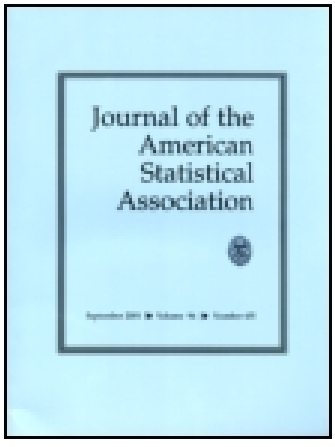

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior

Thomas S. Shively [a] , Robert Kohn [b] & Sally Wood [b]
[a] Department of Management Science and Information Systems , University of Texas , Austin , TX , 78712
[b] Australian Graduate School of Management , University of New South Wales , Sydney , 2052 , Australia
Published online: 17 Feb 2012.

To cite this article: Thomas S. Shively , Robert Kohn & Sally Wood (1999) Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior, Journal of the American Statistical Association, 94:447, 777-794

To link to this article: http://dx.doi.org/10.1080/01621459.1999.10474180

Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior

Thomas S. SHIVELY, Robert KOHN, and Sally WOOD

A hierarchical Bayesian approach is proposed for variable selection and function estimation in additive nonparametric Gaussian regression models and additive nonparametric binary regression models. The prior for each component function is an integrated Wiener process resulting in a posterior mean estimate that is a cubic smoothing spline. Each of the explanatory variables is allowed to be in or out of the model, and the regression functions are estimated by model averaging. To allow variable selection and model averaging, data-based priors are used for the smoothing parameter and the slope at 0 of each component function. A two-step Markov chain Monte Carlo method is used to efficiently obtain the data-based prior and to carry out variable selection and function estimation. It is shown by simulation that significant improvements in the function estimators can be obtained over an approach that estimates all the unknown functions simultaneously. The methodology is illustrated for a binary regression using heart attack data.

KEY WORDS: Bayesian information criterion; Binary regression; Gaussian errors; Model averaging; Posterior probabilities; Time series models.

1. INTRODUCTION

This article addresses the problem of variable selection and function estimation in additive nonparametric Gaussian regression models and additive nonparametric binary regression models. For the binary model, the probit link function is considered in detail, but other link functions can be used in place of the probit link function, as outlined in some earlier work (Wood and Kohn 1998). A hierarchical Bayesian solution is proposed to the foregoing problems in which each component of the regression is allowed a priori to be either in or out of the model, and an integrated Wiener process prior is assumed for each component function. Variable selection is carried out by computing the posterior probability that each function is in the model. The regression functions are estimated by their posterior means taking into account the hierarchical structure of the model. This is sometimes called model averaging in the literature (e.g. Raftery, Madigan, and Hoeting 1997), because the posterior means of the function values are averaged across all possible models using the posterior probabilities of each model as the weight function. We show in Sections 3.3 and 4.4 that better estimates of the regression functions are often obtained using model averaging than by estimating the functions assuming that they all should be included in the model. The computation is carried out using a Markov chain Monte Carlo (MCMC) method.

To illustrate these ideas, consider the following example. Suppose that it is required to model the probability of a heart attack as a function of the risk factors blood pressure (BP), cholesterol ratio (CR), and tobacco consumption (TOB). A popular model for doing so is the probit regression model,

$$\text{pr(heart attack}|\text{BP, CR, TOB)} = \Phi\{f(\text{BP}) + g(\text{CR}) + h(\text{TOB})\}, \quad (1)$$

where Φ is the standard normal cumulative distribution function. The terms $f(\text{BP})$, $g(\text{CR})$, and $h(\text{TOB})$ represent the effects of BP, CR, and TOB. If the functional forms of f, g, and h are known, then (1) is called a parametric model. For example, f, g, and h are often assumed to be linear. If the functions f, g, and h are assumed to be smooth, but otherwise unknown, then (1) is called a nonparametric probit regression model. Variable selection determines which variables should be retained in (1). Using our technique, a variable is retained in the model if its posterior probability is high. For example, if the posterior probability of the tobacco main effect $h(\text{TOB})$ is high, then it is reasonable to assert that it should be included. The functions f, g, and h are estimated using model averaging. For example, our estimate of the function h is equivalent to estimating the posterior mean of h for the eight models obtained by either including or excluding each of BP, CR, and TOB, and then weighting the posterior means by the posterior probabilities associated with the eight models.

In earlier work (Wood, Shively, and Kohn 1996), we gave a Bayesian approach to model selection for the two classes of models considered in this article. Our approach is based on the Bayesian information criterion (BIC) of Schwartz (1978), in which the posterior probability of each model is obtained separately. However, our approach is impractical if the number of variables (p) is moderate to large, because in this case 2^p separate probabilities need to be calculated, which requires $2^p - 1$ runs through the Gibbs sampler.

This article develops a different two-step approach to variable selection and function estimation. The first step uses the Gibbs sampler developed by Wong and Kohn

Thomas S. Shively is Associate Professor, Department of Management Science and Information Systems, University of Texas, Austin, TX 78712. Robert Kohn is Professor and Sally Wood is Ph.D. candidate, Australian Graduate School of Management, University of New South Wales, Sydney 2052, Australia (E-mail: *r.kohn@unsw.edu.au*).

(1996) to obtain the posterior distribution of the regression coefficients and the smoothing parameters assuming a noninformative prior. This posterior distribution is then used to construct a data-based prior for the regression coefficients and the smoothing parameters. The second step carries out variable selection and model averaging using the data-based prior. A Markov chain Monte Carlo sampling scheme is developed to allow the second step of the method to be implemented in a computationally efficient manner. In particular, the time required for the sampling scheme increases linearly with the number of component functions that need to be estimated (e.g., the time required for six functions is double that required for three functions). The data-based prior and sampling scheme are quite general methods for variable selection and model averaging that can be applied to other statistical problems. Section 2.2 shows that the method yields consistent estimators of the true model and is related to the BIC criterion.

We note that although the discussion in this article is restricted to additive nonparametric regression with main effects, our approach to variable selection and function estimation is very general and can be readily extended to handle interactions including bivariate surface estimators, such as thin-plate smoothing splines and interaction splines. A brief discussion is given in Section 2.4. Our approach can also be immediately applied to variable selection and component estimation for time series models. This is discussed in Section 6.

George and McCulloch (1993) used the Gibbs sampler for variable selection in linear regression models with Gaussian errors, and Raftery et al. (1997) discussed model averaging for the same model. Their methods are quite different than ours, however. Few works in the literature consider model averaging and variable selection in nonparametric regression. Two papers that deal with model averaging in the context of nonparametric regression are those of Denison, Mallick, and Smith (1998) and Smith and Kohn (1996). Both articles combine regression splines and model averaging but deal only with Gaussian errors, and the results do not apply to the binary case. Furthermore, their estimators are highly locally adaptive and for smooth functions may not perform as well as the estimators suggested in this article. Hardle and Korostelev (1996) considered variable selection in additive nonparametric regression with Gaussian errors. They used piecewise constant functions, but did not have a data-driven estimator of the bin width.

The article is organized as follows. Section 2 considers a regression model with Gaussian errors, describes the data-based priors for the smoothing parameters, and outlines the MCMC method used for the computation. Section 3 reports on the results of several simulation experiments that study the frequentist properties of the Bayesian approach. Section 4 extends the approach in Section 2 to the binary regression case and reports the results of several simulation experiments for this case. Section 5 analyzes a real example. Section 6 discusses the application of the methodology to time series models, and Section 7 contains some concluding remarks. The Appendix gives further details of the MCMC method.

2. NONPARAMETRIC REGRESSION WITH GAUSSIAN ERRORS

This section considers variable selection and function estimation for an additive nonparametric regression model with Gaussian errors. Section 2.1 contains a description of the nonparametric spline model that we use. Section 2.2 gives the data-based priors on the smoothing parameters and the starting values of the regression function and discusses the motivation underlying these prior distributions. Section 2.3 provides the two sampling schemes required to implement the procedure and discusses the second scheme in detail.

2.1 Nonparametric Regression Model

For simplicity, only the bivariate case is presented in detail, but it is straightforward to generalize the results to a larger number of components. The model that we consider is

$$y_i = \beta_0 + f_1(s_i) + f_2(t_i) + e_i, \qquad i = 1, \ldots, n, \quad (2)$$

where f_1 and f_2 are unknown regression functions in the independent variables s and t. The errors e_i are independent $N(0, \sigma^2)$. Without loss of generality, we assume that $0 \leq s_i \leq 1$ and $0 \leq t_i \leq 1$ for all i. To ensure that f_1 and f_2 are identified, we enforce the restrictions that $f_1(0) = f_2(0) = 0$ and take β_0 as the overall intercept for the regression.

We now discuss the priors for β_0, f_1, f_2, and σ^2 in (2). First, flat priors are used for β_0 and σ^2 throughout the article. The following discussion outlines the prior used for f_1. This function is estimated by its posterior mean. We have two goals in choosing a prior for f_1: First, the posterior mean of f_1 should be smooth, and second, the prior should be flexible enough to allow a wide variety of shapes for f_1 to be estimated properly. Choosing a parametric prior for f_1 (e.g., a linear prior) fulfills the first requirement but not the second. To satisfy both requirements, we follow Wahba (1978) and write the prior for f_1 as

$$f_1(s) = \beta_1 s + g_1(s), \quad g_1(s) = (\tau_1^2)^{1/2} \int_0^s W_1(v)\, dv, \quad (3)$$

where W_1 is a Wiener process with $W_1(0) = 0$ and $\text{var}(W_1(s)) = s$. Thus g_1 is an integrated Wiener process and represents the nonlinear part of f_1. The prior (3) has the following properties for $s \in [0, 1]$:

a. The first derivative of f_1 is continuous, because a Wiener process is continuous.
b. The second derivative of f_1 is diffuse because the first derivative of a Wiener process has infinite variance. This means that there is no prior information about the second derivative of f_1.
c. The posterior mean of f_1 is a cubic smoothing spline. That is, it has a continuous second derivative on [0, 1] and is a cubic on the intervals $(s_{i-1}, s_i), i = 2, \ldots, n$.

The coefficient β_1 is the slope of f_1 at $s = 0$, and the parameter τ_1^2 is called a smoothing parameter because it controls the curvature of f_1. If $\tau_1^2 = 0$, then f_1 is a linear

function of s in (3); otherwise, it is a nonlinear function. If both $\beta_1 = 0$ and $\tau_1^2 = 0$, then f_1 drops out of the regression. Thus the posterior probability $p(\beta_1 = 0, \tau_1^2 = 0|y)$ reflects the information in the data, the model, and the prior that f_1 is not in the regression. Similarly, the posterior probability $p\ (\tau_1^2 = 0|y)$ reflects the information that f_1 is linear.

The prior for f_2 is defined similarly to the prior for f_1:

$$f_2(t) = \beta_2 t + g_2(t), \qquad g_2(t) = (\tau_2^2)^{1/2} \int_0^t W_2(v)\, dv, \quad (4)$$

where W_2 is a Wiener process with $W_2(0) = 0$ and $\text{var}(W_2(t)) = t$. In addition, the priors for f_1 and f_2 are independent.

2.2 Data-Based Prior Distributions for β_1, β_2, τ_1^2, and τ_2^2

We now consider how to set the priors for β_1 and τ_1^2 in (3) and for β_2 and τ_2^2 in (4) to carry out variable selection and model averaging. The following notation is required. Let $J_{1\beta} = 0$ if β_1 is identically 0 and let $J_{1\beta} = 1$ otherwise. Let $J_{1\tau} = 0$ if τ_1^2 is identically 0 and let $J_{1\tau} = 1$ otherwise. Therefore, f_1 is identically 0 if both $J_{1\beta}$ and $J_{1\tau}$ are 0, and f_1 is linear if $J_{1\tau} = 0$. We define the indicator variables $J_{2\beta}$ and $J_{2\tau}$ for β_2 and τ_2^2 in the same way. Finally, let $\mathbf{y} = (y_1, \ldots, y_n)'$.

We first consider the problems that arise in specifying a prior for τ_1^2, with similar discussions holding for τ_2^2, β_1, and β_2. If $J_{1\tau} = 0$, then τ_1^2 is identically 0. If $J_{1\tau} = 1$, then the prior for τ_1^2 should be relatively uninformative. However, the prior for τ_1^2 cannot be improper, because under an improper prior for τ_1^2, the posterior probability that $J_{1\tau} = 0$ is 1; that is, f_1 is always chosen to be linear. However, it is difficult to specify a proper prior for τ_1^2, because τ_1^2 can range from 0 (if f_1 is a straight line) to a very large value (if f_1 is wiggly). To illustrate the problem, suppose that the prior for τ_1^2 is the uniform distribution on $[0, c]$. If c is too large, then it can be shown that $J_{1\tau} = 0$ is always chosen. If c is chosen too close to 0, then the posterior probability that $J_{1\tau} = 0$ will always be approximately $1/2$. The question that arises is what value of c should be chosen so that the posterior probabilities accurately reflect the information in the data and are not overly influenced by the prior due to an arbitrary choice of c. Similar problems in specifying the dispersion of the prior distribution also arise if distributions other than the uniform distribution are used.

To overcome this difficulty in specifying a proper prior for parameters such as τ_1^2 and τ_2^2, a number of authors have suggested-data-based priors; that is, priors depending on $\mathbf{y}$. Berger and Perrichi (1996) proposed a data-based prior based on averaging the posterior distributions of the unknown parameters over a large number of training samples, using a noninformative prior for the unknown parameters for each training sample. Each training sample is chosen just large enough so that the posterior distributions of the unknown parameters are proper. The approach of Berger and Perrichi (1996) appears to be computationally too intensive for setting a proper prior for τ_1^2 and τ_2^2 in our model for two reasons. First, for each training sample the posterior distributions of τ_1^2 and τ_2^2 will need to be evaluated numerically at a large number of points. Second, it will be difficult to generate τ_1^2 and τ_2^2 using this prior as part of a MCMC sampling scheme, because the prior is defined numerically. O'Hagan (1995) proposed using a fractional part of the likelihood as a prior, but this proposal again appears to be computationally too intensive to be practical, because it is difficult to evaluate the likelihood of our model.

