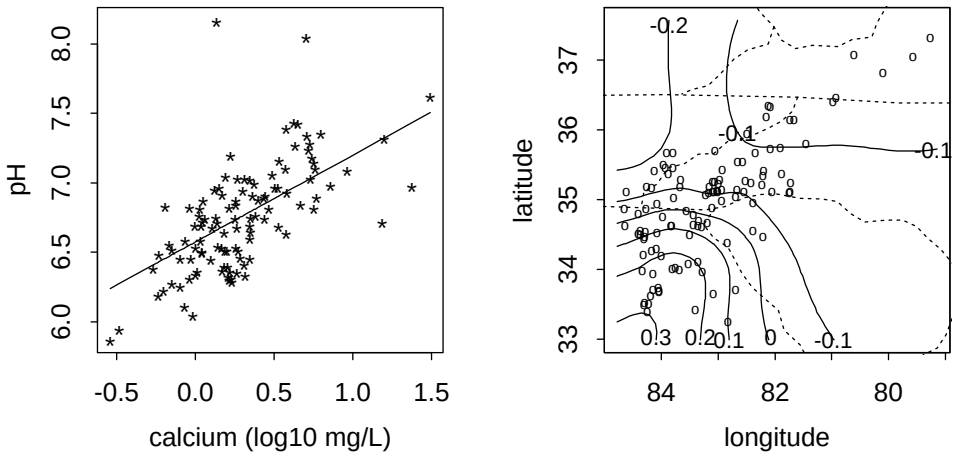

Figure 7.1: Fitted main effects model of Blue Ridge lake acidity. Left: Calcium main effect plus mean, with pH data (*). Right: Contours for geographic main effect. Circles are lakes and dotted lines state borders.

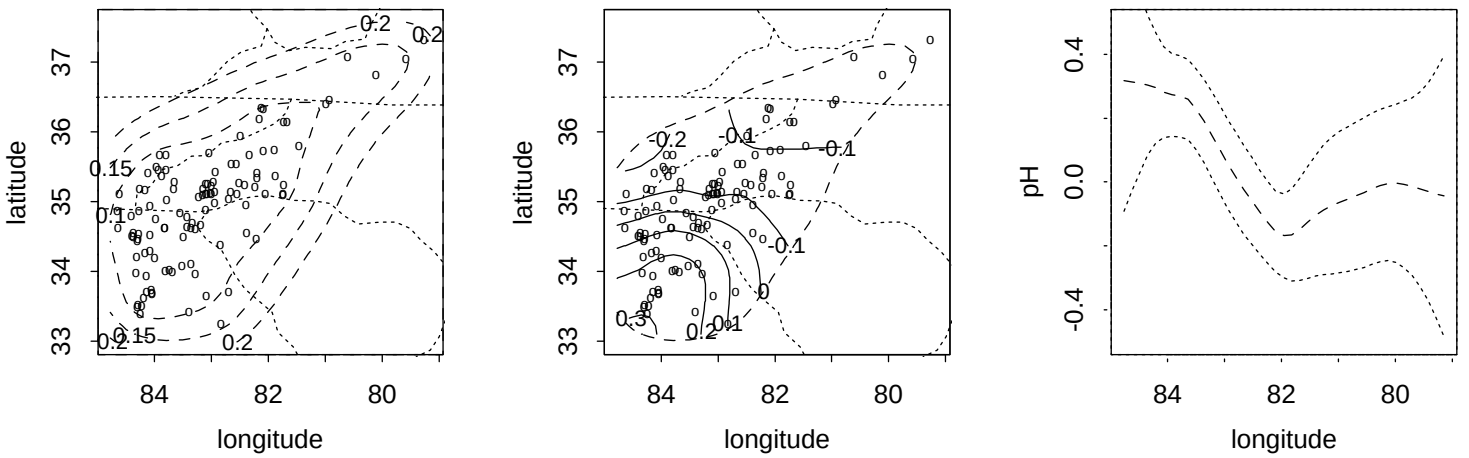

Figure 7.2: Geographic main effect. Left: contour of s_{f_2} (dashed lines); center: contours of estimated geographic main effect (solid lines) confined to small s_{f_2} area; right: estimated geographic main effect (dashed line) and 95% intervals (dotted lines) on lower left to upper right diagonal slice of center frame. In left and center frames, circles are lakes and dotted lines state borders.

