Variable Selection in Bayesian Smoothing Spline ANOVA Models: Application to Deterministic Computer Codes

Brian J. Reich

Department of Statistics
North Carolina State University
Raleigh, NC 27695
(*reich@stat.ncsu.edu*)

Curtis B. Storlie

Department of Mathematics and Statistics
University of New Mexico
Albuquerque, NM 87131

Howard D. Bondell

Department of Statistics
North Carolina State University
Raleigh, NC 27695

With many predictors, choosing an appropriate subset of the covariates is a crucial—and difficult—step in nonparametric regression. We propose a Bayesian nonparametric regression model for curve fitting and variable selection. We use the smoothing splines ANOVA framework to decompose the regression function into interpretable main effect and interaction functions, and use stochastic search variable selection through Markov chain Monte Carlo sampling to search for models that fit the data well. We also show that variable selection is highly sensitive to hyperparameter choice, and develop a technique for selecting hyperparameters that control the long-run false-positive rate. We use our method to build an emulator for a complex computer model for two-phase fluid flow.

KEY WORDS: Bayesian hierarchical modeling; Markov chain Monte Carlo; Nonparametric regression; Smoothing splines ANOVA; Variable selection.

1. INTRODUCTION

Nonparametric regression techniques have become a popular tool for analyzing complex computer model output. Consider, for example, a two-phase fluid flow simulation study (Vaughn et al. 2000) carried out by Sandia National Laboratory as part of the 1996 compliance certification application for a waste isolation pilot plant (WIPP) in New Mexico. The computer model simulates the waste panel's condition 10,000 years after the waste panel has been penetrated by a drilling intrusion. The simulation model uses several input variables describing various environmental conditions. The objectives are to predict waste pressure for new sets of environmental conditions and to determine which environmental factors have the greatest effect on the response. Because the simulation model is computationally intensive, we would like to develop an emulator—a statistical model to replicate the output of the complex computer model—to address these objectives.

The nonparametric regression model for response y_i is $y_i = \mu + f(x_{1i}, \ldots, x_{pi}) + \epsilon_i$, $i = 1, \ldots, N$, where μ is the intercept, f is the unknown function of covariates $x_{1i}, \ldots, x_{pi}$, and ϵ_i is error. With many predictors, choosing an appropriate subset of the covariates is a crucial—and difficult—step in fitting a nonparametric regression model. Several methods are available for curve fitting and variable selection for multiple nonparametric regression. Multivariate adaptive regression splines (MARS; Friedman 1991) is a stepwise procedure that selects variables and knots for a spline basis for each curve. But because stepwise selection is a discrete procedure, it can be unstable and highly sensitive to small changes in the data (Breiman 1995). Therefore, Lin and Zhang (2006) proposed the component selection and smoothing operator (COSSO) for smoothing spline analysis of variance models. The COSSO is a penalization technique for performing variable selection through continuous shrinkage of the norm of each of the functional components.

The Bayesian framework offers several potential advantages for nonparametric regression; for example, missing data and non-Gaussian likelihoods can be easily incorporated in the Bayesian model. Moreover, prediction is improved through Bayesian model averaging, and posterior model probabilities are natural measures of model uncertainty.

A common approach is to model computer output as a Gaussian process; for example, the "blind kriging" method of Joseph, Hung, and Sudjianto (2008) assumes that the response is the sum of a mean trend and a Gaussian process, and variable selection is performed on the mean trend, which is taken to be the sum of second-order polynomials and interactions. But all potential predictors are included in the Gaussian process covariance, and thus blind kriging does not perform any variable selection on the overall model.

In contrast, Linkletter et al. (2006) modeled the regression function f as a p-dimensional Gaussian process with covariance depending on the p covariates. They performed variable selection on the overall model using stochastic search variable

© 2009 American Statistical Association and the American Society for Quality
TECHNOMETRICS, MAY 2009, VOL. 51, NO. 2
DOI 10.1198/TECH.2009.0013

selection through Markov chain Monte Carlo (MCMC) sampling (e.g., Mitchell and Beauchamp 1988; George and McCulloch 1993; Chipman 1996; George and McCulloch 1997) to include/exclude variables from the covariance of the Gaussian process. Although this method of variable selection improves predictions for complex computer model output, interpreting the relative contribution of each covariate or groups of covariates to the p-dimensional fitted surface is difficult. In addition, the covariance function used results in a model that includes all higher-order functional interactions among the important predictors; that is, their model cannot reduce to an additive model in which the response surface is the sum of univariate functions. Therefore, the functional relationship between a predictor and the outcome is always dependent on the value of all of the other predictors included in the model. As a result, although this model is well suited for a complicated response surface, in many cases estimation and prediction can be improved by assuming a simpler model. For instance, Shively, Kohn, and Wood (1999) proposed a model for variable selection in additive nonparametric regression that takes an empirical Bayesian approach and gives each main effect function an integrated Brownian motion prior. Wood et al. (2002) extended the work of Shively, Kohn, and Wood (1999) to nonadditive models, again assuming integrated Brownian motion priors for the main effect functions and model interactions between predictors as two-dimensional surfaces with thin-plate spline priors. But interpreting the relative contributions of the main effect and interaction terms is difficult, because the spans of these terms overlap. To perform model selection, Wood et al. used data-based priors for the parameters that control the prior variance of the functional components. This allows for a Bayesian information criterion (BIC) approximation of the posterior probability of each model under consideration. This approach requires computing posterior summaries of all models under consideration, which is infeasible in situations with many predictors, especially when high-order interaction terms are considered. Gustafson (2000) also included a two-way interaction, but, to ensure identifiability, did not allow main effects in the model simultaneously with interactions and allowed predictors to interact with at most one other predictor. Because complex computer models often have many interaction terms, this is a significant limitation.

In this article we propose a Bayesian model for variable selection and curve fitting for nonparametric multivariate regression. Our model uses the functional analysis of variance framework (Wahba 1990; Wahba et al. 1995; Gu 2002) to decompose the function f into main effects f_j, two-way interactions f_{jk}, and so on, that is,

$$f(x_{1i}, \ldots, x_{pi}) = \sum_{j=1}^{p} f_j(x_{ji}) + \sum_{j<k} f_{jk}(x_{ji}, x_{ki}) + \cdots. \quad (1)$$

The functional ANOVA (BSS–ANOVA) is equipped with stochastic constraints that ensure that each of the components are identified so that their contribution to the overall fit can be studied independently. Rather than confining the regression functions to the span of a finite set of basis functions as in Bayesian splines, we use a more general Gaussian process prior for each regression function.

We perform variable selection using stochastic search variable selection through MCMC sampling to search for models that fit the data well. The orthogonality of the functional ANOVA framework is particularly important when the objective is variable selection. Assume, for example, that two variables have important main effects but their interaction is not needed. If the interaction is modeled haphazardly so that the span of the interaction includes the main effect spaces, then the inclusion probability possibly could be split between the model with main effects and no interaction and the model with the interaction alone, because both models can give the same fit. In this case, inclusion probabilities for the main effects and interaction could be less than $1/2$, and we would fail to identify the important terms. Because of the orthogonality, our model includes only interactions that explain features of the data that cannot be explained by the main effects alone. Moreover, due to the additive structure of our regression function, we are able to easily include categorical predictors, which is problematic for Gaussian process models (although Qian, Wu, and Wu 2008 suggested a way to incorporate categorical predictors into a Gaussian process model). Our model also is computationally efficient, because as we avoid enumerating all possible models and avoid inverting large matrixes at each MCMC iteration. We show that stochastic search variable selection can be sensitive to hyperparameter selection, and overcome this problem by specifying hyperparameters that control the long-run false-positive rate. We use Bayesian model averaging for prediction, which improves predictive accuracy.

The article is organized as follows. Sections 2 and 3 introduce the model. Section 4 describes the MCMC algorithm for stochastic search variable selection. Section 5 presents a brief simulation study comparing our model with other nonparametric regression procedures. Our Bayesian model compares favorably to MARS, COSSO, and the method of Linkletter et al. in terms of predictive performance and selecting important variables in the model. Section 6 analyzes the WIPP data, illustrating the advantages of the Bayesian approach in quantifying variable uncertainty. Section 7 concludes.