We now describe how the data-based prior for $\tau_1^2, \tau_2^2, \beta_1$, and β_2 used in this article is obtained. Our data-based prior is motivated by the BIC criterion, which is a standard method of model selection in many regression problems. For our problem, it is more convenient to use this data-based prior than the data-based priors described earlier, because it can be readily obtained using a MCMC scheme and can then be used in a MCMC scheme to carry out the inference. The simulation results provided herein suggest that this method of deriving the data-based prior works well.

The data-based prior that we propose applies to a wider range of models than just the nonparametric regression model specifically discussed in this article. For this reason, we show how to obtain the data-based prior for a general model, and indicate how it applies specifically to the nonparametric regression model in (2).

Consider observations $\mathbf{y} = (y_1, \ldots, y_n)'$ generated from the density $p(\mathbf{y}|\boldsymbol{\phi})$, where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)$ is an $m \times 1$ vector of parameters. For $i = 1, \ldots, m$, we assume that ϕ_i lies in the interior or on the boundary of a closed bounded interval Φ_i in one-dimensional Euclidean space, and exclude the trivial case where Φ_i is a point. Thus if Φ is the Cartesian product of the Φ_i, then $\boldsymbol{\phi}$ belongs to Φ. We consider model selection for models in which some of the ϕ_i are constrained to be 0, but it is straightforward to extend the approach to models with more general constraints using an approach similar to that of Kohn (1983).

Let $\mathbf{J} = (J_1, \ldots, J_m)$ be a vector of binary variables with $J_i = 0$ if ϕ_i is constrained to be 0 and $J_i = 1$ if ϕ_i is unconstrained. Let $\boldsymbol{\phi}_\mathbf{J}$ be a vector that contains the elements of $\boldsymbol{\phi}$ that are not constrained to be 0 under model $\mathbf{J}$. Therefore, each $\mathbf{J}$ represents a model. Let $\boldsymbol{\phi}_{\text{true}}$ be the true value of $\boldsymbol{\phi}$ and let $\mathbf{J}_{\text{true}}$ be the true model. We say that $\mathbf{J}$ contains the true model if $J_i \geq J_{\text{true},i}$ for all $i = 1, \ldots, m$.

For the nonparametric regression model in (2), $\boldsymbol{\phi} = (\sigma^2, \beta_0, \beta_1, \tau_1^2, \beta_2, \tau_2^2)$, and $\mathbf{J} = (1, 1, J_{1\beta}, J_{1\tau}, J_{2\beta}, J_{2\tau})$. The first two elements of $\mathbf{J}$ are 1, because σ^2 and β_0 are always unconstrained. Suppose that the true regression model is $y_i = f_2(t_i) + \varepsilon_i$. Then $\mathbf{J}_{\text{true}} = (1, 1, 0, 0, 1, 1)$, because the true model has $\beta_1 = 0$ and $\tau_1^2 = 0$. Furthermore, $\mathbf{J}^{(1)} = (1, 1, 0, 0, 1, 1), \mathbf{J}^{(2)} = (1, 1, 1, 0, 1, 1), \mathbf{J}^{(3)} = (1, 1, 0, 1, 1, 1)$, and $\mathbf{J}^{(4)} = (1, 1, 1, 1, 1, 1)$ all contain the true regression model, because the only parameters that need to be estimated in the true model are $\sigma^2, \beta_0, \beta_2$, and τ_2^2, and these are nonzero in models $\mathbf{J}^{(1)} - \mathbf{J}^{(4)}$.

We use the following method to obtain the data-based prior for each model $\mathbf{J}$. Let $\hat{\boldsymbol{\phi}}$ be the posterior mean of $\boldsymbol{\phi}$ and let $\mathbf{A}$ be the posterior covariance matrix of $\boldsymbol{\phi}$ for the full model $\mathbf{J} = (1, 1, 1, 1, 1, 1)$ when a uniform prior on Φ is used for $\boldsymbol{\phi}$, and let $\mathbf{B} = \mathbf{A}^{-1}$. Further, for each model $\mathbf{J}$ under consideration, let $\mathbf{B}_\mathbf{J}$ be the submatrix of $\mathbf{B}$ that

consists of the rows and columns of $\mathbf{B}$ for which $J_i = 1$, and let $\hat{\boldsymbol{\phi}}_{\mathbf{J}}$ contain the elements of $\hat{\boldsymbol{\phi}}$ for which $J_i = 1$. Then the data-based prior that we use for $\boldsymbol{\phi}_{\mathbf{J}}$ is $\mathrm{N}(\hat{\boldsymbol{\phi}}_{\mathbf{J}}, n\mathbf{B}_{\mathbf{J}}^{-1})$.

To justify the use of this data-based prior, we show that the resulting Bayesian model selection procedure based on this prior is similar to Schwartz's (1978) BIC. To show this, let $L_n(\boldsymbol{\phi}_{\mathbf{J}}|\mathbf{J}) = \log p(y|\mathbf{J}, \boldsymbol{\phi}_{\mathbf{J}})$ represent the log of the likelihood function for model $\mathbf{J}$ using a sample of size n, let $L_n(\boldsymbol{\phi}) = \log p(y|\boldsymbol{\phi})$ represent the log of the likelihood function for the full model, let $\tilde{\boldsymbol{\phi}}_{\mathbf{J}}$ be the posterior mean of $\boldsymbol{\phi}_{\mathbf{J}}$ when a proper uniform prior is used for $\boldsymbol{\phi}_{\mathbf{J}}$ in model $\mathbf{J}$, and let

$$\Delta = -\frac{\partial^2 L_n(\hat{\boldsymbol{\phi}})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'}$$

and

$$\Delta_J = -\frac{\partial^2 L_n(\tilde{\boldsymbol{\phi}}_{\mathbf{J}}|\mathbf{J})}{\partial\boldsymbol{\phi}_{\mathbf{J}}\partial\boldsymbol{\phi}_{\mathbf{J}}'}.$$

We assume that $L_n(\boldsymbol{\phi})/n \to \bar{L}(\boldsymbol{\phi})$ uniformly in $\boldsymbol{\phi} \in \Phi$, and that $\bar{L}(\boldsymbol{\phi}_{\text{true}}) > \bar{L}(\boldsymbol{\phi})$ for all $\boldsymbol{\phi} \in \Phi$ if $\boldsymbol{\phi} \neq \boldsymbol{\phi}_{\text{true}}$.

There are two cases to consider:

1. $\boldsymbol{\phi}_{\text{true}}$ lies in the interior of Φ.
2. $\boldsymbol{\phi}_{\text{true}}$ lies on the boundary of Φ.

For example, in the nonparametric regression model in (2), if $\mathbf{J}_{\text{true}} = (1,1,1,1,1,1)$ then $\boldsymbol{\phi}_{\text{true}}$ lies inside Φ because the parameters $\beta_1, \tau_1^2, \beta_2$, and τ_2^2 are all nonzero. If $\mathbf{J}_{\text{true}} = (1,1,0,0,1,1)$, then $\boldsymbol{\phi}_{\text{true}}$ lies on the boundary of Φ, because $\tau_1^2 = 0$. In general, if one of the τ_j^2 values is 0, then $\boldsymbol{\phi}_{\text{true}}$ lies on the boundary of Φ.

Case 1: $\boldsymbol{\phi}_{\text{true}}$ lies in the interior of $\boldsymbol{\Phi}$. If $\boldsymbol{\phi}_{\text{true}}$ lies in the interior of Φ, then, under appropriate regularity conditions, $\hat{\boldsymbol{\phi}} \to \boldsymbol{\phi}_{\text{true}}$ and $\Delta/n - \mathbf{B}/n \to 0$ as $n \to \infty$. Furthermore, if $\mathbf{J}$ contains the true model, then $\tilde{\boldsymbol{\phi}}_{\mathbf{J}} \to \boldsymbol{\phi}_{\mathbf{J},\text{true}}$, where $\boldsymbol{\phi}_{\mathbf{J},\text{true}}$ contains only the elements of $\boldsymbol{\phi}_{\text{true}}$ for which $J_i = 1$, and $\Delta_{\mathbf{J}}/n - \mathbf{B}_{\mathbf{J}}/n \to 0$ as $n \to \infty$. Therefore, for large n and under appropriate regularity conditions, if $\mathbf{J}$ contains the true model, then

$$\begin{aligned} L_n(\boldsymbol{\phi}_{\mathbf{J}}|\mathbf{J}) &\approx L_n(\tilde{\boldsymbol{\phi}}_{\mathbf{J}}|\mathbf{J}) - \frac{1}{2}(\boldsymbol{\phi}_{\mathbf{J}} - \tilde{\boldsymbol{\phi}}_{\mathbf{J}})'\Delta_{\mathbf{J}}(\boldsymbol{\phi}_{\mathbf{J}} - \tilde{\boldsymbol{\phi}}_{\mathbf{J}}) \\ &\approx L_n(\hat{\boldsymbol{\phi}}_{\mathbf{J}}|\mathbf{J}) - \frac{1}{2}(\boldsymbol{\phi}_{\mathbf{J}} - \hat{\boldsymbol{\phi}}_{\mathbf{J}})'\mathbf{B}_{\mathbf{J}}(\boldsymbol{\phi}_{\mathbf{J}} - \hat{\boldsymbol{\phi}}_{\mathbf{J}}). \end{aligned}$$

Hence for large n, and using the data-based prior $\mathrm{N}(\hat{\boldsymbol{\phi}}_{\mathbf{J}}, n\mathbf{B}_{\mathbf{J}}^{-1})$ for $\boldsymbol{\phi}_{\mathbf{J}}$,

$$\begin{aligned} p(\mathbf{y}|\mathbf{J}) &= \int p(\mathbf{y}|J, \boldsymbol{\phi}_{\mathbf{J}}) p_{\text{DB}}(\boldsymbol{\phi}_{\mathbf{J}})\, d\boldsymbol{\phi}_{\mathbf{J}} \\ &\approx (n+1)^{-q_{\mathbf{J}}/2} p(\mathbf{y}|\mathbf{J}, \tilde{\boldsymbol{\phi}}_{\mathbf{J}}) \end{aligned} \tag{5}$$

is very close to the BIC criterion (Schwartz 1978), where $q_{\mathbf{J}}$ is the dimension of $\boldsymbol{\phi}_{\mathbf{J}}$ and $p_{\text{DB}}(\boldsymbol{\phi}_{\mathbf{J}})(= \mathrm{N}(\hat{\boldsymbol{\phi}}_{\mathbf{J}}, n\mathbf{B}_{\mathbf{J}}^{-1}))$ is the data-based prior for $\boldsymbol{\phi}_{\mathbf{J}}$. From this, we can deduce (as in Schwartz 1978) that for large $n, p(\mathbf{y}|\mathbf{J}_{\text{true}}) > p(\mathbf{y}|\mathbf{J})$ for any model $\mathbf{J}$ containing the true model.

If J does not contain the true model, then similarly to the foregoing we can show that $p(\mathbf{y}|\mathbf{J}, \tilde{\boldsymbol{\phi}}_{\mathbf{J}})$ is much less than the $p(\mathbf{y}|\mathbf{J}_{\text{true}}, \tilde{\boldsymbol{\phi}}_{\mathbf{J}_{\text{true}}})$ for large n. Therefore, it follows that $p(\mathbf{y}|\mathbf{J}) < p(\mathbf{y}|\mathbf{J}_{\text{true}})$. That is, using the data-based prior provides a consistent model selection procedure when $\boldsymbol{\phi}_{\text{true}}$ lies in the interior of Φ.

Case 2: $\boldsymbol{\phi}_{\text{true}}$ lies on the boundary of $\boldsymbol{\Phi}$. If $\boldsymbol{\phi}_{\text{true}}$ lies on the boundary of Φ, then we provide a heuristic argument why the Bayesian procedure using a data-based prior provides good results asymptotically. For simplicity, we consider the case of a scalar ϕ, but the argument generalizes in a straightforward way to the case where $\boldsymbol{\phi}$ is a vector. Suppose that $\Phi = [0, c]$ and the model selection choice is between two models M_0: $\phi_{\text{true}} = 0$ and M_1: $0 < \phi_{\text{true}} < c$. If M_1 is true, then ϕ_{true} lies in the interior of Φ; this situation is considered in case 1.

Suppose that M_0 holds. If $\hat{\phi} > 0$, then, as earlier, $p(\mathbf{y}|M_0) \approx p(\mathbf{y}|\hat{\phi})(n+1)^{-1/2} < p(\mathbf{y}|\phi = 0)$ for large n. If $\hat{\phi} = 0$, then for any $\varepsilon > 0$, there exists a δ such that

$$\int_0^{\delta} p_{\text{DB}}(\phi)\, d\phi < \varepsilon$$

and

$$p(y|\phi)/p(y|\phi = 0) < \varepsilon \quad \text{for} \quad \phi > \delta.$$

Therefore,

$$\begin{aligned} p(\mathbf{y}|J = 1) &= \int_0^{\delta} p(\mathbf{y}|J = 1, \phi) p_{\text{DB}}(\phi)\, d\phi \\ &\quad + \int_{\delta}^{c} p(\mathbf{y}|J = 1, \phi) p_{\text{DB}}(\phi)\, d\phi \\ &< p(\mathbf{y}|\phi = 0)\varepsilon + p(\mathbf{y}|\phi = 0)\varepsilon(1 - \varepsilon), \end{aligned}$$

from which we deduce that $p(\mathbf{y}|J = 1) < p(\mathbf{y}|J = 0)$.

