Heteroscedastic BART Using Multiplicative Regression Trees

M. T. Pratola*, H. A. Chipman†, E. I. George‡, and R. E. McCulloch§

September 21, 2017

Abstract

Bayesian additive regression trees (BART) has become increasingly popular as a flexible and scalable non-parametric model useful in many modern applied statistics regression problems. It brings many advantages to the practitioner dealing with large and complex non-linear response surfaces, such as a matrix-free formulation and the lack of a requirement to specify a regression basis a priori. However, while flexible in fitting the mean, the basic BART model relies on the standard i.i.d. normal model for the errors. This assumption is unrealistic in many applications. Moreover, in many applied problems understanding the relationship between the variance and predictors can be just as important as that of the mean model. We develop a novel heteroscedastic BART model to alleviate these concerns. Our approach is entirely non-parametric and does not rely on an a priori basis for the variance model. In BART, the conditional mean is modeled as a sum of trees, each of which determines a contribution to the overall mean. In this paper, we model the conditional variance with a product of trees, each of which determines a contribution to the overall variance. We implement the approach and demonstrate it on a simple low-dimensional simulated dataset, a higher-dimensional dataset of used car prices, a fisheries dataset and data from an alcohol consumption study.

Keywords: Non-parametric, uncertainty quantification, big data, applied statistical inference

*Department of Statistics, The Ohio State University, 1958 Neil Avenue, 404 Cockins Hall, Columbus, OH 43210-1247 (mpratola@stat.osu.edu).

†Department of Statistics, Acadia University.

‡Department of Statistics, The Wharton School, University of Pennsylvania.

§School of Mathematical and Statistical Sciences, Arizona State University.

1 Introduction

In the era of big data, statisticians face the challenging task of developing flexible regression models that can be used to discover important relationships between high-dimensional predictors and responses of interest, predict future behavior or estimate the out-of-sample response, and quantify uncertainties in performing such investigations. There are many popular methods that are emerging to successfully perform these tasks, such as dimension-reduction techniques (Allen et al., 2013; Cook, 2007), statistical model approximations (Sang and Huang, 2012), machine learning techniques such as boosting (Freund and Schapire, 1997; Freidman, 2001), bagging and random forests (Breiman, 2001), and Bayesian tree models (Chipman et al., 2002; Gramacy and Lee, 2008; Chipman et al., 2010; Taddy et al., 2011).

In some applications, simple point predictions are sufficient and many machine learning techniques algorithmically determine a point prediction of $E[Y|\mathbf{x}]$ for the response Y given predictors $\mathbf{x}$ and then choose tuning parameters of the algorithm based on the out of sample prediction error. However, in many applications a sense of the predictive uncertainty is of interest. In particular, for a numeric response, we may have heteroscedasticity: the variability of the response may be predictor dependent. In this paper our goal is to build a model which further allows flexible inference for the conditional variance $Var[Y|\mathbf{x}]$. Our approach is Bayesian, which allows us to fully quantify the predictive uncertainty using the predictive distribution.

The classical linear model approach to handling the issue of heteroscedasticity is to introduce a transformation of the response variable such as a log-transform or more generally the Box-Cox family of transformations (Box and Cox, 1964). When linear regression was the primary widely available modeling tool for practitioners, such approaches seemed justifiable. However, there are a number of limitations. For instance, as Box and Cox (1964) themselves pointed out, these are transformations for normality rather than homoscedasticity, which in practice means that one is not analyzing the data on the original scale which can make interpretation more challenging. The rigid assumed functional form of such transformations is also quite restrictive, requiring a "good" transformation to homoscedasticity to be found with only a single free parameter. In the context of modern regression where the dimension may be large and the sample size could be huge, such strong assumptions seem likely to be counterproductive. The Box-Cox transformations also require a strictly positive response variable. While this has recently been relaxed by the more general class of Yeo-Johnson transformations (Yeo and Johnson, 2000), the inferential challenges remain. Moreover, quantification of transformation uncertainty and investigating the bias/variance tradeoff seem largely ignored.

The Generalized Linear Model (GLM) offers some alternative to the transformation-based approach to modeling by allowing separation of the mean function and the error distribution via certain link functions. However, these models still make strong parametric assumptions which are unduly restrictive in the high-dimensional settings that motivate our methodology.

In the high-dimensional scenarios, even when $n > p$, the ability to flexibly account for heteroscedasticity and explore the relationship between variance and predictors is, to the best of our knowledge, very limited. Yet there is broad recognition of the importance of this problem. For example, Daye et al. (2012) note that the problem of heteroscedasticity has largely been ignored in the analysis of high-dimensional genomic data even though it is known to be a common feature of biological data. They propose a basis expansion of the mean and log variance as in Carroll (1988), and a penalized maximum likelihood algorithm for estimation similar to the LASSO (Tibshirani, 1996) to introduce sparsity in the solution. However, this approach depends on knowing a useful regression basis a priori, which is not always feasible in the high-dimensional setting and is a more challenging framework in which to quantify uncertainties.

Bayesian treed models nicely capture heteroscedasticity by building a single tree and letting the model be different in each bottom node of the tree (Gramacy and Lee, 2008; Taddy et al., 2011). This identifies subregions of the predictor space where different models are needed. However, there are advantages to the ensemble approach in which models are built up of many trees. Using a single tree may be overly simplistic and the ensemble approach enables a dramatically different search of model space. Bleich and Kapelner (2014) introduced an ensemble of trees regression model supporting heteroscedasticity, however while the mean model used an ensemble of trees to provide a flexible non-parametric form, the variance model proposed was a linear parametric model similar to Daye et al. (2012).

Bayesian additive regression trees (BART) (Chipman et al., 2010) use a Bayesian ensemble approach to modeling the conditional mean. In this paper, both the mean and the variance of an observed response are modeled using an ensemble-of-trees representation. In BART, the conditional mean is modeled as a "sum of trees", where each tree determines a mean and the overall mean is the sum of the contributions from the many trees. In this paper, we elaborate on the BART approach by also modeling the conditional variance with a "product of trees", where each tree determines a variance (or standard deviation) and the overall variance is the product of the contributions from the many trees. This takes advantage of the flexible tree-basis to form a response surface defined product-wise which allows the model to be conditionally conjugate, a desirable feature in some applications (e.g. financial models), which avoids the common yet somewhat artificial use of the log transform.

To motivate our modeling scenario of interest, consider the very simple simulated dataset shown in Figure 1 which we will later analyze in Section 4.1. This figure simply visualizes the inference from our model. First, the true mean function and the mean function $\pm$ twice the true predictor-dependent error standard deviation are shown using dashed lines. Second, the estimates from our model are displayed with solid lines. Figure 1 displays the basic goal of the paper: we seek to flexibly estimate both the predictor-dependent mean function *and* a predictor-dependent variance function.

In Figure 1 the predictor space is simplistic, being only one-dimensional. Our methodology will be demonstrated on four examples in Section 4, with our subsequent examples illustrating how our model works for higher-dimensional predictor spaces. Our second example in Section 4.2 is an extended data

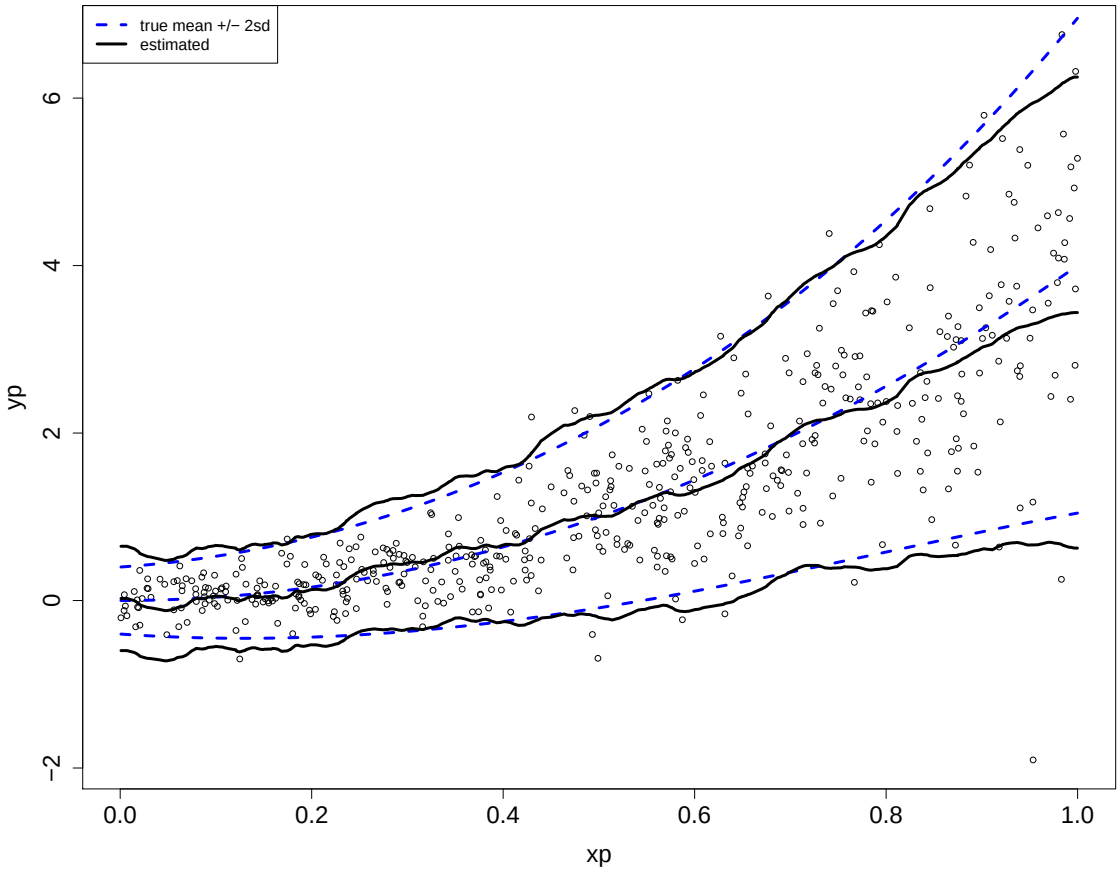

Figure 1: Simulated example. The simulation model is depicted by the dashed curves. The middle dashed curve is $f(x) = E[Y|\mathbf{x}]$ and the upper and lower dashed curves are draw at $f(x) \pm 2\,s(x)$ where $s(x) = \sqrt{Var[Y|\mathbf{x}]}$. The middle solid curve is $\hat{f}(x)$ and the upper and lower solid curves are draw at $\hat{f}(x) \pm 2\,\hat{s}(x)$ where $\hat{f}(x)$ and $\hat{s}(x)$ are estimates from our model. The symbols are plotted at the simulated (x_i, y_i) pairs.

analysis showing how our model can be used. In this example, the dependent variable is the price of a used car and the 15 dimensional predictor captures characteristics of the cars. In Section 4.3 we present two additional examples on fishing and alcohol consumption where our model is clearly not appropriate, but argue that it gives a very reasonable approximate solution in a reasonably simple way. In the former, the response variable is the amount of fish caught by a commercial fishing vessel, and there are 25 predictor variables. In the latter, the response is the amount of alcohol consumed and there are 35 predictor variables.

