Efficient sampling schemes for Bayesian MARS models with many predictors

Article *in* Statistics and Computing · April 2005

DOI: 10.1007/s11222-005-6201-x · Source: DBLP

3 authors, including:

Hiep Duc
Office of Environment and Heritage
99 PUBLICATIONS **322** CITATIONS

SEE PROFILE

Statistics and Computing **15**: 93–101, 2005

Efficient sampling schemes for Bayesian MARS models with many predictors

DAVID J. NOTT*, ANTHONY Y.C. KUK† and HIEP DUC‡

**Department of Statistics, University of New South Wales, Sydney NSW 2052 Australia*
D.Nott@unsw.edu.au
†*Associate Professor, Department of Statistics and Applied Probability, The National University of Singapore, Singapore*
stakuka@nus.edu.sg
‡*Atmospheric Scientist, New South Wales Environmental Protection Authority, P.O. Box 29, Lidcombe, NSW 1825 Australia*

Received October 2002 and accepted October 2004

Multivariate adaptive regression spline fitting or MARS (Friedman 1991) provides a useful methodology for flexible adaptive regression with many predictors. The MARS methodology produces an estimate of the mean response that is a linear combination of adaptively chosen basis functions. Recently, a Bayesian version of MARS has been proposed (Denison, Mallick and Smith 1998a, Holmes and Denison, 2002) combining the MARS methodology with the benefits of Bayesian methods for accounting for model uncertainty to achieve improvements in predictive performance. In implementation of the Bayesian MARS approach, Markov chain Monte Carlo methods are used for computations, in which at each iteration of the algorithm it is proposed to change the current model by either (a) Adding a basis function (birth step) (b) Deleting a basis function (death step) or (c) Altering an existing basis function (change step). In the algorithm of Denison, Mallick and Smith (1998a), when a birth step is proposed, the type of basis function is determined by simulation from the prior. This works well in problems with a small number of predictors, is simple to program, and leads to a simple form for Metropolis-Hastings acceptance probabilities. However, in problems with very large numbers of predictors where many of the predictors are useless it may be difficult to find interesting interactions with such an approach. In the original MARS algorithm of Friedman (1991) a heuristic is used of building up higher order interactions from lower order ones, which greatly reduces the complexity of the search for good basis functions to add to the model. While we do not exactly follow the intuition of the original MARS algorithm in this paper, we nevertheless suggest a similar idea in which the Metropolis-Hastings proposals of Denison, Mallick and Smith (1998a) are altered to allow dependence on the current model. Our modification allows more rapid identification and exploration of important interactions, especially in problems with very large numbers of predictor variables and many useless predictors. Performance of the algorithms is compared in simulation studies.

Keywords: Markov chain Monte Carlo, multivariate adaptive regression splines, nonparametric regression, Bayesian inference, high-dimensional regression

1. Introduction

A powerful general approach to nonparametric regression involves modelling the mean of the response in terms of a linear combination of adaptively chosen basis functions. Papers using this approach include Biller (2000), Denison, Mallick and Smith (1998b), DiMatteo, Genovese and Kass (2001), Friedman and Silverman (1989), Kohn, Smith and Chan (2001) and Smith and Kohn (1996). In problems involving many predictors where there are interactions involving just a small number of the predictors, the multivariate adaptive regression spline (MARS) approach of Friedman (1991) is very powerful. In MARS, the mean response is represented as a linear combination of basis terms which are tensor products of linear spline basis functions.

In the original MARS approach and subsequent variants (Kooperberg, Bose and Stone 1997) the model is built up in

a stepwise fashion, adding basis functions in a greedy search to produce an overfitted model and then pruning back the overfitted model by deletion of terms to arrive at the final model, which is chosen via minimization of a penalized likelihood criterion. In the forward stage of the stepwise search, an interaction term can only be added by taking one of the existing basis functions in the model and multiplying by a univariate function of one of the predictors. This idea of building up higher order interactions from lower order ones greatly reduces the complexity of the search for good basis functions to add. The MARS methodology can also be extended to classification problems (Holmes and Denison 2002, Kooperberg, Bose and Stone 1997).

Recently, a Bayesian approach to MARS has been proposed (Denison, Mallick and Smith 1998a, Holmes and Denison 2002) which combines the original MARS idea with Bayesian methods for accounting for model uncertainty to achieve improvements in predictive performance. A full probability model is set up for all the unknown parameters in the MARS representation of the mean response, and then the posterior distribution on the parameters is explored using the reversible jump Markov chain Monte Carlo (MCMC) method of Green (1995).

In the MCMC scheme of Denison, Mallick and Smith (1998a), at each iteration of the algorithm we consider one of three types of changes to the current model: (a) Addition of a new basis function (birth step), (b) Deletion of a basis function (death step) or (c) Change of an existing basis function (change step). When a new basis function is added to the model in a birth step, the type of the basis function (how many predictors are involved and which ones) is simulated from the prior distribution on basis function types, which is uniform. This procedure works well in problems with small numbers of predictors, leads to a simple form for the Metropolis-Hastings acceptance probabilities, and is easy to program. However, when there are large numbers of predictors the number of different possible interactions is huge, even if interactions are restricted to second order, and it becomes very difficult to find interesting interaction structure by simulating proposals from the prior.