The data-based prior approach just outlined is based on computing the posterior mean and standard deviation of ϕ under a flat prior for the largest model. The data-based prior for each model of interest is then computed from this posterior distribution. If only a small number of models are considered, then an alternative data-based prior for $\boldsymbol{\phi}_{\mathbf{J}}$ is $\mathrm{N}(\tilde{\boldsymbol{\phi}}_{\mathbf{J}}, n\mathbf{A}_{\mathbf{J}})$, where $\tilde{\boldsymbol{\phi}}_{\mathbf{J}}$ is the posterior mean of $\boldsymbol{\phi}_{\mathbf{J}}$ and $\mathbf{A}_{\mathbf{J}}$ is the posterior covariance matrix of $\boldsymbol{\phi}_{\mathbf{J}}$ using a proper uniform prior for $\boldsymbol{\phi}_{\mathbf{J}}$. Using the former approach is important if model selection is to be carried out for a large number of models, because only a single run of the Gibbs sampler is required to obtain the data-based priors for all of the models. If the latter approach is used, then the Gibbs sampler must be run for each model individually to obtain the data-based prior for that model.

We conjecture that the proposed data-based prior approach to model selection is superior to using BIC, because BIC corresponds to using both the data-based prior and to approximating the likelihood by a Gaussian, whereas the data-based prior approach does not require that the likelihood be approximated by a Gaussian. This is especially important if the parameters of interest are variances, such as τ_1^2 and τ_2^2.

As a practical matter, we have found that taking $\mathbf{A}$ as a diagonal matrix containing the posterior variances of the elements of ϕ under a proper uniform prior works well in practice, and simplifies the computations substantially.

2.3 Markov Chain Monte Carlo Sampling

The MCMC sampling schemes described herein are used to carry out function estimation and variable selection. (For a discussion of Bayesian inference using MCMC methods, see Casella and George 1993; Gelfand and Smith 1990; Tierney 1994.)

The following notation is used to describe the sampling scheme. Let $\mathbf{e} = (e_1, \ldots, e_n)'$, $\mathbf{g}_1 = (\mathbf{g}_1(s_1), \ldots, \mathbf{g}_1(s_n))'$, $\mathbf{g}_2 = (g_2(t_1), \ldots, g_2(t_n))'$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$. Let $\mathbf{Z}$ be the $n \times 3$ matrix with 1s in the first column, $(s_1, \ldots, s_n)'$ in the second column, and $(t_1, \ldots, t_n)'$ in the third column. Then (2) can be written in matrix notation as

$$\mathbf{y} = \mathbf{Z}\beta + \mathbf{g}_1 + \mathbf{g}_2 + e. \tag{6}$$

The following Gibbs sampling scheme is used to obtain the data-based priors for τ_1^2 and τ_2^2. (A detailed discussion of this sampling scheme and its implementation has been given in Wong and Kohn 1996.) A closely related sampling scheme for additive nonparametric regression was given by Erkanli and Gopalan (1996). In this sampling scheme, we use a noninformative uniform prior on $[-c, c]$ for β_1 and β_2 and a noninformative uniform prior on $[0, c]$ for τ_1^2 and τ_2^2, where $c = 10^{14}$. We experimented using values of $c = 10^{10}$ and 10^{12}, and the results were the same as for $c = 10^{14}$.

Sampling Scheme 1.

0. Start with some initial values $\beta_2^{[0]}, \mathbf{g}_2^{[0]}$, and $(\sigma^2)^{[0]}$ for $\beta_2, \mathbf{g}_2$, and σ^2.
1. Generate β_0, β_1, and $\mathbf{g}_1$ as a block conditional on $\mathbf{y}, \tau_1^2, \beta_2, \mathbf{g}_2$, and σ^2.
2. Generate τ_1^2 conditional on $\mathbf{g}_1$.
3. Generate β_2 and $\mathbf{g}_2$ as a block conditional on y, τ_2^2, $\beta_0, \beta_1, \mathbf{g}_1$, and σ^2.
4. Generate τ_2^2 conditional on $\mathbf{g}_2$.
5. Generate σ^2 conditional on $\mathbf{y}, \boldsymbol{\beta}, \mathbf{g}_1$, and $\mathbf{g}_2$.

The next sampling scheme is used to carry out variable selection and model averaging using the data-based priors for $\tau_1^2, \tau_2^2, \beta_1$, and β_2.

Sampling Scheme 2.

0. Start with some initial values $\beta_2^{[0]}, \mathbf{g}_2^{[0]}$, and $(\sigma^2)^{[0]}$ for $\beta_2, \mathbf{g}_2$, and σ^2.
1. Generate $J_{1\tau}, J_{1\beta}$, and τ_1^2 as a block in the following order, conditional on $y, \beta_2, \mathbf{g}_2$, and σ^2:
   a. Generate $J_{1\tau}$ and $J_{1\beta}$ jointly conditional on y, β_2, $\mathbf{g}_2$, and σ^2.
   b. Generate τ_1^2 conditional on $J_{1\tau}, J_{1\beta}, \mathbf{y}, \beta_2, \mathbf{g}_2$, and σ^2 from a distribution that approximates $p(\tau_1^2 | J_{1\tau}, J_{1\beta}, \mathbf{y}, \beta_2, \mathbf{g}_2, \sigma^2)$.
   c. Perform a Metropolis–Hastings acceptance step using the generated values of $J_{1\tau}, J_{1\beta}$, and τ_1^2 as the proposal.
2. Generate β_0, β_1, and $\mathbf{g}_1$ as a block by first generating β_0 and β_1 from $p(\beta_0, \beta_1 | J_{1\beta}, \tau_1^2, \mathbf{y}, \beta_2, \mathbf{g}_2, \sigma^2)$ with $\mathbf{g}_1$ integrated out, and then generating g_1 from $p(\mathbf{g}_1 | \tau_1^2, \mathbf{y}, \boldsymbol{\beta}, \mathbf{g}_2, \sigma^2)$.
3. Generate $J_{2\tau}, J_{2\beta}$, and τ_2^2 as a block conditional on $\mathbf{y}, \beta_1, \mathbf{g}_1$, and σ^2, similarly to step 1.
4. Generate β_2 and $\mathbf{g}_1$ as a block conditional on $J_{2\beta}, \tau_2^2$, $\mathbf{y}, \beta_0, \beta_1, \mathbf{g}_1$, and σ^2, similarly to step 2.
5. Generate σ^2 conditional on $\mathbf{y}, \boldsymbol{\beta}, \mathbf{g}_1$, and $\mathbf{g}_2$.

Sampling scheme 2 is irreducible and aperiodic, because it is readily checked that in one step the sampling scheme can reach any point in the parameter space from any other point. Therefore, sampling scheme 2 converges to the posterior distribution of Tierney (1994).

Steps 1–5 of sampling schemes 1 and 2 are repeated many times and in two stages. The first stage is called a warmup period, and it is assumed that at the end of this period, the sampling scheme generates iterates from the posterior distribution. The second stage is called the sampling period and iterates generated from this stage are used for inference.

We now describe briefly how sampling scheme 2 is implemented. (Further details are given in the Appendix.) In step 1a, the discrete probability $p\ (J_{1\tau} = 1, J_{1\beta} = 1 | \mathbf{y}, \beta_2, \mathbf{g}_2, \sigma^2)$ is obtained up to a multiplicative constant by integrating out β_0, β_1, and $\mathbf{g}_1$ analytically and then integrating out τ_1^2 numerically using Gaussian quadrature. The conditional probabilities $p(J_{1\tau}, J_{1\beta} | \mathbf{y}, \beta_2, \mathbf{g}_2, \sigma^2)$ for the other three combinations of $J_{1\tau}$ and $J_{1\beta}$ are obtained similarly (up to the same multiplicative constant). The conditional probability of $J_{1\tau}$ and $J_{1\beta}$ is then obtained by normalization, and $J_{1\tau}$ and $J_{1\beta}$ are generated. In step 1b, if $J_{1\tau}$ is generated as 0, then τ_1^2 is set to 0. If $J_{1\tau} = 1$, then $\theta_1 = \log(\tau_1^2)$ is generated from a Gaussian approximation to the conditional density of θ_1 obtained by using the first two moments of the conditional density of θ_1. If $J_{1\beta} = 1$, then these moments are computed with β_0, β_1, and $\mathbf{g}_1$ integrated out analytically, whereas if $J_{1\beta} = 0$, then the moments are computed with β_0 and $\mathbf{g}_1$ integrated out analytically and β_1 set to 0. The two moments are obtained using a Gaussian quadrature procedure similar to the one used to integrate out τ_1^2 in step 1a. However, the numerical integrations required to obtain these two moments require almost no additional calculations, because the required function evaluations are done in step 1a when $J_{1\beta}$ and $J_{1\tau}$ are generated. Step 1c is a Metropolis-Hastings acceptance step performed because the numerical integration undertaken in step 1a is only approximate; and because θ_1 (and thus τ_1^2) is generated from an approximate distribution if $J_{1\tau} = 1$. In step 2, if $J_{1\beta} = 0$, then β_1 is set to 0, and only β_0 is generated. If $J_{1\beta} = 1$, then both β_0 and β_1 are generated. Similarly, if $J_{1\tau} = 0$, then the vector $\mathbf{g}_1$ is set to 0, whereas if $J_{1\tau} = 1$, then $\mathbf{g}_1$ is generated following Wong and Kohn (1996). Steps 3 and 4 are carried out similarly to steps 1 and 2. In step 5, σ^2 is generated from its conditional density, which is inverse gamma.

2.4 Inference

Sampling scheme 2 can be used to select the independent

variables that should be in the regression and to estimate the component regression functions using model averaging. We use the posterior probability of a variable to decide whether it should be included in the model. For example, the variable s is included in the model if $p\ (J_1 = 1|\mathbf{y})$ exceeds some prescribed threshold, which here we take as .5, where $J_1 = \max\{J_{1\beta}, J_{1\tau}\}$. Let $J_{1\beta}^{[j]}, J_{1\tau}^{[j]}, j = 1, \ldots, M$, be the iterates of $J_{1\beta}$ and $J_{1\tau}$ in the sampling period and let $J_1^{[j]} = \max\{J_{1\beta}^{[j]}, J_{1\tau}^{[j]}\}$. Then the sample mean of the $J_1^{[j]}$ is an estimate of the posterior probability that $J_1 = 1$. The estimate of $p\ (J_2 = 1|\mathbf{y})$ is obtained similarly. Variable selection requires one run using sampling scheme 1 to obtain the data-based prior for τ_1^2 and τ_2^2, and then a second run using sampling scheme 2 to select the variables using the data-based priors.

The second problem addressed is the estimation of the intercept β_0 and the functions f_1 and f_2. Three approaches are compared. The first approach estimates the components by model averaging. The second approach first selects the appropriate variables for inclusion as outlined earlier and then estimates only the included components by their posterior means. The third approach estimates all the components without constraint using flat priors for $\beta_0, \beta_1, \beta_2, \tau_1^2$, and τ_2^2. Most current methods for nonparametric estimation of an additive regression use unconstrained estimation. The first approach requires one run using sampling scheme 1 to obtain the data-based prior for τ_1^2 and τ_2^2, and then a second run using sampling scheme 2 to obtain the posterior means using model averaging. The second approach requires a third run using sampling scheme 1 to estimate the components selected in the second run. The third approach requires a single run of sampling scheme 1.

It is sufficient to discuss the estimation of the posterior means of $\beta_0, f_1(s) = \beta_1 s + g_1(s)$ and $f_2(s) = \beta_2 s + g_2(s)$ for the first approach. Similar estimates are used for the second and third approaches, although in the second approach only the functions for the selected variables are estimated. The estimates of the posterior means of β_0 and $\mathbf{g}_1$ are given by the following expressions:

$$\hat{\beta}_0 = \frac{1}{M} \sum_{j=1}^{M} E(\beta_0|\mathbf{y}, \mathbf{g}_1^{[j]}, \mathbf{g}_2^{[j]}, \beta_1^{[j]}, \beta_2^{[j]}, \sigma^{[j]})$$

and

$$\hat{\mathbf{g}}_1 = \frac{1}{M} \sum_{j=1}^{M} E(\mathbf{g}_1|\mathbf{y}, \mathbf{g}_2^{[j]}, \boldsymbol{\beta}^{[j]}, (\tau_1^2)^{[j]}, \sigma^{[j]}).$$

Similar expressions provide estimates for the posterior means of β_1, β_2, and $\mathbf{g}_2$.

2.5 Modeling Interactions

It is straightforward to incorporate into the model the multiplicative interaction $s \times t$ of s and t by adding a third component $f_3(s_i t_i)$ to (2) and treating $u = st$ as a third variable. More generally, the approach in this article can handle general bivariate interactions using thin-plate splines and interaction splines by making use of their correspondence to Gaussian priors (as in Wahba 1990, chaps. 2 and 10).

3. SIMULATION RESULTS

3.1 Introduction

A number of simulation experiments were carried out to study the frequentist properties of the Bayesian approach to variable selection and function estimation. Results for variable selection are reported first, followed by those for function estimation. The following functions are used in the simulation experiments: flat function $f(s) = 0$, linear function $f(s) = 2s - 1$, exponential function $f(s) = \exp(1.1s^3) - 2$, and sine function $f(s) = \sin(4\pi s)$. These four functions were chosen because they are representative of many of the regression functions found in the literature. The flat function is the null function. The linear function is used frequently in regression. The exponential function is monotonic and can be well estimated by an estimator using a single bandwidth. The sine function is not monotonic but can also be well estimated by an estimator using a single bandwidth. To make it easier to interpret the variable selection results and the estimation results, all of the regression functions listed here except for the flat function are normalized to have a range of two.

In all of the simulation experiments described herein, the warmup and sampling periods were 2,000 iterations. Extensive simulation work suggested that these warmup and sampling periods were adequate in that time series plots of the estimated posterior probabilities showed that they had converged before 2,000 iterations were completed. In addition, the same results were obtained from different starting values.