The paper proceeds as follows. In Section 2 we review the standard homoscedastic treed regression modeling scenario and in particular BART. In Section 3, we develop our novel flexible heteroscedastic

regression tree model which builds on BART's additive representation for the mean component with a multiplicative component for the variance model. We analyze the examples in Section 4. Finally, we conclude in Section 5.

2 Homoscedastic Treed Regression

The typical assumed scenario in the vast majority of statistical regression and machine learning techniques is modeling a process $Y(\mathbf{x})$ that is formed by an unknown mean function, $E[Y|\mathbf{x}] = f(\mathbf{x})$, and an unknown constant variance, $Var[Y|\mathbf{x}] = \sigma^2$, along with a stochastic component arising from independent random perturbations Z. This process is assumed to be generated according to the relation

$$Y(\mathbf{x}) = f(\mathbf{x}) + \sigma Z \tag{1}$$

where $Z \sim \mathrm{N}(0, 1)$ and $\mathbf{x} = (x_1, \ldots, x_d)$ is a d-dimensional vector of predictor variables. Such a process is known as a homoscedastic process.

In Bayesian treed regression models, the unknown mean function, $f(\mathbf{x})$, is modeled using a Bayesian regression tree. Regression trees are an elegant way of non-parametrically specifying adaptive regression bases, where the form of the bases are themselves learned from the observational data. The Bayesian approach models the data using a stochastic binary tree representation that is made up of interior nodes, $\mathbf{T}$, and a set of maps, $\mathbf{M}$, associated with the terminal nodes. Each interior tree node, η_i, has a left and right child, denoted $l(\eta_i)$ and $r(\eta_i)$. In addition, all nodes also have one parent node, $p(\eta_i)$, except for the tree root. One may also refer to a node by a unique integer identifier i, counting from the root using in-order traversal. For example, the root node η_1 is node 1. One can also label a subtree starting at node η_i simply as T_i. Figure 2 summarizes our notation.

Internal nodes of regression trees have split rules depending on the predictors and "cutpoints" which are the particular values of the predictors that the internal nodes split at. This modeling structure is encoded in $\mathbf{T}$, which accounts for the split rules at each internal node of a tree and the topological arrangement of nodes and edges forming the tree. Given the design matrix $\mathbf{X}$ of predictors having dimension $n \times d$, each column represents a predictor variable $v, v = 1, \ldots, d$ and each row $\mathbf{x}$ corresponds to the observed settings of these predictors. At a given internal node, the split rule is then of the form $x_v < c$ where x_v is the chosen split variable and c is the chosen cutpoint c for split variable x_v.

The Bayesian formulation proceeds by specifying discrete probability distributions on the split variables v taking on a value in $\{1, \ldots, d\}$ and specifying discrete probability distributions on the cutpoint value, taking on a value in $\{0, \frac{1}{n_v-1}, \ldots, \frac{n_v-2}{n_v-1}, 1\}$ where n_v is the total number of discrete cutpoints available for variable v. For a discrete predictor, n_v will equal one less the number of levels the predictor has while for continuous predictors, a choice of $n_v = 100$ is common (Chipman et al., 2010). The internal modeling structure of a tree, $\mathbf{T}$, could then be expressed as $\mathbf{T} = \{(v_1, c_1), (v_2, c_2), \ldots\}$.

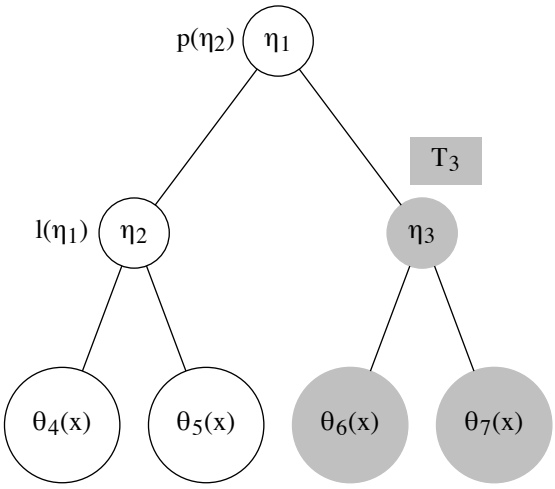

Figure 2: Labeling for a single regression tree $\mathbf{T}$. Nodes are denoted by circles and labeled using the symbol η. Lines denote branches connecting the nodes. Nodes can also be identified as left and right children (e.g. $\eta_2 = l(\eta_1)$) or as parent (e.g. $\eta_1 = p(\eta_2)$). Terminal nodes have no branches below them and contain maps represented as $\theta(\mathbf{x})$. A sub-tree uses the T symbol with subscript given by the node index (e.g. the subtree including η_3 and all its children is T_3). Note that later in the paper $\mathbf{T}_j$ will also index one member of an ensemble of trees, its use will be clear from context.

The Bayesian formulation is completed by specifying prior distributions on the maps at the terminal nodes. For $n^g = |\mathbf{M}|$ terminal nodes in a given tree, the corresponding maps are $\mathbf{M} = \{\theta_1, \ldots, \theta_{n^g}\}$. Taken all together, the Bayesian regression tree defines a function $g(\mathbf{x}; \mathbf{T}, \mathbf{M})$ which maps input $\mathbf{x}$ to a particular response θ_j, $j \in 1 \ldots n^g$.

2.1 BART

In the BART methodology, the model specifies the use of an ensemble of regression trees by combining a sum of the basic binary regression tree arrangement outlined above under the constant variance assumption,

$$y(\mathbf{x}_i) = \sum_{j=1}^{m} g(\mathbf{x}_i; \mathbf{T}_j, \mathbf{M}_j) + \sigma Z_i, \quad Z_i \sim N(0,1)$$

where the observation y collected at predictor setting $\mathbf{x}_i$ is modeled as the sum of m Bayesian regression trees and σ^2 is the variance of the observational process. Since the trees are modeling the mean response in BART, one usually describes the bottom node maps as $\theta \equiv \mu$. Then, conjugate normal priors are

specified for the bottom node maps,

$$\pi(\mu_{jk}) \sim N(0, \tau^2)$$

where μ_{jk} is the kth bottom node for tree j, along with a conjugate inverse chi-squared prior for the variance,

$$\sigma^2 \sim \chi^{-2}(\nu, \lambda).$$

where $\chi^{-2}(\nu, \lambda)$ denotes the distribution $(\nu\lambda)/\chi^2_\nu$. The model is regularized by penalizing tree complexity through a depth prior. The prior is specified by describing how a tree is drawn from the prior. A node at depth d spawns children with probability

$$\alpha(1 + d)^{-\beta},$$

for $\alpha \in (0, 1)$ and $\beta \geq 1$. As the tree grows, d gets bigger so that a node is less likely to spawn children and more likely to be remain a bottom node. Details on specifying the parameters of the prior distributions are discussed in detail in Chipman et al. (2010), while typically the choice $m = 200$ trees appears to be reasonable in many situations.

The conjugate normal priors on the terminal node μ's lead to a standard Gibbs sampler, as does the conjugate prior on the constant variance. Selecting the split variables and cutpoints of internal tree nodes is performed using a Metropolis-Hastings algorithm by growing and pruning each regression tree in a sequential Bayesian backfitting algorithm. The growing/pruning are performed by birth and death proposals which either split a current terminal node in $\mathbf{M}$ on some variable v at some cutpoint c, or collapse two terminal nodes in $\mathbf{M}$ to remove a split. For complete details of the MCMC algorithm, the reader is referred to Chipman et al. (1998); Denison et al. (1998); Chipman et al. (2010); Pratola (2016).

3 Heteroscedastic BART Model

As suggested earlier, real world data does not always follow the simple constant-variance model of BART. Instead, our interest is in modeling a process $Y(\mathbf{x})$ that is formed by an unknown mean function, $E[Y|\mathbf{x}] = f(\mathbf{x})$, *and* an unknown variance function, $Var[Y|\mathbf{x}] = s^2(\mathbf{x})$, along with a stochastic component arising from independent random perturbations Z. This process is assumed to be generated according to the relation

$$Y(\mathbf{x}) = f(\mathbf{x}) + s(\mathbf{x})Z \tag{2}$$

where again $Z \sim \mathrm{N}(0, 1)$ and $\mathbf{x} = (x_1, \ldots, x_d)$ is a d-dimensional vector of predictor variables which, for simplicity of exposition, are assumed to be common to both $f(\mathbf{x})$ and $s(\mathbf{x})$, although this is not strictly necessary in the proposed approach. This heteroscedastic process is the natural generalization of the homoscedastic "Normal-errors with constant variance" assumption of Equation (1) that is pervasive in the statistics and machine learning literatures. The types of inference in this setting are broader in scope

than the usual case. That is, in addition to predicting the mean behavior of the process by estimating $f(\mathbf{x})$ and investigating the importance of predictor variables on the mean response, one is also interested in inferring the variability of the process by estimating $s(\mathbf{x})$ and investigating which predictors are related to the process variability.