2. A BAYESIAN SMOOTHING SPLINE ANOVA MODEL

2.1 Simple Nonparametric Regression

For ease of presentation, we introduce the nonparametric model in the single-predictor case here and then extend it to the multiple-predictor case in Section 2.2. The simple nonparametric regression model is

$$y_i = \mu + f(x_i) + \epsilon_i, \quad (2)$$

where f is an unknown function of a single covariate $x_i \in [0, 1]$ and $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$. The regression function f is typically restricted to a particular class of functions. We consider the subset of Mth-order Sobolev space that includes only functions that integrate to zero and have M proper derivatives, that is, $f \in \mathcal{F}_M$, where

$$\mathcal{F}_M = \left\{ g \middle| g, \ldots, g^{(M-1)} \text{ are absolutely continuous}, \int_0^1 g(s)\, ds = 0, g^{(M)} \in L^2[0, 1] \right\}. \quad (3)$$

To ensure that each draw from f's prior is a member of this space, we select a Gaussian process prior with $\mathrm{cov}(f(s), f(t)) = \sigma^2\tau^2 K_1(s,t)$, with the kernel defined as

$$K_1(s,t) = \sum_{m=1}^{M+1} c_m B_m(s) B_m(t) + \frac{(-1)^M}{(2(M+1))!} B_{2(M+1)}(|s-t|),$$

$c_m > 0$ are known constants, and B_m is the mth Bernoulli polynomial. Wahba (1990) showed that each draw from this Gaussian process resides in $\mathcal{F}_M$ and that the posterior mode assuming $c_m = \infty$ for $m \le M$ and $c_{M+1} = 1$ is the $(M+1)$st-order smoothing spline. Steinberg and Bursztyn (2004) discussed an additional interpretation of this kernel, demonstrating that for the same choices of c_m, this Gaussian process model is equivalent to a Bayesian trigonometric regression model with diffuse priors for the low-order polynomial trends, a proper Gaussian prior for the $(M+1)$st-degree polynomial, and independent Gaussian priors for the trigonometric basis function's coefficients with variances depending on the frequencies of the trigonometric functions.

Throughout, we select $M = 1$ and set $c \equiv c_1 = \cdots = c_{M+1}$. Therefore, draws from the prior are continuously differentiable with path properties of integrated Brownian motion. As discussed in Section 3, performing variable selection requires that $c < \infty$. In this kernel, the term $K_P(s,t) = \sum_{m=1}^{M+1} B_m(s)B_m(t)$ controls the variability of the $(M+1)$st-degree polynomial trend, and $K_N(s,t) = \frac{(-1)^M}{(2(M+1))!} B_{2(M+1)}(|s-t|)$ is the stationary covariance of the deviation from the polynomial trend. In our analyses in Sections 5 and 6, the constant c is set to 100 to give vague, yet proper priors for the linear and quadratic trends. Thus, our model essentially fits a quadratic response surface regression plus a remainder term that is a mean-0 stationary Gaussian process constrained to be orthogonal to the quadratic trend. We have intentionally overparameterized with τ^2 and σ^2 for reasons that we make clear in Section 3.

2.2 Multiple Regression

The nonparametric multiple regression model for response y_i is $y_i = \mu + f(x_{1i}, \ldots, x_{pi}) + \epsilon_i$, where $x_{1i}, \ldots, x_{pi} \in [0,1]$ are covariates, $f \in \mathcal{F}$ is the unknown function, and $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$. To perform variable selection, we use the ANOVA decomposition of the space $\mathcal{F}$ into orthogonal subspaces, that is,

$$\mathcal{F} = \left\{ \bigoplus_{j=1}^{p} \mathcal{F}_j \right\} \oplus \left\{ \bigoplus_{k<l}^{p} (\mathcal{F}_k \otimes \mathcal{F}_l) \right\} \oplus \cdots, \tag{4}$$

where $\oplus$ the direct sum, $\otimes$ is the direct product, and each $\mathcal{F}_j$ is given by (3) (see Wahba 1990 or Gu 2002 for more detail). Assume that each f_j is a Gaussian process with covariance $\sigma^2\tau_j^2 K_1(x_{ji}, x_{ji'})$, and that each f_{kl} is a Gaussian process with covariance $\sigma^2\tau_{kl}^2 K_2(x_{ki}, x_{ki'}, x_{li}, x_{li'})$, where

$$\begin{aligned} K_2(x_{ki}, x_{ki'}, x_{li}, x_{li'}) &= (K_P(x_{ki}, x_{ki'}) + K_N(x_{ki}, x_{ki'})) \\ &\quad \times (K_P(x_{li}, x_{li'}) + K_N(x_{li}, x_{li'})) \\ &\quad + (c-1) K_P(x_{ki}, x_{ki'}) K_P(x_{li}, x_{li'}). \end{aligned} \tag{5}$$

For large c, the final term $(c-1)K_P(x_{ki}, x_{ki'})K_P(x_{li}, x_{li'})$ gives a vague prior to the low-order bivariate polynomial trend. Using this kernel, $f_j \in \mathcal{F}_j$ and $f_{kl} \in \mathcal{F}_k \otimes \mathcal{F}_l$. This ensures that each draw from this space will satisfy $\int_0^1 f_j(s)\,ds = 0$, $j = 1, \ldots, p$, to identify the intercept. This also identifies the main effects by forcing the interactions to satisfy $\int_0^1 f_{kl}(s,t)\,ds = \int_0^1 f_{kl}(s,t)\,dt = 0$, $k < l = 2, \ldots, p$. These constraints allow for a straightforward interpretation of each term's effect.

Higher-order interactions also can be included; however, these terms are difficult to interpret. Thus we combine all higher-order interactions into a single process. Let the higher-order interaction space be

$$\mathcal{F}_o = \mathcal{F} \cap \left\{ \bigoplus_{j=1}^{p} \mathcal{F}_j \oplus \bigoplus_{k<l}^{p} (\mathcal{F}_k \otimes \mathcal{F}_l) \right\}^{\mathcal{C}}, \tag{6}$$

where $A^{\mathcal{C}}$ is the compliment of A. The covariance of the Gaussian process $f_o \in \mathcal{F}_o$ is

$$\begin{aligned} &\mathrm{cov}(f_0(\mathbf{x}_i), f_0(\mathbf{x}_{i'})) \\ &\quad = \sigma^2\tau_0^2 \left[\prod_{j=1}^{p} (1 + K_1(x_{ji}, x_{ji'})) \right. \\ &\qquad \left. - 1 - \sum_{j=1}^{p} K_1(x_{ji}, x_{ji'}) - \sum_{j<k}^{p} K_2(x_{ji}, x_{ji'}, x_{ki}, x_{ki'}) \right]. \end{aligned} \tag{7}$$

Defining the covariance in this manner ensures that f_0 will be orthogonal to each main effect and interaction term.

The finite-dimensional model for the vector of observations $\mathbf{y} = (y_1, \ldots, y_n)'$ is

$$\mathbf{y} = \boldsymbol{\mu} + \sum_{j=1}^{p} \mathbf{f}_j(\mathbf{x}_j) + \sum_{k<l} \mathbf{f}_{kl}(\mathbf{x}_k, \mathbf{x}_l) + \mathbf{f}_0(\mathbf{x}_1, \ldots, \mathbf{x}_p) + \boldsymbol{\varepsilon}, \tag{8}$$

where $\boldsymbol{\mu} = (\mu, \ldots, \mu)'$ is the intercept, $\mathbf{x}_j = (x_{j1}, \ldots, x_{jn})'$ is a vector of observations for the jth covariate, $\mathbf{f}_j(\mathbf{x}_j)$ is the jth main effect function evaluated at the n observations, $\mathbf{f}_{kl}(\mathbf{x}_k, \mathbf{x}_l)$ is the vectorized interaction, $\mathbf{f}_0(\mathbf{x}_1, \ldots, \mathbf{x}_p)$ captures higher-order interactions, and $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2 I_n)$. We assume that the intercept μ has a flat prior and that $\sigma^2 \sim \mathrm{InvGamma}(a/2, b/2)$. The priors for the main effect and interaction functions are defined through the kernels in (4) as

$$\mathbf{f}_j(\mathbf{x}_j) \sim \mathrm{N}(0, \sigma^2\tau_j^2\Sigma_j), \tag{9}$$

$$\mathbf{f}_{kl}(\mathbf{x}_k, \mathbf{x}_l) \sim \mathrm{N}(0, \sigma^2\tau_{kl}^2\Sigma_{kl}), \tag{10}$$

$$\mathbf{f}_0(\mathbf{x}_1, \ldots, \mathbf{x}_p) \sim \mathrm{N}(0, \sigma^2\tau_0^2\Sigma_0), \tag{11}$$

where the (i, i') component of the covariance matrix Σ_j is $K_1(x_{ji}, x_{ji'})$, the (i, i') component of the covariance matrix Σ_{kl} is $K_2(x_{ki}, x_{ki'}, x_{li}, x_{li'})$, Σ_0 is defined similarly following (7), and τ_j, τ_{kl}, and τ_0 are unknown with priors given in Section 3. To help specify priors for τ_j, τ_{kl}, and τ_0, we rescale Σ_j, Σ_{kl}, and Σ_0 to have trace n. After this standardization, $\sigma\tau_j$ ($\sigma\tau_{kl}$) can be viewed as the typical prior standard deviation of an element of $\mathbf{f}_j$ ($\mathbf{f}_{kl}$).