3.2 Variable Selection

Variable selection is first considered in the univariate case. The performance of the Bayesian approach is studied in detail using three noise levels ($\sigma = .5, 1.0$, and 2.0). Thus, for the nonconstant functions, σ takes the values of one-quarter the range of the function, one-half the range of the function, and the range of the function. We refer to these as the low-, medium-, and high-noise cases. The prior probability that the variable is in is .5. Table 1 reports the percentage of times in 50 replications in which the variable is selected. Only sample size $n = 100$ was used, but it is possible to deduce results for larger sample sizes by noting that decreasing the error standard deviation by a factor of two is approximately the same as increasing the sample size by a factor of four. The table shows that variable selection works well in the univariate case, except for the high-noise case. For $n = 400$, variable selection will also work well for the high-noise case ($\sigma = 2.0$), because the results will be similar to those for $n = 100$ with medium noise ($\sigma = 1.0$).

Table 2 reports the results of five simulation experiments in the multiple regression case. Each experiment consists of 50 replications using a sample size of $n = 400$, and the prior probability that the ith variable is in the regression is .5. In experiments 1, 2, and 3 are three explanatory variables, which we call s, t, and u. The values of s, t, and u were obtained by generating 100 triplets (s, t, u) uniformly on (0, 1) and independent of each other, and then replicating

Table 1. Variable Selection Results for Univariate Nonparametric Regression

Function	Sample size n	Error standard deviation σ	Percentage of times variable is selected
Flat	100	0.50	0
Flat	100	1.00	0
Flat	100	2.00	0
Linear	100	0.50	100
Linear	100	1.00	100
Linear	100	2.00	48
Exponential	100	0.50	100
Exponential	100	1.00	98
Exponential	100	2.00	32
Sine	100	0.50	100
Sine	100	1.00	98
Sine	100	2.00	14

NOTE: The percentage of times that the variable is selected in 50 replications is reported. The prior probability of inclusion is .5.

each triplet four times. Four hundred values of the dependent variables were generated using the functions described in Table 2. The results for experiments 1–3 of Table 2 are interpreted as follows. In experiment 1, the error standard deviation is $\sigma = 1.0$ and the three regression functions are flat, linear, and exponential. In 50 replications, the first variable was selected 0% of the time, the second variable was selected 100% of the time, and the third variable was selected 100% of the time; that is, in every replication the variable selection method chose the correct combination of variables. The results suggest that for the cases studied in the simulation, variable selection works well.

Experiments 4 and 5 consider three explanatory variables s, t, and u, together with their two-way multiplicative interactions $s \times t, s \times u$, and $t \times u$, so that

$$y = \beta_0 + f_1(s) + f_2(t) + f_3(u) + f_4(st) + f_5(su) + f_6(tu) + e.$$

The explanatory variables s, t, and u were generated independently as in experiments 1–3. The error standard deviation $\sigma = .5$ is used for experiment 4, whereas $\sigma = 1.0$ is used for experiment 5. The results of the variable selection are again excellent for both experiments 4 and 5.

The variable selection procedure can also be used to determine whether a linear or nonlinear function is appropriate for modeling each relationship by determining whether $\tau^2 = 0$ for each function. Thus one computes the posterior probability $p\ (J_\tau = 0|y) = p\ (\tau^2 = 0|y)$ for each function. If $p\ (J_\tau = 0|y) > .5$, then the information in the data combined with the information in the priors for J_τ and τ^2 suggests the relationship can be modeled linearly. Conversely, if $p\ (J_\tau = 1|y) \geq .5$, then the function should be modeled nonlinearly. Table 3 reports the results of three simulation experiments (denoted experiments 6–8). These experiments use the same regressor variables as experiments 1–3. The results for experiments 6–8 are interpreted as follows. In experiment 6, the error standard deviation is $\sigma = 1.0$ and the functions are linear, exponential, and sine. In 50 replications, the first variable was specified to be nonlinear 2% of the time, the second variable was specified to be nonlinear 96% of the time, and the third variable was specified to be nonlinear 100% of the time; that is, in almost every replication the model selection procedure chose the correct functional forms. The results suggest that for the cases studied in the simulation the selection procedure works well.

3.3 Function Estimation

Section 2.4 discussed three methods for estimating the components of the regression model. These three methods are compared in this section using the designs in experiments 1–5 described in Section 3.2. Performance is measured in terms of the root mean integrated squared error (RMISE) for the individual components and for the aggregate regression function. For the function f_1, the RMISE is defined as

$$\text{RMISE}(f_1) = \left\{ \frac{1}{400} \sum_{i=1}^{400} (\hat{f}_1(s_i) - f_1(s_i))^2 \right\}^{1/2},$$

where $s_1, \ldots, s_{400}$ are the generated values of $s, f_1(s_i)$ is the true value of f_1 at s_i, and $\hat{f}_1(s_i)$ is the estimated value. The RMISE is defined similarly for the other components. For experiments 1–3, the overall RMISEs for the aggregate regression functions are defined as

$$\text{RMISE}(\beta_0, f_1, f_2, f_3) = \left\{ \frac{1}{400} \sum_{i=1}^{400} (\hat{\beta}_0 + \hat{f}_1(s_i) + \hat{f}_2(t_i) + \hat{f}_3(u_i) - \beta_0 - f_1(s_i) - f_2(t_i) - f_3(u_i))^2 \right\}^{1/2}.$$

Table 2. Variable Selection Results for a Nonparametric Regression Model With Gaussian Errors

Experiment number	Sample size n	Error standard deviation σ	Regression function					
			Function 1	Function 2	Function 3	Function 4	Function 5	Function 6
1	400	1.0	Flat	Linear	Exponential	NA	NA	NA
			0	100	100	NA	NA	NA
2	400	1.0	Flat	Flat	Exponential	NA	NA	NA
			2	0	100	NA	NA	NA
3	400	1.0	Flat	Flat	Flat	NA	NA	NA
			0	2	2	NA	NA	NA
4	400	0.5	Exponential	Sine	Flat	Flat	Flat	Flat
			100	100	0	0	0	0
5	400	1.0	Exponential	Sine	Flat	Flat	Flat	Flat
			100	100	0	0	0	0

NOTE: The percentage of times the variable is selected in 50 replications is reported. The prior probability of inclusion is .5. NA means that the entry is not applicable.

Table 3. Model Selection Results for Linearity for a Nonparametric Regression Model With Gaussian Errors

Experiment number	Sample size n	Error standard deviation σ	Regression function		
			Function 1	Function 2	Function 3
6	400	1.0	Linear	Exponential	Sine
			2	96	100
7	400	1.0	Linear	Linear	Exponential
			2	2	96
8	400	1.0	Linear	Linear	Linear
			4	2	0

NOTE: The percentage of times the function is specified to be nonlinear in 50 replications is reported. The prior probability of a linear relationship is .5.

The overall RMISEs for experiments 4 and 5 are defined similarly, but with the three interaction terms included.

Figure 1 provides boxplots of the RMISEs for all three components and the aggregate regression function for experiment 2. In each panel of Figure 1 are three boxplots corresponding to, from left to right, the estimators obtained by model averaging, variable selection followed by function estimation, and unrestricted function estimate. Each boxplot represents the 50 replicates of the RMISEs for a particular estimator. Figures 2 and 3 show the corresponding results for experiments 3 and 5. The three figures demonstrate that substantial improvement is obtained in estimating null components by using model averaging, or by first carrying out variable selection and then estimating the functions remaining in the model, over an approach that estimates all of the components in an unrestricted fashion. Furthermore, little or no loss (in terms of RMISE) results from estimating nonnull components by model averaging, or variable selection followed by function estimation, compared to an approach that estimates all of the components in an unrestricted fashion. Finally, for the overall regression function, when there are some null components a smaller, but still substantial, decrease in the RMISE results when the components are estimated by model averaging, or variable selection followed by function estimation, compared to the RMISE obtained when all of the components are estimated in an unrestricted fashion. The performance of function estimation by model averaging is similar to that of function estimation after variable selection. We prefer function estimation by model averaging because it requires one less run of sampling scheme 1. The same conclusions apply to experiments 1 and 4.

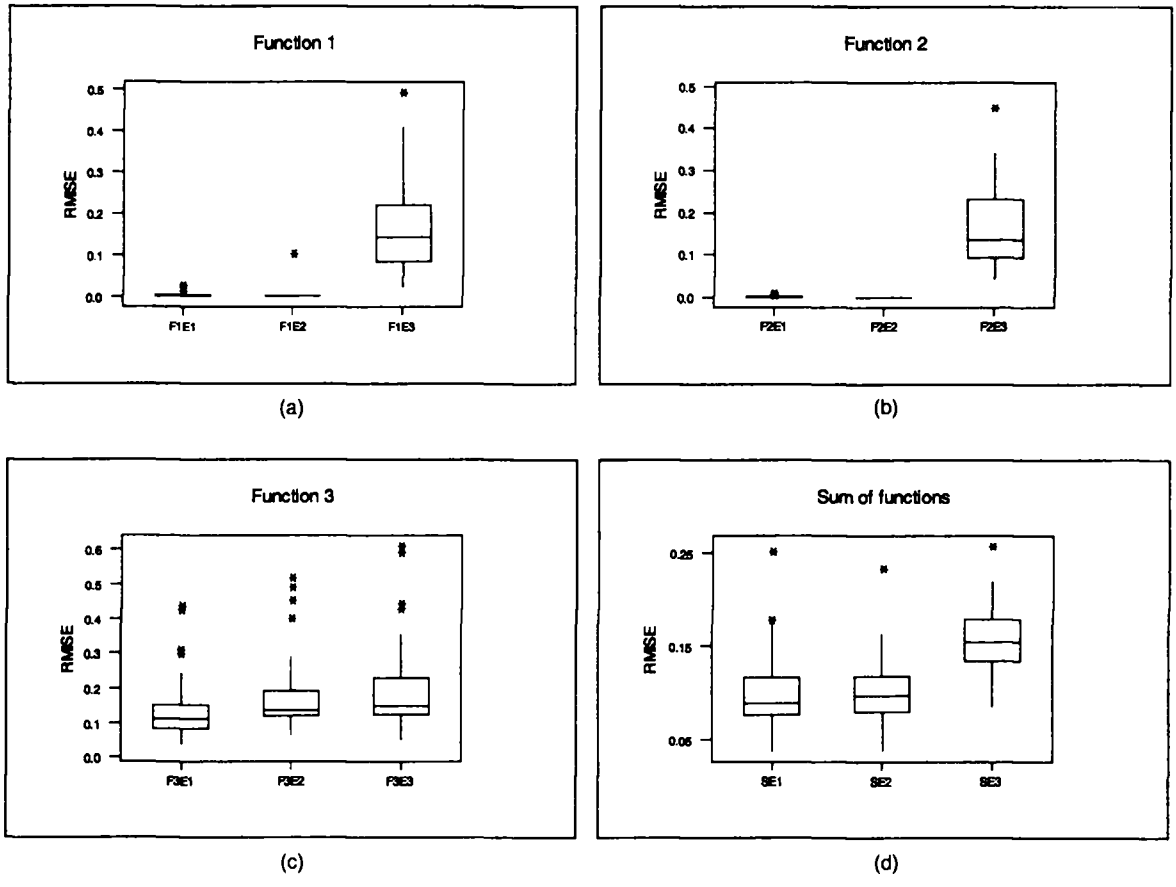

Figure 1. Boxplots of the RMISEs From Experiment 2 in the Gaussian Error Case. (a) Function 1; (b) function 2; (c) function 3; (d) sum of functions. In each panel the three boxplots correspond to, from left to right, the estimators obtained by model averaging, variable selection followed by function estimation, and unconstrained estimation.

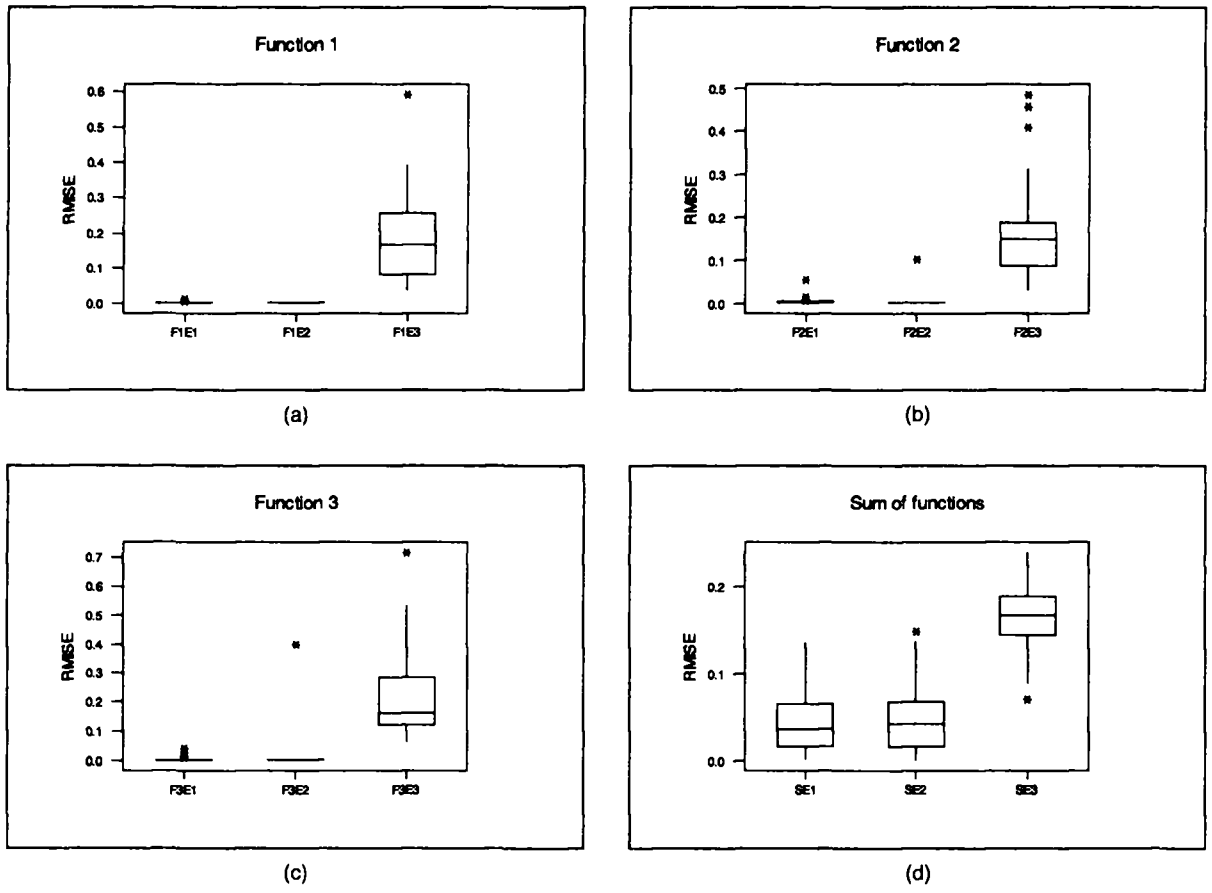

Figure 2. Boxplots of the RMISEs From Experiment 3 in the Gaussian Error Case. (a) Function 1; (b) function 2; (c) function 3; (d) sum of functions. In each panel the three boxplots correspond to, from left to right, the estimators obtained by model averaging, variable selection followed by function estimation, and unconstrained estimation.