The proposed methodology will model the unknown mean function, $f(\mathbf{x})$, and the unknown variance function, $s^2(\mathbf{x})$, using ensembles of Bayesian regression trees. Our approach uses an additive regression tree model for the mean as in BART,

$$f(\mathbf{x}) = \sum_{j=1}^{m} g(\mathbf{x}; \mathbf{T}_j, \mathbf{M}_j),$$

and a multiplicative regression tree model for the variance component,

$$s^2(\mathbf{x}) = \prod_{l=1}^{m'} h(\mathbf{x}; \mathbf{T}'_l, \mathbf{M}'_l). \tag{3}$$

In this model, $\mathbf{T}_j$ encodes the structure of the j^{th} tree for the *mean* and $\mathbf{M}_j = \{\mu_{j,1}, \ldots, \mu_{j,n_j^g}\}$ are the $n_j^g = |\mathbf{M}_j|$ scalar terminal-node parameters for the mean in each tree. Similarly, $\mathbf{T}'_l$ encodes the structure of the l^{th} tree for the *variance* and in this case we represent the bottom node maps as $\theta \equiv s^2$ so that $\mathbf{M}'_l = \{s^2_{l,1}, \ldots, s^2_{l,n_l^h}\}$ are the $n_l^h = |\mathbf{M}'_l|$ scalar terminal-node parameters for the variance in each tree. In other words, $s^2(\mathbf{x}_i)$ is modeled as a product of Bayesian regression trees.

The posterior of the proposed heteroscedastic BART model is factored as

$$\pi(\mathbf{T}, \mathbf{M}, \mathbf{T}', \mathbf{M}') \propto L(\mathbf{y}|\mathbf{T}, \mathbf{M}, \mathbf{T}', \mathbf{M}') \prod_{j=1}^{m} \pi(\mathbf{T}_j)\pi(\mathbf{M}_j|\mathbf{T}_j) \prod_{l=1}^{m'} \pi(\mathbf{T}'_l)\pi(\mathbf{M}'_l|\mathbf{T}'_l)$$

where

$$\pi(\mathbf{M}_j|\mathbf{T}_j) = \prod_{k=1}^{n^g} \pi(\mu_{jk})$$

and

$$\pi(\mathbf{M}'_l|\mathbf{T}'_l) = \prod_{k=1}^{n^h} \pi(s^2_{lk}).$$

This specification assumes a priori independence of terminal node parameters for both the mean and variance, as well as independence of the mean trees and variance trees. This approach allows for easy specification of priors for the mean and variance model components and straightforward use of conditional conjugacy in implementing the MCMC sampler, which eases computations.

The proposed heteroscedastic regression tree model outlined above is fitted using a Markov Chain Monte Carlo (MCMC) algorithm. The basic steps of the algorithm are given in Algorithm 1. The algorithm is a Gibbs sampler in which we draw each $(\mathbf{T}_j, \mathbf{M}_j)$ and $(\mathbf{T}'_j, \mathbf{M}'_j)$ conditional on all other parameters and the

data. To draw $(\mathbf{T}_j, \mathbf{M}_j)$ we integrate out $\mathbf{M}_j$ and draw $\mathbf{T}_j$ and then $\mathbf{M}_j|\mathbf{T}_j$. The conditionally conjugate prior specifcations in Section 3.1 allow us to do the integral and draw easily. Our product of trees model and prior specifications (Section 3.3) allow us to use the same strategy to draw $(\mathbf{T}'_j, \mathbf{M}'_j)$. The draws of $\mathbf{T}_j|\cdot$ and $\mathbf{T}'_j|\cdot$ are done using Metropolis-Hastings steps as in Chipman et al. (2010) and Pratola (2016).

Data: $y_1, \ldots, y_N$; $\mathbf{x}_1, \ldots, \mathbf{x}_N$
Result: Approximate posterior samples drawn from $\pi(\mathbf{T}, \mathbf{M}, \mathbf{T}', \mathbf{M}'|y_1, \ldots, y_N; \mathbf{x}_1, \ldots, \mathbf{x}_N)$
for N_{mcmc} *iterations* **do**
  for $j = 1, \ldots, m$ **do**
    i. Draw $\mathbf{T}_j|\cdot$
    ii. Draw $\mathbf{M}_j|\mathbf{T}_j, \cdot$
  end
  for $j = 1, \ldots, m'$ **do**
    iii. Draw $\mathbf{T}'_j|\cdot$
    iv. Draw $\mathbf{M}'_j|\mathbf{T}'_j, \cdot$
  end
end

Algorithm 1: MCMC steps for the proposed heteroscedastic BART model.

Next we outline the mean and variance models in detail as well as the full conditionals required for implementing Algorithm 1.

3.1 Mean Model

Viewed as a function of μ_{jk}, the heteroscedastic BART likelihood in the k^{th} terminal node of the j^{th} mean tree is

$$L(\mu_{jk}|\cdot) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}s(\mathbf{x}_i)} exp\left(\frac{(r_i - \mu_{jk})^2}{s^2(\mathbf{x}_i)}\right)$$

where n is the number of observations mapping to the particular terminal node and

$$r_i = y_i - \sum_{q \neq j} g(\mathbf{x}_i; \mathbf{T}_q, \mathbf{M}_q).$$

As in BART, the conjugate prior distribution for the mean component is

$$\pi(\mu_{jk}) \sim N(0, \tau^2), \ \forall j, k.$$

Then the full conditional for the mean component is

$$\pi(\mu_{jk}|\cdot) \sim N\left(\frac{\sum_{i=1}^{n}\frac{r_i}{s^2(\mathbf{x}_i)}}{\frac{1}{\tau^2}+\sum_{i=1}^{n}\frac{1}{s^2(\mathbf{x}_i)}}, \frac{1}{\frac{1}{\tau^2}+\sum_{i=1}^{n}\frac{1}{s^2(\mathbf{x}_i)}}\right) \tag{4}$$

and (ignoring terms which cancel) the integrated likelihood is

$$\int L(\mu_{jk}|\cdot)\pi(\mu_{jk})d\mu_{jk} \propto \left(\tau^2\sum_{i=1}^{n}\frac{1}{s^2(\mathbf{x}_i)}+1\right)^{-1/2} exp\left(\frac{\frac{\tau^2}{2}\left(\sum_{i=1}^{n}\frac{r_i}{s^2(\mathbf{x}_i)}\right)^2}{\tau^2\sum_{i=1}^{n}\frac{1}{s^2(\mathbf{x}_i)}+1}\right) \tag{5}$$

which depend on the data only via the sufficient statistic

$$\sum_{i=1}^{n}\frac{r_i}{s^2(\mathbf{x}_i)}.$$

These forms are nearly identical to those of the homoscedastic BART model, with the only change arising from replacing a scalar variance s^2 with a vector variance $s^2(\mathbf{x}_i), i = 1, \ldots, n$. Conditional on a drawn realization of the variance component at the n observation sites, steps i and ii of Algorithm 1 are analogous to the BART sampler, where now instead of needing to calculate the sample means of data mapping to each tree's bottom node, one need calculate the variance-normalized means of the data to perform the Gibbs update of equation (4) and for updating the mean tree structures using the marginal likelihood of equation (5).

3.2 Calibrating the Mean Prior

As in BART, the specified prior for μ_{jk} implies a prior on the mean function, $f(\mathbf{x}) \sim \mathrm{N}(0, m\tau^2)$. This prior assumes mean-centered observations so that the mean of the prior on μ_{jk} is simply 0 as shown above. Calibrating the variance of the prior proceeds using a weakly data-informed approach by taking the minimum and maximum response values from the observed data, y_{min}, y_{max} and assigning a high probability to this interval, i.e. setting

$$\tau = \frac{y_{max} - y_{min}}{2\sqrt{m}\kappa}$$

where, for instance, $\kappa = 2$ species a 95% prior probability that $f(\mathbf{x})$ lies in the interval (y_{min}, y_{max}).

Essentially, the hyperparameter κ controls the bias-variance tradeoff: the higher is κ, the greater the probability that the mean function accounts for the range of observed data implying a smaller variance,

while the smaller is κ, the less probability that the mean function accounts for the range of observed data implying a larger variance. In homoscedastic BART, the recommended default setting is $\kappa = 2$ but for the proposed heteroscedastic BART model we have found that a higher default is preferable, such as $\kappa = 5$ or $\kappa = 10$. Ideally, we recommend carefully selecting κ in a prescribed manner in practice. In the examples of Section 4, a simple graphical approach and a cross-validation approach to selecting κ will be explored.