2.3 Categorical Predictors

Complex models often have categorical variables that represent different states or point to different submodels to be used in the analysis. The BSS–ANOVA framework also is amendable to these unordered categorical predictors. Assume that $x_i \in \{1, 2, \ldots, G\}$ is categorical and that $f(x_i) = \theta_{x_i}$, where $\theta_g \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2\tau^2)$, $g = 1, \ldots, G$. To identify the intercept, we enforce the sum-to-zero constraint $\sum_{g=1}^{G} \theta_g = 0$. This model also can be written in the kernel framework by taking f to be a mean-0 Gaussian process with singular covariance $\mathrm{cov}(f(s), f(t)) = \sigma^2\tau^2 K_1(s,t)$, where the kernel is defined as $K_1(s,t) = K_N(s,t)$,

$$K_N(s,t) = \frac{G-1}{G} I(s=t) - \frac{1}{G} I(s \neq t), \tag{12}$$

and $I(\cdot)$ is the indicator function. Note that with unordered categorical predictors, we exclude the low-order polynomial trend, that is, $K_P(s,t) = 0$ for all s and t.

Interactions including categorical predictors with the kernel given in (12) are handled no differently than interactions between continuous predictors. Assume, for example, that $x_1 \in \{1, \ldots, G\}$ is categorical and $x_2 \in [0,1]$ is continuous. The kernel-based interaction is equivalent to the model $f_{1,2}(x_1, x_2) = h_{x_1}(x_2)$ for some $h_{x_1} \in \mathcal{F}_M$; that is, the effect of x_2 is different within each level of x_1. An attractive feature of this kernel is that it enforces the restrictions $\int h_{x_1}(x_2)\,dx_2 = 0$ for all $x_1 \in \{1, \ldots, G\}$ and $\sum_g h_g(x_2) = 0$ for all $x_2 \in [0,1]$ to separate the interaction from the main effects.

3. VARIABLE SELECTION

In variable selection, it is common to represent the subset of covariates included in the model with indicator variables γ_j and γ_{jk}, where γ_j is 1 if the main effect for $\mathbf{x}_j$ is in the model and 0 otherwise and γ_{jk} is 1 if the interaction for $\mathbf{x}_j$ and $\mathbf{x}_k$ is in the model and 0 otherwise. To avoid enumerating all possible models, stochastic search variable selection (e.g., Mitchell and Beauchamp 1988; George and McCulloch 1993; Chipman 1996; George and McCulloch 1997) assigns priors to the binary indicators and computes model probabilities using MCMC sampling. To perform variable selection in the nonparametric setting, we specify priors for the standard deviations τ_j and τ_{kl} in terms of indicators γ_j and γ_{kl}, to give priors with positive mass at 0. Given that τ_j (τ_{kl}) is 0 and c is finite, the curve f_j (f_{kl}) is equal to 0, and the term is removed from the model. This approach is slightly different than the original formulation of George and McCulloch (1993), who gave small (but nonzero) variance to negligible variables; in contrast to their approach, setting the variance precisely to 0 completely removes variables from the model.

Parameterization, identification, and prior selection for the hypervariances in Bayesian hierarchical models are notoriously problematic and are topics of active research. In a comparative study of several commonly used priors, Gelman (2006) recommended either a uniform or half-Cauchy prior on the standard deviation. Following this recommendation, we assume that $\tau_j = \gamma_j \eta_j$, where $\gamma_j \overset{\text{iid}}{\sim} \mathrm{Bern}(0.5)$ and $\eta_j \overset{\text{iid}}{\sim} \mathrm{HC}(\rho)$, where ρ is the median of the half-Cauchy prior. The interaction standard deviations τ_{kl} are modeled similarly.

Variable selection can be sensitive to the prior standard deviation. To illustrate the effect of the prior standard deviation on model selection, first consider the simpler case of multiple linear regression with orthogonal covariates, that is, $y_i = \sum_{j=1}^{p} \gamma_j X_{ij} \beta_j + \epsilon_i$, where $\mathbf{X}'\mathbf{X} = I_n$, $\gamma_j \overset{\text{iid}}{\sim} \mathrm{Bern}(0.5)$, $\boldsymbol{\beta} \sim \mathrm{N}(0, \sigma^2\tau^2 I_n)$, and $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2 I_n)$. Assuming that $\sigma^2 \sim \mathrm{InvGamma}(a/2, b/2)$, the marginal posterior log odds of $\gamma_j = 1$ are approximately

$$\log \frac{p(\gamma_j = 1 | \mathbf{y}, \tau)}{p(\gamma_j = 0 | \mathbf{y}, \tau)} \approx -\frac{1}{2}\log(1+\tau^2) + \frac{t_j^2 \tau^2}{2(1+\tau^2)}, \tag{13}$$

where $t_j^2 = \hat{\beta}_j^2 / \hat{\sigma}^2$, $\hat{\beta} = \mathbf{X}'\mathbf{y}$ is the least squares estimate of $\boldsymbol{\beta}$, and $\hat{\sigma}^2 = (\mathbf{y}'\mathbf{y} + b)/(n+a)$ is σ^2's posterior mode. If $\tau = 0$, then the log odds are 0 for any value of t_j^2; as τ goes to infinity, the log odds decline to negative infinity for any value of t_j^2. Therefore, choosing the prior of τ haphazardly can result in the influence of the data being completely overwhelmed by the prior standard deviation.

The subtle relationship between τ and the posterior of γ_j complicates the choice of a prior for τ that accurately depicts our prior model uncertainty. To alleviate this difficulty, we select priors for the standard deviations to give desirable long-run false-positive rates. The marginal log odds for the univariate nonparametric model given in Section 2.1 [analogous to (13) for linear regression] are approximately

$$\log \frac{p(\gamma = 1 | \mathbf{y}, \tau)}{p(\gamma = 0 | \mathbf{y}, \tau)} \approx -\frac{1}{2}\log|\tau^2\Sigma + I| + \frac{1}{2}\mathbf{y}'(\Sigma^{-1}/\tau^2 + I)^{-1}\mathbf{y}. \tag{14}$$

In Appendix A.1 we show that under the null distribution $\mathbf{y} \sim \mathrm{N}(0, \sigma I)$,

$$\mathrm{E}\left[\log\left(\frac{p(\gamma = 1 | \mathbf{y}, \tau)}{p(\gamma = 0 | \mathbf{y}, \tau)}\right)\right] \approx -n\tau^2, \tag{15}$$

where the expected value is taken with respect to $\mathbf{y}$. This suggests that the prior of τ should be scaled by $\sqrt{n}$; for example, we take $\tau \sim \mathrm{HC}(\lambda/\sqrt{n})$. This is similar to the unit-information prior of Kass and Wasserman (1995), which uses $\sqrt{n}$-scaling, and also to the approach of Ishwaran and Rao (2005), which uses $\sqrt{n}$-scaling for the Bayesian linear regression model to give desirable frequentist properties. It is important to note that because our prior depends on the sample size n, the procedure is not technically fully Bayesian; however, it could be easily modified to be fully Bayesian by incorporating reliable prior information for τ.

To select λ, we randomly generate 10,000 $\mathbf{y}$ for various n assuming that $\mathbf{y} \sim \mathrm{N}(0, I_n)$. For each simulated data set, we compute $E(\pi|\mathbf{y})$. Because it is common to select a variable if $E(\pi|\mathbf{y}) > 0.5$ (e.g., Barbieri and Berger 2004), Figure 1 shows the proportion of the 10,000 data sets that give $E(\pi|\mathbf{y}) > 0.5$ for each n and λ. After the prior of τ is tuned to depend on n, the false-positive rate remains stable for $n \geq 50$ and is around 0.05 for $\lambda = 2$. Although this result applies to the univariate model, we also use half-Cauchy priors with $\lambda = 2$ for each standard deviation in the multivariate model. The simulation study reported in Section 5 verifies that this prior controls the false-positive rate in the multiple-predictor setting as well, even in the presence of correlated predictors.