The simulation results of Ansley, Kohn, and Tharm (1991) showed that in the univariate regression case the performance of smoothing splines, with the smoothing parameter estimated by marginal likelihood, compares favorably in terms of RMISE with the performance of smoothing splines when the smoothing parameter is estimated by generalized cross-validation. It can be shown empirically that the third Bayesian estimator described in Section 2.4 (which uses unconstrained function estimation combined with flat priors for β and the smoothing parameters) has performance similar to that of the marginal likelihood estimator. This provides a connection between the estimators in this article and some previous estimators in the nonparametric literature.

4. BINARY REGRESSION

4.1 Introduction

This section extends the results in Section 2 to the binary regression case using the data augmentation approach of Albert and Chib (1993). For conciseness, only the case of two explanatory variables is discussed, but it is straightforward to extend the results to more than two variables.

Let w be a binary variable taking the values 0 and 1, and suppose that w depends on the explanatory variables s and t. Using probit regression, we model $p\ (w = 1|s,t)$ as

$$p(w = 1|s,t) = \Phi(\beta_0 + f_1(s) + f_2(t)), \tag{7}$$

where the link function Φ is the standard normal cumulative distribution function. More general link functions can be handled as in our earlier work (Wood and Kohn 1998).

To carry out inference on the model (7), we introduce the latent variables $y_1, \ldots, y_n$ such that

$$y_i = \beta_0 + f_1(s) + f_2(t) + e_i, \tag{8}$$

with the e_i independent N(0, 1) random variables. Let $w_i = 1$ if $y_i > 0$ and let $w_i = 0$ otherwise. Following Albert and Chib (1993), it is straightforward to show that if y_i is defined by (8) and w_i is defined in terms of y_i as earlier, then (7) holds. Conversely, if $w_i = 1$, then y_i is normally distributed with mean $\beta_0 + f_1(s_i) + f_2(t_i)$ and is constrained to be positive. If $w_i = 0$, then y_i is normally distributed with the same mean and variance as earlier, but is constrained to be negative.

To carry out a Bayesian analysis on the model (7) with the probit link function, we assume the same priors on β_0 and the functions f_1 and f_2 as in Section 2.

4.2 Markov Chain Monte Carlo Sampling

Sampling schemes 1 and 2 in Section 2.2 are modified for probit nonparametric regression as follows. Let $\mathbf{w} = (w_1, \ldots, w_n)'$.

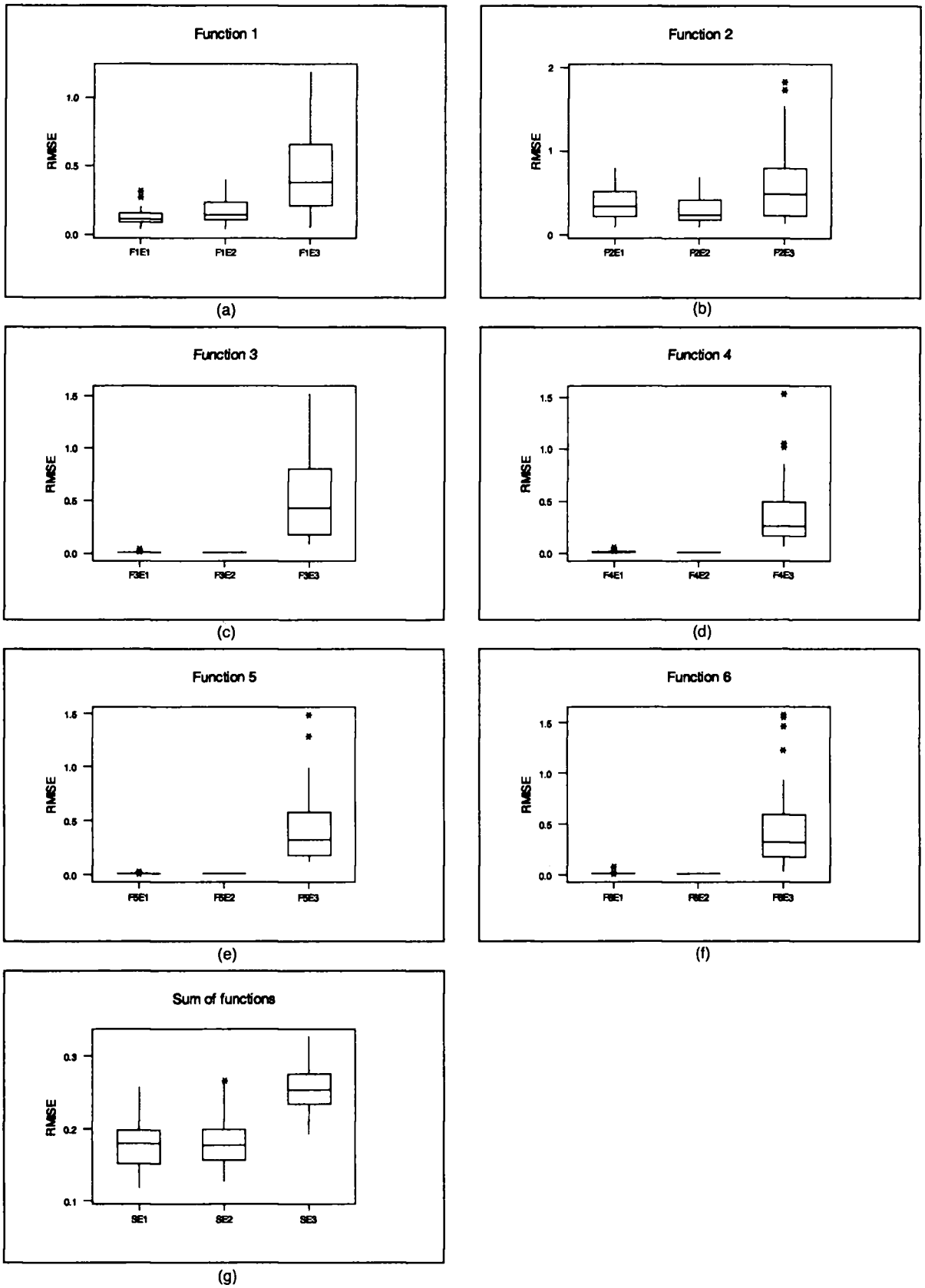

Figure 3. Boxplots of the RMISEs From Experiment 5 in the Gaussian Error Case. (a) Function 1; (b) function 2; (c) function 3; (d) function 4; (e) function 5; (f) function 6; (g) sum of functions. In each panel the three boxplots correspond to, from left to right, the estimators obtained by model averaging, variable selection followed by function estimation, and unconstrained estimation.

Sampling scheme 3 uses the same noninformative uniform priors for τ_1^2, τ_2^2, and β as in sampling scheme 1; it is used to obtain the data-based priors for τ_1^2 and τ_2^2. A detailed discussion of this sampling scheme and its implementation has been given earlier (Wood and Kohn 1998).

Sampling Scheme 3.

0. Start with some initial values $\mathbf{g}_1^{[0]}, \mathbf{g}_2^{[0]}$, and $\beta^{[0]}$ for $\mathbf{g}_1, \mathbf{g}_2$, and β.

1. For $i = 1, \ldots, n$ generate y_i conditional on $\mathbf{w}, \mathbf{g}_1, \mathbf{g}_2$, and β as follows:

   a. If $w_i = 1$, then y_i is Gaussian with mean $\beta_0 + f_1(s_i) + f_2(t_i)$ and variance 1, and is constrained to be positive.
   b. If $w_i = 0$, then y_i is Gaussian with the same mean and variance, but is constrained to be negative.

Steps 2–5 are the same as steps 1–4 sampling scheme 1.

Sampling scheme 4 is used to carry out variable selection and model averaging using the data-based priors for $\tau_1^2, \tau_2^2, \beta_1$, and β_2.

Sampling Scheme 4. Steps 0 and 1 of sampling scheme 4 are the same as in sampling scheme 3, whereas steps 2–5 are the same as steps 1–4 in sampling scheme 2.

Given sampling scheme 4, variable selection and function estimation for probit nonparametric regression are carried out in exactly the same way as outlined in Section 2.4.

4.3 Simulation Results for Variable Selection

Table 4 reports the results of four simulation experiments whose design is the same as that of experiments 1–4 in Section 3.2, except that in the binary case the error standard deviation is not part of the design. The variable selection approach again worked well in all four experiments.

4.4 Simulation Results for Function Estimation

The three methods of estimating unknown functions described in Section 2.4 are now compared for the probit regression case using the designs in experiments 1–4. The quality of the estimators of the individual regression functions and the aggregate regression function is measured in terms of RMISE as described in Section 3.3. Because the observations are binary two other measures of performance are used for the aggregate regression function in addition to the RMISE. The second measure is the integrated Kullback–Leibler distance (IKLD), which is defined as

$$\begin{aligned}\text{IKLD} = \sum_{i=1}^{n} & p(w_i = 1|s_i, t_i) \\ & \times \log\{p(w_i = 1|s_i, t_i)/\hat{p}(w_i = 1|s_i, t_i)\} \\ & + p(w_i = 0|s_i, t_i) \\ & \times \log\{p(w_i = 0|s_i, t_i)/\hat{p}(w_i = 0|s_i, t_i)\}, \quad (9)\end{aligned}$$

where $\hat{p}(w = 1|s, t)$ is the estimate of $p(w = 1|s, t)$. We note that the ith summand in (9) is the Kullback–Leibler distance between $p(w_i = 1|s_i, t_i)$ and $\hat{p}(w_i = 1|s_i, t_i)$. By Rao (1973, pp. 58–59), this measure is nonnegative and is 0 only if $p(w_i = 1|s_i, t_i) = \hat{p}(w_i = 1|s_i, t_i)$. Hence the IKLD is always nonnegative, and is equal to 0 only if $p(w_i = 1|s_i, t_i) = \hat{p}(w_i = 1|s_i, t_i)$ for all $i = 1, \ldots, n$.

The last measure is an integrated goodness-of-fit criterion (IGFC), defined as

$$\text{IGFC} = \sum_{i=1}^{n} \frac{(\hat{p}(w_i = 1|s_i, t_i) - p(w_i = 1|s_i, t_i))^2}{p(w_i = 1|s_i, t_i)(1 - p(w_i = 1|s_i, t_i))}. \quad (10)$$

In (10) the numerator in each summand is the square of the difference between $\hat{p}(w_i|s_i, t_i)$ and $p(w_i|s_i, t_i)$, whereas the denominator is equal to $\text{var}\{\hat{p}(w_i|s_i, t_i) - p(w_i|s_i, t_i)\}$ and thus standardizes the numerator.

Figures 4, 5, and 6 give boxplots of the RMISE for 50 replications for the individual components and the aggregate regression function in experiments 2, 3, and 4. The figures also give boxplots of the IKLD and the IGFC for the aggregate regression function. These figures show that there is a substantial advantage in terms of RMISE in estimating null components by model averaging, or by first carrying out variable selection followed by function estimation, over an approach that estimates all the components in an unrestricted fashion. Furthermore, when there are some null components in the regression, there are smaller, but still substantial, decreases in RMISE, IKLD, and IGFC for the overall regression function obtained using model averaging, or variable selection followed by function estimation, over an approach that estimates all of the functions in an unrestricted fashion.

Table 4. Variable Selection Results for a Nonparametric Probit Regression Model

Experiment number	Sample size n	Error standard deviation σ	Regression function					
			Function 1	Function 2	Function 3	Function 4	Function 5	Function 6
1	400	1.0	Flat	Linear	Exponential	NA	NA	NA
			0	100	100	NA	NA	NA
2	400	1.0	Flat	Flat	Exponential	NA	NA	NA
			6	0	100	NA	NA	NA
3	400	1.0	Flat	Flat	Flat	NA	NA	NA
			2	2	0	NA	NA	NA
4	400	1.0	Exponential	Sine	Flat	Flat	Flat	Flat
			96	100	2	8	0	0

NOTE: The percentage of times the variable is selected in 50 replications is reported. The prior probability of inclusion is .5. NA means that the entry is not applicable.