3.3 Variance Model

Viewed as a function of s^2_{lk}, the heteroscedastic BART likelihood in the k^{th} terminal node of the l^{th} variance tree is

$$L(s^2_{lk}|\cdot) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} s_{lk}} exp\left(-\frac{e_i^2}{2s^2_{lk}}\right),$$

where n is the number of observations mapping to the particular terminal node and

$$e_i^2 = \frac{\left(y(\mathbf{x}_i) - \sum_{j=1}^{m} g(\mathbf{x}_i; \mathbf{T}_j, \mathbf{M}_j)\right)^2}{s^2_{-l}(\mathbf{x}_i)},$$

where

$$s^2_{-l}(\mathbf{x}_i) = \prod_{q \neq l} h(\mathbf{x}_i; \mathbf{T}'_q, \mathbf{M}'_q).$$

We specify a conjugate prior distribution for the variance component as

$$s^2_{lk} \sim \chi^{-2}(\nu', \lambda'), \quad \forall l, k.$$

Then, it is readily shown that the full conditional for the variance component is

$$s^2_{lk}|\cdot \sim \chi^{-2}\left(\nu' + n, \frac{\nu'\lambda'^2 + \sum_{i=1}^{n} \frac{e_i^2}{s^2_{-l}(\mathbf{x}_i)}}{\nu' + n}\right) \tag{6}$$

so that the terminal nodes of the m' variance component trees in the product decomposition (3) are easily updated using Gibbs sampler steps. The integrated likelihood is also available in closed form,

$$\int L(s^2_{lk}|\cdot)\pi(s^2_{lk})ds^2_{lk} = \frac{\Gamma(\frac{\nu'+n}{2})\left(\frac{\nu'\lambda'^2}{2}\right)^{\nu'/2}}{(2\pi)^{n/2}\prod_{i=1}^{n} s_{-l}(\mathbf{x}_i)\Gamma(\nu'/2)\left(\nu'\lambda'^2 + \sum_{i=1}^{n} e_i^2\right)^{\frac{\nu'+n}{2}}} \tag{7}$$

which depends on the data only via the sufficient statistic,

$$\sum_{i=1}^{n} e_i^2.$$

This closed-form solution allows for easily exploring the structure of variance trees $\mathbf{T}'_1, \ldots, \mathbf{T}'_{m'}$ using Metropolis-Hastings steps in the same way that the mean model trees are explored. That is, the conjugate chi-squared prior leads to a Gibbs step drawing from a chi-squared full conditional when updating the components of $\mathbf{M}'$ using equation (6). Sampling the tree structure $\mathbf{T}'_j$ is performed via Metropolis-Hastings steps via the marginal likelihood (7). This is made possible as the integrated likelihood is analytically tractable with the heteroscedastic model specified.

The interpretation of this model form follows as the mean being factored into a sum of weakly informative components (as usual in BART) while the variance is factored into the product of weakly informative components. This latter factoring is indexed by the predictor $\mathbf{x}_i$ where each tree $h(\mathbf{x}_i; \mathbf{T}'_l, \mathbf{M}'_l)$ contributes a small component of the variance at $\mathbf{x}_i$ with the product of the $l = 1, \ldots, m'$ trees modeling the overall variance.

For the variance model we again specify discrete uniform priors on the split variables and cutpoints. Note that the number of variance component trees, m', need not equal the number of mean component trees, m. A default value that has worked well in the examples explored is $m' = 40$. Since the trees making up the mean model are different from the trees making up the variance model, the number of bottom nodes in the lth variance component tree, n_l^h, is unrelated to the number of bottom nodes in the jth mean component tree, n_j^g. This means, for instance, that the complexity of the variance function may be different from that of the mean function, and the predictors that are important for the variance function may also differ from those that are important for the mean.

Similar to the mean model component, a penalizing prior is placed on the depth of variance component trees, with the probability of a node spawning children equal to $\alpha'(1+d)^{-\beta'}$ where d is the depth of the node. Typically, the specification of this prior is chosen similar to that used in the mean model components, i.e. $\alpha' = 0.95, \beta' = 2$ specifies a prior preference for shallow trees having a depth of 2 or 3.

3.4 Calibrating the Variance Prior

The simplicity of the prior for $f(\mathbf{x})$ (Section 3.2) is a major strength of BART. Our prior for $s(\mathbf{x})$ is more complicated. We show in this section that a simple strategy for assessing the prior gives very reasonable results.

From Section 3.3 our prior is

$$s(\mathbf{x})^2 \sim \prod_{l=1}^{m'} s_l^2, \quad \text{with } s_l^2 \sim \chi^{-2}(\nu', \lambda'), \text{ i.i.d.}$$

As in the case of $f(\mathbf{x})$, the prior for $s(\mathbf{x})$, does not depend on $\mathbf{x}$.

Selecting the prior parameters ν', λ' for the variance components may be done in the following way. We suppose we start with a prior for a single variance in the context of the homoscedastic BART model $Y = f(\mathbf{x}) + \sigma Z$ with

$$\sigma^2 \sim \chi^{-2}(\nu, \lambda).$$

Chipman et al. (2010) discuss strategies for choosing the parameters (ν, λ) in this case. We can choose a prior in the heteroscedastic model to match the prior in the homoscedastic case by matching the prior means. We have

$$E[\sigma^2] = \frac{\nu\lambda}{\nu - 2},$$

and

$$E[s(\mathbf{x})^2] = \prod_{l=1}^{m'} E[s_l^2] = \lambda^{m'} \left(\frac{\nu'}{\nu' - 2}\right)^{m'}.$$

We then match the means by separately matching the "λ piece" and the "ν piece" giving

$$\lambda' = \lambda^{\frac{1}{m'}}, \qquad \nu' = \frac{2}{1 - \left(1 - \frac{2}{\nu}\right)^{1/m'}}.$$

Figure 3 illustrates this procedure. For a problem in which the response is the price of (very expensive) used cars we elicited a prior on σ with $\nu = 10$ and $\lambda = 26000^2$. The resulting prior density for σ is plotted with a solid line in Figure 3. Using the formulas above with $m' = 40$ we get $\nu' = 360$ and $\lambda' = 1.66$. The resulting prior density for $s(\mathbf{x})$ is plotted with a dashed line in Figure 3. In both cases, we simply took a large number of draws from the prior and the reported and density smooth of the draws. These priors are remarkably similar both in their location *and* in their general shape. Using this approach, prior specification is no more complicated that in the homoscedastic case.

4 Examples

We now demonstrate the proposed methodology on several examples. Our first example (Section 4.1) is a simulated example with a one dimensional x. With a one-dimensional x we can use simple graphics to display the results obtained using our model. The second example (Section 4.2) is an extended real example in which we develop a model to predict the price of a used car using car characteristics. Graphical examination of the data shows that both nonlinearity and heteroscedasticity are present. Our model successfully captures both of these. We then briefly present two real examples (Section 4.3) with $\mathbf{x}$ having dimension 25 and 35. In these examples, the basic $Y = f(\mathbf{x}) + s(\mathbf{x})Z$ structure is suspect but we argue that our model finds a reasonable approximation that is relatively interpretable.

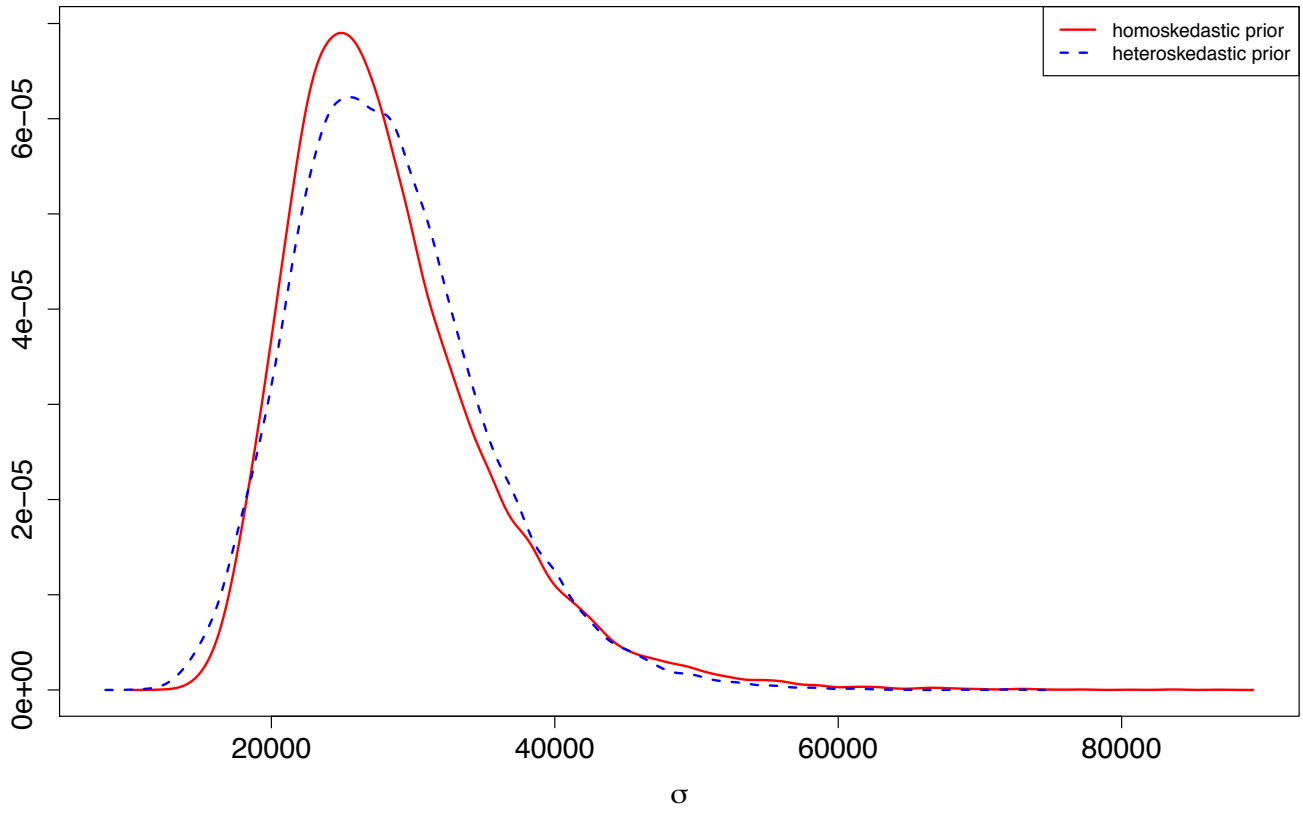

Figure 3: Choosing the prior for the error variance. The plot is on the standard deviation scale. The solid line is a prior choice for σ. The dashed line is the (approximate) density curve for a matching prior for $s(\mathbf{x})$ calculated via simulation.