The contribution of this paper is to suggest Metropolis-Hastings proposal distributions in Bayesian MARS that incorporate some of the intuition of the original MARS stepwise search algorithm by allowing dependence of the proposal on what variables and interactions currently appear in the model. In problems with large numbers of variables and many useless predictors it may be possible to identify which variables are important fairly quickly, even if the exact form of the relationship between the response and the important predictor variables is complex. Our proposal distributions focus more effort on birth steps involving variables currently in the model, while still allowing birth steps involving other variables with a smaller probability. Also, our method adapts the way that birth steps are done according to the interaction order of terms currently appearing in the model, so for near additive relationships our method will propose mostly main effects terms. While for general purpose use in small problems the sampling scheme of Denison, Mallick and Smith (1998a) works well, our modifications of the Bayesian MARS sampling algorithm can produce large improvements in problems with large numbers of predictors where many of the predictors are useless. Nonparametric regression problems involving large numbers of predictors and many useless predictors are of great interest in recent data mining applications.

We now describe the Bayesian MARS model in detail. First we introduce some notation. Let $y = (y_1, \ldots, y_n)^T$ be a vector of n observations of a response variable and let X be an $n \times p$ matrix where the ith row $x_i = (x_{i1}, \ldots, x_{ip})^T$ is an observation of the values of p predictor variables $(X_1, \ldots, X_p)^T$ for the ith response. We assume the data are generated as

$$y_i = f(x_i) + \epsilon_i$$

where $f(\cdot)$ is a mean function to be estimated from the data and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is a collection of zero mean independent Gaussian errors with variance σ^2. In the Bayesian MARS approach of Denison, Mallick and Smith (1998a), the following representation is adopted for $f(\cdot)$:

$$f(x) = \beta_0 + \sum_{i=1}^{m} \beta_i \psi_i(x)$$

where $\psi_i(x)$, $i = 1, \ldots, m$ are adaptively chosen basis functions, β_0 is an intercept term and β_i is the coefficient of the ith basis function. The basis function $\psi_i(x)$ has the representation

$$\psi_i(x) = \prod_{j=1}^{R_i} \left[g_{ji} \left(x_{c_{ji}} - \xi_{ji} \right) \right]_+$$

where $[z]_+ = \max(0, z)$, R_i is the interaction order of $\psi_i(x)$ (how many of the predictor variables appear in $\psi_i(x)$), g_{ji} is a sign parameter equal to one or negative one, the $c_{ji} \in \{1, \ldots, p\}$ give the indices of the predictors appearing in the ith basis function (constrained to be distinct) and the ξ_{ji} are knot points.

The unknown parameters in the MARS representation are the number of basis functions m, the coefficients $\beta = (\beta_0, \ldots, \beta_m)^T$, the error variance σ^2 and the parameters in each of the basis functions. For the basis function parameters we write $c_i = (c_{1i}, \ldots, c_{R_i i})^T$, $g_i = (g_{1i}, \ldots, g_{R_i i})^T$ and $\xi_i = (\xi_{1i}, \ldots, \xi_{R_i i})^T$ for the parameters defining basis function i. Also, write $R = (R_1, \ldots, R_m)^T$, $c = (c_1^T, \ldots, c_m^T)^T$, $g = (g_1^T, \ldots, g_m^T)^T$ and $\xi = (\xi_1^T, \ldots, \xi_m^T)^T$.

Our priors on the unknown parameters are similar to those in Denison, Mallick and Smith (1998a), although we have incorporated some slight modifications following Holmes and Denison (2002). In the paper by Holmes and Denison (2002) two class classification problems are considered. A probit model is suggested where computations can be conveniently carried out by writing the model in an equivalent form involving the introduction of appropriate latent Gaussian variables (Albert and Chib 1993). Conditional on the latent Gaussian variables their model is just a Bayesian MARS regression model with Gaussian errors and $\sigma^2 = 1$, so that the discussion of Holmes and Denison is relevant to the current setting.

The prior has a hierarchical structure, and after assuming appropriate conditional independence relationships between parameters we can write our prior as

$$p(\lambda, m, R, c, g, \xi, \sigma^2, \beta) = p(\lambda)p(m \mid \lambda)p(R \mid m)p(c \mid R, m) \times p(g|c, R, m)p(\xi \mid c, R, m)p(\sigma^2)p(\beta \mid m, \sigma^2)$$

where λ is a hyperparameter in the prior on m. For the prior on m, we use a truncated Poisson distribution,

$$p(m) \propto \frac{\exp(-\lambda)\lambda^m}{m!}$$

$m = 0, \ldots, m_{\max}$ where $m_{\max}$ is the maximum allowable number of basis functions. We follow Denison, Mallick and Smith (1998a) in giving the hyperparameter λ a Gamma(10,10) prior. For the prior on R, we assume independent priors on the components R_i. The prior on R_i is uniform on $\{1, \ldots, I\}$ where I is the maximum allowable interaction order for a basis function. A common choice for I is 2, restricting interactions to second order, and in all that follows we set $I = 2$. Setting $I = 1$ would result in an additive model.

For the prior on c, we assume the components c_i are independent *a priori*, and that the prior for c_i is uniform,

$$p(c_i \mid R_i, m) = \binom{p}{R_i}^{-1}.$$

For the prior on g, we assume an independent uniform prior on all components. For the prior on ξ, we assume independent priors for all components, and the prior for ξ_{ji} is uniform on $\{x_{1c_{ji}}, \ldots, x_{nc_{ji}}\}$ (the set of observed values for predictor c_{ji}). Our prior on σ^2, following Denison, Mallick and Smith (1998a), is a fairly diffuse inverse gamma distribution,

$$p(\sigma^{-2}) \sim \text{Gamma}(10^{-3}, 10^{-3}).$$

Finally, for our prior on β, we assume components are independent *a priori* with β_i given a normal prior with mean 0 and variance $\mu^{-1}\sigma^2$ where μ is chosen to be small (making the prior diffuse).

