Bayesian Regression Structure Discovery

Hugh A. Chipman, Edward I. George, Robert E. McCulloch *

January 11, 2012

Abstract

The general problem of statistical regression is concerned with the discovery of a relationship between y and a set of potential predictors $x_1, \ldots, x_p$. Because y may be related only to an unknown subset of the potential predictors, especially when p is large, variable selection is also an inherent part of this problem. In this chapter, we describe two very different Bayesian approaches to this general problem. In one case, variable selection takes place in the context of a tightly specified parametric model. Here, priors may be used to guide the model search and posterior quantities of interest are evident. In the other case, a far more flexible model, essentially nonparametric, allows for the opportunity to discover richer structure in the data, but requires more subtle methods for inference. With simple examples, we show how this second approach allows for model-free variable selection, and further for model-free interaction detection, the discovery of when variables work together to influence the response.

KEY WORDS: stochastic search, interaction detection, model selection, variable selection.

*Hugh Chipman is Professor and Canada Research Chair in Mathematical Modeling, Department of Mathematics and Statistics, Acadia University, Wolfville, Nova Scotia, B4P 2R6, hugh.chipman@acadiau.ca. Edward I. George is Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, 3730 Walnut St, 400 JMHH, Philadelphia, PA 19104-6304, edgeorge@wharton.upenn.edu. Robert E. McCulloch is Professor of Statistics, IROM Department, 1 University Station, B6500, Austin, TX 78712-1175, robert.mcculloch1@gmail.com. The authors are grateful to Adam Kapelner for valuable suggestions.

Contents

1 Introduction

The general problem of statistical regression is concerned with the discovery of a relationship between a variable of interest y and a set of potential predictors $x_1, \ldots, x_p$. It is usually realistic, especially when p is large, to consider that y may be related only to an unknown subset of the potential predictors, thus making variable selection an inherent part of the problem. In this paper, we describe two very different Bayesian approaches to this general problem. The feasible implementation of both of these approaches has been made possible by Markov chain Monte Carlo (MCMC) Bayesian posterior simulation, Gelfand & Smith (1990) and Tierney (1994). In particular, variations of the Gibbs sampler and the Metropolis-Hastings algorithms have allowed for the exploration of the otherwise intractable posteriors via simulation.

One approach, which dovetails with classical parametric approaches to variable selection, begins with an assumption that the relationship between y and $x_1, \ldots, x_p$ can be described by a full parametric model within which the actual subset model is nested. The most popular form used here is the normal linear model, in large part because of its appealing analytical tractability and because of its usefulness as an approximation to other forms, possibly after suitable transformations. A structured hierarchical mixture prior that captures all sources of remaining uncertainty is then used to obtain a posterior distribution which allocates more probability to the more promising subset models.

In contrast to the parametric approach, our second approach does not require an initial assumption about the nature of all the relationships between y and the subsets of $x_1, \ldots, x_p$. Nonparametric in nature, it begins with a very rich over parametrized functional form, a sum-of-trees model, that approximates a wide class of functions from R^p to R. However, with this more complex model it becomes more challenging to formulate useful priors and extract information about the relationship between y and x. A strong regularization prior over the multitude of parameters of the a sum-of-trees model is used to obtain a posterior distribution over the possible relationships between y and $x_1, \ldots, x_p$. A variety of useful inferential summaries can be obtained by MCMC sampling from this posterior. In particular, by keeping track of how often each predictor is used in the sum-of-trees model, this approach allows for model-free variable selection, and further for model-free interaction detection, the discovery of when variables work together to influence the response.

2 Parametric Bayesian Structure Discovery

To illustrate the parametric Bayesian approach to structure discovery, we focus on the classical version of the problem which begins with the assumption of a normal linear model for every subset model, namely

$$Y = X_\gamma \beta_\gamma + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I) \tag{1}$$

where Y is the $n \times 1$ vector of y observations, X_γ is the $n \times q_\gamma$ matrix whose columns correspond to the γth subset of $x_1, \ldots, x_p$, and β_γ is the $q_\gamma \times 1$ vector of unknown regression coefficients. For convenience, we have indexed each of the 2^p possible subset choices by

$$\gamma = (\gamma_1, \ldots, \gamma_p)', \tag{2}$$

where $\gamma_i = 1$ or 0 according to whether predictor x_i is included or excluded, respectively. The size (number of covariates) of the γth subset is thus $q_\gamma \equiv \gamma' 1$. The variable selection problem may then be regarded as how to use the data to choose γ. Particular Bayesian treatments of this formulation yield analytical reductions that allow for faster calculations as well as clearer insights how the machinery works. Such Bayesian treatments also extend naturally to likelihoods that are a function of X_γ only through $X_\gamma \beta_\gamma$. There is by now a vast literature on Bayesian analyses for this formulation. See, for example, George & McCulloch (1997), Chipman, George & McCulloch (2001) Clyde & George (2004), and the references therein.

It should be noted that assumption (1) for every possible submodel γ is a strong assumption. Its strength is that it effectively turns the the variable selection problem into a model selection problem which can be treated using variations of standard Bayesian parametric formulations. Its weakness is that a subset of predictors may be rejected because a normal linear submodel is inadequate rather than because Y is unrelated to the subset.

2.1 Prior formulations

The parametric problem formulation in (1), provides a likelihood $L(\beta_\gamma, \sigma, \gamma \mid Y)$. Thus, a Bayesian analysis proceeds with the choice of prior forms for

$$p(\beta_\gamma, \sigma, \gamma) = p(\beta_\gamma, \sigma \mid \gamma) p(\gamma). \tag{3}$$

For the specification of the model space prior $p(\gamma)$, many Bayesian variable selection implementations have used simple independence priors of the form

$$p(\gamma) = w^{q_\gamma}(1-w)^{p-q_\gamma}, \tag{4}$$

with a prespecified value for w, the expected proportion of $x_i's$ in the submodel. Under this prior, each x_i enters the submodel independently with probability $p(\gamma_i = 1) = 1 - p(\gamma_i = 0) = w$. To avoid the fact that any such prior will be informative about the size of the model, a reasonable alternative is to margin out w in (4) with respect to a Beta prior to obtain

$$p(\gamma) = \frac{B(\alpha + q_\gamma, \beta + p - q_\gamma)}{B(\alpha, \beta)}, \tag{5}$$

a special case of the more general form

$$p(\gamma) = \binom{p}{q_\gamma}^{-1} h(q_\gamma) \tag{6}$$

which is uniform over the set of submodels of a given size q_γ. See George & McCulloch (1993), Cui & George (2008) and Scott & Berger (2010).