In all our examples we use two plots to visualize the effectiveness of heteroscedastic BART and compare it to the homoscedastic version. In the *H-evidence* plot we display posterior intervals for $s(\mathbf{x}_i)$ sorted by the values of $\hat{s}(\mathbf{x}_i)$. As such, the H-evidence plot is a simple graphic to detect evidence in favor of heteroscedasticity. Using this plot we can quickly see if the conditional variance is predictor dependent. In the *predictive qq-plot* we start with a sample of observations $(\mathbf{x}_i, y_i)$ (in or out of sample). For each $\mathbf{x}_i$ we compute quantiles of y_i based on the predictive distribution $Y|\mathbf{x}_i$ obtained from our model. If the model is correct, the quantiles should look like draws from the uniform distribution. We use qq-plots to compare our quantiles to uniform draws.

In the cars example we demonstrate the use of cross-validation to choose prior settings. Typically, cross-validation involves the choice of a simple measure of predictive performance such as RMSE (root mean square error). However, the goal of heteroscedastic BART is to get a better feeling for the conditional uncertainty than one typically obtains from statistical/machine-learning methodologies. Hence we use the *e-statistic* (Székely and Rizzo, 2004) as a measure of quality of the predictive qq-plot as our target in assessing out-of-sample performance.

We use the following model and prior specifications unless otherwise stated. For the number of trees in our ensembles, $m = 200$ for the mean model and $m' = 40$ for the variance model. The tree prior uses $\alpha = 2$ and $\beta = .95$ for both $\{\mathbf{T}_j\}$, the mean trees, and $\{\mathbf{T}'_j\}$, the variance trees. The mean prior parameter is $\kappa = 2$ with the consequent choice for τ discussed in Section 3.2. We use $\nu = 10$ and λ equal to the sample variance of y to specify a σ prior and then use the approach of Section 3.4 to specify the ν' and λ' for the heteroscedastic s prior.

4.1 Simulated Example

We simulated 500 observations from the model

$$Y_i = 4x^2 + .2\, e^{2x}\, Z_i.$$

so that $Y_i = f(x_i) + s(x_i)\, Z_i$ with $f(x) = 4x^2$ and $s(x) = .2\, e^{2x}$. Each x is drawn independently from the uniform distribution on (0,1) and each Z_i is drawn independently from the standard normal distribution. We then simulated an independent data set in the exact same to serve as out-of-sample data.

The out-of-sample simulated data, simulation model, and point estimates are depicted in Figure 1. The dashed curves represent the model with the center line being x vs. $f(x)$ and the two outer lines are drawn at $f(x) \pm 2\, s(x)$. The points are plotted at the simulated (x_i, y_i) pairs.

We ran the MCMC for 1,000 burn-in draws and kept 2,000 subsequent draws to represent the posterior. The solid lines in Figure 1 represent estimates of $\hat{f}(x)$ and $\hat{f}(x) \pm 2\hat{s}(x)$ obtained by averaging $\{f_j(x)\}$ and $\{s_j(x)\}$ where j indexes post burn-in MCMC draws. Although very similar results are obtained using the default $\kappa = 2$. we used $\kappa = 5$. This is appropriate given the very smooth nature of the true f.

Figure 4 informally examines the performance of the MCMC chain by displaying sequences of post burn-in draws for certain marginals. The top panel displays the draws of σ from the homoscedastic BART model $Y = f(x) + \sigma Z$. For homoscedastic BART, this plot is a simple way to get a feeling for the performance of the MCMC. We can see that the draws vary about a fixed level with an appreciable but moderate level of autocorrelation. For heteroscedastic BART, there is no simple summary of the overall error level comparable to the homoscedastic σ. The middle panel plots draws of $s(x)$ for $x = .12, .42, .63, .79, .91$. The bottom panel plots the draws of $\bar{s} = \frac{1}{n}\sum s(x_i)$, the average s value for each MCMC draw of s. Here the x_i are from the test data. In all plots, the MCMC appears to be reasonably well behaved.

Figure 5 displays the uncertainty associated with our inference for f and s. The left panel displays inference for f and the right panel for s. In each panel, the dashed panel is the true function, the solid line is the estimated function (the posterior mean estimated by the average of MCMC draws) and the dot-dash line represents point-wise 95% posterior intervals for each $f(x_i)$ (left panel) and $s(x_i)$ (right panel). The intervals are estimated by the quantiles of the MCMC draws of $\{f(x_i)\}$ and $\{s(x_i)\}$ and the x_i are from the test data.

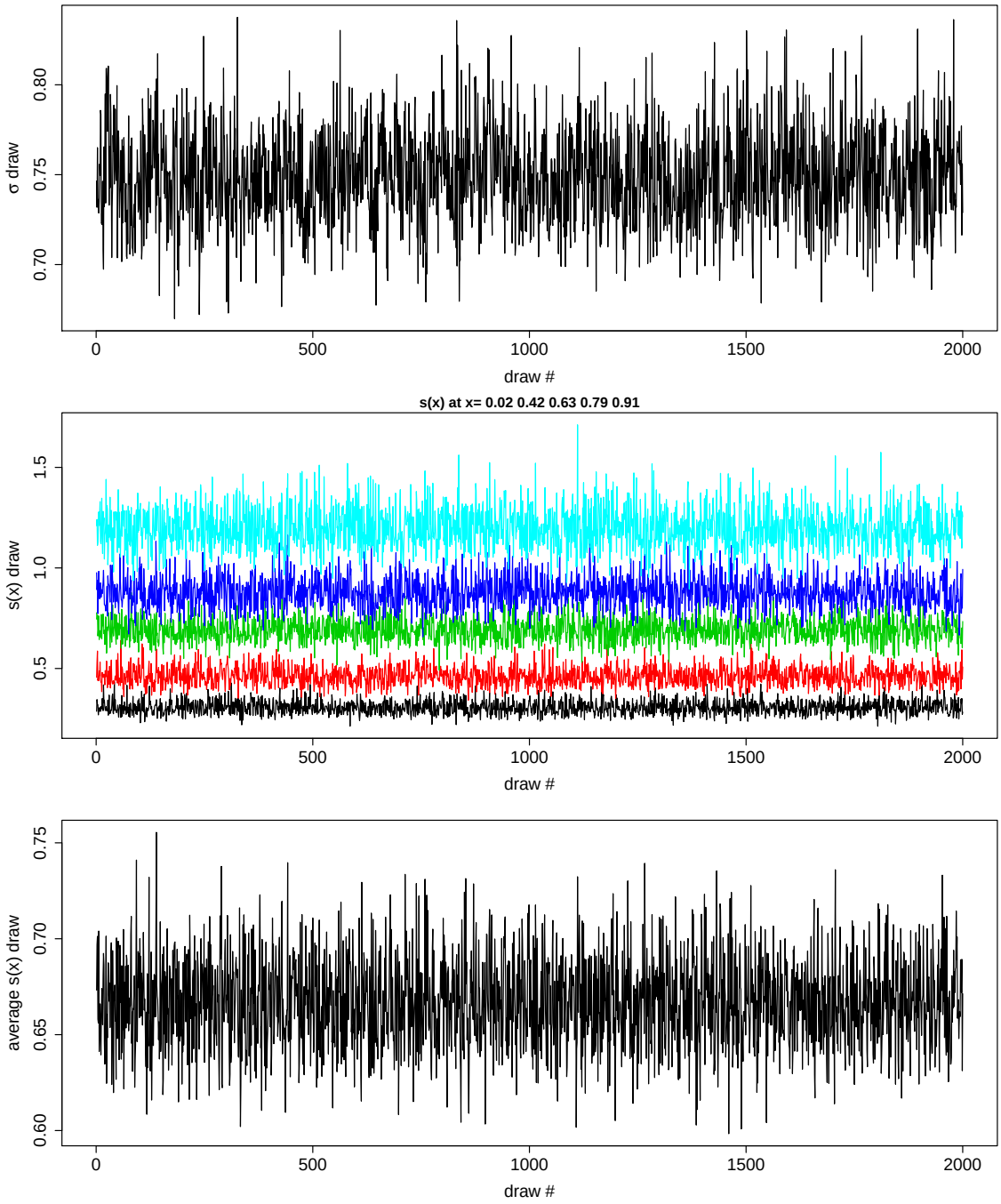

Figure 4: Simulated example. Top panel: MCMC draws of σ in homoscedastic BART. Middle panel: MCMC draws of $s(x)$ for five different x. Bottom panel: MCMC draws of $\bar{s}$ the average of $s(x_i)$ for each MCMC draw.