Now that we have established the model, we describe our MCMC sampling algorithm.

2. Efficient sampling schemes for Bayesian MARS models

The posterior distribution on the unknown parameters in the Bayesian MARS model is very complex, and posterior means and other summary quantities of interest cannot be obtained analytically. We use reversible jump Markov chain Monte Carlo methods (Green 1995) for exploring the posterior distribution. For a general overview of MCMC methods see Gilks *et al.* (1996).

We describe our sampling algorithm first and then discuss how it differs from the sampling scheme of Denison, Mallick and Smith (1998a). Write $\theta \in \Theta$ for the set of unknown parameters, where Θ denotes the parameter space. Our MCMC algorithm generates a Markov chain on Θ,

$$\left\{\theta^{(t)} : t \geq 0\right\}$$

where $\theta^{(t)} = (\lambda^{(t)}, m^{(t)}, R^{(t)}, c^{(t)}, g^{(t)}, \xi^{(t)}, \sigma^{2(t)}, \beta^{(t)})$ denotes the value at iteration t for $(\lambda, m, R, c, g, \xi, \sigma^2, \beta)$. The chain is constructed so that the equilibrium distribution is the posterior distribution $p(\theta \mid y)$. Reversible jump MCMC provides the appropriate methodology (Green 1995), generalizing the ordinary Metropolis-Hastings algorithm to problems of Bayesian model selection (see, for instance, Chib and Greenberg 1995, or Tierney 1994, for introductory accounts of the Metropolis-Hastings algorithm). An alternative computational methodology for handling problems of Bayesian model selection is given by Carlin and Chib (1995).

Our algorithm for sampling the posterior has the same basic form as that of Denison, Mallick and Smith (1998a), but with some important differences which are described later. At each step of the algorithm we can either (a) Add a new basis function to the model (a birth step) (b) Delete a basis function from the model (a death step) or (c) Change an existing basis function (a change step). When there are m basis functions in the current model, write b_m, d_m and c_m for the probabilities of choosing birth, death and change steps respectively at a given iteration. In the examples we choose b_m, d_m and c_m so that the available move types have equal probability. The general form of the algorithm of Denison, Mallick and Smith (1998a) and of our own algorithm is given below, with more detail on the steps involved following.

1. Initialize $\theta^{(0)}$. Set $t = 0$.
2. With respective probabilities b_m, d_m and c_m propose either a birth, death or change step, generating the proposal values for $R^{(t+1)}$, $c^{(t+1)}$, $g^{(t+1)}$ and $\xi^{(t+1)}$.
3. After generating the proposal values in 2, compute the Metropolis-Hastings acceptance probability and determine if the proposal values are accepted or the old values retained.
4. Generate $\lambda^{(t+1)}$, $\sigma^{2(t+1)}$ and $\beta^{(t+1)}$ by sampling from the full conditional distributions for these parameters.
5. Increment t. Go to step 2 if $t \leq T$ where T is the number of iterations to run the chain.

To describe our birth, death and change moves in step 2 of our algorithm, we need some more notation. Writing $I(A)$ for the indicator function which is 1 when event A occurs and zero otherwise, we define for a given model M a set of weights

$$z_l^*(M) = \delta + \sum_{i=1}^{m} \sum_{j=1}^{R_i} I(l == c_{ji})$$

$l = 1, \ldots, p$. $z_l^*(M)$ reflects the importance of predictor l in model M, since the second term above is the number of basis functions in which x_l appears. δ is the weight given to a predictor

not appearing in model M. Define

$$z_l(M) = \frac{z_l^*(M)}{\sum_{m=1}^{p} z_m^*(M)}$$

$l = 1, \ldots, p$ so that $z_l(M)$ is a probability function. The probability function $z_l(M)$ will be used in defining proposal distributions for the birth step of our Metropolis-Hastings algorithm. Our algorithm also attempts to adjust the proportion of main effects proposals versus interaction term proposals according to how nearly additive the current model is. To give our proposal we define weights reflecting the relative importance of terms of different interaction orders in the model. For a given model M, let

$$I_l^*(M) = \gamma + \sum_{i=1}^{m} I(R_i = l)$$

$l = 1, 2$. The second term in the above expression is simply the number of terms of interaction order l. γ is the weight assigned to a given interaction order when there are no terms of that order in model M. Also, let

$$I_l(M) = \frac{I_l^*(M)}{I_1^*(M) + I_2^*(M)}$$

$l = 1, 2$ so that $I_l(M)$ is a probability function.

In a birth move in step 2 of our algorithm a new basis function (call it $\psi_{m+1}(x)$) is added to the model as follows. First, we generate the interaction order R_{m+1} of $\psi_{m+1}(x)$ by sampling from $I_l(M^c)$, where M^c denotes the current model. If $R_{m+1} = 1$, we determine which predictor is involved in $\psi_{m+1}(x)$ by sampling c_{1m+1} uniformly from $\{1, \ldots, p\}$. If $R_{m+1} = 2$, we first choose x_l as one of the pair of predictors in the basis function with probability $z_l(M^c)$, and then given that x_l was chosen first x_q $(q \neq l)$ is chosen with probability

$$\frac{z_q(M^c)}{1 - z_l(M^c)}.$$

Without regard to order, x_l and x_q are chosen as the pair of predictors involved in $\psi_{m+1}(x)$ with probability

$$z_{lq}(M^c) = \frac{z_l(M^c)z_q(M^c)}{1 - z_l(M^c)} + \frac{z_q(M^c)z_l(M^c)}{1 - z_q(M^c)}.$$

The way that the predictors involved in an interaction are generated gives more weight to variables currently appearing in the model. Once a predictor appears in a main effects term, it will have an increased probability of being proposed for inclusion in interaction terms. Predictors with significant main effects are more likely to be involved in interactions.