For the specification of the parameter prior $p(\beta_\gamma, \sigma \,|\, \gamma) = p(\beta_\gamma \,|\, \sigma^2, \gamma)\, p(\sigma^2 \,|\, \gamma)$, an especially convenient choice is the conjugate normal-inverse-gamma prior

$$p(\beta_\gamma \,|\, \sigma^2, \gamma) = N_{q_\gamma}(0, \sigma^2 \Sigma_\gamma), \tag{7}$$

$$p(\sigma^2 \,|\, \gamma) = p(\sigma^2) = IG(\nu/2, \nu\lambda/2). \tag{8}$$

($p(\sigma^2)$ here is equivalent to $\nu\lambda/\sigma^2 \sim \chi^2_\nu$). A valuable feature of this prior is its analytical tractability; β_γ and σ^2 can be eliminated by routine integration to yield

$$p(Y \,|\, \gamma) \propto |X'_\gamma X_\gamma + \Sigma_\gamma^{-1}|^{-1/2} |\Sigma_\gamma|^{-1/2} (\nu\lambda + S^2_\gamma)^{-(n+\nu)/2} \tag{9}$$

where

$$S^2_\gamma = Y'Y - Y'X_\gamma (X'_\gamma X_\gamma + \Sigma_\gamma^{-1})^{-1} X'_\gamma Y. \tag{10}$$

The use of these closed form expressions can substantially speed up posterior evaluation and MCMC exploration, as we will see.

For choosing the prior covariance matrix Σ_γ that controls $p(\beta_\gamma \,|\, \sigma^2, \gamma)$, specification is substantially simplified by setting $\Sigma_\gamma = c\, V_\gamma$, where c is a scalar and V_γ is a preset form such as $V_\gamma = (X'_\gamma X_\gamma)^{-1}$ (as in the Zellner (1986) g-prior) or $V_\gamma = I_{q_\gamma}$, the $q_\gamma \times q_\gamma$ identity matrix. Having fixed V_γ, the goal is then to choose c large enough so that $p(\beta_\gamma \,|\, \sigma^2, \gamma)$ is relatively flat

over the region of plausible values of β_γ, thereby reducing prior influence. At the same time it is important to avoid excessively large values of c because the Bayes factors will eventually put increasing weight on the null model as $c \to \infty$, the Bartlett-Lindley paradox. For practical purposes, a rough guide is to choose c so that $p(\beta_\gamma \mid \sigma^2, \gamma)$ assigns substantial probability to the range of all plausible values for β_γ. A recent alternative of interest, are the hyper-g priors for β_γ which effectively integrate out c with respect to Beta prime distributions, Cui and George (2008), Liang, Paulo, Molina, Clyde & Berger (2008) and Maruyama & George (2011).

In choosing values for the hyperparameters that control $p(\sigma^2)$, λ may be thought of as a prior estimate of σ^2, and ν may be thought of as the prior sample size associated with this estimate. Alternatively, one might use the data informally to choose λ and ν as follows. Let σ^2_{FULL} and σ^2_Y denote the traditional estimates of σ^2 based on the saturated and null models respectively. Treating σ^2_{FULL} and σ^2_Y as rough under- and over-estimates of σ^2, one might choose λ and ν so that $p(\sigma^2)$ assigns substantial probability to the interval $(\sigma^2_{FULL}, \sigma^2_Y)$. This should at least avoid gross misspecification. As a third option, the explicit choice of λ and ν can be avoided by using $p(\sigma^2) \propto 1/\sigma^2$, the limit of the inverse-gamma prior as $\nu \to 0$.

2.2 Posterior Exploration and Information Extraction

The previous conjugate prior formulations allow for analytical margining out of β and σ^2 from $p(Y, \beta, \sigma^2 \mid \gamma)$ to yield a computable, closed form expression

$$g(\gamma) \propto p(Y \mid \gamma) p(\gamma) \propto p(\gamma \mid Y) \tag{11}$$

that can greatly facilitate posterior calculation and exploration. For example, when $\Sigma_\gamma = c\,(X'_\gamma X_\gamma)^{-1}$, we can obtain

$$g(\gamma) = (1+c)^{-q_\gamma/2} (\nu\lambda + Y'Y - (1+1/c)^{-1} W'W)^{-(n+\nu)/2} p(\gamma) \tag{12}$$

where $W = T'^{-1} X'_\gamma Y$ for upper triangular T such that $T'T = X'_\gamma X_\gamma$ (obtainable by the Cholesky decomposition). This representation allows for fast updating of T, and hence W and $g(\gamma)$, when γ is changed one component at a time, requiring $O(q^2_\gamma)$ operations per update, where γ is the changed value.

The availability of $g(\gamma) \propto p(\gamma \mid Y)$ allows for the flexible construction of MCMC algorithms that simulate a Markov chain

$$\gamma^{(1)}, \gamma^{(2)}, \gamma^{(3)}, \ldots \tag{13}$$

converging (in distribution) to $p(\gamma \mid Y)$. A variety of such MCMC algorithms can be conveniently obtained by applying the Gibbs sampler with $g(\gamma)$. For example, by generating each γ component from the full conditionals

$$p(\gamma_i \mid \gamma_{(i)}, Y) \tag{14}$$

($\gamma_{(i)} = \{\gamma_j : j \neq i\}$) where the γ_i may be drawn in any fixed or random order. The generation of such components can be obtained rapidly as a sequence of Bernoulli draws using simple functions of the ratio

$$\frac{p(\gamma_i = 1, \gamma_{(i)} \mid Y)}{p(\gamma_i = 0, \gamma_{(i)} \mid Y)} = \frac{g(\gamma_i = 1, \gamma_{(i)})}{g(\gamma_i = 0, \gamma_{(i)})}. \tag{15}$$

The availability of such closed form $g(\gamma)$ also facilitates the use of MH algorithms. Because $g(\gamma)/g(\gamma') = p(\gamma \mid Y)/p(\gamma' \mid Y)$, these are of the form:

1. Simulate a candidate γ^* from a transition kernel $q(\gamma^* \mid \gamma^{(j)})$.
2. Set $\gamma^{(j+1)} = \gamma^*$ with probability
$$\alpha(\gamma^* \mid \gamma^{(j)}) = \min\left\{\frac{q(\gamma^{(j)} \mid \gamma^*)}{q(\gamma^* \mid \gamma^{(j)})}\frac{g(\gamma^*)}{g(\gamma^{(j)})}, 1\right\}. \tag{16}$$
3. Otherwise, set $\gamma^{(j+1)} = \gamma^{(j)}$.