Figure 6 displays the inference for $s(x)$ in a way that will also work for higher dimensional x. We sort the observations according to the values of $\hat{s}(x)$. We plot $\hat{s}(x)$ on the horizontal axis and posterior intervals

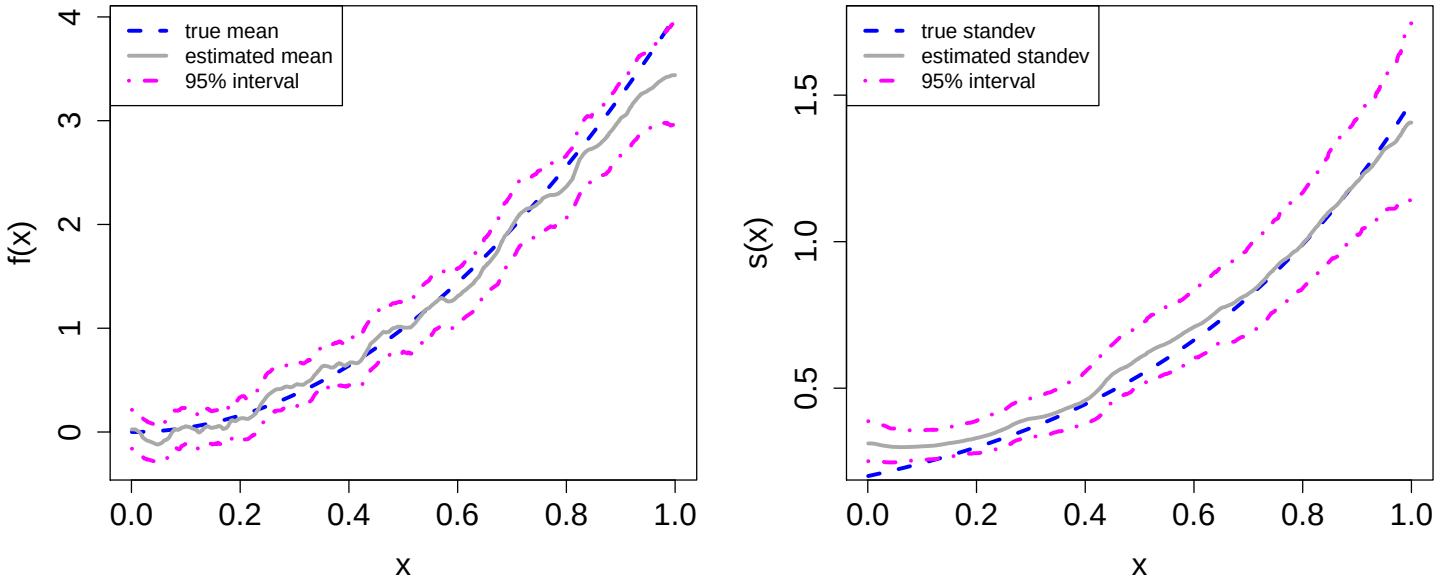

Figure 5: Simulated example. The left panel displays inference for f and the right panel displays inference for s. In each panel the dashed line is the true function and the solid line is the estimated function. The dot-dash lines give point-wise 95% posterior intervals for f and s.

for $s(x)$ on the vertical axis. The solid horizontal line is drawn at the estimate of σ obtained from the homoscedastic version of BART (see the top panel of Figure 4). We can clearly see that the posterior intervals are cleanly separated from the horizontal line indicating that our discovery of heteroscedasticity is "significant". In our real applications, where the units of σ and $s(x)$ are the same as the units of the response, we are able to assess the *practical significance* of the estimated departure from constant variance. We have found this plot useful in many applications. We will use it in our other examples and call it the H-evidence plot.

Figure 7 assesses the fit of our model by looking at qq-plots (quantile-quantile plots) based on the predictive distribution obtained from our model. For each i we obtain draws from $p(y \mid x_i)$ and then compute the percentile of the observed y_i in these draws. If the model is correct, these percentiles should look like draws from the uniform distribution on (0,1). We use the qq-plot to compare these percentiles to draws from the uniform. Since the (x_i, y_i) pairs are from the test data so that we are evaluating the out-of-sample predictive performance.

The left panel of Figure 7 shows the qq-plot obtained from our heteroscedastic model. A "45 degree" line is drawn with intercept 0 and slope 1. We see that our predictive percentiles match the uniform draws very well. In the right panel, we do the same exercise, but this time our predictive draws are obtained from the homoscedastic model. The failure of the homoscedastic BART model is striking.

Note that in our real applications we will be able to use the formats of Figures 6 and 7 to visualize the

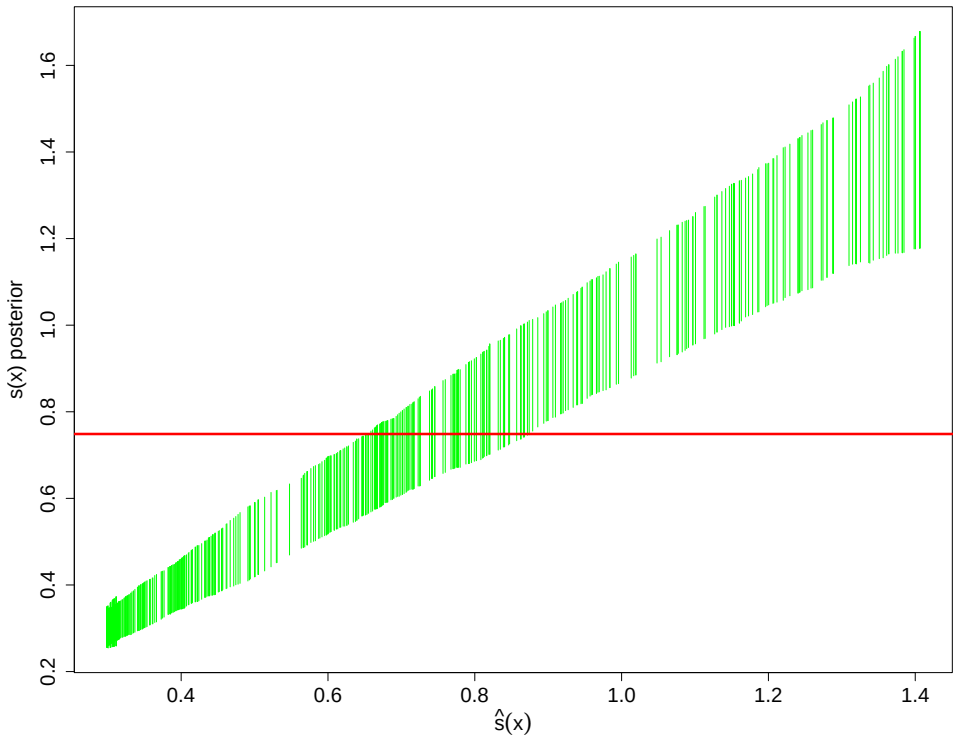

Figure 6: Simulated example. H-evidence plot. Posterior intervals for $s(x_i)$ sorted by $\hat{s}(x_i)$. The solid horizontal line is drawn at the estimate of σ obtained from fitting homoscedastic BART.

inference of heteroscedastic BART with high dimensional $\mathbf{x}$.

4.2 Cars

Perhaps one of the universally desired and undesired consumer expenditures in modern society is the purchase of a new or used car. These large transactions are not only a challenge for consumers but also for the car sales industry itself. Factors such as the commission-based nature of car sales, brand reliability, model reliability, warranty period, projected maintenance costs, and broader macro-economic conditions such as gasoline prices and job security all weigh on a consumers mind when purchasing a vehicle while trying to extract the most value for their dollar. At the same time, these same variables weigh on car dealers as they try to extract the greatest profit from their inventory. When important variables - such as gasoline prices - change, the effect can be profound. For instance, if gasoline prices suddenly rise over a short time period, a consumer who recently purchased a large, expensive but fuel inefficient vehicle may find their economic assumptions change for the worse, while a dealership with a large inventory of such vehicles may suddenly be facing abnormally large losses rather than the normally small profits.

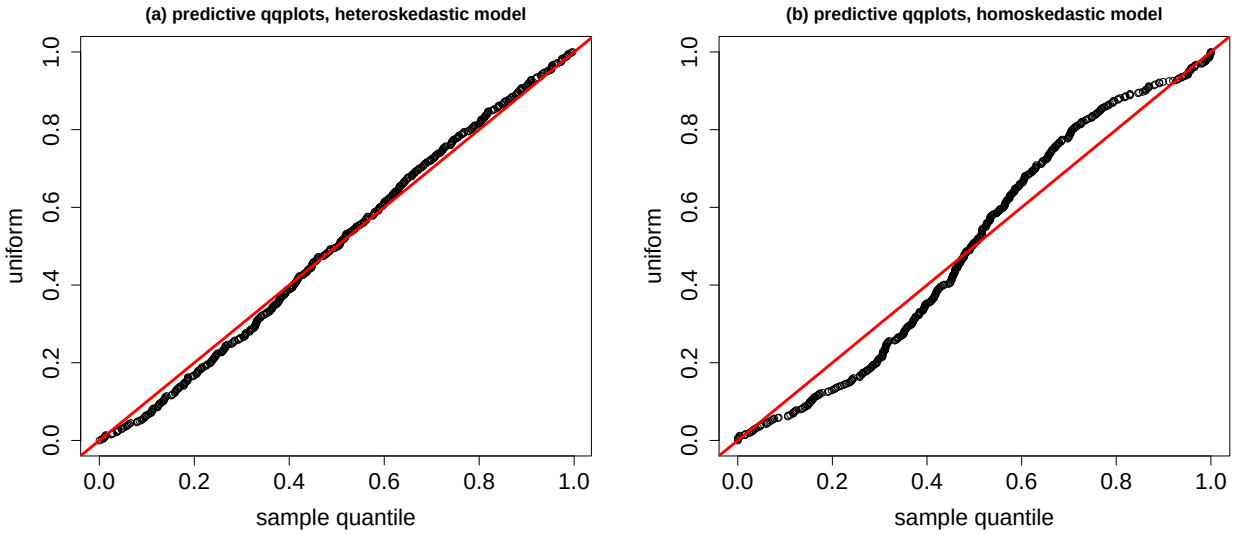

Figure 7: Simulated example. Predictive qq-plots. Left panel: heteroscedastic BART. Right panel: homoscedastic BART.

Variable	Levels	Description
`trim`	430,500,550,other	Higher trim corresponds to higher-end vehicle.
`color`	black,silver,white,other	Color of vehicle.
`displacement`	4.6, 5.5, other	Larger displacement corresponds to more powerful gas engine.
`isOneOwner`	true, false	Has vehicle had a single owner.