After determining the predictors involved in $\psi_{m+1}(x)$, we need to generate the parameters ξ_{m+1} and g_{m+1}. We generate these from their priors. The coefficients β can be sampled from the full conditionals, but this turns out to be unnecessary in step 2 of the algorithm: with the full conditional as the proposal for the coefficients (described later), the Metropolis-Hastings acceptance probability does not depend on the values of the coefficients, only on the posterior modes (cf. Holmes and Denison 2002). While this independence of the acceptance probability from the value of β occurs quite generally with the full conditional as proposal, it may be difficult to get an explicit expression for the acceptance probability unless we can calculate the full conditional distribution for β (which we can do here because of the normal likelihood and normal prior on β). Use of the full conditional as proposal is reminiscent of the related model space method of Carlin and Chib (1995).

Given that a birth step has been chosen, the probability of choosing the proposed model M^p is

$$b(M^c \to M^p) = \begin{cases} I_1(M^c)\left(\dfrac{1}{2pn}\right) & \text{if } R_{m+1} = 1 \\ I_2(M^c)z_{lq}(M^c)\left(\dfrac{1}{2n}\right)^2 & \text{if } R_{m+1} = 2 \text{ and } x_l \\ & \quad \text{and } x_q \text{ are in } \psi_{m+1}(x) \end{cases}$$

In the above probability the factors of $1/n$ come from the choice of knots and the factors of $1/2$ come from the choice of sign parameters. In the death move defined in a moment (which reverses the birth move) we propose deleting a basis function, and the basis function to delete is chosen uniformly from the current basis functions in the model. We can now write down the acceptance probability for our birth step. The acceptance probability is $\min\{1, \alpha\}$, where

$$\alpha = \text{Prior Ratio} \times \text{Likelihood Ratio} \times \text{Proposal Ratio}.$$

In the general reversible jump Metropolis-Hastings acceptance probability (see Green, 1995) there is also a Jacobian term in the expression for α given above, but this disappears here because of the way that proposals are generated for the coefficients unconstrained by current values.

The prior ratio is

$$\frac{p(m+1)}{p(m)} \times \frac{(1/2)^{(m+1)}}{(1/2)^m} \times \frac{\prod_{i=1}^{m+1}\binom{p}{R_i}^{-1}}{\prod_{i=1}^{m}\binom{p}{R_i}^{-1}} \times \frac{(m+1)!}{m!}$$

$$\times \frac{(1/2n)^{\sum_{i=1}^{m+1} R_i}}{(1/2n)^{\sum_{i=1}^{m} R_i}} \times \frac{(2\pi\mu^{-1}\sigma^2)^{-(m+2)/2}\exp(-\frac{\mu}{2\sigma^2}\beta'^T\beta')}{(2\pi\mu^{-1}\sigma^2)^{-(m+1)/2}\exp(-\frac{\mu}{2\sigma^2}\beta^T\beta)}$$

The terms in the above ratio represent (from left to right) the ratio of priors for m, the ratio of priors for R, the ratio of priors for c, a term which accounts for the different ways of ordering the basis functions, the ratio of priors for g and ξ and the ratio of priors for the coefficients.

To write down the likelihood ratio we first introduce some notation. Let ψ be the $n \times (m+1)$ design matrix with the first column a column of ones, and $\psi_{ij} = \psi_{j-1}(x_i)$ for $i = 1, \ldots, n$, $j = 2, \ldots, (m+1)$ where $\psi_j(x)$, $j = 1, \ldots, m$ are the basis functions in the current model (so that column j of ψ, $2 \leq j \leq m+1$ gives the values of the basis function $\psi_{j-1}(x)$ at

the observed predictor values). Also, let ψ' be the corresponding design matrix for the proposed model. Given the design matrix ψ, the Bayesian MARS model is just a linear model, and writing β for the coefficients in the current model and β' for the coefficients in the proposed model, the likelihood ratio can then be written as

$$\frac{(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(y-\psi'\beta')^T(y-\psi'\beta')\right)}{(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(y-\psi\beta)^T(y-\psi\beta)\right)}.$$

To give the proposal ratio, we also need some more notation. Denote the posterior mode of the coefficients in the current model as $\hat{\beta} = (\psi^T\psi + \mu I)^{-1}\psi^T y$, and let $\hat{\beta}' = (\psi'^T\psi' + \mu I)^{-1}\psi'^T y$ denote the posterior mode for the coefficients in the proposed model. The proposal distribution used for the coefficients in a birth move is just the full conditional distribution: the proposal distribution for β is thus normal, with mean $\hat{\beta}$ and covariance matrix $\sigma^2(\psi^T\psi + \mu I)^{-1}$. The proposal ratio is

$$\frac{d_{m+1}}{b_m(m+1)b(M^c \to M^p)} \times \frac{(2\pi\sigma^2)^{-\frac{m+1}{2}}|\psi^T\psi + \mu I|^{1/2}}{(2\pi\sigma^2)^{-\frac{m+2}{2}}|\psi'^T\psi' + \mu I|^{1/2}} \times \frac{\exp\left(-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^T(\psi^T\psi+\mu I)(\beta-\hat{\beta})\right)}{\exp\left(-\frac{1}{2\sigma^2}(\beta'-\hat{\beta}')^T(\psi'^T\psi'+\mu I)(\beta'-\hat{\beta}')\right)}.$$

The first term represents the model proposal ratio and the remainder is the proposal ratio for the coefficients.