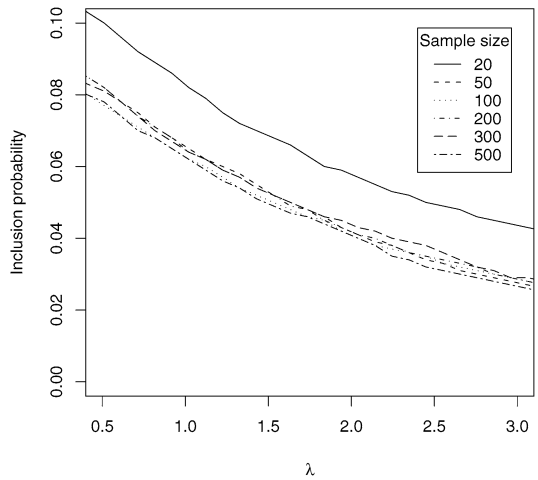

Figure 1. Plot of the probability (with respect to the null distribution of $\mathbf{y}$) that $E(\pi|\mathbf{y},\lambda) > 0.5$ by λ.

4. MCMC ALGORITHM

Here we present the algorithm used to draw MCMC samples from the posterior of the models defined in Sections 2 and 3. Gibbs sampling is used for μ and σ^2. The full conditionals for these parameters are

$$\mu|\text{rest} \sim N\big(\mathbf{1}'[\mathbf{y} - \mathbf{f}(\mathbf{x}_1,\ldots,\mathbf{x}_p)]/n, \sigma^2/n\big), \tag{16}$$

$$\sigma^2|\text{rest} \sim \text{InvGamma}\big(\big[n(p + p(p-1)/2 + 1) + a\big]/2, [SSE + SSM + b]/2\big), \tag{17}$$

where

$$SSE = (\mathbf{y} - \mathbf{f}(\mathbf{x}_1,\ldots,\mathbf{x}_p))'(\mathbf{y} - \mathbf{f}(\mathbf{x}_1,\ldots,\mathbf{x}_p)), \tag{18}$$

$$SSM = \sum_{j=0}^{p} f_j(\mathbf{x}_j)'\Sigma_j^{-1} f_j(\mathbf{x}_j)/\tau_j^2 + \sum_{k<l} f_{kl}(\mathbf{x}_k,\mathbf{x}_l)'\Sigma_{kl}^{-1} f_{kl}(\mathbf{x}_k,\mathbf{x}_l)/\tau_{kl}^2. \tag{19}$$

In cases with categorical predictors, the covariance matrixes are singular, and we use the generalized inverses.

Define all of the parameters in the model other than the jth main effect parameters $\mathbf{f}_j(\mathbf{x}_j)$ and τ_j as Θ_j. Draws from $p(\mathbf{f}_j(\mathbf{x}_j), \tau_j|\Theta_j)$ are made by first integrating over $\mathbf{f}_j(\mathbf{x}_j)$ and making a draw from $p(\tau_j|\Theta_j)$, and then sampling $\mathbf{f}_j(\mathbf{x}_j)$ given τ_j and Θ_j. Integrating over $\mathbf{f}_j(\mathbf{x}_j)$ gives

$$p(\tau_j|\Theta_j) \propto \begin{cases} \exp\left(-\dfrac{\mathbf{z}_j'\mathbf{z}_j}{2\sigma^2}\right) & \text{if } \tau_j = 0 \\ |\Delta_j|^{1/2}\exp\left(-\dfrac{\mathbf{z}_j'\Delta_j\mathbf{z}_j}{2\sigma^2}\right) g(\tau_j|\lambda) & \text{if } \tau_j > 0, \end{cases} \tag{20}$$

where $\Delta_j = I_n - (I_n + \Sigma_j^{-1}/\tau_j^2)^{-1}$, $\mathbf{z}_j = \mathbf{y} - \mathbf{f}(\mathbf{x}_1,\ldots,\mathbf{x}_p) + \mathbf{f}_j(\mathbf{x}_j)$, and $g(\tau_j|\lambda)$ is the half-Cauchy density function. Samples are drawn from $p(\tau_j|\Theta_j)$ using adaptive-rejective sampling, with candidates taken from the prior of τ_j. Note that we do not directly sample γ_j or η_j, but instead directly sample the standard deviation $\tau_j = \gamma_j\eta_j$, assuming its zero-inflated half-Cauchy prior. Given τ_j, the main effect curve has full conditional,

$$\mathbf{f}_j(\mathbf{x}_j)|\tau_j, \Theta_j \sim \text{N}(\Delta_j\mathbf{z}_j, \sigma^2\Delta_j), \tag{21}$$

and is updated using Gibbs sampling. We also use this approach to update the interaction curves.

Inverting the $n \times n$ matrix $I_n + \Sigma_j^{-1}/\tau_j^2$ at each MCMC iteration can be cumbersome for large data sets. Matrix inversion can be avoided by computing the spectral decomposition Σ_j outside of the MCMC algorithm. Let $\Sigma_j = \Gamma_j D_j \Gamma_j'$, where Γ_j is the $n \times n$ orthonormal eigenvector matrix and D_j is the diagonal matrix of eigenvalues $d_{j1} \geq \cdots \geq d_{jn}$. Then $\mathbf{f}_j(\mathbf{x}_j)$ can be updated by drawing $\mathbf{r}_j \sim \text{N}(0, \sigma^2[I_n + \tau_j^2 D_j]^{-1})$ and setting

$$\mathbf{f}_j = \Gamma[(I_n + \tau_j^2 D_j)^{-1}\Gamma'\mathbf{z}_j + \mathbf{r}_j]. \tag{22}$$

This sampling procedure requires inversion of only the diagonal matrix, $I_n + \tau_j^2 D_j$.

In practice, retaining all n eigenvector/eigenvalue pairs in the spectral decomposition of Σ_j may be unnecessary. A reduced model replaces $\Sigma_j = \Gamma_j D_j \Gamma_j'$ with $\Sigma_j^* = \Gamma_j^* D_j^* \Gamma_j^{*\prime}$, where Γ_j^* is the first K rows of Γ_j and D_j^* is the diagonal matrix with diagonal elements $d_{j1},\ldots,d_{jK}$. Analogous simplifications may be used for the interaction curves.

MCMC sampling is carried out in the freely available software package R (R Development Core Team 2006). We generate 20,000 samples from the posterior and discard the first 5,000 of these. Convergence is monitored by inspecting trace plots of the deviance and several of the variance parameters. For each MCMC iteration, our model is on the order of (number of terms in the model) $* K^2$; therefore, computation becomes increasingly time-consuming as the number of interactions grows. For the WIPP data reported in Section 6, running the two-way interaction model takes a few hours on an ordinary PC.

We compare models using the deviance information criterion (DIC) of Speigelhalter et al. (2002), defined as $DIC = \bar{D} + p_D$, where $\bar{D}$ is the posterior mean of the deviance, $p_D = \bar{D} - \hat{D}$ is the effective number of parameters, and $\hat{D}$ is the deviance evaluated at the the posterior mean of the parameters in the likelihood. This model's fit is measured by $\bar{D}$, and its complexity is captured by p_D. Models with smaller values of DIC are preferred.

5. SIMULATION STUDY

In this section we report a simulation study conducted to compare the BSS–ANOVA model described in Section 2 with MARS, COSSO, and the Gaussian process model of Linkletter et al. (2006). MARS analyses are done in R using the "polymars" function in the "polspline" package. The Gaussian process model of Linkletter et al. assumes that $\mathbf{f}(\mathbf{x}_1,\ldots,\mathbf{x}_p)$ is multivariate normal with mean μ and covariance

$$\text{cov}(f(x_{1i},\ldots,x_{pi}), f(x_{1i'},\ldots,x_{pi'})) = \tau^2 \prod_{j=1}^{p} \rho_j^{4(x_{ji}-x_{ji'})^2}. \tag{23}$$

The prior of the correlation parameter ρ_j is a mixture of a Uniform(0, 1) and a point mass at 1, with the point mass at 1 having prior probability 0.25. If $\rho_k = 1$, then $\mathbf{x}_k$ does not appear in the covariance and is essentially removed from the model.