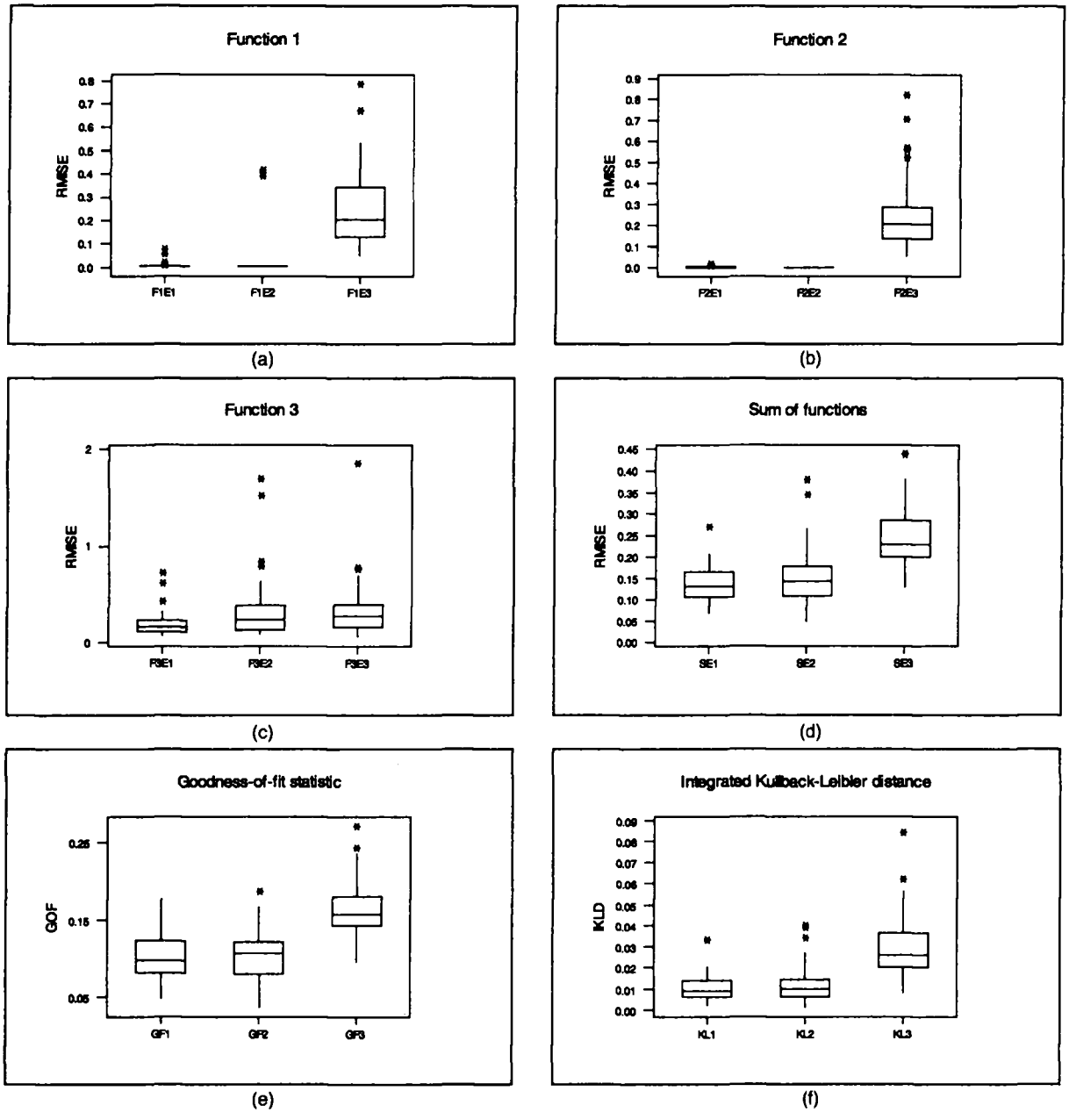

Figure 4. Boxplots of the RMISEs, IGFC, and IKLD From Experiment 2 for the Probit Regression Case. (a) Function 1; (b) function 2; (c) function 3; (d) sum of functions; (e) goodness-of-fit statistic; (f) integrated Kullback-Leibler distance. In each panel the three boxplots correspond to, from left to right, the estimators obtained by model averaging, variable selection followed by function estimation, and unconstrained estimation.

Earlier (Wood and Kohn 1998) showed by simulation that the third Bayesian estimator, which estimates the components without constraints, compares favorably with the estimators for binary nonparametric regression proposed by Wahba, Wang, Gu, Klein, and Klein (1995). This suggests that when one or more of the functions are flat, then model averaging, or variable selection followed by function estimation, should compare favorably to the methods of Wahba et al. (1995).

5. ANALYSIS OF HEART ATTACK DATA

We now use the results of Section 4 to analyze the incidence of heart attack in a group of 463 subjects as a function of four risk factors: systolic blood pressure (BP); tobacco consumption (TOB), a measure of lifetime tobacco consumption; cholesterol ratio (CR); and type A (TYPEA), a measure of psychosocial stress. The purpose of the analysis is to determine, using variable selection, which variables should be in the regression, and to estimate nonparametrically the unknown component functions using model averaging. These heart attack data were previously analyzed by Hastie and Tibshirani (1987).

We fit the additive probit nonparametric regression model,

$$p(\text{heart attack}|\text{BP, TOB, CR, TYPEA}) = \Phi(\beta_0 + f_1(\text{BP}) + f_2(\text{TOB}) + f_3(\text{CR}) + f_4(\text{TYPEA})), \tag{11}$$

where the functions $f_1, \ldots, f_4$ are estimated nonparametri-

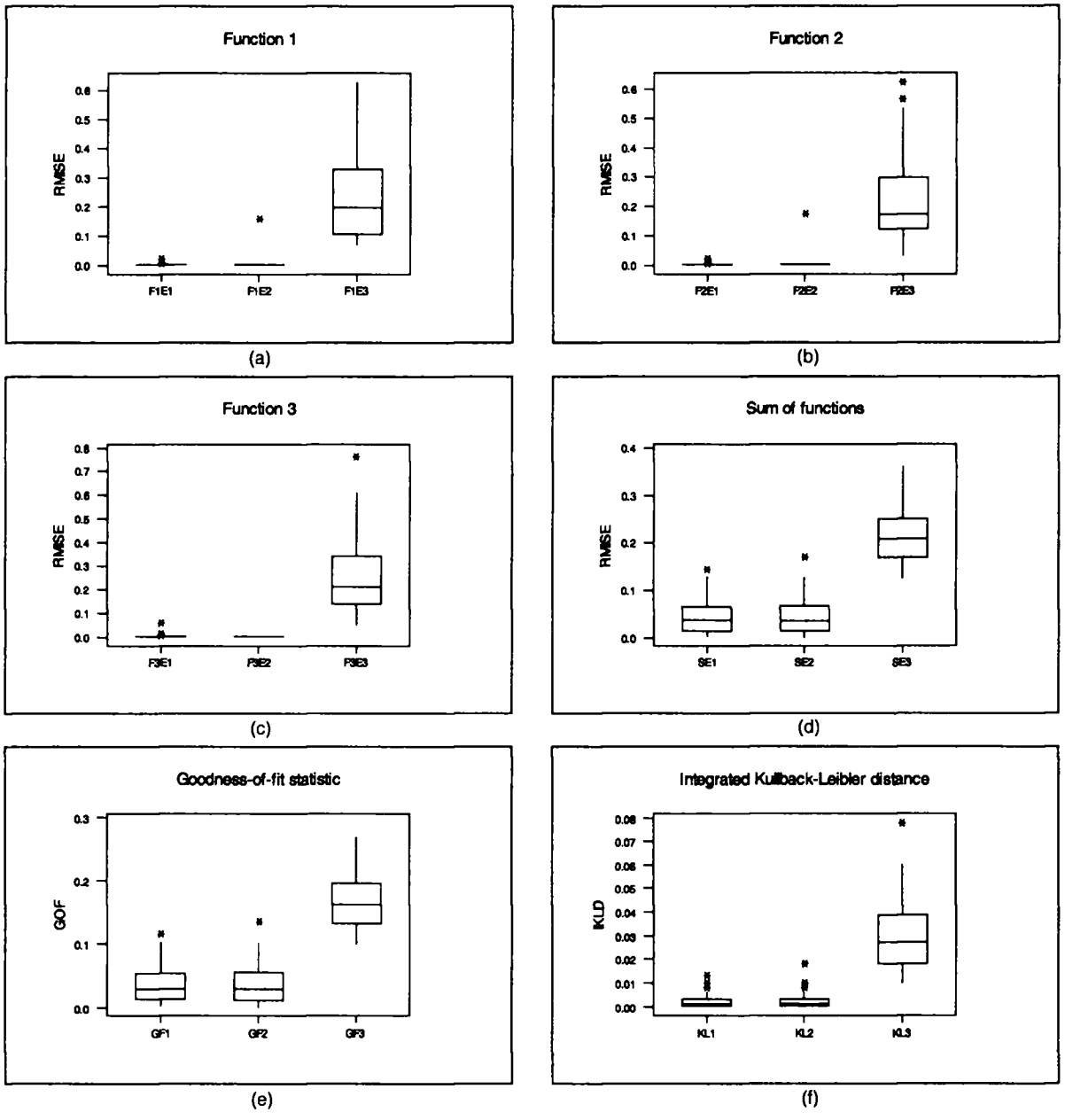

Figure 5. Boxplots of the RMISEs, IGFC, and IKLD From Experiment 3 for the Probit Regression Case. (a) Function 1; (b) function 2; (c) function 3; (d) sum of functions; (e) goodness-of-fit statistic; (f) integrated Kullback-Leibler distance. In each panel the three boxplots correspond to, from left to right, the estimators obtained by model averaging, variable selection followed by function estimation, and unconstrained estimation.

cally and variable selection is carried out on all four explanatory variables. The marginal posterior probabilities of inclusion for the explanatory variables are as follows: blood pressure, .17; tobacco, 1.00; cholesterol ratio, 1.00; and type A, 0.20. The results suggest that tobacco and cholesterol ratio should be included in the model, whereas blood pressure and type A should not be included. We also computed the joint posterior probability for each of the $2^4 = 16$ combinations of functions; that is, if $J_k = 0$ when f_k is identically 0 and $J_k = 1$ when f_k is not identically 0, then we computed $\text{pr}(J_1, \ldots, J_4|y)$ for each combination of $(J_1, \ldots, J_4)$. The highest posterior probability was $\text{pr}\{(J_1, \ldots, J_4) = (0, 1, 1, 0)|y\} = .64$, which also indicates that the tobacco and cholesterol ratios should be included in the model. The next two highest posterior probabilities were $\text{pr}\{(J_1, \ldots, J_4) = (0, 1, 1, 1)|y\} = 0.17$ and $\text{pr}\{(J_1, \ldots, J_4) = (1, 1, 1, 0)|y\} = 0.13$.

Figure 7 plots the estimates of $f_1, \ldots, f_4$ obtained using model averaging as well as one standard deviation confidence bands. The number of iterations used in the warmup and sampling periods were 20,000 and 50,000 for both sampling schemes. Time series plots of the estimated coefficients and posterior probabilities indicated that the sampling schemes had converged after this many iterations. Also, different sets of starting values provided the same results.

The marginal posterior probabilities that each function is nonlinear are blood pressure, .09; tobacco, .18; cholesterol

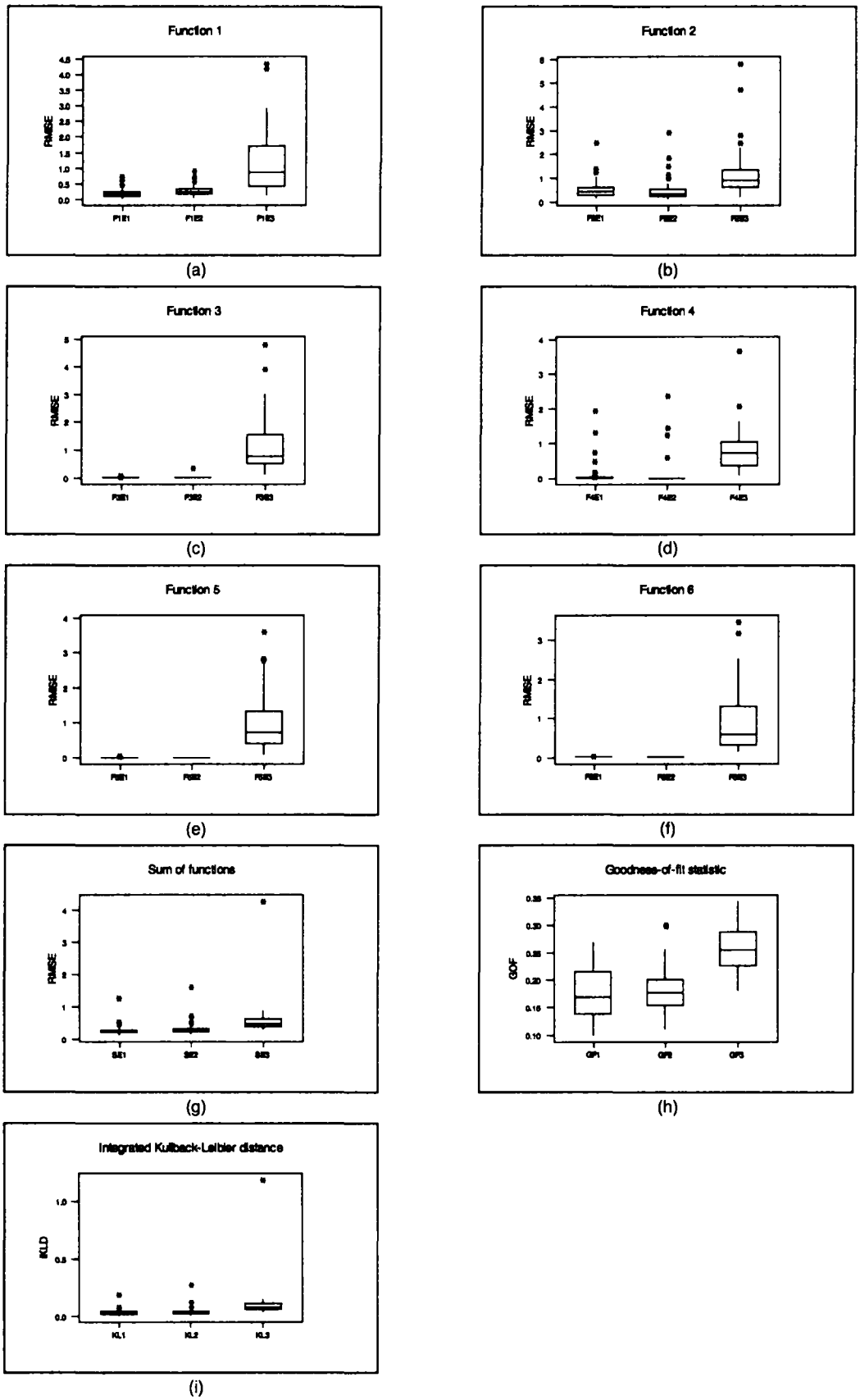

Figure 6. Boxplots of the RMISEs, IGFC, and IKLD From Experiment 4 for the Probit Regression Case. (a) Function 1; (b) function 2; (c) function 3; (d) function 4; (e) function 5; (f) function 6; (g) sum of functions; (h) goodness-of-fit statistic; (i) integrated Kullback-Leibler distance. In each panel the three boxplots correspond to, from left to right, the estimators obtained by model averaging, variable selection followed by function estimation, and unconstrained estimation.

ratio, .54; and type A, .09. This suggests that Cholesterol ratio should be estimated nonparametrically and the remaining functions can be estimated using linear functions. The number of iterations for both sampling schemes were the same as in the variable selection procedure discussed above.