Combining the prior ratio, proposal ratio and likelihood ratio and cancelling terms we get

$$\alpha = \frac{p(m+1)}{p(m)} \times \left(\frac{1}{2}\right) \times \binom{p}{R_{m+1}}^{-1} \times \left(\frac{1}{2n}\right)^{R_{m+1}} \times \frac{d_{m+1}}{b_m} \times \frac{1}{b(M^c \to M^p)} \times \frac{\mu^{1/2}|\psi^T\psi + \mu I|^{1/2}\exp\left(-\frac{1}{2\sigma^2}\hat{\beta}^T(\psi^T\psi+\mu I)\hat{\beta}\right)}{|\psi'^T\psi' + \mu I|^{1/2}\exp\left(-\frac{1}{2\sigma^2}\hat{\beta}'^T(\psi'^T\psi'+\mu I)\hat{\beta}'\right)}.$$

Note that the acceptance probability does not depend on β or β'.

The reverse move of the birth move we have just described in step 2 of our algorithm is the death move: in this move, we choose one of the current basis functions and delete this from the model. The basis function proposed for deletion is chosen uniformly from the basis functions in the current model. Without loss of generality (since we can relabel the basis functions) suppose that it is the mth basis function that is proposed for deletion. Then the acceptance probability for the death move is $\min\{1, \alpha\}$, where

$$\alpha = \text{Prior Ratio} \times \text{Likelihood Ratio} \times \text{Proposal Ratio}.$$

Following a similar argument to the one for the birth step, we obtain an expression for α which is the inverse of the expression for α in the birth step:

$$\alpha = \frac{p(m-1)}{p(m)} \times \binom{p}{R_m} \times 2 \times (2n)^{R_m} \times \frac{b_{m-1}}{d_m} \times b(M^p \to M^c) \times \frac{\mu^{-1/2}|\psi^T\psi + \mu I|^{1/2}\exp\left(-\frac{1}{2\sigma^2}\hat{\beta}^T(\psi^T\psi+\mu I)\hat{\beta}\right)}{|\psi'^T\psi' + \mu I|^{1/2}\exp\left(-\frac{1}{2\sigma^2}\hat{\beta}'^T(\psi'^T\psi'+\mu I)\hat{\beta}'\right)}$$

where as before ψ and $\hat{\beta}$ are the design matrix and posterior mode of coefficients in the current model and ψ' and $\hat{\beta}'$ are the design matrix and posterior mode of coefficients in the proposed model. Finally, our change move at step 2 of the algorithm and the updates from the full conditional distributions in step 4 are the same as in Denison, Mallick and Smith (1998a) and we do not discuss these further.

Before moving on to some examples we briefly outline the difference between our algorithm and that of Denison, Mallick and Smith (1998a). The difference lies in our definition of the birth step. In the birth step, we have used information about currently active variables and the interaction order of current basis functions in generating the basis function proposed for addition. In the algorithm of Denison, Mallick and Smith (1998a), when a new basis function $\psi_{m+1}(x)$ is added to the current model, then what they call the type of the basis function (essentially which predictors are involved in the basis function) is generated from the prior. This is simple to program, works well in problems with small numbers of predictors and leads to a simple form for the Metropolis-Hastings acceptance probability. However, when there are large numbers of predictors with many useless predictors it will be very difficult to find interesting interaction structure using this approach of simulating from the prior, even if interactions are restricted to second order, because of the very large number of possible interactions. It may be easier to find the important main effects terms, and then predictors with important main effects are more likely to be involved in important interactions. Our sampling algorithm attempts to exploit this heuristic. Our algorithm also attempts to adjust the proportion of main effects proposals versus interaction term proposals according to how nearly additive the current model is. These alterations can lead to substantial efficiency gains in some problems.

3. Simulation studies

In what follows we compare our proposed sampling scheme to a sampling scheme similar to that of Denison, Mallick and Smith (1998a). Denison, Mallick and Smith (1998a) define the type of a basis function, which is essentially the set of predictors involved in a basis function. They have a uniform prior on the type, which tends to favour higher order interactions since there tend to be more interaction type basis functions than main effects type basis functions. In the birth step of their algorithm, they simulate

the type of the basis function from the prior (in our notation when adding a basis function $\psi_{m+1}(x)$ this determines R_{m+1} and c_{m+1}). The uniform prior on the type is altered in Holmes and Denison (2002), where a prior similar to the one used in the present paper is discussed. That is, a hierarchical prior is used for the basis function type, with an explicit prior on interaction order, and then a prior on the variables involved in a basis function given interaction order. For comparison with our own method we alter the original sampling scheme of Denison, Mallick and Smith (1998a) by sampling from this altered prior in the birth step, similarly to Holmes and Denison (2002). When we talk about the method of Denison, Mallick and Smith (1998a) in the examples which follow we are actually referring to this slightly altered version of their original algorithm. In our proposed algorithm, we set the tuning parameters γ and δ in our proposal to be $\lambda/2$ and $1/4$ respectively for the simulations below. Setting $\gamma = \lambda/2$ means that in a model involving λ terms where all terms have the same interaction order a probability of $3/4$ is assigned to proposals involving terms of that interaction order. Setting $\delta = 1/4$ results in an odds of 5 for a predictor currently appearing in one basis term compared to a predictor currently appearing in none in the distribution $z_l(M)$ determining the predictors in proposals involving interactions. One could set these parameters to be more aggressive in focusing on variables in the current model, and in this situation it might be helpful to consider a hybrid algorithm where we choose randomly between one of our proposal moves and a proposal move of the kind considered by Denison, Mallick and Smith (1998a), a suggestion made to us by an anonymous reviewer.