5.1 Setting

We generate data assuming the underlying models given in Table 1. We use 50 simulated data sets for each simulation scenario. Following Li and Zhang (2006), we specify models using four building block functions (plotted in Figure 2):

- $g_1(t) = t$
- $g_2(t) = (2t-1)^2$
- $g_3(t) = \sin(2\pi t)/(2-\sin(2\pi t))$
- $g_4(t) = 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin^2(2\pi t) + 0.4\cos^3(2\pi t) + 0.5\sin^3(2\pi t)$.

The covariates $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are generated on the interval [0, 1] using three covariance structures: independence, compound symmetry (CS), and autoregressive (AR). For the independence case, the covariates are generated as independent Uniform(0, 1). To draw covariates with a compound symmetric covariance, we sample $\mathbf{w}_0, \ldots, \mathbf{w}_p$ as independent Uniform(0, 1) variables and define $\mathbf{x}_j = (\mathbf{w}_j + t\mathbf{w}_o)/(1+t)$, to give $\mathrm{cov}(\mathbf{x}_j, \mathbf{x}_j') = t^2/(1+t^2)$ for any pair (j, j'). We generate the AR covariates by sampling $\mathbf{w}_1, \ldots, \mathbf{w}_p$ as independent Normal(0, 1) variables and defining $\mathbf{x}_1 = \mathbf{w}_1$ and $\mathbf{x}_j = \rho\mathbf{x}_{j-1} + \sqrt{1-\rho^2}\mathbf{w}_j$ for $j > 1$. The covariates are trimmed on $[-2.5, 2.5]$ and scaled to [0, 1].

We compare the methods in terms of prediction accuracy and variable selection. For each data set and method, we compute

$$MSE = \frac{1}{1000}\sum_{j=1}^{1000}(f(z_{1j}, \ldots, z_{pj}) - \hat{f}(z_{1j}, \ldots, z_{pj}))^2, \quad (24)$$

where f is the true mean curve; $\hat{f}$ is the estimated value (the posterior mean, averaged over all models, for Bayesian methods); x_i, $i = 1, \ldots, n$, are the observed design points; and z_j, $j = 1, \ldots, 1000$, are unobserved locations drawn independently from the covariate distribution.

We also record the true-positive and false-positive rates for each model. The true- (false-) positive rate is computed by recording the proportion of the important (unimportant) variables included for each data set and averaging over all simulated data sets. A variable is deemed to be included in the BSS–ANOVA model if the posterior inclusion probability exceeds 0.5. Linkletter et al. (2006) used an added-variable method to select important variables at a given type I error level. For computational convenience, we assume that a covariate is in the model if the posterior median of ρ_k is <0.5. This gives a type I error of approximately 0.05 for the simulation reported later. We also tune the generalized cross-validation penalty of MARS to give a type I error near 0.05 (gcv = 2.5 for design 1 and gcv = 2 for designs 2 and 3).

Main effects–only models are used for design 1 for the MARS, COSSO, and BSS–ANOVA models; all possible two-way interactions are included as candidates for the other designs. The f_0 component is included for all BSS–ANOVA fits.

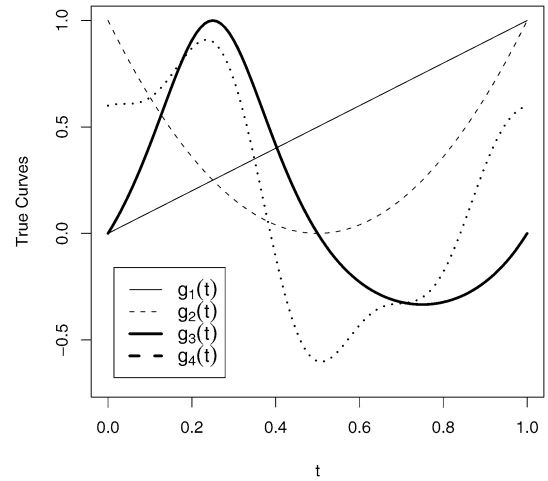

Figure 2. Plots of the true functions used in the simulation study.

5.2 Results

For each simulation design, the Bayesian mean squared error (MSE) is significantly smaller than the MSE for MARS and COSSO [Table 2(a)]. Although MARS is able is mimic many of the important features of the true curves, its piecewise linear fit does not match the smooth true curves shown in Figure 2. As was also shown by Lin and Zhang (2006), the COSSO improves on MARS. Although the fitted curves from COSSO are often similar to those of the Bayesian model, the Bayesian model achieves smaller MSE through model averaging. It also may be possible to improve the performance of the frequentist methods using non-Bayesian model averaging, such as bagging (Brieman 1996).

For each simulation design, the BSS–ANOVA model also maintains the nominal false-positive rate; for all simulations, 3%–10% of the truly uninformative variables are included in the model [Table 2(b)], supporting the choice of hyperparameters in Section 3. (Note that inclusion rates are not given for design 2, because it is does not fall within our BSS–ANOVA model, which does not include three-way interactions.) To further support the hyperparameter selection, we also simulated 50 data sets from the null model with $p = 10$ unimportant predictors and $\sigma = 2.28$ (not shown in Table 2). The false-selection rate was no more than 7.5% for independent, CS, or AR(1) covariates. Moreover, despite the potential effects of concurvity (Gu 1992, 2004), a nonparametric analog of multicollinearity, the BSS–ANOVA is able to identify truly important predictors at a high rate even with correlated predictors.

The BSS–ANOVA model also outperforms the method of Linkletter et al. for designs 1 and 3. These designs exclude some of the interactions involving the important main effects, and therefore the Linkletter full-interaction model is not appropriate. The method of Linkletter et al. does perform well for

Table 1. Simulation study design

Design	n	p	σ	$\mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_p)$
1	100	10	2.28	$5g_1(\mathbf{x}_1) + 3g_2(\mathbf{x}_2) + 4g_3(\mathbf{x}_3) + 6g_4(\mathbf{x}_4)$
2	100	4	2.28	$5g_1(\mathbf{x}_1) + 3g_2(\mathbf{x}_2) + 4g_3(\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3)$
3	100	6	2.28	$5g_1(\mathbf{x}_1) + 3g_2(\mathbf{x}_2) + 4g_3(\mathbf{x}_3) + 6g_4(\mathbf{x}_4) + 4g_3(\mathbf{x}_1\mathbf{x}_2)$

Table 2. PMSE is reported as the mean (standard error) over the simulated data sets for each simulation setting. The true- (false-) positive rate is computed by recording the proportion of the important (unimportant) variables included for each data set and then averaging over all simulated data sets. A variable is deemed to be included in the Bayesian models if the posterior inclusion probability exceeds 0.5

Design	Correlation of the predictors	MARS	COSSO	BSS–ANOVA	Linkletter
(a) Prediction MSE					
1	Ind	3.23 (0.28)	2.33 (0.13)	1.67 (0.08)	3.50 (0.12)
1	CS ($t = 1$)	7.60 (0.83)	6.08 (0.40)	4.11 (0.27)	7.39 (0.24)
1	AR ($\rho = 0.5$)	5.86 (0.44)	5.37 (0.33)	3.72 (0.18)	6.38 (0.22)
2	Ind	2.26 (0.08)	1.68 (0.04)	1.63 (0.04)	1.40 (0.05)
3	Ind	5.03 (0.38)	4.79 (0.23)	2.72 (0.09)	4.50 (0.08)
(b) Inclusion percentage for variables not in the true model					
1	Ind	0.04	0.06	0.03	0.03
1	CS ($t = 1$)	0.04	0.13	0.03	0.08
1	AR ($\rho = 0.5$)	0.03	0.12	0.05	0.06
2	Ind	–	–	–	–
3	Ind	0.04	0.13	0.10	0.11
(c) Inclusion percentage for variables in the true model					
1	Ind	0.78	0.91	0.91	0.81
1	CS ($t = 1$)	0.74	0.83	0.79	0.80
1	AR ($\rho = 0.5$)	0.75	0.82	0.78	0.80
2	Ind	–	–	–	–
3	Ind	0.67	0.77	0.82	0.89

NOTE: Results of the simulation study.

design 2, which includes a three-way interaction and no variables not included in the three-way interaction. This indicates that the method of Linkletter et al. is preferred if the response surface is a complicated function of high-order interactions between all of the significant predictors, whereas the proposed method is likely to perform well if the response surface is the sum of simple univariate and bivariate functions.