Table 1: Summary of categorical predictor variables in the cars dataset.

Studying pricing data for car sales is therefore a problem of great interest. However, besides understanding the mean behavior of price in response to changes in predictor variables, in such complex markets where profit margins are minimal, changes in the *variability* of prices could be equally important to understand for consumers and dealers alike. In this section, a dataset of $n = 1,000$ observations of used car sales data taken between 1994-2013 is investigated. Notably, this dataset covers the 2007-2008 financial crisis and the subsequent recovery. The dataset consists of the response variable `price` (USD), 2 continuous predictor variables, `mileage` (miles) and `year`, and 4 categorical predictor variables, `trim`, `color`, `displacement` and `isOneOwner`. The categorical variables are summarized in Table 1.

Expanding the categorical variables into binary dummy variables results in a total of 15 predictor variables. The relationship between active categorical predictors and the response variable `price` and continuous predictors `mileage` and `year` are summarized in Figure 8. Note that the categorical predictor `color` does not appear in this figure as it has little marginal effect on the response.

This plot provides some notable summary information on how the categorical predictors marginally affect

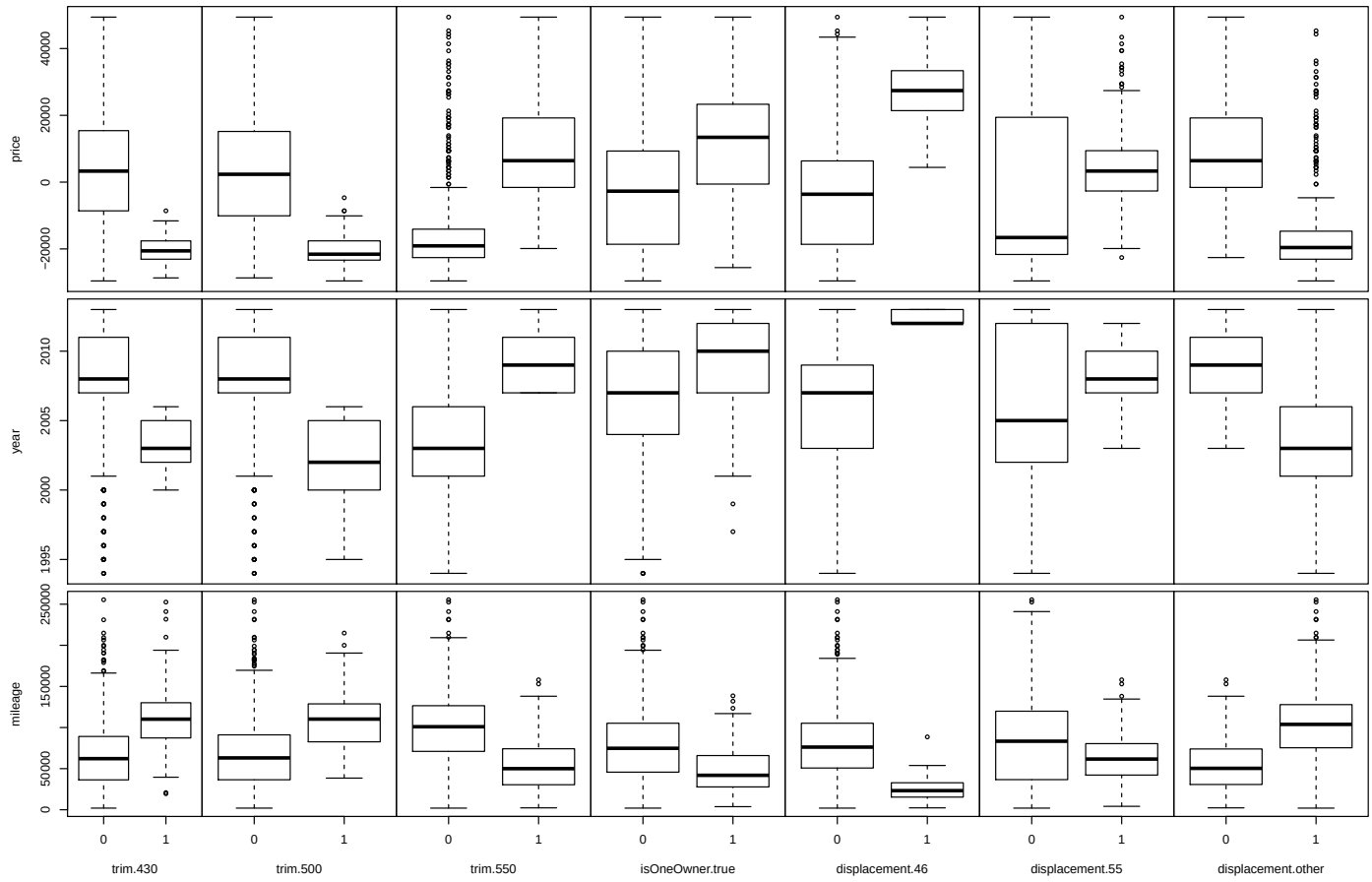

Figure 8: Used cars example. Summary of response variable `price` and continuous predictors `mileage` and `year` by the levels of the important categorical predictors `trim`, `isOneOwner` and `displacement`.

car price:

- `trim.430` and `trim.500` have lower median price, while `trim.550` has higher median price;
- `displacement.46` and `displacement.55` has higher median price, while `displacement.other` has lower median price, but note that `displacement.55` is more or less located in the middle range of prices.

There is also evidence of collinearity between the categorical predictors and continuous predictor variables:

- `trim.430` and `trim.500` have higher median mileage, while `trim.550` has lower median mileage;
- `trim.430` and `trim.500` have lower median year (older cars), while `trim.550` has higher median year (younger cars);

- `displacement.46` and `displacement.55` has lower median mileage and higher median year (younger cars) while `displacement.other` has higher median mileage/lower median year (older cars);
- `isOneOwner.true` tends to correspond to younger cars with lower median mileage.

We can better understand these complex relationships by plotting the continuous variables color coded by each important categorical variable as shown in Figures 9 and 10. These figures provide added insight. For instance, there is a clear curvilinear relationship between `mileage` and `price`, with higher `mileage` implying lower `price`. There is also a curvilinear relationship between `year` and `price` with higher `year` (younger car) implying higher `price`. However `mileage` and `year` are strongly negatively correlated (Pearson correlation of -0.74), so the amount of additional information for including one of these predictors after the other may be relatively small.

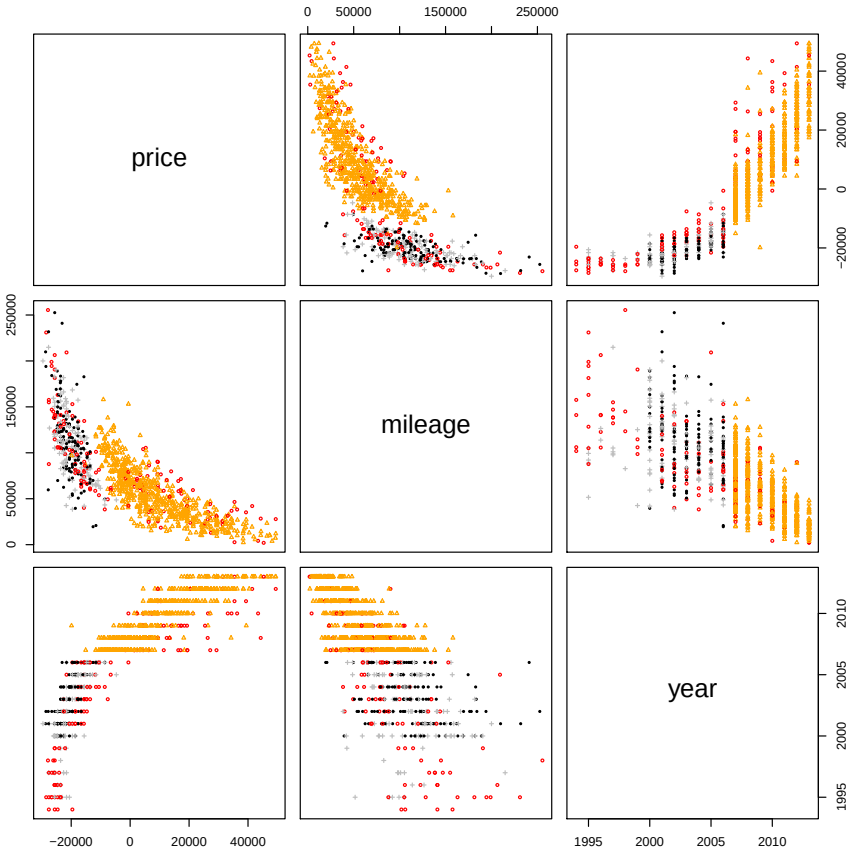

Figure 9: Used cars example. Summary of continuous variables coded by level of `trim`. `trim.430` shown by black solid dots, `trim.500` by grey '+', `trim.550` by orange triangles and `trim.other` by red 'o'.