3.1. *Simulated examples with 20 predictors*

We first consider some simulated examples with 20 predictors for five different test functions. These examples were introduced by Hwang *et al.* (1994) and were discussed by Denison, Mallick and Smith (1998a) in comparing Bayesian MARS with the original MARS algorithm of Friedman (1991).

For each test function, 225 observations are generated. The test functions are two dimensional, and the predictor values $(x_{i1}, x_{i2})^T$, $i = 1, \ldots, 225$ are generated independently and uniformly on $[0, 1]$. Since we are interested in this paper in problems involving large numbers of predictors and many useless predictors, we also simulate eighteen additional useless predictors (which do not affect the mean response) $(x_{i3}, \ldots, x_{i20})^T$ $i = 1, \ldots, 225$, also independently and uniform on $[0, 1]$. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{i20})^T$. The responses are generated as

$$y_i = f(\mathbf{x}_i) + \epsilon_i$$

$i = 1, \ldots, 225$, where the ϵ_i are independent normal errors with mean 0 and standard deviation 0.25. In the notation of Denison, Mallick and Smith (1998a), the five mean functions $f(\mathbf{x})$ considered are

$$f^{(1)}(\mathbf{x}) = 10.391((x_1 - 0.4)(x_2 - 0.6) + 0.36)$$

$$f^{(2)}(\mathbf{x}) = 24.234(r^2(0.75 - r^2)) \quad \text{where } r^2 = (x_1 - 0.5)^2 + (x_2 - 0.5)^2$$

$$f^{(3)}(\mathbf{x}) = 42.659\big(0.1 + \hat{x}_1\big(0.05 + \hat{x}_1^4 - 10\hat{x}_1^2\hat{x}_2^2 + 5\hat{x}_2^4\big)\big) \quad \text{where } \hat{x}_1 = x_1 - 0.5 \text{ and } \hat{x}_2 = x_2 - 0.5$$

$$f^{(4)}(\mathbf{x}) = 1.3356(1.5(1 - x_1) + \exp(2x_1 - 1)\sin(3\pi(x_1 - 0.6)^2) + \exp(3(x_2 - 0.5))\sin(4\pi(x_2 - 0.9)^2))$$

$$f^{(5)}(\mathbf{x}) = 1.9(1.35 + \exp(x_1)\sin(13(x_1 - 0.6)^2) \times \exp(-x_2)\sin(7x_2))$$

As in Denison, Mallick and Smith (1998a), we refer to $f^{(1)}(\mathbf{x})$ as the simple interaction function, $f^{(2)}(\mathbf{x})$ as the radial function, $f^{(3)}(\mathbf{x})$ as the harmonic function, $f^{(4)}(\mathbf{x})$ as the additive function and $f^{(5)}(\mathbf{x})$ as the complex interaction function.

Our simulation studies show a substantial improvement for our proposed sampling algorithm for several of the test functions. Before describing the results, we discuss how the sampling schemes were compared. Because of the complexity of the model space, we find that both our sampling scheme and that of Denison, Mallick and Smith are only exploring local modes of the posterior distribution, and different results are obtained in different runs (although all runs are able to identify MARS models with excellent predictive performance). For each of the simulated examples described above, we did four different runs of 50 000 iterations for each of the sampling schemes, discarding the first 40 000 iterations as burn in. Different random seeds were used for each run, and for each run the initial model contains just an intercept. In all examples, after examining a plot of the residual sum of squares against iteration number as suggested by Denison, Mallick and Smith (1998a), the sampling schemes appeared to have converged to a local mode of the posterior for all test functions and all runs after 40 000 iterations. The last 10 000 iterations are used to estimate the mean function, by averaging the means for the models at each iteration after the burn in period (so-called Bayesian model averaging, see Hoeting *et al.* 1999). Denoting the fitted mean function by $\hat{f}(\mathbf{x})$, following Denison, Mallick and Smith (1998a) we can calculate the fraction of variance unexplained (FVU) for each run. Let $X = (X_1, \ldots, X_{20})^T$ denote a vector of independent uniform random variables on $[0, 1]$. Then we define the fraction of variance unexplained as

$$\text{FVU} = \frac{E_X((\hat{f}(X) - f(X))^2)}{E_X((f(X) - E_X(f(X)))^2)}.$$

Where E_X denotes the expectation with respect to X. As in Denison, Mallick and Smith (1998a), we estimate the FVU based on a test set of 10 000 simulated values for the predictors. The average value of the FVU over four runs of each of the

Table 1. *FVU, $\bar{m}$, $\bar{m}_1$, $\bar{m}_2$, $\bar{d}$ and acceptance rate for sampling scheme of Denison, Mallick and Smith (DMS) and proposed sampling scheme*