6. ANALYSIS OF THE WASTE ISOLATION PILOT PLANT DATA

In this section we analyze the WIPP data set mentioned in Section 1. The outcome variable of interest here is cumulative brine flow (m3) into a waste repository at 10,000 years for a drilling intrusion at 1000 years that penetrates the repository and an underlying region of pressurized brine (an E1 intrusion in the terminology of Helton et al. 2000). The four main pathways by which brine enters the repository are flow through the anhydrite marker beds, drainage from the disturbed rock zone, flow down the intruding borehole from overlying formations, and brine flow up the borehole from a pressurized brine pocket. There is a total of $n = 300$ observations, and we include $p = 11$ possible predictors. The predictors involved in the two-phase fluid flow model relate to various environmental conditions and are described briefly in Appendix A.2 and in detail by Vaughn et al. (2000). All of the predictors are continuous except the pointer variable for microbial degradation of cellulose (WMICDFLG), which has three levels: (a) no microbial degradation of cellulose, (b) microbial degradation only of cellulose, and (c) microbial degradation of cellulose, plastic, and rubber.

We compare the BSS–ANOVA model with two-way interactions with the model of Linkletter et al. (2006). Although incorporating categorical predictors in Linkletter et al.'s Gaussian process model is difficult, to facilitate the comparison, we order the three categories of microbial degradation of cellulose by their within-level mean response and treat the ordered variable as continuous. The inclusion probabilities given in Table 3 for the BSS–ANOVA model's main effects are fairly similar to those for the model of Linkletter et al. Five variables have posterior inclusion probabilities equal to 1.00 for both models: anhydrite permeability (ANHPRM), borehole permeability (BHPRM), bulk compressibility of the brine pocket (BPCOMP), halite porosity (HALPOR), and WMICDFLG. This set of important variables is consistent with previous analyses of this model using stepwise regression approaches (Helton et al. 2000; Storlie and Helton 2007).

The posterior mean curves from the BSS–ANOVA model for several predictors are plotted in Figure 3. Note that because of the BSS–ANOVA decomposition, the estimates of the main effect curves are interpretable by themselves; there is no need to numerically integrate over the other predictors, as in partial dependence plots (Hastie, Tibshirani, and Friedman 2001). The effects for BPCOMP and BHPERM are positive; increasing BPCOMP increases the amount of brine that leaves the brine pocket for each unit drop in pressure, and increasing BHPRM both reduces the pressure in the repository and reduces resistance to flow between the brine pocket and the repository. These effects result in increased brine flow into the repository through the borehole. Positive effects also are indicated for ANHPRM and HALPOR, which reduce resistance to flow in the anhydrite and halite, thereby increasing brine flow from the marker beds. Also note the flat effect from ANHPRM for the first half of its range, which occurs because it needs to exceed a threshold before the permeability becomes sufficient to counteract the pressure in the repository and allow for brine to flow from the marker beds. There is also an overall negative effect when moving from level 1 to level 2 and then to level 3 for the microbial

Table 3. Comparison of variable importance for the WIPP data set

	BSS–ANOVA		Linkletter	BSS–ANOVA best model
	Inc. Prob.	L2 norm	Inc. Prob.	L2 norm
ANHPRM	1.00	(0.43, 1.10)	1.00	(0.41, 0.89)
BHPERM	1.00	(1.59, 3.13)	1.00	(1.81, 2.88)
BPCOMP	1.00	(0.78, 1.67)	1.00	(0.83, 1.45)
BPPRM	0.08	(0.00, 0.03)	0.01	–
HALPOR	1.00	(0.56, 1.67)	1.00	(0.58, 1.09)
HALPRM	0.46	(0.00, 0.10)	0.03	–
SHPRMCLY	0.28	(0.00, 0.07)	0.01	–
SHPRMSAP	0.12	(0.00, 0.03)	0.07	–
SHPRNHAL	0.11	(0.00, 0.04)	0.00	–
SHRBRSAT	0.66	(0.00, 0.14)	0.03	(0.01, 0.13)
WMICDFLG	1.00	(0.66, 1.55)	1.00	(0.80, 1.57)
BPCOMP × WMICDFLG	1.00	(0.41, 0.99)	–	(0.44, 0.92)
BPCOMP × BHPERM	0.93	(0.00, 0.21)	–	(0.04, 0.34)
SHPRMSAP × WMICDFLG	0.85	(0.00, 0.18)	–	(0.02, 0.15)
SHRGSSAT × SHPRNHAL	0.60	(0.00, 0.10)	–	(0.01, 0.12)

NOTE: "Inc. Prob" refers to the posterior inclusion probability; "L2 norm" to the posterior 95% interval of $\int_0^1 f_j^2(s)\,ds$.

degradation flag (WMICDFLG), as shown in Figure 4(b); this is because the more microbial gas that is generated, the higher the repository pressure, which discourages brine inflow.

The inclusion probabilities for the remaining variables are <0.10 using the model of Linkletter et al. (2006). Our BSS–ANOVA model identifies an additional main effect—

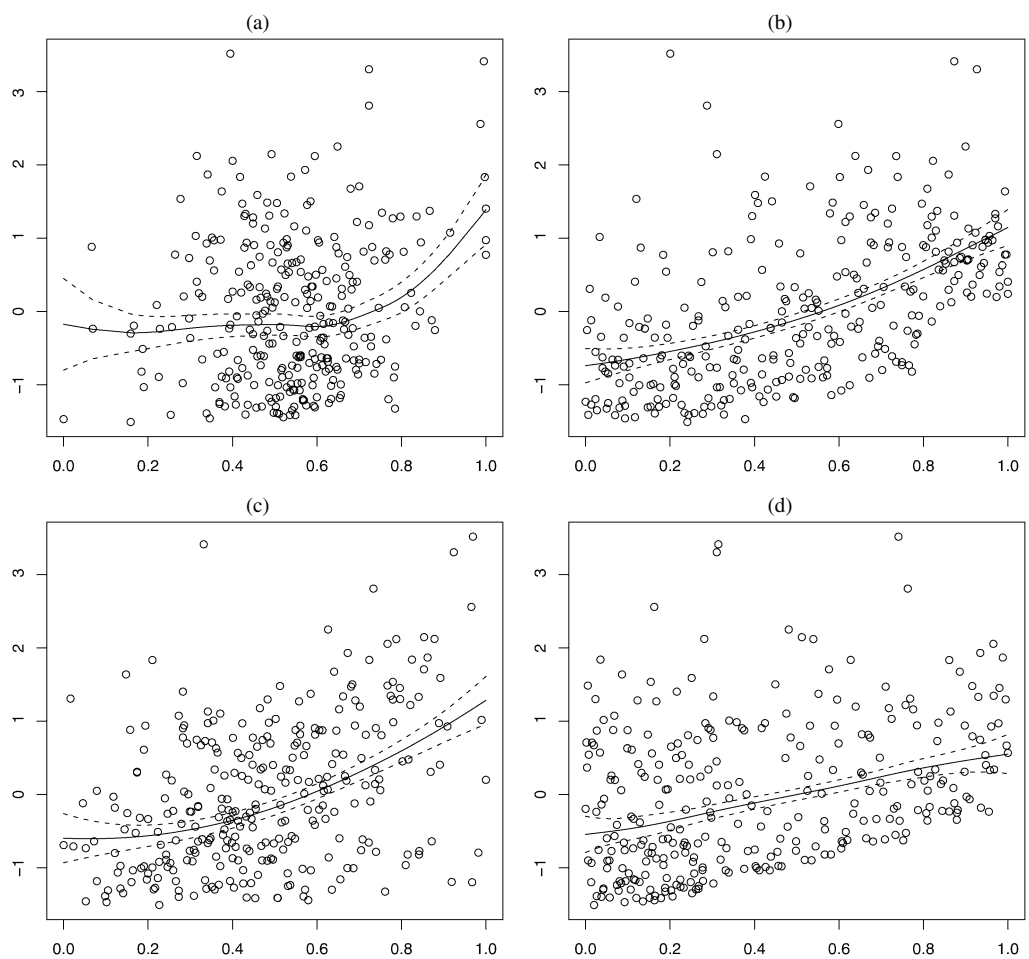

Figure 3. Raw data versus main effect curves [i.e., $f_j(x)$] for the WIPP data: (a) ANHPRM, (b) BHPERM, (c) BPCOMP, and (d) HALPOR. The solid lines are the medians, and the dashed lines are 95% intervals.