When available, fast updating schemes for $g(\gamma)$ can be exploited in all these MCMC algorithms.

The simulated Markov chain sample $\gamma^{(1)}, \ldots, \gamma^{(K)}$ contains valuable information about the posterior $p(\gamma \mid Y)$. Empirical frequencies provide consistent estimates of individual model probabilities or characteristics such as $p(\beta_i \neq 0 \mid Y)$. When closed form $g(\gamma)$ are available, we can do better. For example, the exact relative probability of any two values γ^0 and γ^1 is obtained as $g(\gamma^0)$ / $g(\gamma^1)$ in the sequence of simulated values. Such $g(\gamma)$ also facilitates estimation of the normalizing constant $p(\gamma \mid Y) = C\, g(\gamma)$. Let A be a preselected subset of γ values and let $g(A) = \sum_{\gamma \in A} g(\gamma)$ so that $p(A \mid Y) = C\, g(A)$. Then, a consistent estimate of C is

$$\hat{C} = \frac{1}{g(A)K}\sum_{k=1}^{K} I_A(\gamma^{(k)}) \tag{17}$$

where $I_A(\,)$ is the indicator of the set A. This yields alternative estimates of the probability of individual γ values $\hat{p}(\gamma \mid Y) = \hat{C}\, g(\gamma)$, as well as an estimate of the total visited probability $\hat{p}(B \mid Y) = \hat{C}\, g(B)$, where B is the set of visited γ values.

3 Nonparametric Bayesian Structure Discovery

To illustrate the nonparametric Bayesian approach to structure discovery, we focus on an approach we call BART (Bayesian Additive Regression Trees), (Chipman, George & McCulloch (2010)) which assumes only that y is related to $x = (x_1, \ldots, x_p)$ via a flexible sum-of-trees model of the form[1]

$$Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2), \tag{18}$$

where each T_j is a binary regression tree with a set M_j of associated terminal node constants μ_{ij}, and $g(x; T_j, M_j)$ is the function which assigns $\mu_{ij} \in M_j$ to x according to the sequence of decision rules in T_j. These decision rules are binary splits of the predictor space of the form $\{x \in A\}$ vs $\{x \notin A\}$ where A is a subset of the range of x. When $m = 1$, (18) reduces to the single tree model used by Chipman, George & McCulloch (1998) for Bayesian CART.

Under (18), $E(Y \mid x)$ equals the sum of all the terminal node μ_{ij}'s assigned to x by the $g(x; T_j, M_j)$'s. As these can be any values, it is easy to see that the sum-of-trees model (18) is a very flexible representation capable of representing a wide class of functions from R^n to R, especially when the number of trees m is large. Note also that the sum-of-trees representation is composed of many simple functions from R^p to R, namely the $g(x; T_j, M_j)$, rendering it much more manageable than a representation with more complicated basis elements such as multidimensional wavelets or multidimensional splines.

3.1 A regularization prior

We complete the BART model specification by imposing a prior over all the parameters of the sum-of-trees model, namely $(T_1, M_1), \ldots, (T_m, M_m)$ and σ. Note that these parameters entail all the bottom node parameters as well as the tree structures and decision rules, a very large number of parameters, especially when m is large. We do this using a prior that effectively regularizes the fit by keeping the individual tree effects from being unduly influential. Without such a regularizing influence, large tree components would overwhelm the rich structure of (18), thereby limiting its scope of approximation.

To begin with we simplify our prior specification task by restricting attention to prior for-

[1] Note that here we use Y as a random scalar rather than an $n \times 1$ random vector as in Section 2.

mulations of the form

$$p((T_1, M_1), \ldots, (T_m, M_m), \sigma) = \left[\prod_j \left(\prod_i p(\mu_{ij} \mid T_j) \right) p(T_j) \right] p(\sigma), \tag{19}$$

where $\mu_{ij} \in M_j$. These independence restrictions simplify prior specification the choice of prior forms for $p(T_j), p(\mu_{ij} \mid T_j)$ and $p(\sigma)$, and to simplify matters further we consider for all of these, the same prior forms as those proposed by Chipman et al. (1998) for Bayesian CART. These forms are controlled by just a few interpretable hyperparameters which can be calibrated using the data to yield effective default specifications for regularization of the sum-of-trees model.

For $p(T_j)$, we use the Chipman et al. (1998) tree-generating process which is specified by three aspects: (i) the probability that a node at depth d ($= 0, 1, 2, \ldots$) is nonterminal, given by

$$\alpha(1+d)^{-\beta}, \qquad \alpha \in (0,1), \beta \in [0, \infty), \tag{20}$$

(ii) the distribution on the splitting variable assignments at each interior node, and (iii) the distribution on the splitting rule assignment in each interior node, conditional on the selected splitting variable. For (ii) and (iii) we use the simple defaults in Chipman et al. (1998) , namely a uniform prior on each set of possibilities

For $p(\mu_{ij} \mid T_j)$, we use the conjugate normal distribution $N(\mu_\mu, \sigma_\mu^2)$ which allows μ_{ij} to be margined out, greatly simplifying MCMC posterior calculations. Note that under this choice the prior distribution of $E(Y \mid x)$ is $N(m\,\mu_\mu, m\,\sigma_\mu^2)$, (because $E(Y \mid x)$ is the sum of m independent μ_{ij}'s under the sum-of-trees model). Thus, it is highly probable that $E(Y \mid x)$ is between y_{min} and y_{max}, the observed minimum and maximum of y in the data, a fact which we can use to guide the specification of the hyperparameters μ_μ and σ_μ. The essence of our informal strategy is then to choose μ_μ and σ_μ so that $N(m\,\mu_\mu, m\,\sigma_\mu^2)$ assigns substantial probability to the interval (y_{min}, y_{max}). This can be conveniently done by choosing μ_μ and σ_μ so that $m\,\mu_\mu - k\,\sqrt{m}\,\sigma_\mu = y_{min}$ and $m\,\mu_\mu + k\,\sqrt{m}\,\sigma_\mu = y_{max}$ for some preselected value of k such 1,2 or 3. For example, $k = 2$ would yield a 95% prior probability that $E(Y \mid x)$ is in the interval (y_{min}, y_{max}). The goal of this specification strategy for μ_μ and σ_μ is to ensure that the implicit prior for $E(Y \mid x)$ is in the right "ballpark" in the sense of assigning substantial probability to the entire region of plausible values of $E(Y \mid x)$ while avoiding overconcentration and overdispersion. As long as this goal is met, BART seems to be very robust to the variations of these specifications.

