

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

By: Brian J. Reich<sup>1</sup>, Curtis B. Storlie<sup>2</sup>, and Howard D. Bondell<sup>1</sup>

<sup>1</sup> North Carolina State University

<sup>2</sup> University of New Mexico

E-mail:reich@stat.ncsu.edu

# Motivation

- ▶ A common problem for statisticians is a sensitivity analysis of a complex computer model.
- ▶ For example, we consider a two-phase fluid flow simulation study carried out by Sandia National Labs
- ▶ 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 largest effect on the response.

# NP regression for complex computer models

- ▶ Because computer models are governed by complex differential equations, linear regression is not an adequate approach for these data.
- ▶ The standard approach is nonparametric regression:  
$$y = f(x_1, \dots, x_p) + \epsilon.$$
- ▶  $f : \mathcal{R}^p \rightarrow \mathcal{R}$  is an unknown function that relates the predictors to  $y$ 's mean.
- ▶ Typically  $f$  is modeled as a Gaussian process.
- ▶  $\epsilon \sim N(0, \sigma^2)$  is error.

## SS-ANOVA models

- ▶ Modeling  $f$  as a Gaussian process is very flexible, but from this model it can be difficult to identify the effect of an individual covariate.
- ▶ When the goal is to identify the effect of each covariate on the response a more structured model may be preferred.
- ▶ The smoothing splines ANOVA (SS-ANOVA) model decomposes the  $p$ -dimensional surface  $f$  into the sum of one-dimensional main effect curves main effects  $f_j$ , two-dimensional interaction curves  $f_{lk}$ , etc.

$$f(x_1, \dots, x_p) = \beta_0 + \sum_{j=1}^p f_j(x_j) + \sum_{l < k} f_{lk}(x_l, x_k) + \dots$$

- ▶ Our goal is to develop a computationally efficient method for selecting important main effects and interactions when  $p$  is fairly large.

In this talk we will...

- ▶ Develop a stochastic search variable selection model to identify important main effect and interaction terms in the SS-ANOVA framework.
- ▶ Propose a method for selecting prior to give desirable long-run properties.
- ▶ Apply our approach to the Sandia Labs data.

## Model for the main effect curves

- ▶ The regression functions  $f_j$  is typically restricted to a particular class of functions.
- ▶ We consider the subset of  $M^{\text{th}}$ -order Sobolev space that includes only functions that integrate to zero and have  $M$  proper derivatives, i.e.,  $f_j \in \mathcal{F}_M$  where

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

- ▶ We restrict the main effects to integrate to zero to identify the overall intercept  $\beta_0$ .
- ▶ We select  $M = 1$  so that draws from the prior are continuously differentiable with path properties of integrated Brownian motion.

## Prior for $f_j$

- ▶ We assume  $f_j$  is a Gaussian process with  $E(f_j(s)) = 0$  and covariance  $\text{Cov}(f_j(s), f_j(t)) =$

$$\sigma^2 \tau_j^2 \left[ \sum_{m=1}^2 c_m B_m(s) B_m(t) + \frac{1}{4!} B_5(|s - t|) \right],$$

where  $c_m > 0$  are known constants and  $B_m$  is the  $m^{\text{th}}$  Bernoulli polynomial.

- ▶ The first term  $\sum_{m=1}^2 c_m B_m(s) B_m(t)$  controls the covariance of the quadratic trend.
- ▶ The second term controls the covariance of the deviance from the quadratic trend.

## Prior for $f_j$ (cont.)

- ▶ Complex models often have categorical variables that represent different states or point to different submodels to be used in the analysis.
- ▶ Assume  $x_j \in \{1, 2, \dots, G\}$  is categorical and  $f(x_j) = \theta_{x_j}$ , where  $\theta_g \stackrel{iid}{\sim} N(0, \sigma^2 \tau_j^2)$ ,  $g = 1, \dots, G$ .
- ▶ To identify the intercept we enforce the sum-to-zero constraint  $\sum_{g=1}^G \theta_g = 0$ .
- ▶ This model can also be written in the kernel framework by taking  $f$  to be a mean-zero Gaussian process with singular covariance

$$\text{cov}(f(s), f(t)) = \sigma^2 \tau_j^2 \left[ \frac{G-1}{G} I(s=t) - \frac{1}{G} I(s \neq t) \right]$$

## Prior for $f_j$ (cont.)

- ▶ We take  $c_m = 100$  to give a vague prior for the quadratic trend.
- ▶  $\tau_j^2$  controls the overall variance, relative to error variance  $\sigma^2$ .
- ▶ Interactions are model similarly.
- ▶  $f_{lk}$  has mean zero and covariance proportional to  $\sigma^2 \tau_{lj}^2$
- ▶  $f_{lk}$ 's covariance places a vague prior on the second-order interaction trends.
- ▶ We do not include higher-order interactions.
- ▶ Interactions with categorical predictors are handled no different than other interactions.