Function	Sampling scheme	FVU	$\bar{m}$	$\bar{m}_1$	$\bar{m}_2$	$\bar{d}$	Acceptance rate
$f^{(1)}$	DMS	0.0038	5.7	2.0	3.7	6.2	23.3
	Proposed	0.0033	6.7	1.9	4.8	6.8	19.4
$f^{(2)}$	DMS	0.0389	11.0	5.1	5.9	6.7	20.3
	Proposed	0.0206	12.1	4.9	7.1	7.5	14.1
$f^{(3)}$	DMS	0.0208	20.6	8.4	12.1	10.1	15.1
	Proposed	0.0083	24.3	8.4	15.9	8.5	9.4
$f^{(4)}$	DMS	0.0056	11.7	9.2	2.5	6.1	16.8
	Proposed	0.0058	11.9	9.2	2.8	6.5	12.7
$f^{(5)}$	DMS	0.1652	10.9	4.3	6.6	5.9	19.3
	Proposed	0.1420	14.6	3.6	11.0	6.8	12.2

The values given are averages over four different runs, each run of length 50 000 with a burn in period of 40 000.

sampling schemes for each of the five test functions is shown in Table 1. This shows that while both sampling schemes appear to converge in all the examples, it is still the case that better local modes are being found (as measured by out of sample predictive performance through the FVU) for our proposed sampling scheme, particularly for the test functions which have complex interaction structure. Also shown in Table 1 is the mean value of m over four runs of each of the sampling schemes for each of the five test functions (a measure of how parsimonious the models are). We also have the mean number of first order terms ($\bar{m}_1$), the mean number of second order terms ($\bar{m}_2$), the mean number of distinct predictors ($\bar{d}$) and the acceptance rates for the sampling schemes. The averages are calculated based on the last 10 000 iterations for each run.

Table 1 shows that the acceptance rates for moves on model space are higher for the method of Denison, Mallick and Smith (1998a) than for our proposed method. This initially surprised us, but the explanation lies in the effect of the proposal ratio on the Metropolis-Hastings acceptance probability. Although we have found that proposals in the birth step of our new sampling scheme tend to produce stronger posterior ratios in general, the proposal ratio tends to be smaller leading to slightly lower acceptance probabilities than for the algorithm of Denison, Mallick and Smith (1998a).

3.2. *Simulated examples with 50 predictors*

As a further experiment we took the test function where our sampling scheme performed worst in the 20 dimensional case (the additive test function) and the sampling scheme where it performed best (the complex interaction test function) and added an additional 30 spurious predictors to the simulated data sets that had been generated for these test functions. The new predictors were generated in the same way as before. This results in a high dimensional situation with 50 predictors being considered. For these data sets we also considered varying the algorithm interaction parameters γ and δ with every combination of the levels $\gamma = \lambda/2$ and $\lambda/4$ and $\delta = 1/4$ and $1/9$ being considered. The values $\gamma = \lambda/4$ and $\delta = 1/9$ result in more aggressive proposals, with the predictors and interaction orders in the current model being favoured more in birth proposals. For each choice of the parameters and each test function we ran four chains of our algorithm. Again we ran each chain for 50 000 iterations, discarding the first 40 000 iterations as burn in. We also ran 16 chains of the algorithm of Denison, Mallick and Smith (1998a). Table 2 shows the average FVU values for our algorithm and that of Denison, Mallick and Smith (1998a) in the different cases.

Using a more agressive proposal strategy with respect to the predictor variables ($\delta = 1/9$) seems helpful for the complex interaction test function. As might be expected, our sampling scheme does not provide any real improvement for the additive test function—adapting the interaction order of proposals does not seem to have much benefit and of course since we only make use of information about predictors in the current model in the interaction type proposals there is not much benefit in our sampling scheme for this case. Of course, we

Table 2. *FVU for sampling scheme of Denison, Mallick and Smith (DMS) and proposed sampling scheme with different values for algorithm parameters and 50 predictors*

Function	Sampling scheme	γ	δ	FVU
$f^{(4)}$	DMS	–	–	0.0067
	Proposed	$\lambda/2$	1/4	0.0059
	Proposed	$\lambda/2$	1/9	0.0055
	Proposed	$\lambda/4$	1/4	0.0115
	Proposed	$\lambda/4$	1/9	0.0064
$f^{(5)}$	DMS	–	–	0.4872
	Proposed	$\lambda/2$	1/4	0.1920
	Proposed	$\lambda/2$	1/9	0.1817
	Proposed	$\lambda/4$	1/4	0.1933
	Proposed	$\lambda/4$	1/9	0.1698

The FVU values given are averages over four different runs for the proposed sampling scheme and over 16 different runs for DMS for each test function, each run of length 50 000 with a burn in period of 40 000.