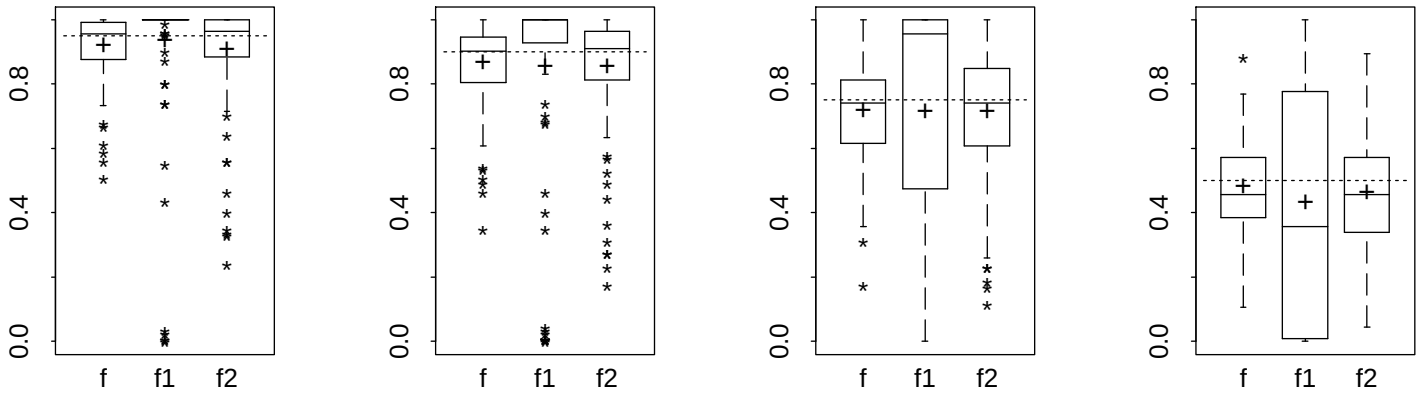

Figure 7.3: Coverage percentages in lake acidity simulations. Left to right: nominal 95%; nominal 90%; nominal 75%; nominal 50%. Plusses: sample means; dotted lines: nominal coverages.

of Figure 7.2 presents a cross section of the 95% intervals for f_2 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, roughly the location of Mt. Mitchell, the highest point in NC, at the high point of the crest of the Blue Ridge mountains.

To check how trustworthy the estimate and the intervals were, we simulated data on the same design points with the above fitted main-effect-only model as the truth and the associated variance estimate $\sigma^2 = .0655$ as the variance of the additive Gaussian noise. Similar to simulations in §6, 100 replicates were evaluated. The coverage percentages were collected, and are summarized in Figure 7.3. There were 3 replicates in which the fits interpolated the data, and these cases are omitted from Figure 7.3. As seen in Figure 7.1 the model fitted to the real data has a strong but linear calcium main effect which is in the null space of the penalty, and as a consequence the intervals for f_1 in the simulations demonstrated a fairly clear all-or-none coverage behavior. Based on the first replicate in the simulation, Figure 7.4 presents a parallel to Figure 7.2 but with the truth superimposed in the right frame as the solid line.