For $p(\sigma)$, we also use a conjugate prior, here the inverse chi-square distribution $\sigma^2 \sim \nu\,\lambda/\chi^2_\nu$, the same form we used for $p(\sigma)$ for the parametric variable selection problem previously. Here again, we use a data-informed prior approach, to guide the specification of the hyper-parameters ν and λ, in this case to assign substantial probability to the entire region of plausible values of σ while avoiding overconcentration and overdispersion. Essentially, we calibrate the prior df ν and scale λ using a "rough data-based overestimate" $\hat{\sigma}$ of σ. Two natural choices of where $\hat{\sigma}$ are 1) a "naive" specification, the sample standard deviation of Y, or 2) a "linear model" specification, the residual standard deviation from a least squares linear regression of Y on all the predictors. We then pick a value of ν between 3 and 10 to get an appropriate shape, and a value of λ so that the qth quantile of the prior on σ is located at $\hat{\sigma}$, that is $P(\sigma < \hat{\sigma}) = q$. We consider values of q such as 0.75, 0.90 or 0.99 to center the distribution below $\hat{\sigma}$.

3.2 Posterior Calculation and Information Extraction

Combing the regulation prior with the likelihood, $L((T_1, M_1), \ldots, (T_m, M_m), \sigma \mid y)$ induces a posterior distribution

$$p((T_1, M_1), \ldots, (T_m, M_m), \sigma | \, y) \tag{21}$$

over the full sum-of-trees model parameter space. Fortunately, the following backfitting MCMC algorithm can be used to simulate samples from this posterior.

We begin with a Gibbs sampler at the outer level. Let $T_{(j)}$ be the set of all trees in the sum *except* T_j, and similarly define $M_{(j)}$, so that $T_{(j)}$ will be a set of $m-1$ trees, and $M_{(j)}$ the associated terminal node parameters. A Gibbs sampling strategy for sampling from (21) is obtained by m successive draws of (T_j, M_j) conditionally on $(T_{(j)}, M_{(j)}, \sigma)$:

$$(T_j, M_j)|T_{(j)}, M_{(j)}, \sigma, y, \tag{22}$$

$j = 1, \ldots, m$, followed by a draw of σ from the full conditional:

$$\sigma|T_1, \ldots T_m, M_1, \ldots, M_m, y. \tag{23}$$

The draw of σ in (23) is simply a draw from an inverse gamma distribution and so can be easily obtained by routine methods. More subtle is the implementation of the m draws of (T_j, M_j) in (22). This can be done by taking advantage of the following reductions. First,

observe that the conditional distribution $p(T_j, M_j | T_{(j)}, M_{(j)}, \sigma, y)$ depends on $(T_{(j)}, M_{(j)}, y)$ only through

$$R_j \equiv y - \sum_{k \neq j} g(x; T_k, M_k), \tag{24}$$

the $n-$vector of partial residuals based on a fit that excludes the jth tree. Thus, the m draws of (T_j, M_j) given $(T_{(j)}, M_{(j)}, \sigma, y)$ in (22) are equivalent to m draws from

$$(T_j, M_j)|R_j, \sigma, \tag{25}$$

$j = 1, \ldots, m$. Because we have used a conjugate prior for M_j, $p(T_j|R_j, \sigma)$ can be obtained in closed form up to a norming constant. This allows us to carry out each draw from (25) in two successive steps as

$$T_j|R_j, \sigma \tag{26}$$

$$M_j|T_j, R_j, \sigma. \tag{27}$$

The draw of T_j in (26), although somewhat elaborate, can be obtained using the Metropolis-Hastings (MH) algorithm of Chipman et al. (1998). The draw of M_j in (27) is simply a set of independent draws of the terminal node μ_{ij}'s from a normal distribution. The draw of M_j enables the calculation of the subsequent residual R_{j+1} which is critical for the next draw of T_j.

We initialize the chain with m simple single node trees, and then iterations are repeated until satisfactory convergence is obtained. Fortunately, this backfitting MCMC algorithm appears to mix very well as we have found that different restarts give remarkably similar results even in difficult problems. At each iteration, each tree may increase or decrease the number of terminal nodes by one, or change one or two decision rules. The sum-of-trees model, with its abundance of unidentified parameters, allows for "fit" to be freely reallocated from one tree to another. Because each move makes only small incremental changes to the fit, we can imagine the algorithm as analogous to sculpting a complex figure by adding and subtracting small dabs of clay.

For inference based on our MCMC sample, we rely on the fact the our backfitting algorithm is ergodic. Thus, the induced sequence of sum-of-trees functions

$$f^*(\cdot) = \sum_{j=1}^{m} g(\cdot; T_j^*, M_j^*), \tag{28}$$

for the sequence of draws $(T_1^*, M_1^*), \ldots, (T_m^*, M_m^*)$, is converging to $p(f \,|\, y)$, the posterior distribution on the "true" $f(\cdot)$. Thus, by running the algorithm long enough after a suitable

burn-in period, the sequence of f^* draws, say $f_1^*, \ldots, f_K^*$, may be regarded as an approximate, dependent sample of size K from $p(f \mid y)$. Bayesian inferential quantities of interest can then be approximated with this sample as indicated below.

To estimate $f(x)$ or predict Y at a particular x, in-sample or out-of-sample, a natural choice is the average of the after burn-in sample $f_1^*, \ldots, f_K^*$,

$$\frac{1}{K} \sum_{k=1}^{K} f_k^*(x), \tag{29}$$

which approximates the posterior mean $E(f(x) \mid y)$. Posterior uncertainty about $f(x)$ may be gauged by the variation of $f_1^*(x), \ldots, f_K^*(x)$. For example, a natural and convenient $(1-\alpha)\%$ posterior interval for $f(x)$ is obtained as the interval between the upper and lower $\alpha/2$ quantiles of $f_1^*(x), \ldots, f_K^*(x)$.