We also fit a model using the four variables in the previous analysis as well as the additional variables of age (AGE); total energy (TE), a measure of the total energy expended in leisure time and occupational activities; and family history (FH), a 0–1 variable, with 1 indicating that a family member has had a heart attack and 0 indicating that a family member has not had a heart attack. Hastie and Tibshirani (1987) used these seven variables in their analysis. The probit model using these variables is

$$p(\text{heart attack}|\text{BP, TOB, CR, TYPEA, AGE, TE, FH})$$
$$= \Phi(\beta_0 + f_1(\text{BP}) + f_2(\text{TOB}) + f_3(\text{CR}) + f_4(\text{TYPEA}) + f_5(\text{AGE}) + f_6(\text{TE}) + \beta_7(\text{FH}), \quad (12)$$

where the functions $f_1, \ldots, f_6$ are estimated nonparametrically and variable selection is carried out on all seven explanatory variables. The marginal posterior probabilities of inclusion for the explanatory variables are blood pressure, 0.10; tobacco, 0.36; cholesterol ratio, .32; type A, .80; energy, .41; age, 1.00; and family history, 0.99. The results suggest strong evidence that type A, age, and family history should be included in the model, moderate evidence that tobacco, cholesterol ratio, and energy should be included, and no evidence that blood pressure should be included. The highest joint posterior probabilities for the $2^7 = 128$ combinations of functions was $\text{pr}\{(J_1, \ldots, J_7) = (0,0,0,1,1,0,1)|y\} = 0.17$, which also indicates that type A, age, and family history should be included in the model. The next two highest posterior probabilities were $\text{pr}\{(J_1, \ldots, J_7) = (0,0,0,1,1,1,1)|y\} = .15$ and $\text{pr}\{(J_1, \ldots, J_7) = (0,1,0,1,0,1,1)|y\} = .11$. Figure 8 plots the estimates of $f_1, \ldots, f_6$ and one standard deviation confidence bands obtained by model averaging.

To check the quality of the variable selection results, two simulations were done using the models in (11) and (12). For the model in (11), the functions for blood pressure, tobacco, and cholesterol ratio were set to be the linear function $f(s) = 2(s/s_{\max}) - 1$, and the function for type A was set to be the exponential function $f(s) = \exp(1.1^*(s/s_{\max})^3) - 2$, where $s_{\max}$ is the maximum value of the regressor. The regressor values are divided by $s_{\max}$, so that the function values will have a maximum range of two. Ten runs of the simulation were done using 20,000 iterations for the warmup and sampling periods for both sampling schemes. Of the 10 runs, all four variables were selected in four of the runs and three variables were selected in four of the other runs. One variable and two variables were selected

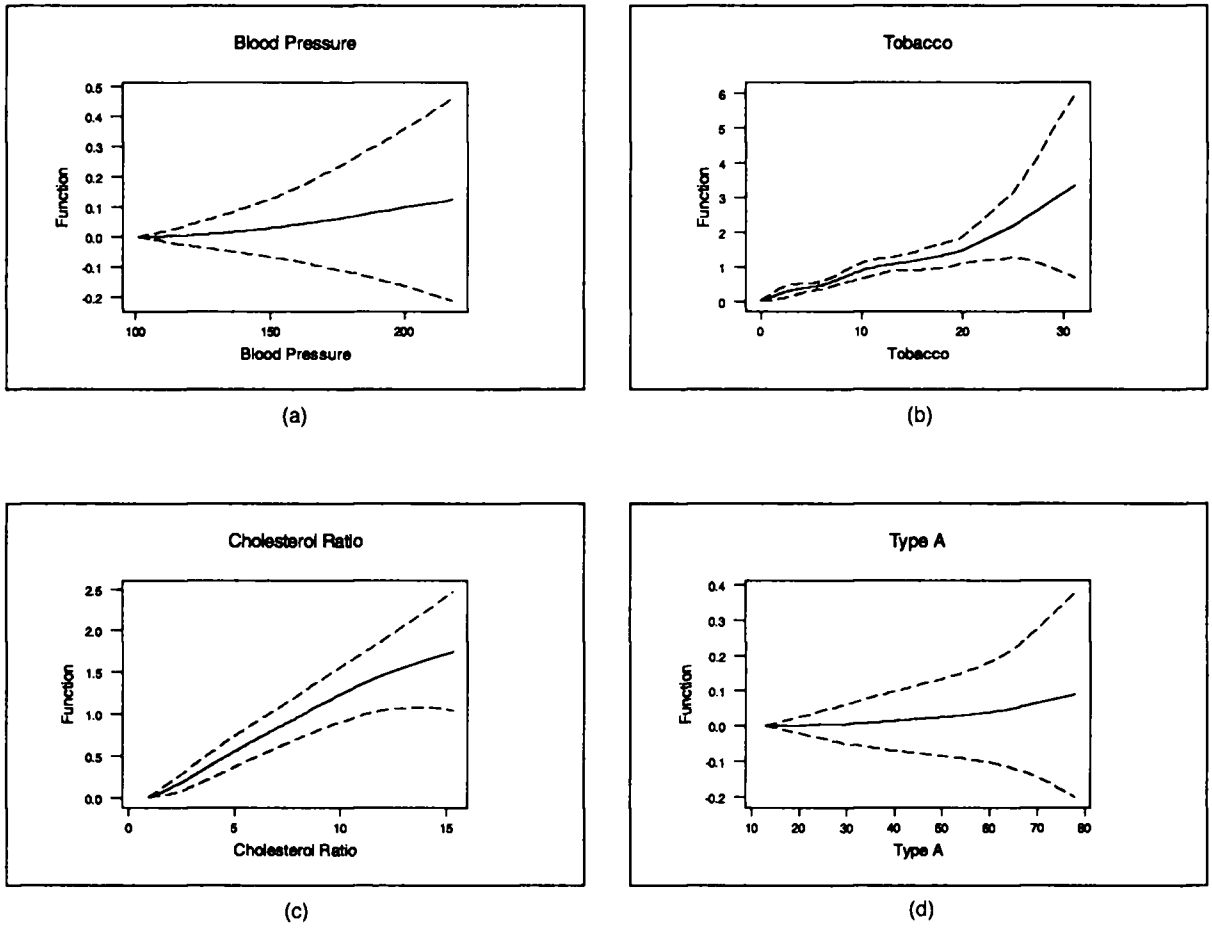

Figure 7. The Heart Attack Data Using Model Averaging (Method 1) With the Four Variables (a) Blood Pressure, (b) Tobacco, (c) Cholesterol Ratio, and (d) Type A in the Model. The solid lines represent the function estimates; the dashed lines, one standard deviation confidence bands.

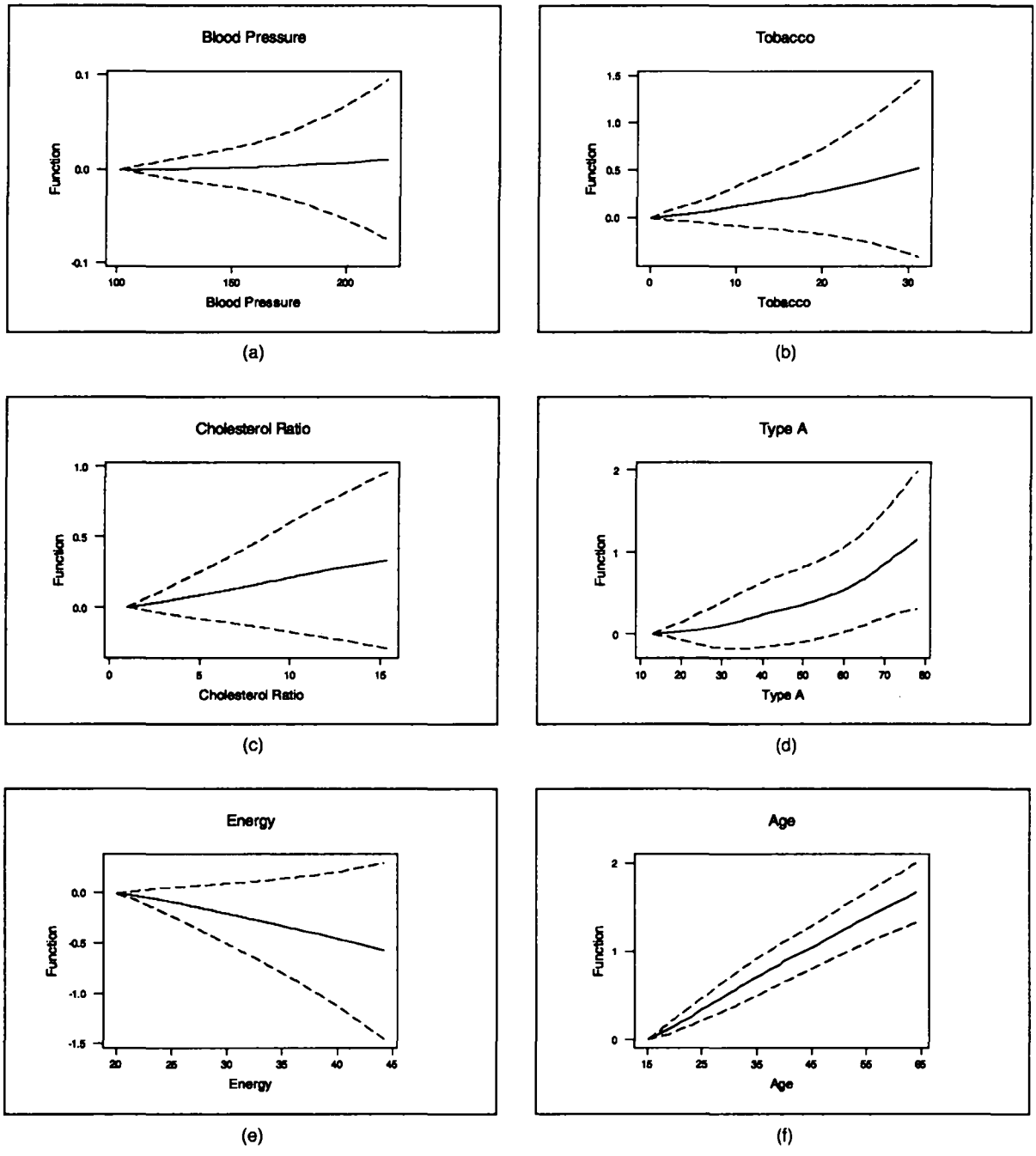

Figure 8. The Heart Attack Data Using Model Averaging (Method 1) With All Seven Variables in the Model: (a) Blood Pressure; (b) Tobacco; (c) Cholesterol Ratio; (d) Type A; (e) Energy; and (f) Age. The solid lines represent the function estimates; the dashed lines, one standard deviation confidence bands.

in the other two runs. Thus the procedure selected at least three of the four variables correctly 80% of the time.

For the model in (12), the functions for blood pressure, tobacco, cholesterol ratio, and family history were set to be the linear function $f(s) = 2(s/s_{\max}) - 1$, and the functions for type A, total energy, and age were set to be the exponential function $f(s) = \exp(1.1^*(s/s_{\max})^3) - 2$. Ten runs of the simulation were done using 20,000 iterations for the warmup and sampling periods for both sampling schemes. Of the 10 runs, six variables were selected once, five variables were selected five times, four variables were selected once, three variables were selected twice, and two variables were selected once.

6. CONCLUSIONS

This article has presented an empirical Bayes approach for variable selection and function estimation in Gaussian and probit nonparametric regression models. To carry out the inference, it is necessary to impose a proper data-based prior on the smoothing parameters and the starting values of each function. We propose a prior that is motivated by BIC but is more convenient to work with for the class of problems that we consider than previous data-based proposals. Our approach to data-based priors and variable selection can also be applied to more general statistical problems.

The variable selection and function estimation procedure given herein is a two-step process. The first step uses a

sampling scheme developed by Wong and Kohn (1996) to provide the data-based proper priors for the smoothing parameter and the starting value for the first derivative of each function in the model. The data-based priors are then used in the second stage of the procedure, in which the posterior probabilities that each variable should be included in the model are computed. The second stage also provides function estimates using model averaging. This article provides an efficient MCMC sampling scheme to carry out the computations in the second stage of the procedure. Extensive simulation results show that both the variable selection and model averaging procedures work well in a practical sense.

It is straightforward to extend the approach in this article to variable selection and model averaging for a wide range of time series models. Examples include stochastic coefficient regression models with Gaussian errors (e.g., Min and Zellner 1993), structural time series models (e.g., Harvey 1989), and binary stochastic coefficient regression models. In general, the approach of this article can be applied to any time series model that can be written in state–space form.