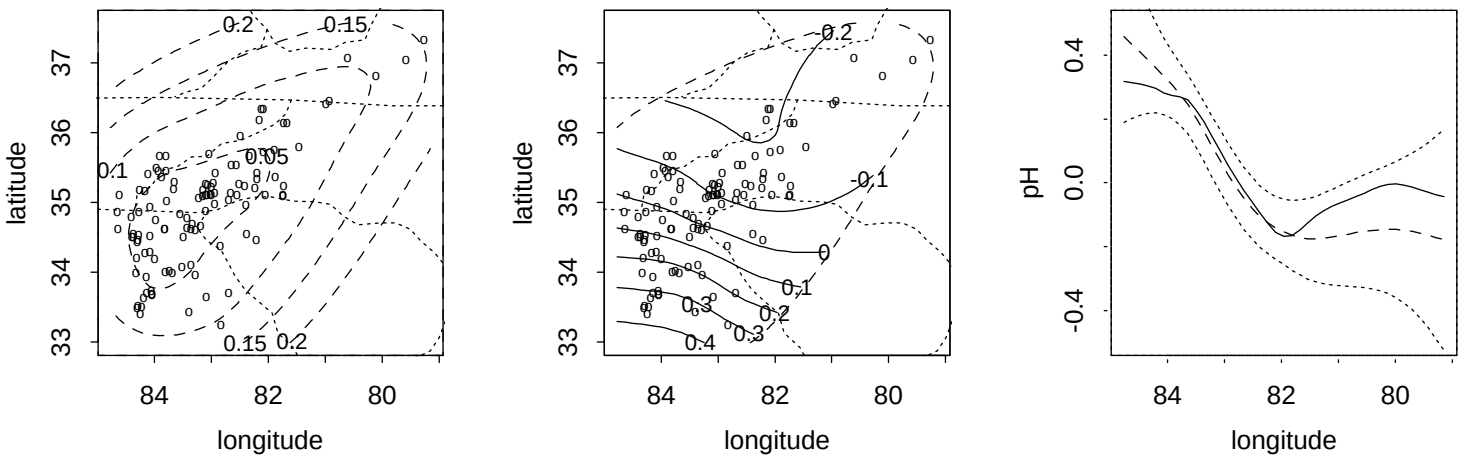

Figure 7.4: Geographic main effect based on simulated data. Left: contour of s_{f_2} (dashed lines); center: contour of $\hat{f}_2$ (solid lines) confined to small s_{f_2} area; right: $\hat{f}_2$ (dashed line) and 95% intervals (dotted lines) on diagonal slice of center frame. Circles and dotted lines are as in Figure 7.2. Solid line in right frame is the truth.

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: $Q_\lambda(\boldsymbol{s},\boldsymbol{t}) = E((f_\lambda(\boldsymbol{s}) - f(\boldsymbol{s}))(f_\lambda(\boldsymbol{t}) - f(\boldsymbol{t}))|Y = y)$ is given by:

$$\begin{aligned}\frac{1}{b}Q_\lambda(\boldsymbol{s},\boldsymbol{t}) &= (\phi_1(\boldsymbol{s}),\cdots,\phi_M(\boldsymbol{s}))(S'M^{-1}S)^{-1}\begin{pmatrix}\phi_1(\boldsymbol{t})\\ \vdots \\ \phi_M(\boldsymbol{t})\end{pmatrix}\\ &\quad -(\phi_1(\boldsymbol{s}),\cdots,\phi_M(\boldsymbol{s}))(S'M^{-1}S)^{-1}S'M^{-1}\begin{pmatrix}R(\boldsymbol{t},\boldsymbol{t}(1))\\ \vdots \\ R(\boldsymbol{t},\boldsymbol{t}(n))\end{pmatrix}\\ &\quad -(\phi_1(\boldsymbol{t}),\cdots,\phi_M(\boldsymbol{t}))(S'M^{-1}S)^{-1}S'M^{-1}\begin{pmatrix}R(\boldsymbol{s},\boldsymbol{t}(1))\\ \vdots \\ R(\boldsymbol{s},\boldsymbol{t}(n))\end{pmatrix}\end{aligned}$$

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

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),

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

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

$$\begin{aligned} & \phi'(\boldsymbol{s})[\eta I - \eta S'(\eta SS' + n\lambda I)^{-1}\eta S]\phi(\boldsymbol{t}) \\ - \quad & \eta\phi'(\boldsymbol{s})S'(\eta SS' + M)^{-1}R(\boldsymbol{t}) \\ - \quad & R(\boldsymbol{s})'(\eta SS' + M)^{-1}\eta S\phi(\boldsymbol{t}) \\ + \quad & R(\boldsymbol{s},\boldsymbol{t}) - R(\boldsymbol{s})'(\eta SS' + M)^{-1}R(\boldsymbol{t}). \end{aligned} \tag{A.3}$$

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

$$\lim_{\eta\to\infty} \eta I - \eta S'(\eta SS' + M)^{-1}S\eta \;\; = \;\; (S'M^{-1}S)^{-1}$$

$$\lim_{\eta\to\infty} \eta S'(\eta SS' + M)^{-1} = (S'M^{-1}S)^{-1}S'M^{-1}$$
$$\lim_{\eta\to\infty} (\eta SS' + M)^{-1} = M^{-1} - M^{-1}S(S'M^{-1}S)^{-1}S'M^{-1}. \quad \text{(A.4)}$$

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.

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., 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. D. (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.

de Boor, C. and Lynch, R. E. (1966). On splines and their minimum properties. *J. Math. Mech.*, **15** 953 – 969.

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.

Friedman, J. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1 – 141.

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

—— (1992). Penalized likelihood regression: A Bayesian analysis. *Statist. Sin.*, **2** 255 – 264.

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. (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.

—— (1991c). Smoothing Spline ANOVA with Component-Wise Bayesian "confidence intervals". Technical Report 881(rev.), Dept. of Statistics, University of Wisconsin, Madison.

—— (1993). Semiparametric ANOVA with tensor product thin plate splines. *J. Roy. Statist. Soc. Ser. B*, **55** 000 – 000. (Technical Report 90-61, Dept. of Statistics, Purdue University.)

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). *Generalized 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.

Moody, J. and Utans, R. (1991) Principled architecture selection for neural networks: Application to corporate bond rating prediction. In J. Moody, S. Hanson, and R. Lippman, editors, *Advances in Neural Information Processing Systems 4*. Kaufmann, San Mateo, 1991.

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.

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.

—— (1991). Multivariate Regression Splines. Technical Report 317, 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.

—— (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.