Finally, BART provides a new approach to variable selection and interaction detection by identifying those variables or combination of variables that appear most often in the fitted sum-of-trees models. Interestingly, the variable selection strategy does not seem to work well when m is large because the redundancy offered by so many trees allows many irrelevant predictors to be mixed in with the relevant ones. However, as m is decreased and that redundancy is diminished, BART tends to heavily favor relevant predictors for its fit. In a sense, when m is small the predictors compete with each other to improve the fit. In contrast, interaction detection seems to work well with large m.

This model-free approach to variable selection is accomplished by observing what happens to the x component usage frequencies in a sequence of MCMC samples $f_1^*, \ldots, f_K^*$ as the number of trees m is set smaller and smaller. More precisely, for each simulated sum-of-trees model f_k^*, let z_{ik} be the proportion of all splitting rules that use the ith component of x. Then

$$v_i \equiv \frac{1}{K} \sum_{k=1}^{K} z_{ik} \tag{30}$$

is the average use per splitting rule for the ith component of x. As m is set smaller and smaller, the sum-of-trees models tend to more strongly favor inclusion of those x components which improve prediction of y and exclusion of those x components that are unrelated to y. In effect, smaller m seems to create a bottleneck that forces the x components to compete for entry into the sum-of-trees model. As we shall see in Section 4.2, the x components with the larger v_i's will then be those that provide the most information for predicting y. A BART approach to model-free interaction detection proceeds in analogous fashion, for example, let z_{ijk} be the proportion of all trees in which both the ith and jth components of x appear.

4 Information Extraction: Details and Examples

In Section (3) we outlined the BART model and discussed, in general terms, how it can be used to extract information about the relationship between y and x. In Section (4.1) we provide additional detail on the CART MCMC, highlighting the crucial aspects of our prior and algorithm that enable us to find structure. In Section (4.2) we give examples of information extraction. We show how we can extract information about what variables are important (using equation (30)) and which pairs of variables work together generating an "interaction" effect.

4.1 Stochastic Search in CART Models

Perhaps the crucial model search step in Section (3) is the draw given in (26): $T_j|R_j, \sigma$. Our Gibbs sampler MCMC structure allows us to focus on one tree so that we are back to a CART problem. It is in this step, that we actually modify the structure of a tree. It is in this step, that a new variable may be introduced to our model.

This draw is done using a Metropolis-within-Gibbs proposal. The CART algorithm given in Chipman et al. (1998) uses several types of proposals (see also Wu, Tjelmeland & West (2007) for additional MCMC strategies).

The essential proposals[2] are a complementary BIRTH/DEATH pair of moves. In a BIRTH proposal, a bottom node of the current tree is chosen and we propose to give it a pair of children. A *nog* node of a tree is a tree which has children, but no grandchildren. Thus, both children of a nog node are bottom nodes. In a DEATH proposal, we choose a nog node from the current tree and we propose "killing its children". In order to make our general discussion more concrete and document some of the details, we give the acceptance probability for a BIRTH proposal.

The CART algorithm assumes a discrete set of possible split values for each component of x and integrates out the bottom node μ_{ij} so that our Metropolis search in tree space is over a large but discrete set of possible models. Let T_0 denote the *current* tree and T^* denote the *proposed* tree. Thus, T^* differs from T_0 only in that one of the bottom nodes of T_0 has given birth to a pair of children in T^*.

[2]The other two proposals in CMG98 are CHANGE and SWAP.

Since we are traversing a discrete space, we accept the proposal with Metropolis-Hastings probability

$$\alpha = \min\{1, \frac{P(T^*)\, P(T^* \to T_0)}{P(T_0)\, P(T_0 \to T^*)}\} \tag{31}$$

where $P(T_0)$ and $P(T^*)$ are the posterior probabilities of trees T_0 and T^* respectively, $P(P \to T_0)$ is the probability of proposing T_0 while at T^* (a DEATH), and $P(T_0 \to T^*)$ (a BIRTH) is the probability of proposing T^* while at T_0. $P(T_0)$ and $P(T^*)$ will depend on both the likelihood and our prior, while the transition probabilities depend on the mechanics of our proposal.

First we discuss the likelihood contribution. Let y_i denote the observed y in the i^{th} bottom node given a tree T. Because the μ_{ij} are iid in our prior we have:

$$p(y \,|\, T) = \Pi\, p(y_i \,|\, \text{T}). \tag{32}$$

Thus the contribution of the likelihood to the ratio $P(P)/P(T_0)$ is just

$$\frac{p(y_l, y_r \,|\, T^*)}{p(y_{lr} \,|\, T_0)} = \frac{p(y_l \,|\, T^*)\, p(y_r \,|\, T^*)}{p(y_{lr} \,|\, T_0)} \tag{33}$$

where y_l denotes the observations in the new left child in T^*, y_r denotes the observation in the new right child in T^*, and y_{lr} denotes $\{y_l, y_r\}$. All other contributions to the likelihoods cancel out because of the product form of (32). Note that all three terms in the right hand side of (33) are just the predictive densities for a normal mean problem with known variance and normal prior on the mean.

As with the likelihood, much of the prior contributions to the posterior ratio cancel out since there is only place where the trees differ and our stochastic tree growing prior draws tree components independently at different "places" of the tree. Hence the prior contribution to the $P(T^*)/P(T_0)$ ratio is

$$\frac{(PG)\,(1 - PGl)\,(1 - PGr)\,P(rule)}{(1 - PG)}, \tag{34}$$

where

- PG: prior probability of growing at chosen bottom node of T_0.
- PGl: prior probability of growing at new left child in T^*.
- PGr: prior probability of growing at new right child in T^*.

- $P(rule)$: prior probability of choosing the rule defining the new children in T^*.

Each of the PG quantities is obtained from (20). The prior $P(rule)$ places a uniform distribution on variables and then a uniform distribution on the discrete set of split values associated with the drawn variable.

Finally, the ratio $P(T^* \to T_0)/P(T_0 \to T^*)$, is given by

$$\frac{(PD)\,(Pnog)}{(PB)\,(Pbot)\,P(rule)}, \tag{35}$$

where

- PD: probability of choosing the death proposal at tree T^*.
- $Pnog$: probability of choosing the nog node that gets you back T_0.
- PB: probability of choosing a birth proposal at T_0.
- $Pbot$: probability of choosing the T_0 bottom node such that a birth gets you to T^*.
- $P(rule)$: probability of drawing the new splitting rule to generate T^*'s children.

Our proposal draw of the new rule generating the two new bottom nodes is a draw from the prior. It is in this draw that variable selection (or, perhaps, variable proposal) occurs! Note that since our proposal for the $rule$ is a draw from the prior, it cancels out in the ratio (31).