# Stochastic search variable selection (SSVS)

- ▶ One way to do variable selection would be to fit all possible models and pick the one with the smallest AIC, BIC, DIC, etc.
- ▶ If  $p = 11$  there are 11 main effects, 55 interactions, and  $2^{66}$  possible models.
- ▶ We use stochastic search variable selection to avoid enumerating all possible models.
- ▶ SSVS treats the models as a random variable and uses MCMC to search for model that fit the data well.

# SSVS for linear regression

- ▶ Model:  $y_i = \beta_0 + \sum_{j=1}^p x_j \beta_j + \text{error}$ , where:
  - ▶  $\beta_j = \gamma_j \alpha_j$
  - ▶  $\gamma_j \sim \text{Bern}(\pi_j)$
  - ▶  $\alpha_j \sim \text{N}(0, \tau_j)$
- ▶  $(\gamma_1, \dots, \gamma_p)'$  is the “model”, and it treated as an unknown random variable.
- ▶ The prior probability that  $x_j$  is included in the model is  $P(\beta_j \neq 0) = \pi_j$ .
- ▶ Inference is based on the posterior probability that  $x_j$  is included in the model,  $P(\beta_j \neq 0|y)$ .
- ▶ It is common to pick the final model to include all variables with  $P(\beta_j \neq 0|y) > 0.5$ .

# SSVS for nonparametric regression

- ▶ How do we extend this idea to nonparametric regression?
- ▶ The main effect curve  $f_j$  is equal to zero everywhere and removed from the model if  $\tau_j^2 = 0$ .
- ▶ So we use a mixture prior for the standard deviations.
- ▶  $\tau_j = \gamma_j \eta_j$  where  $\gamma_j \sim \text{Bern}(0.5)$  and  $\eta_j \sim \text{Half Cauchy}(\rho)$ .
- ▶ If  $\gamma_j = 0$  then  $f_j$  is removed from the model.
- ▶ If  $\gamma_j = 1$  then  $f_j$  is included in the model and its prior median is  $\rho$ .
- ▶ Interactions are model similarly.

# Tuning the prior

- ▶ Results can be sensitive to the hyperprior  $\rho$ .
- ▶ To alleviate this issue, we select priors for the standard deviations to give desirable long-run false positive rates.
- ▶ Consider the single-predictor model  $\mathbf{y} \sim N(\mathbf{f}, \sigma^2 I)$  with  $\mathbf{f} \sim N(0, \sigma^2 \gamma \eta \Sigma)$ , and  $\gamma \sim \text{Bern}(0.5)$ .
- ▶ Then the posterior log odds of including  $\mathbf{f}$  in the model are

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

- ▶ Under the null distribution  $\mathbf{y} \sim N(0, \sigma^2 I)$ ,

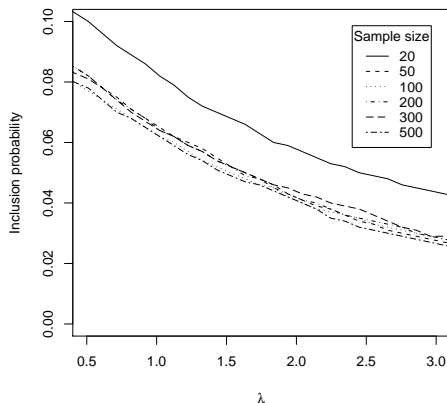
$$E \left[ \log \left( \frac{p(\gamma = 1 | \mathbf{y}, \eta)}{p(\gamma = 0 | \mathbf{y}, \eta)} \right) \right] \approx -m\eta^2,$$

# Root- $n$ scaling

- ▶ Therefore we take  $\eta \sim \text{HC}(\lambda/\sqrt{n})$ .
- ▶ Kass and Wasserman (1995) and Ishwaran and Rao (2005) also use  $\sqrt{n}$ -scaling for the Bayesian linear regression model to give desirable frequentist properties.
- ▶ How to pick  $\lambda$ ?
- ▶ We randomly generate 10,000  $\mathbf{y}$  under the null for various  $n$ .
- ▶ The next slide shows the proportion of data sets with  $E(\pi|\mathbf{y}) > 0.5$  for each  $\lambda$ .
- ▶ This leads us to select  $\lambda = 2$ .
- ▶ We also use  $\lambda = 2$  for multiple regression.

# Calibration plot

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



## Application to the Sandia Labs data

- ▶ The outcome variable of interest here is cumulative brine flow (m<sup>3</sup>) into waste repository at 10,000 years for a drilling intrusion at 1000 year that penetrates the repository and an underlying region of pressurized brine.
- ▶ There are  $n = 300$  observations and we include  $p = 11$  possible predictors.
- ▶ The predictors involved in the Two-Phase Fluid Flow model describe various environmental conditions.
- ▶ 10 predictors are continuous, one predictors has three categories.