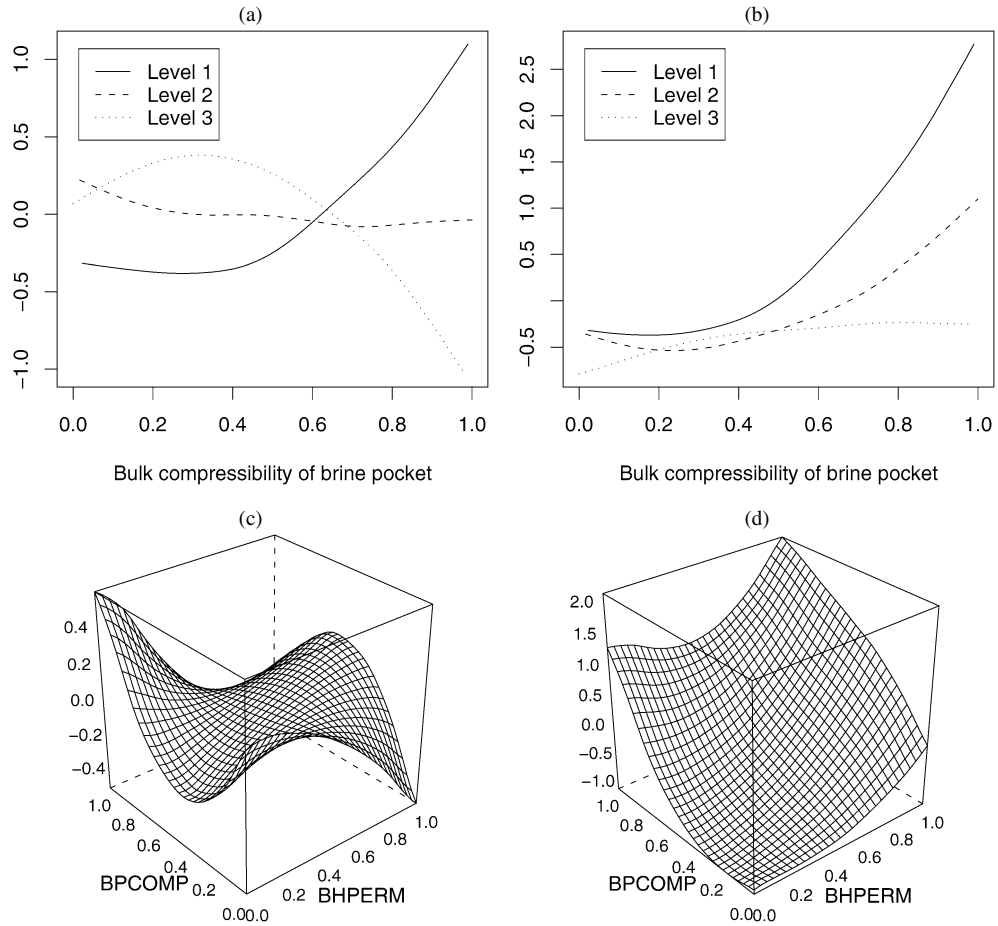

Figure 4. Interaction plots for BPCOMP × WMICDFLG [panels (a) and (b)] and BPCOMP × BHPERM [panels (c) and (d)]. Panels (a) and (c) give the posterior mean of $f_{jk}(x_j, x_k)$, and (b) and (d) give the posterior means of $f_j(x_j) + f_k(x_k) + f_{jk}(x_j, x_k)$.

residual brine saturation in the shaft (SHRBRSAT)—with inclusion probability 0.66. This association is somewhat surprising, because the shaft seals are quite effective, and so the flow is unlikely to go down the shaft. This is a topic of ongoing study.

The inclusion probabilities for the BSS–ANOVA model's main effects are equal to or greater than those of the model of Linkletter et al. for each predictor. This may be due to the fact that when a variable is included in the Gaussian process model, all interactions must be included, whereas the additive model can simply add a main effect curve. The posterior mean curves in Figure 3 are fairly smooth, suggesting that low-order polynomial fits are adequate. Because the priors for these low-order polynomials are vague under the BSS–ANOVA model, the model is able to essentially reduce to quadratic regression for these predictors. This fit is very different than that of Linkletter et al.'s model, which for most draws is a full Gaussian process in these five dimensions. For these data, DIC prefers the BSS–ANOVA decomposition ($DIC = 362$; $p_D = 68.4$) over that of Linkletter et al.'s model ($DIC = 375$; $p_D = 65.3$). Note that we do not use DIC for variable selection; rather, we use the Bayesian variable selection algorithm described in Section 3. We use DIC to compare the fits of the nonnested BSS–ANOVA and Gaussian process models, both of which average over several models defined by the binary inclusion indicators.

Of the 55 possible two-way interactions in the BSS–ANOVA model, 4 have an inclusion probability > 0.5 (Table 3). Moreover, the f_0 term for higher-order terms is included only 7% of the time. The interaction with the highest inclusion probability (1.00) is the interaction between BPCOMP and WMICDFLG. Figures 4(a) and (b) plot the fitted values (posterior mean, averaging over all models) of the interaction effect for this pair of predictors. Figure 4(a) clearly demonstrates the constraints of the BSS–ANOVA model for interactions. The curve for each level of WMICDFLG integrates to 0, and the sum of the three curves equals 0 for each value of BPCOMP. Figure 4(b) illustrates the sum of the interaction and main effect curves. It shows an increasing trend for BPCOMP for each level of WMICDFLG; however, the trend is nearly flat when WMICDFLG equals level 3, implying microbial degradation of cellulose, plastic, and rubber. This is a reasonable implication, because the gas produced by the degradation could produce sufficient pressure to make brine inflow negligable for this range of BPCOMP values. Figures 4(c) and (d) plot the fitted values for the interaction between BPCOMP and BHPERM, which has the second-largest inclusion probability (0.93). Note that in the upper right corner, the interaction indicates a decrease in brine inflow from the additive effects. This is very interesting, because at high values of BHPERM and BPCOMP, so much brine flows down the borehole that the repository saturates and rises

to hydrostatic pressure, which reduces brine inflow from the brine pocket. Although these important interactions have not been studied in previous analyses of this problem, they could provide insight into and/or confirmation of models.

We measure variable importance with the posterior 95% interval of the L2 norms of f_j and f_{kl}, $\int_0^1 f_j(s)^2\,ds$ and $\int_0^1\int_0^1 f_{kl}(s_1, s_2)^2\,ds_1\,ds_2$, respectively. The L2 norms are proportional to the proportion of variation in the model explained by each term. We approximate these integrals by taking the sum at the $n = 300$ design points. The L2 norm intervals given in Table 3 show that of the predictors included with probability 1.00, BHPRM generally explains the largest proportion of the variance in the fitted function. In addition, even though there are interactions with probability > 0.5, these terms explain less variation in the fitted surface than the important main effects. This sensitivity analysis accounts for variable selection uncertainty; that is, the L2 norm is computed at every MCMC iteration, even at iterations that exclude the variable. Another common approach to sensitivity analysis is to first select the important variables, and then compute the L2 norms for the important variables using the model including only the selected variables. To illustrate how these approaches differ, we refit the BSS–ANOVA model using only the variables with inclusion probability > 0.5. The resulting L2 norms are given in Table 3 (under "BSS–ANOVA best model"). The intervals for this model are generally narrower than the intervals from the full model, demonstrating that accounting for variable selection uncertainty in sensitivity analysis gives wider, more realistic posterior intervals.

In Section 3 we develop a method for selecting the hyperparameter λ, which controls the strength of the priors of the variances. Based on these results, we recommend using $\lambda = 2$. But for these data, the posterior inclusion probabilities are robust to the selection of λ. Consequently, we refit the model with $\lambda \in \{1, 2, 3\}$; the posterior mean number of variables in the model [i.e., the posterior mean of $\sum_{j=0}^{p} I(\tau_j > 0) + \sum_{k<l} I(\tau_{kl} > 0)$] is 16.4 with $\lambda = 1$, 14.8 with $\lambda = 2$, and 13.4 with $\lambda = 3$. In addition, for all three choices of λ the same subset of terms with inclusion probability > 0.5 are identified, with the sole exception that HALPRM is included in the model with inclusion probability 0.51 with $\lambda = 1$, compared with 0.46 with $\lambda = 2$.