The formulas given above correspond very closely to the source code in the BayesTree package in R. However, there are still many details omitted. For example, a quantity PGl might be zero if we keep track of which variables in x have been "used up" in that no further splits are possible.

4.2 Variable Selection and Interaction Detection Using BART

In this section we illustrate two forms of information extraction. The first is the variable selection approach given in (30). The second, interaction detection, uncovers which pairs of variables interact in analogous fashion by keeping track of the percentage of trees in the sum in which both variables occur. This exploits the fact that a sum-of-trees model captures an interaction between x_i and x_j by using them both for splitting rules in the same tree.

We illustrate the use of these methods in a simulated example and a real data example. Both of examples are "old chestnuts". Since our goal is interpretation, rather than prediction, we hope the use of familiar examples eases the path of the reader.

4.2.1 The Friedman Simulation Setup

We simulate $n = 500$ observations from our basic model

$$y = f(x) + \sigma Z, \quad Z \sim N(0, 1),$$

with x ten dimensional and

$$f(x_1, x_2, \ldots, x_{10}) = 10 \sin(\pi x_1 x_2) + 20 (x_3 - .5)^2 + x_4 + x_5.$$

The x_i are iid uniform on $(0, 1)$ and $\sigma = 1$.

Friedman (1991) originally suggested this simulation setup to study the efficacy of nonlinear regression techniques. However, the setup is perfect for illustrating variable selection and the discovery of interaction. Only the first five of the ten x components matter. With ten x's there are 45 possible interaction pairs. Our simulated data has just one of these possibilities present: only x_1 and x_2 interact. In a real application it would be of tremendous interest to know that only these two variables interact, even without having further knowledge of the functional form.

Results for one simulated data set are displayed in Figure (1). In panel (a) we have variable selection results. This panel corresponds closely to Figure (5) of Chipman et al. (2010). For each variable, we plot the posterior mean of the percentage of rules (across all m tree) which use that variable. With $m = 20$, we very clearly identify the first five variables as being important.

Panel (b) gives the interaction detection results. With ten variables, there are $\binom{10}{2} = 45$ possible variable pairs. For each pair, we plot the posterior mean of the percent of trees (out of m) which use both of the variables in splitting rules. We normalize the $m = 20$ and $m = 200$ results by dividing by each set of 45 posterior means by the maximum. Thus, the largest value displayed in each case is one. With both $m = 20$ and $m = 200$ we clearly identify the first pair (x_1 and x_2) as being of interest. With two variables involved, a pair is less likely to come in inconsequentially, so that the identification of interesting pairs is less sensitive to the choice of m than in the case of variable selection.

4.2.2 The Boston Housing Data

For an example with real data we turn to our second "old chestnut", the Boston housing data. The data where obtained from the R-package `mlbench` (R Development Core Team (2011)).

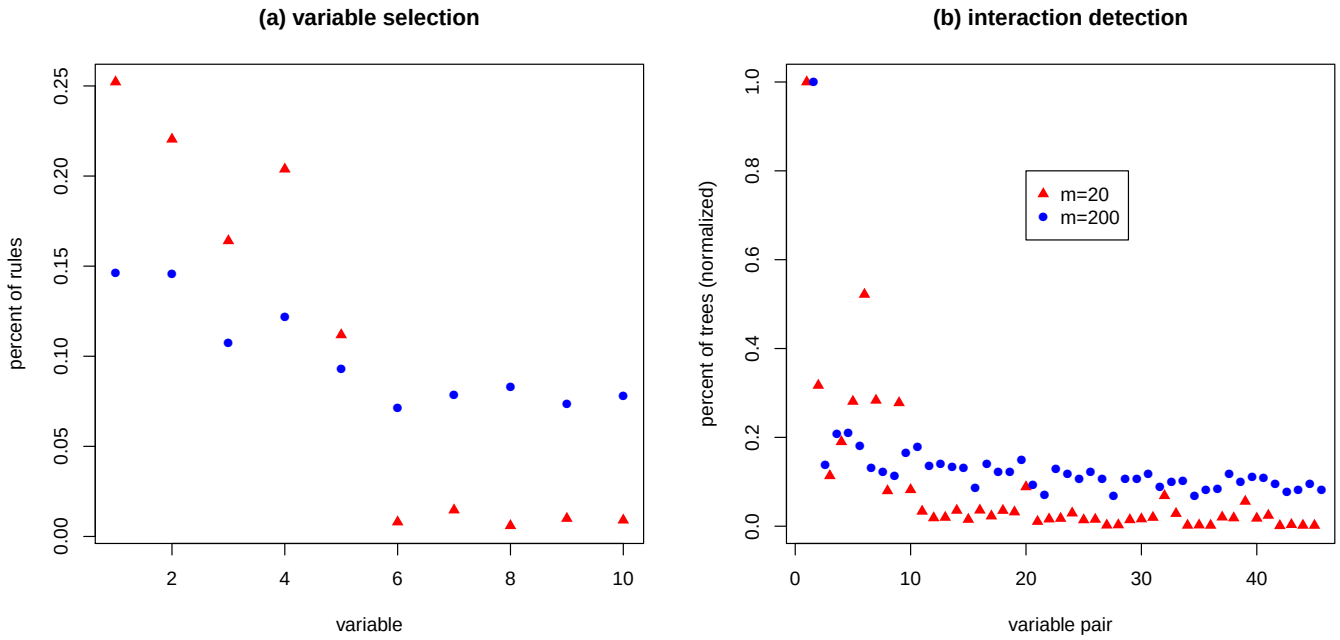

Figure 1: In panel (a) we correctly identify the first five variables as being important. In panel(b) we correctly identify the first interaction, which corresponds to variables x_1 and x_2.

There are 506 observations. Each observation corresponds to neighborhood. The response is the median house price in the neighborhood. There are 13 explanatory variables measuring characteristics of the neighborhoods. We did a preliminary variable selection (using the approach illustrated in the previous section) and tossed out three of the x's. Fitted values (from BART) with and without the three x's are very similar.

Figure (2) displays the results of the interaction detection. The format is the same as in panel (b) of Figure (1). Several pairs of interest are identified. Our real data has more interesting structure than our simulated data! We will investigate the pair `dis` and `lstat` simply because these variables are more easily understood. `dis` is the "weighted distances to five Boston employment centers". `lstat` is the "percentage of lower status of the population".