APPENDIX: IMPLEMENTING SAMPLING SCHEME 2

Here we outline how to implement steps 1–4 in sampling scheme 2. For step 1, let $\tilde{\mathbf{y}} = \mathbf{y} - \beta_2\mathbf{t} - \mathbf{g}_2$, so, by (6),

$$\tilde{\mathbf{y}} = \beta_0\boldsymbol{\iota} + \beta_1\mathbf{s} + \mathbf{g}_1 + \mathbf{e}, \tag{A.1}$$

where $\boldsymbol{\iota} = (1, \ldots, 1)', \mathbf{g}_1 \sim \mathrm{N}(0, \tau_1^2\mathbf{V}_1)$ and $\mathbf{V}_1$ is a positive definite matrix with ijth element $V_1(i,j) = (t_i^2/2)(t_j - t_i/3)$ (see Wahba 1990, p. 30) and $\mathbf{e} \sim \mathrm{N}(0, \sigma^2 I)$. Let $\mathbf{V}_1 = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1'$ be the spectral decomposition of $\mathbf{V}_1$, where $\mathbf{P}_1$ is an $n \times n$ matrix whose columns contain the eigenvectors of $\mathbf{V}_1$ and $\mathbf{D}_1$ is an $n \times n$ diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{V}_1$. Multiplying both sides of (A.1) by $\mathbf{P}_1'$ gives

$$\tilde{\mathbf{y}} = \beta_0\tilde{\boldsymbol{\iota}} + \beta_1\tilde{\mathbf{s}} + \tilde{\mathbf{g}}_1 + \tilde{\mathbf{e}},$$

where $\tilde{\mathbf{y}} = \mathbf{P}_1'\tilde{\mathbf{y}}, \tilde{\boldsymbol{\iota}} = \mathbf{P}_1'\boldsymbol{\iota}, \tilde{\mathbf{s}} = \mathbf{P}_1'\mathbf{s}, \tilde{\mathbf{g}}_1 = \mathbf{P}_1'\mathbf{g}_1 \sim \mathrm{N}(0, \tau_1^2\mathbf{D}_1)$, and $\tilde{\mathbf{e}} = \mathbf{P}_1'\mathbf{e} \sim \mathrm{N}(0, \sigma^2 I_n)$. The spectral decomposition $\mathbf{V}_1 = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1'$ needs to be done only once before the sampling scheme begins.

Step 1a: Generate $J_{1\tau}$, $J_{1\beta}|\mathbf{y}$, β_2, $\mathbf{g}_2$, and σ^2

To generate $J_{1\tau}$ and $J_{1\beta}$ given $\mathbf{y}, \beta_2, \mathbf{g}_2$, and σ^2, it is necessary to evaluate the discrete distribution $p(J_{1\tau}, J_{1\beta}|\mathbf{y}, \beta_2, \mathbf{g}_2, \sigma^2) = p(J_{1\tau}, J_{1\beta}|\tilde{\mathbf{y}}, \sigma^2) \propto p(\tilde{\mathbf{y}}|J_{1\tau}, J_{1\beta}, \sigma^2)p(J_{1\tau})p(J_{1\beta})$. We show how to evaluate the likelihood $p(\tilde{\mathbf{y}}|J_{1\tau}, J_{1\beta}, \sigma^2)$ for $J_{1\tau} = 1$ and $J_{1\beta} = 1$. The likelihood for the other three combinations of $J_{1\tau}$ and $J_{1\beta}$ are evaluated similarly, but more simply. Let $\theta_1 = \log(\tau_1^2)$. Then

$$\begin{aligned} &p(\tilde{\mathbf{y}}|J_{1\tau} = 1, J_{1\beta} = 1, \sigma^2) \\ &\quad = \int\int\int p(\tilde{\mathbf{y}}|J_{1\tau} = 1, J_{1\beta} = 1, \beta_0, \beta_1, \theta_1, \sigma^2) \\ &\qquad \times p(\theta_1|J_{1\tau} = 1)p(\beta_0, \beta_1|J_{1\beta} = 1)\, d\beta_0\, d\beta_1\, d\theta_1 \end{aligned} \tag{A.2}$$

and

$$\begin{aligned} &p(\tilde{\mathbf{y}}|J_{1\beta} = 1, J_{1\tau} = 1, \beta_0, \beta_1, \theta_1, \sigma^2) \\ &\quad = (2\pi)^{-n/2}|\sigma^2\mathbf{I} + e^{\theta_1}\mathbf{D}_1|^{-1/2} \exp\left\{\frac{1}{2}(\tilde{\mathbf{y}} - \beta_0\tilde{\boldsymbol{\iota}} - \beta_1\tilde{\mathbf{s}})'\right. \\ &\qquad \left. \times (\sigma^2\mathbf{I} + e^{\theta_1}\mathbf{D}_1)^{-1}(\tilde{\mathbf{y}} - \beta_0\tilde{\boldsymbol{\iota}} - \beta_1\tilde{\mathbf{s}})\right\}. \end{aligned}$$

β_0 and β_1 can be integrated out analytically, whereas θ_1 can be integrated out numerically using a Gaussian quadrature procedure. Given the diagonal eigenvalue matrix $\mathbf{D}_1$, the function evaluations required for the numerical integration in (A.2) can be done in $O(n)$ operations. We integrate with respect to $\theta_1 = \log(\tau_1^2)$ rather than τ_1^2, because the function in (A.2) is approximately normal with respect to θ_1.

Step 1b: Generate $\tau_1^2|J_{1\tau}$, $J_{1\beta}$, $\mathbf{y}$, β_2, $\mathbf{g}_2$, and σ^2

If $J_{1\tau} = 0$, then τ_1^2 is set to 0. If $J_{1\tau} = 1$, then $\theta_1 = \log\tau_1^2$ is generated using a normal approximation to the distribution of $\theta_1|\tilde{\mathbf{y}}, \sigma^2, J_{1\tau} = 1, J_{1\beta}$. The density $p(\theta_1|\tilde{\mathbf{y}}, \sigma^2, J_{1\tau} = 1, J_{1\beta})$ is approximated by a Gaussian density with mean $E(\theta_1|\tilde{\mathbf{y}}, \sigma^2, J_{1\tau} = 1, J_{1\beta})$ and variance $\mathrm{var}(\theta_1|\tilde{\mathbf{y}}, \sigma^2, J_{1\tau} = 1, J_{1\beta})$. The expectations $E(\theta_1^j|\tilde{\mathbf{y}}, \sigma^2, J_{1\tau} = 1, J_{1\beta}), j = 1, 2$, used to obtain these moments are evaluated using a Gaussian quadrature procedure similar to the one used to evaluate the integrals in step 1a. The numerical integrations required to obtain $E(\theta_1^j|\tilde{\mathbf{y}}, \sigma^2, J_{1\tau} = 1), j = 1, 2$, require almost no additional calculations, because the function evaluations are done in step 1a.

Step 1c: Implement the Metropolis–Hastings step

To implement the Metropolis–Hastings step, let $q(J_{1\tau}, J_{1\beta})$ be the discrete distribution from which $(J_{1\tau}, J_{1\beta})$ is generated and let $q(\theta_1|J_{1\tau}, J_{1\beta})$ be the distribution from which θ_1 is generated. Let $J_{1\tau}^P, J_{1\beta}^P$, and θ_1^P be the generated values of $J_{1\tau}, J_{1\beta}$, and θ_1, and let $J_{1\tau}^C, J_{1\beta}^C$, and θ_1^C be the current (i.e., pregeneration) values. Form the ratio

$$\begin{aligned} R = &\frac{p(\tilde{\mathbf{y}}|J_{1\tau}^P, J_{1\beta}^P, \theta_1^P)p(\theta_1^P|J_{1\tau}^P, J_{1\beta}^P)p(J_{1\tau}^P, J_{1\beta}^P)}{p(\tilde{\mathbf{y}}|J_{1\tau}^C, J_{1\beta}^C, \theta_1^C)p(\theta_1^C|J_{1\tau}^C, J_{1\beta}^C)p(J_{1\tau}^C, J_{1\beta}^C)} \\ &\times \frac{q(J_{1\tau}^C, J_{1\beta}^C)q(\theta_1^C|J_{1\tau}^C, J_{1\beta}^C)}{q(J_{1\tau}^P, J_{1\beta}^P)q(\theta_1^P|J_{1\tau}^P, J_{1\beta}^P)}. \end{aligned}$$

Then, with probability $\min\{1, R\}$, we accept $J_{1\tau}^P, J_{1\beta}^P$, and θ_1^P as the new values of $J_{1\tau}, J_{1\beta}$, and θ_1; otherwise, we retain the current values.

Step 2: Generate β_0, β_1, $\mathbf{g}_1|J_{1\beta}$, τ_1^2, $\mathbf{y}$, β_2, $\mathbf{g}_2$, and σ^2

If $J_{1\beta} = 1$, then β_0 and β_1 are generated from

$$\begin{aligned} &p(\beta_0, \beta_1|\mathbf{y}, J_{1\beta} = 1, \tau_1^2, \beta_2, \mathbf{g}_2, \sigma^2) \\ &\quad \propto p(\mathbf{y}|J_{1\beta} = 1, \tau_1^2, \boldsymbol{\beta}, \mathbf{g}_2, \sigma^2)p(\beta_0, \beta_1|\mathbf{y}, J_{1\beta} = 1, \tau_1^2, \sigma^2), \end{aligned}$$

which is Gaussian in β_0 and β_1. Note that $\mathbf{g}_1$ is integrated out analytically. If $J_{1\beta} = 0$, then β_1 is set to 0, and only β_0 is generated.

To generate $\mathbf{g}_1$, if $J_{1\tau} = 0$, then the vector $\mathbf{g}_1$ is set to 0, whereas if $J_{1\tau} = 1$, then $\mathbf{g}_1$ is generated following Wong and Kohn (1996).

Steps 3 and 4 are carried out similarly to steps 1 and 2.

[Received February 1997. Revised November 1998.]

REFERENCES

Albert, J., and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88, 669–679.

Ansley, C. F., Kohn, R., and Tharm, D. (1991), "The Performance of Cross-Validation and Maximum Likelihood Estimators of Spline Smoothing Parameters," *Journal of the American Statistical Association*, 86, 1042–1050.

Berger, J. O., and Pericchi, L. R. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.
Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174.
Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian Curve Fitting," *Journal of the Royal Statistical Society*, Ser. B, 80, 331–350.
Erkanli, A., and Gopalan, R. (1996), "Bayesian Nonparametric Regression: Smoothing Using Gibbs Sampling," in *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner*, eds. D. Berry, K. Chaloner, and J. Geweke, New York: Wiley, pp. 267–277.
Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.
George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.
Hardle, W., and Korostelev, A. (1996), "Search for Significant Variables in Nonparametric Additive Regression," *Biometrika*, 83, 541–549.
Harvey, A. C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge, U.K.: Cambridge University Press.
Hastie, T. J., and Tibshirani, R. J. (1987), "Generalized Additive Models: Some Applications," *Journal of the American Statistical Association*, 82, 371–386.
Kohn, R. (1983), "Consistent Estimation of Minimal Model Dimension," *Econometrica*, 51, 367–376.
Min, C., and Zellner, A. (1993), "Bayesian and non-Bayesian Methods for Combining Models and Forecasts With Applications to Forecasting International Growth Rates," *Journal of Econometrics*, 56, 89–118.
O'Hagan, A. (1995), "Fractional Bayes Factors for Model Comparison" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 99–138.
Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.
Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: Wiley.
Schwartz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.
Smith, M., and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–344.
Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1762.
Wahba, G. (1978), "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression," *Journal of the Royal Statistical Society*, Ser. B, 40, 364–372.
——— (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, Philadelphia: SIAM.
Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995), "Smoothing Spline ANOVA for Exponential Families, With Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy," *The Annals of Statistics*, 23, 1865–1895.
Wong, C., and Kohn, R. (1996), "A Bayesian Approach to Additive Semiparametric Regression," *Journal of Econometrics*, 74, 209–235.
Wood, S., and Kohn, R. (1998), "A Bayesian Approach to Robust Nonparametric Binary Regression," *Journal of the American Statistical Association*, 93, 203–213.
Wood, S., Shively, T. S., and Kohn, R. (1996), "Model Selection in Spline Nonparametric Regression," working paper, Australian Graduate School of Management.

Comment

Babette A. BRUMBACK, David RUPPERT, and M. P. WAND

1. INTRODUCTION

This article by Tom Shively, Robert Kohn, and Sally Wood provides a very effective solution to the difficult model selection problem in multiple predictor semiparametric regression, adding to a long list of impressive work of this type by Kohn and coauthors. As usual, the authors have done a thorough job, with lots of simulation testing to ensure good performance of their proposed strategy.

As the article's title suggests, the model and fitting procedure can be broken into two components: (a) variable selection through Bayesian modeling and Markov chain Monte Carlo (MCMC) schemes, and (b) function estimation. Concerning (a), we certainly have less expertise about computational Bayesian methods than the authors, and there is little we can add to their MCMC algorithm. Most of our research involves function estimation, and most of our comments are on this topic, though Section 4 briefly discusses non-Bayesian methods of variable selection. Also, for simplicity, our discussion deals only with the Gaussian case.

2. FUNCTION ESTIMATION

The function estimation component consists of a model expressed as a line plus an integrated Wiener process (with arbitrary variance) and with fitting achieved using state-space formulation and application of the Kalman filter.

2.1 Complexity

Additive models continue to gain popularity in applied research as a flexible and interpretable regression technique. For example, they are now used routinely in environmental epidemiology studies conducted at the Harvard School of Public Health. The input is a spreadsheet of numbers; one column representing measurements on the response variable, the remainder being measurements on each of the covariates. The output is a set of curves and coefficients, along with variability estimates, which describe the effects of each

Babette A. Brumback is Postdoctoral Research Fellow and M. P. Wand is Associate Professor, Department of Biostatistics, School of Public Health, Harvard University, Boston, Massachusetts 02115. David Ruppert is Professor, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853. These comments benefited from a conversation with Jim Hobert. This research was supported by National Science Foundation grant DMS-9804058 (Ruppert) and by U.S. Environmental Protection Agency grant R 824757 (Wand).

Journal of the American Statistical Association
September 1999, Vol. 94, No. 447, Theory and Methods