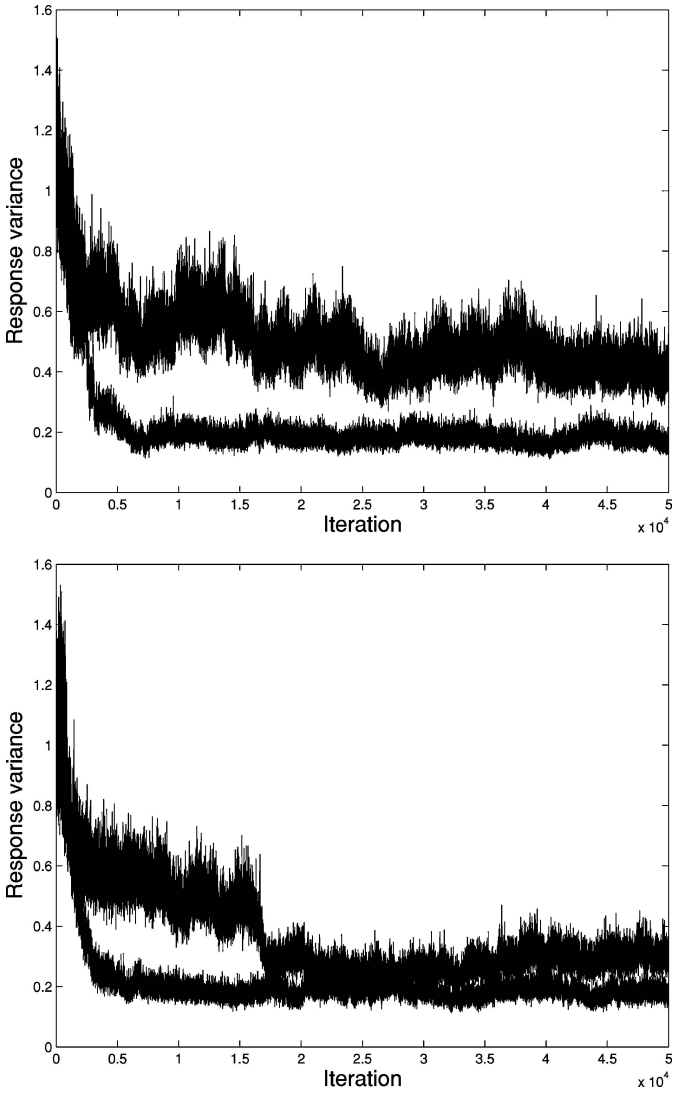

Fig. 1. *Plot of σ^2 against iteration number for a run of the sampling scheme of Denison, Mallick and Smith (DMS) and our proposed sampling scheme for $\gamma = \lambda/4$, $\delta = 1/9$ (top) and $\gamma = \lambda/2$, $\delta = 1/4$ (bottom). In each plot the uppermost line is for DMS, the lowermost line is our proposed scheme*

could use information from the current model in constructing a proposal for main effects terms but the algorithm of Denison, Mallick and Smith (1998a) is able to find significant main effects reasonably quickly, and there is a danger that adapting our sampling scheme in this way may lead to a proposal that performs poorly in the early stages of sampling before important variables are identified which might hinder convergence.

Figure 1 shows plots of the value of σ^2 against iteration number for the two sampling schemes for the complex interaction test function with the runs for our sampling scheme having $\gamma = \lambda/2$, $\delta = 1/4$ and $\gamma = \lambda/4$, $\delta = 1/9$ (the least and most aggressive proposal strategies). As can be seen from the plots, convergence is much more rapid for our scheme, and convergence is to better local modes. Of course, running both sampling schemes longer may result in different local modes being found.

4. Discussion

A new sampling scheme for exploring the posterior distribution in Bayesian MARS models has been suggested which can give substantial improvements over the sampling scheme of Denison, Mallick and Smith (1998a) in problems with large numbers of predictors and many useless predictors. Our algorithm was a simple attempt to build some of the intuition behind the original MARS model selection procedure of Friedman (1991) into Metropolis-Hastings proposals in the Bayesian MARS sampling algorithm.

Although we are able to obtain large improvements over the original sampling scheme of Denison, Mallick and Smith (1998a) in some problems, we found that their sampling scheme worked surprisingly well in many cases. Simple as it may be, the device of simulating proposals from the prior seems to be reasonably effective, particularly if there are not strong interaction effects. We believe that large improvements of our method over the method of Denison, Mallick and Smith (1998a) occur when there are one or more interactions of a complex form, and when there are large numbers of predictors and many useless predictors. This kind of situation is common enough in many recent applications of nonparametric regression methods for our algorithm to be useful.

References

Albert J.H. and Chib S. 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88: 669–679.

Biller C. 2000. Adaptive Bayesian regression splines in semiparametric generalized linear models. Journal of Computational and Graphical Statistics 9: 122–140.

Carlin B.P. and Chib S. 1995. Bayesian model choice via Markov chain Monte Carlo methods. Journal of the Royal Statistical Society, Ser. B 57: 473–484.

Chib S. and Greenberg E. 1995. Understanding the Metropolis-Hastings algorithm. The American Statistician 40: 327–335.

Denison D.G.T., Mallick B.K. and Smith A.F.M. 1998a. Bayesian MARS. Statistics and Computing 8: 337–346.

Denison D.G.T., Mallick B.K. and Smith A.F.M. 1998b. Automatic Bayesian curve fitting. Journal of the Royal Statistical Society, Ser. B. 60: 333–350.

DiMatteo I., Genovese C.R. and Kass R.E. 1998. Bayesian curve fitting with free knot splines. Biometrika 88: 1055–1073.

Friedman J.H. 1991. Multivariate adaptive regression splines. The Annals of Statistics 19: 1–141.

Friedman J.H. and Silverman B.W. 1989. Flexible parsimonious smoothing and additive modelling. Technometrics 31: 3–39.

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

Hoeting J.A., Madigan D., Raftery A.E., and Volinsky C. 1999. Bayesian model averaging: A tutorial (with discussion). Statistical Science 14: 382–417.

Holmes C.C. and Denison D.G.T. 2002. A Bayesian MARS classifier. Machine Learning, to appear.

Hwang J.-N., Lay S.-R., Maechler M., Martin D., and Schmiert J. 1994. Regression modelling in back-propagation and projection pursuit learning. IEEE Transations on Neural Networks 5: 342–353.

Kohn R., Smith M., and Chan D. 2001. Nonparametric regression using linear combinations of basis functions. Statistics and Computing 11: 313–322.

Kooperberg C., Bose S., and Stone C.J. 1997. Polychotomous regression. Journal of the American Statistical Association 93: 117–127.

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. Annals of Statistics 22: 1701–1728.