In Figure (3) we attempt to graphically see the interaction between `dis` and `lstat` suggested by Figure (2). In panel (a) we plot `dis` vs. `lstat`. Four subsets of points are identified depending on whether `dis` and `lstat` are "low" or "high". In the (b) panel we plot the fitted values from the BART run with $m = 200$. Before fitting we subtracted off the average response so the vertical axis is actually the amount the median value for a neighborhood is above the average. The four boxplots correspond to the four data subsets indicated in panel (a).

So, for example, the first boxplot displays the fitted prices when both `ds` and `lstat` are low. The observations included here correspond to those highlighted in the bottom left corner of panel (a). The label "`dL_lL`" indicates that `ds` is Low and `lstat` is Low. Similarly, the third boxplot is labeled "`dH_lL`", indicating that `ds` is High and `lstat` is Low.

The first pair of boxplots indicate the effect of increasing `lstat` when `ds` is low. The second pair of boxplots indicate the effect of increasing `lstat` when `ds` is high. Clearly, the boxplots indicate a strong interaction. For low `dl`, the effect of a the change in `lstat` is much more pronounced. A nice neighborhood close to the city center is highly desirable whereas a bad neighborhood close to the city center may be very bad.

5 Discussion and Beyond

The discovery of regression structure is an important and difficult problem in for all approaches to data analysis. With our modern computational tools it has become even more important. However, in a sense, it has also become more difficult, as we struggle to grapple with complex, high-dimensional models.

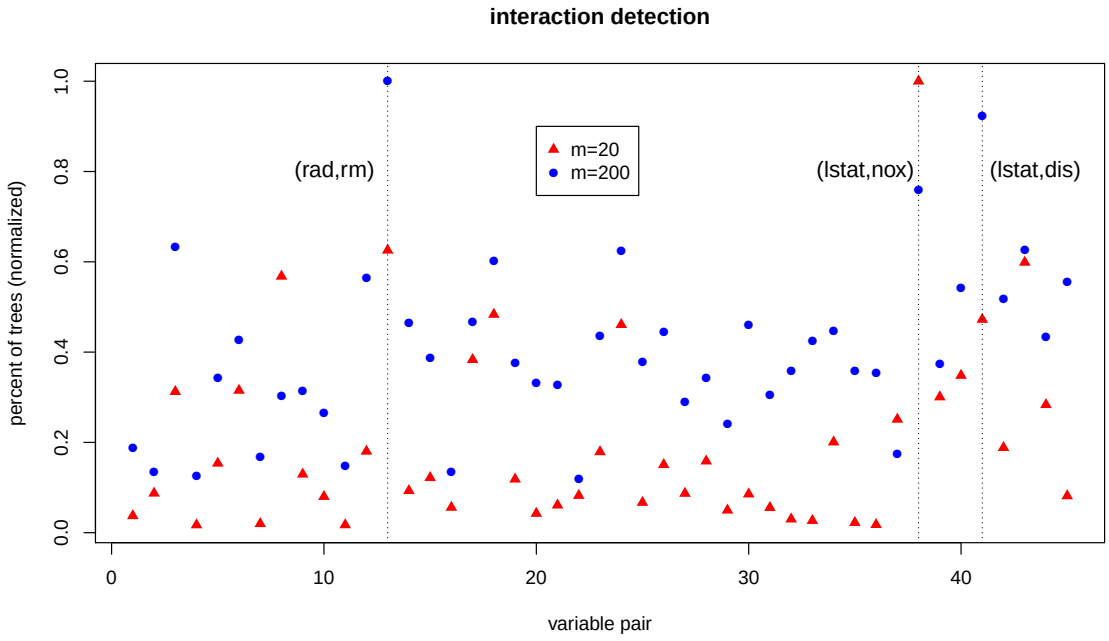

Figure 2: Interaction detection for the Boston housing data with ten explanatory variables.

In this article, we describe and contrast two different Bayesian approaches that illustrate the vast potential of Bayesian methods to extract information hidden in high dimensional data. The first is based on the classical parametric form of the normal linear model, while the second is based on a rich overparameterized sum-of-trees model, nonparametric in nature. In our examples, we show that even though the overall BART sum-of-trees model is complex, the simple structure of the individual tree components enables us to uncover structure with inferential posterior summaries. In particular, we have shown how BART provides a novel approach to model-free variable selection, the search for interesting variables, and model-free interaction detection, the search for interesting pairs of variables. Going beyond what we have presented here, the companion pieces by Clyde and Iverson (2011) and Gramacy (2011) in this volume, shed new light on directions for the development of Bayesian methods for grappling with model uncertainty.

As laid out by Clyde and Iverson (2011), the general Bayesian formulations for dealing with model uncertainly includes our parametric Bayesian formulation for variable selection as a special case. An important issue there is whether the class of models under consideration includes the actual data generating model. When it does, the so-called $\mathcal{M}$-closed setting, the implicit posterior model probabilities make sense. This will be the case in our parametric