7. DISCUSSION

In this article we present a fully Bayesian procedure for variable selection and curve fitting for nonparametric regression. Our model uses the smoothing splines ANOVA decomposition and selects components through stochastic search variable selection. We tune the model to have a desired false-positive rate. Our simulation study demonstrates that the Bayesian model has advantages over other nonparametric variable selection models in terms of both prediction accuracy and variable selection. The model is used to build an emulator for complex computer model output.

Another challenge in the analysis of complex computer model output is to jointly model computer model output and actual field data. A common approach to this is to model both the true response and the bias between field and simulated data with separate Gaussian processes. In this case, our approach could be used to identify important variables for both Gaussian processes, that is, to identify conditions that affect the true process and identify potentially different variables that predict a discrepancy between simulated and real data. Moreover, although here we applied our method to the deterministic WIPP model, our simulation study suggests that our BSS–ANOVA model also is adept in estimating the mean response for data with random errors.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Science Foundation (DMS-0354189 to B.R.; DMS-0705968 to H.D.) and Sandia National Laboratories (SURP Grant 22858). The authors thank Dr. Hao Zhang of North Carolina State University for providing code to run the COSSO model, and Jon Helton for his help with the analysis of the two-phase flow model. The authors also thank the reviewers, associate editor, and co-editors for their most constructive comments, many of which are incorporated in the current version of this article.

APPENDIX

A.1 Approximate Expected log Odds of $\pi = 1$

For the univariate nonparametric model presented in Section 2.1, integrating over $\mathbf{f}$ and σ^2 gives

$$\frac{p(\gamma = 1|\mathbf{y}, \tau)}{p(\gamma = 0|\mathbf{y}, \tau)} = |\tau^2\Sigma + I|^{-1/2}\left(1 - \frac{\mathbf{y}'(\Sigma^{-1}/\tau^2 + I)^{-1}\mathbf{y}}{\mathbf{y}'\mathbf{y} + b}\right)^{-(n+a)/2}. \quad (25)$$

Assuming that the data are standardized, so that $\mathbf{y}'\mathbf{y} = n$, and assuming that $a = b$, for large n, we have

$$\log\frac{p(\gamma = 1|\mathbf{y}, \tau)}{p(\gamma = 0|\mathbf{y}, \tau)} \approx -\frac{1}{2}\log|\tau^2\Sigma + I| + \frac{1}{2}\mathbf{y}'(\Sigma^{-1}/\tau^2 + I)^{-1}\mathbf{y}. \quad (26)$$

Taking the expected value with respect to $\mathbf{y} \sim \mathrm{N}(0, I)$ gives

$$\mathrm{E}\left[\log\frac{p(\gamma = 1|\mathbf{y}, \tau)}{p(\gamma = 0|\mathbf{y}, \tau)}\right] = -\frac{1}{2}\log|\tau^2\Sigma + I| + \frac{1}{2}\operatorname{trace}[(\Sigma^{-1}/\tau^2 + I)^{-1}] \quad (27)$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\log(1 + \tau^2 d_i) + \frac{1}{2}\sum_{i=1}^{n}\frac{\tau^2 d_i}{1 + \tau^2 d_i}, \quad (28)$$

where $d_1, \ldots, d_n$ are the eigenvalues of Σ. Recalling that Σ is scaled so that $\operatorname{trace}(\Sigma) = \sum_{i=1}^{n} d_i = n$, a first-order Taylor series at $\tau^2 = 0$ gives $\mathrm{E}[\log\frac{p(\gamma=1|\mathbf{y},\tau)}{p(\gamma=0|\mathbf{y},\tau)}] \approx -n\tau^2$.

A.2 Variable Descriptions for the Two-Phase Fluid Flow Example

ANHPRM: Logarithm of anhydrite permeability (m^2).

BHPRM: Logarithm of borehole permeability (m^2).

BPCOMP: Bulk compressibility of the brine pocket (Pa^{-1}).

BPPRM: Logarithm of intrinsic brine pocket permeability (m^2).

HALPOR: Halite porosity (dimensionless).

HALPRM: Logarithm of halite permeability (m^2).

SHPRMSAP: Logarithm of permeability (m^2) of the asphalt component of the shaft seal (m^2).

SHPRMCLY: Logarithm of permeability (m^2) for the clay components the of shaft.

SHPRMHAL: Pointer variable (dimensionless) used to select permeability in the crushed salt component of the shaft seal at different times.

SHRBRSAT: Residual brine saturation in the shaft (dimensionless).

WMICDFLG: Pointer variable for microbial degradation of cellulose. WMICDFLG = 1, 2, and 3 implies no microbial degradation of cellulose; microbial degradation of only cellulose; and microbial degradation of cellulose, plastic, and rubber.

[Received June 2007. Revised August 2008.]

REFERENCES

Barbieri, M., and Berger, J. (2004), "Optimal Predictive Model Selection," *Annals of Statistics*, 32, 870–897.

Brieman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

——— (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140.

Chipman, H. (1996), "Bayesian Variable Selection and Related Predictors," *Canadian Journal of Statistics*, 24, 17–36.

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

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models" (with comment by Browne and Draper), *Bayesian Analysis*, 1, 515–534.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

——— (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Gu, C. (1992), "Diagnostics for Nonparametric Regression Models With Additive Terms," *Journal of the American Statistical Association*, 87, 1051–1058.

——— (2002), *Smoothing Spline ANOVA Models*, New York: Springer.

——— (2004), "Model Diagnostics for Smoothing Spline ANOVA Models," *The Canadian Journal of Statistics*, 32, 347–358.

Gustafson, P. (2000), "Bayesian Regression Modeling With Interactions and Smooth Effects," *Journal of the American Statistical Association*, 95, 745–763.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

Helton, J. C., Bean, J. E., Economy, K., Garner, J. W., MacKinnon, R. J., Miller, J., Schreiber, J. D., and Vaughn, P. (2000), "Uncertainty and Sensitivity Analysis for Two-Phase Flow in the Vicinity of the Repository in the 1996 Performance Assessment for the Waste Isolation Pilot Plant: Disturbed Conditions," *Reliability Engineering and System Safety*, 69, 263–304.

Ishwaran, H., and Rao, J. S. (2005), "Spike and Lab Variable Selection: Fand Bayesian Strategies," *Annals of Statistics*, 33, 730–773.

Joseph, V. R., Hung, Y., and Sudjianto, A. (2008), "Blind Kriging: A New Method for Developing Metamodels," *Journal of Mechanical Design*, 130, 031102.

Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

Lin, Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models," *Annals of Statistics*, 34, 2272–2297.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., and Ye, K. Q. (2006), "Variable Selection for Gaussian Process Models in Computer Experiments," *Technometrics*, 48, 478–490.

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression" (with discussion), *Journal of the American Statistical Association*, 83, 1023–1036.

Qian, P. Z., Wu, H., and Wu, C. F. J. (2008), "Gaussian Process Models for Computer Experiments With Qualitative and Quantitative Factors," *Technometrics*, 50, 383–396.

R Development Core Team (2006), "R: A Language and Environment for Statistical Computing," available at *http://www.R-project.org*.

Shively, T., Kohn, R., and Wood, S. (1999), "Variable Selection and Function Estimation in Nonparametric Regression Using a Data-Based Prior" (with discussion), *Journal of the American Statistical Association*, 94, 777–806.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion and rejoinder), *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.

Steinberg, D. M., and Bursztyn, D. (2004), "Data Analytic Tools for Understanding Random Field Regression Models," *Technometrics*, 46, 411–420.

Storlie, C. S., and Helton, J. C. (2007), "Multiple Predictor Smoothing Methods for Sensitivity Analysis: Example Results," *Reliability Engineering and System Safety*, 93, 55–77.

Vaughn, P., Bean, J. E., Helton, J. C., Lord, M. E., MacKinnon, R. J., and Schreiber, J. D. (2000), "Representation of Two-Phase Flow in the Vicinity of the Repository in the 1996 Performance Assessment for the Waste Isolation Pilot Plant," *Reliability Engineering and System Safety*, 69, 205–226.

Wahba, G. (1990), *Spline Models for Observational Data. CBMS–NSF Regional Conference Series in Applied Mathematics*, Vol. 59, Philadelphia, PA: SIAM.

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995), "Smoothing Spline ANOVA for Exponential Families, With Application to the WESDR," *Annals of Statistics*, 23, 1865–1895.

Wood, S., Kohn, R., Shively, T., and Jiang, W. (2002), "Model Selection in Spline Nonparametric Regression," *Journal of the Royal Statistical Society*, Ser. B, 64, 119–140.