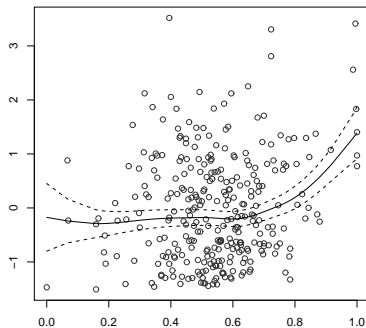
# Posterior inclusion probabilities

Main effect	$P(\mathbf{f}_j \neq 0)$
Anhydrite permeability	1.00
Borehole permeability	1.00
Bulk compressibility of brine pocket	1.00
Halite porosity	1.00
Microbial degradation of cellulose	1.00
Residual brine saturation in shaft	0.66
Halite permeability	0.46
Permeability of asphalt component of shaft seal	0.28
Permeability for clay components of shaft	0.12
Permeability in crushed salt component of shaft seal	0.11
Intrinsic brine pocket permeability	0.08

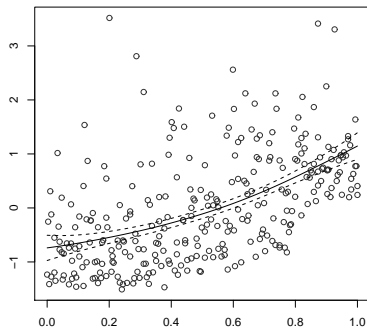
# Posterior 95% interval of L2-norm = $\int_0^1 f_j^2(s) ds$

Main effect	L2-norm
Anhydrite permeability	(0.43, 1.10)
Borehole permeability	(1.59, 3.13)
Bulk compressibility of brine pocket	(0.78, 1.67)
Halite porosity	(0.56, 1.67)
Microbial degradation of cellulose	(0.66, 1.55)
Residual brine saturation in shaft	(0.00, 0.14)
Halite permeability	(0.00, 0.10)
Permeability of asphalt component of shaft seal	(0.00, 0.07)
Permeability for clay components of shaft	(0.00, 0.03)
Permeability in crushed salt component of shaft seal	(0.00, 0.04)
Intrinsic brine pocket permeability	(0.00, 0.03)

# Data vs posterior mean main effect curves $f_j$



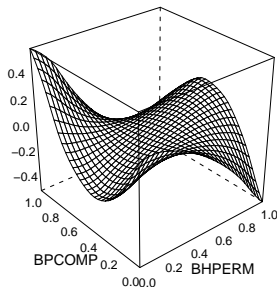
Anhydrite permeability



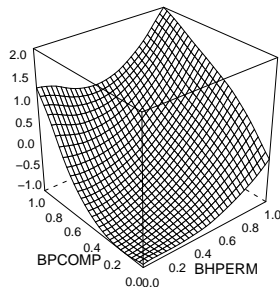
Borehole permeability

# Interaction plots

There were also three interactions included at least half of the time, including the interaction between compressibility of brine pocket (BPCOMP)  $\times$  intrinsic brine pocket permeability (BHPERM).



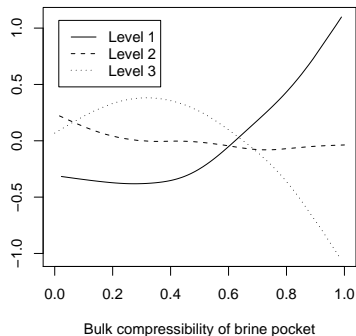
Without main effects



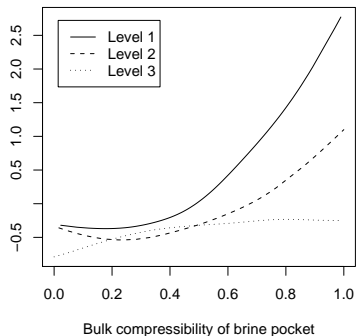
With main effects

## Interaction plots (cont.)

One of the significant interactions involved the categorical predictor microbial degradation of cellulose and the continuous predictor bulk compressibility of brine pocket.



Without main effects



With main effects

- ▶ Our model for variable selection in nonparametric regression identifies important main effect and interaction terms in a computationally-efficient manner using SSVS.
- ▶ This model can also handle categorical predictors.
- ▶ Motivated by an application with the EPA, we are currently extending this approach to *stochastic* computer models.
- ▶ Stochastic computer models return a distribution, rather than a scalar, for each set of inputs.
- ▶ Here we embed SSVS in Bayesian density estimation.
- ▶ **Special thanks to David Steinberg and *Technometrics* for the invitation to speak.**