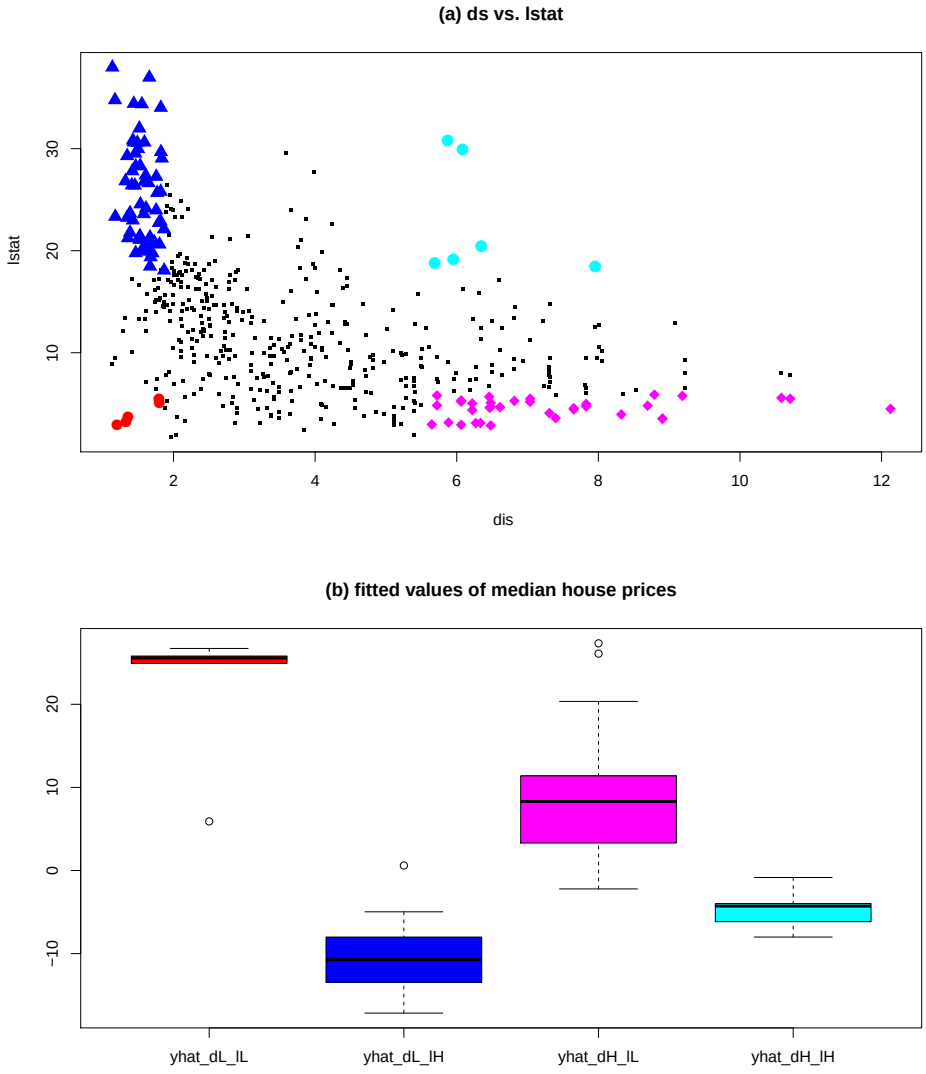

Figure 3: In panel (a) we identify four subsets of our data by whether each of `ds` and `lstat` are low or high. In panel (b) the boxplots display the fitted values (median house values) for the observations in the four subsets. The average of the dependent variable was subtracted off so that the vertical axis is the amount the median value of a neighborhood is above average. The first pair of boxplots both have low values of `dis`. The first box has low values of `lstat` and the second box has high values of `lstat`. The second pair of boxplots again compare low and high `lstat` but now `ds` is high.

Bayesian variable selection framework when it is valid to assume that the complete data was generated by normal linear model (with possibly some zero coefficients). However, it may often be more realistic to allow that the unknown actual model is outside the class under considerations, the so-called $\mathcal{M}$-open setting. This is precisely the linear model assumption limitation which we alluded to earlier, that a subset of actually related predictors may ignored simply because the relationship is not linear. Recognizing that the general limitation of using conventional Bayesian machinery for the $\mathcal{M}$-open setting, Clyde and Iverson consider alternatives to obtaining the weights corresponding to the conventional setting's posterior model probabilities. For this purpose they propose a principled decision theoretic cross-validation approaches for selection of weights that optimize model averaged predictions.

A key message in Gramacy (2011) is the central role played by the prior formulation in Bayesian variable selection. Indeed, the structured hierarchical mixture prior of Section 2.1 for the normal linear model, and the regularization prior of Section 3.1 for BART are essential for the effectiveness of these approaches. In both cases, it is necessary to use sensible hyperparameter values that balance "null-versus-alternative" possibilities in a way that allows the variable selection information in the data to emerge. Gramacy speaks to this in discussing the prior allocations that must be balanced in a variety of structured hierarchical mixture priors for the linear model, priors with coefficient marginals that are *both* concentrated near 0 and heavy-tailed, similar in the spirit to the lasso. By concentrating prior probability near zero, strong input from the data is needed to escape a neighborhood about zero. However, once the estimate has escaped from zero, the heavy tails allow it to wander far. Going further, Gramacy shows how latent variable formulations allow extension of the approaches to non-normal errors and binary observations. Convenient and efficient Gibbs sampling algorithms for posterior computation are detailed throughout allowing Gramacy to argue persuasively that these Bayesian approaches are powerful tools in our modern data rich environment.

References

Chipman, H. A., George, E. I. & McCulloch, R. E. (1998), 'Bayesian CART model search', *Journal of the American Statistical Association* **93**, 935–948.

Chipman, H. A., George, E. I. & McCulloch, R. E. (2001), The practical implementation of Bayesian model selection, *in* P. Lahiri, ed., 'Model Selection', Vol. 38 of *IMS Lecture Notes - Monograph Series*, Institute of Mathematical Statistics, Beachwood, OH, pp. 65–116.

Chipman, H. A., George, E. I. & McCulloch, R. E. (2010), 'BART: Bayesian additive regression trees', *Annals of Applied Statistics* **4**(1), 266–298.

Clyde, M. & George, E. I. (2004), 'Model uncertainty', *Statistical Science* **19**, 81–94.

Cui, W. & George, E. I. (2008), 'Empirical bayes vs. fully bayes variable selection', *Statist. Plann. Inference* **138**, 888–900.

Friedman, J. H. (1991), 'Multivariate adaptive regression splines (Disc: P67-141)', *The Annals of Statistics* **19**, 1–67.

Gelfand, A. E. & Smith, A. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Association* **85**, 398–409.

George, E. I. & McCulloch, R. E. (1993), 'Variable selection via Gibbs sampling', *Journal of the American Statistical Association* **88**, 881–889.

George, E. I. & McCulloch, R. E. (1997), 'Approaches for Bayesian variable selection', *Statistica Sinica* **7**, 339–374.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2008), 'Mixtures of g-priors for bayesian variable selection', *Journal of the American Statistical Association* **103**, 410–423.

Maruyama, Y. & George, E. I. (2011), 'Fully bayes factors using a generalized g-prior', *Ann. Statist.*

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Scott, J. G. & Berger, J. O. (2010), 'Bayes and empirical bayes multiplicity adjustment in the variable selection problem', *Ann. Statist* **38**, 2587–2619.

Tierney, L. (1994), 'Markov chains for exploring posterior distributions', *Ann. Statist* **22**, 1701–1762.

Wu, Y., Tjelmeland, H. & West, M. (2007), 'Bayesian CART: Prior specification and posterior simulation', *Journal of Computational and Graphical Statistics* **16**, 44–66.

Zellner, A. (1986), On assessing prior distribuitions and bayesian regression analysis with g-prior distributions, *in* 'Bayesian inference and decision techniques, Stud. Bayesian Econometrics Statist', Vol. 6, North-Holland, Amsterdam, pp. 233–243.