Figure 9 shows that `trim.550` explains much of the "jump" seen in this figure for `year`$\geq$ 2007, but not all as there are some `trim.other` cars that spread the entire range of years in the dataset. There also appears

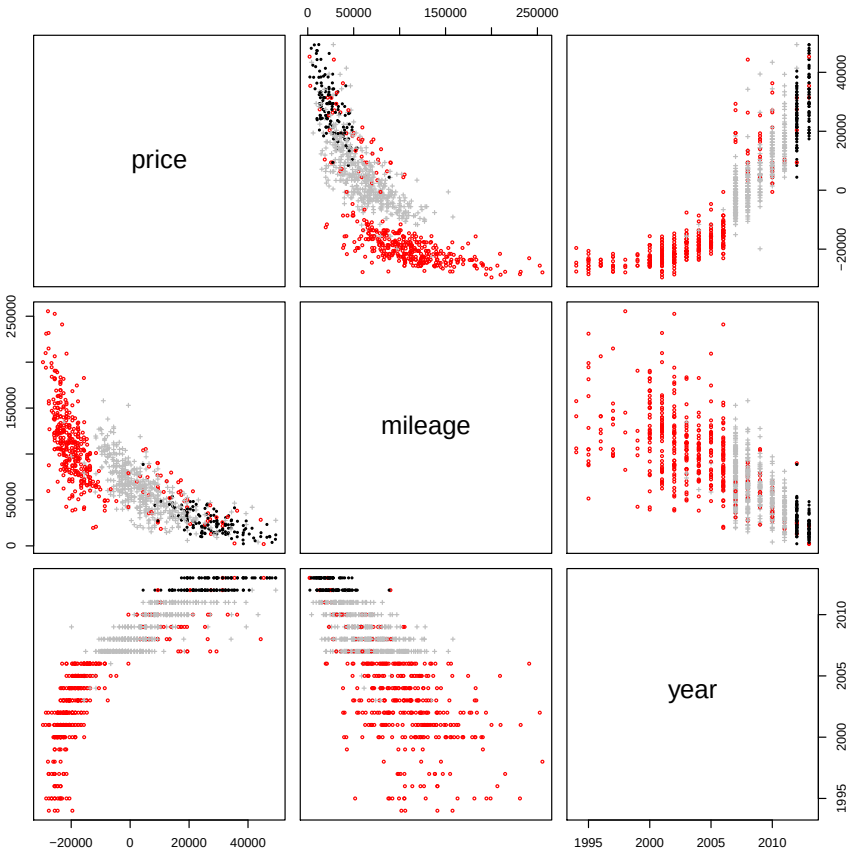

Figure 10: Used cars example. Summary of continuous variables coded by level of `displacement`. `displacement.46` shown by black solid dots, `displacement.55` by grey '+' and `displacement.other` by red 'o'.

to be a notable change in the spread of prices from 2007 onwards that does not seem likely to be explained by a mean function of the predictors. Rather, it suggests evidence for heteroscedasticity. It also appears that the spread changes in an abrupt manner around the year 2007 across a wide range of price values, which suggests a simple approach such as log transforming the data to make the spread appear more constant might ignore potential insights that can be gained from analyzing this dataset. For instance, looking at the log transform of `price` coded by levels of `trim` shown in Figure 11, a non-constant spread in `price` still appears evident across many years and many trim levels. Therefore, taking a simplistic transformation approach will not allow us to extract all the information available from the variables on their natural scale, and makes it more difficult to interpret the data. In addition, Box and Cox (1964) note that such power transformations alter both the variance and distribution of the data, making it difficult to separate second-moment corrections from higher-moment corrections, which aim to correct for normality rather than provide a concerted attempt at modeling and inferring heteroscedasticity.

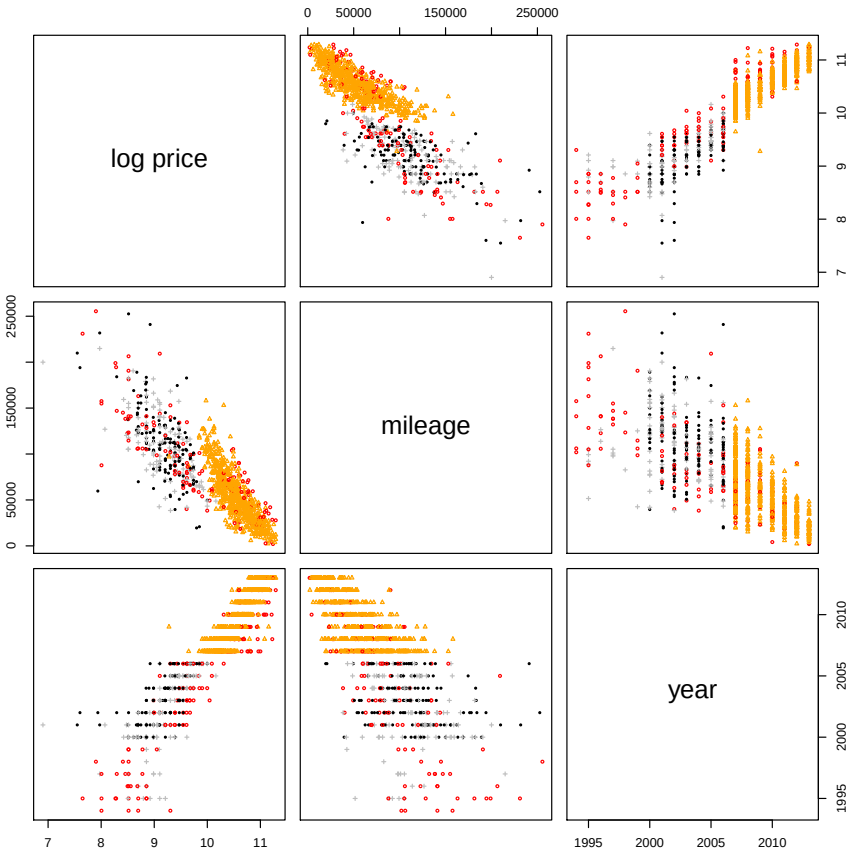

Figure 11: Used cars example. Summary of log(`price`) and other continuous variables coded by level of `trim`. `trim.430` shown by black solid dots, `trim.500` by grey '+', `trim.550` by orange triangles and `trim.other` by red 'o'.

Both variants of the BART model were applied to analyze this dataset. When analyzing a real-world dataset, it is desirable to have a principled approach for tuning the settings of important hyperparameters. In the case of the proposed heteroscedastic BART model, it is important to judiciously select the prior hyperparameter κ in specifying the prior mean model. This parameter essentially controls whether BART will tend to fit smoother mean functions with a preference for greater heteroscedasticity (large κ) or fit more complex mean functions with a preference for homoscedasticity (small κ). However, selecting κ using cross-validation based on MSE, for instance, is not adequate in our setting since we are interested in matching the first and second moments of our dataset rather than just the mean.

Instead, we build on the idea of the qq-plots shown in section 4.1 which did allow us to compare the fit of models from a *distributional* perspective, rather than a simplistic MSE perspective. Rather than viewing a potentially large number of qq-plots, we use a measure of distance between distributions to compare

percentiles calculated as in Figure 7 to the uniform distribution using a 1-number summary. Our approach makes use of the so-called energy, or e-statistic proposed by Székely and Rizzo (2004), although many alternative metrics of distributional distance are available in the literature.

For both the usual homoscedastic model and the proposed heteroscedastic model, 10-fold cross-validation was performed to select the value of the hyperparameter κ based on the e-statistic. That is, for each (x, y) in the held-out fold, we compute the percentile of y in the predictive distribution of $Y|x$ computed from the other folds. We then use the e-statistic to compare these percentiles (one for each observation in the held-out fold) to the uniform distribution. This gives us one e-statistic comparison for each of the 10 folds. This procedures is repeated for a small selection of plausible values of κ. The result of this cross-validation is summarized in Figure 12. The cross-validation results suggest using a value of $\kappa = 2$ (or smaller) for the homoscedastic model while greater emphasis on smoothing the mean model is suggested with a cross-validated value of $\kappa = 10$ for the heteroscedastic model. Besides the suggested settings for κ, note the overall much smaller values of e-distance for the heteroscedastic model, implying that this model is much better at capturing the *overall distributional pattern* of the observations rather than simply the mean behavior. Generally, the e-statistic is much smaller for the heteroscedastic model.

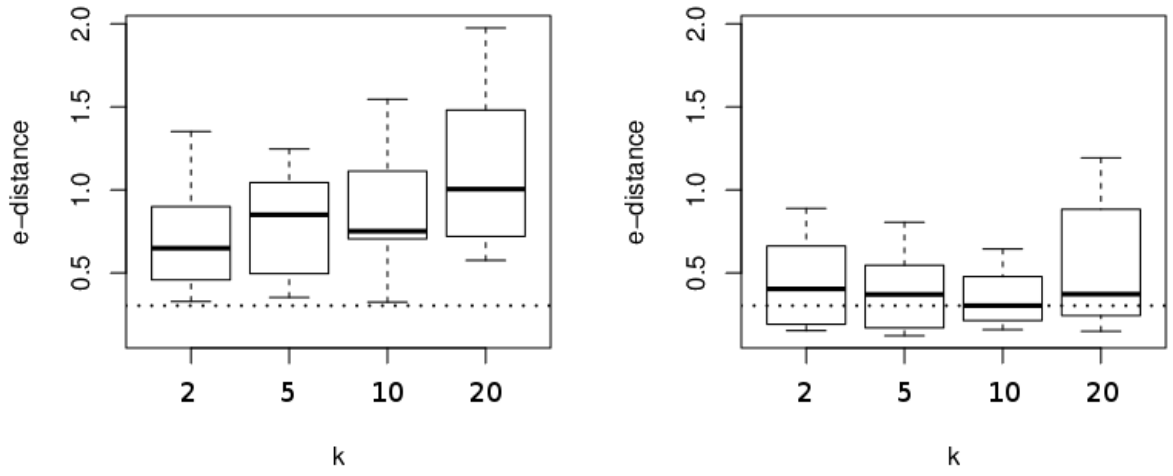

Figure 12: Used cars example. Boxplots of e-distance from 10-fold cross-validation of homoscedastic BART (left pane) and heteroscedastic BART (right pane) in tuning the prior mean hyperparameter κ. Horizontal dotted line corresponds to the median e-distance for $\kappa = 10$ with the heteroscedastic BART model.

The corresponding predictive out-of-sample qq-plots for the homoscedastic model trained on 60% of the full dataset with $\kappa = 2$ and the heteroscedastic model with $\kappa = 10$ are shown in Figure 13. This serves

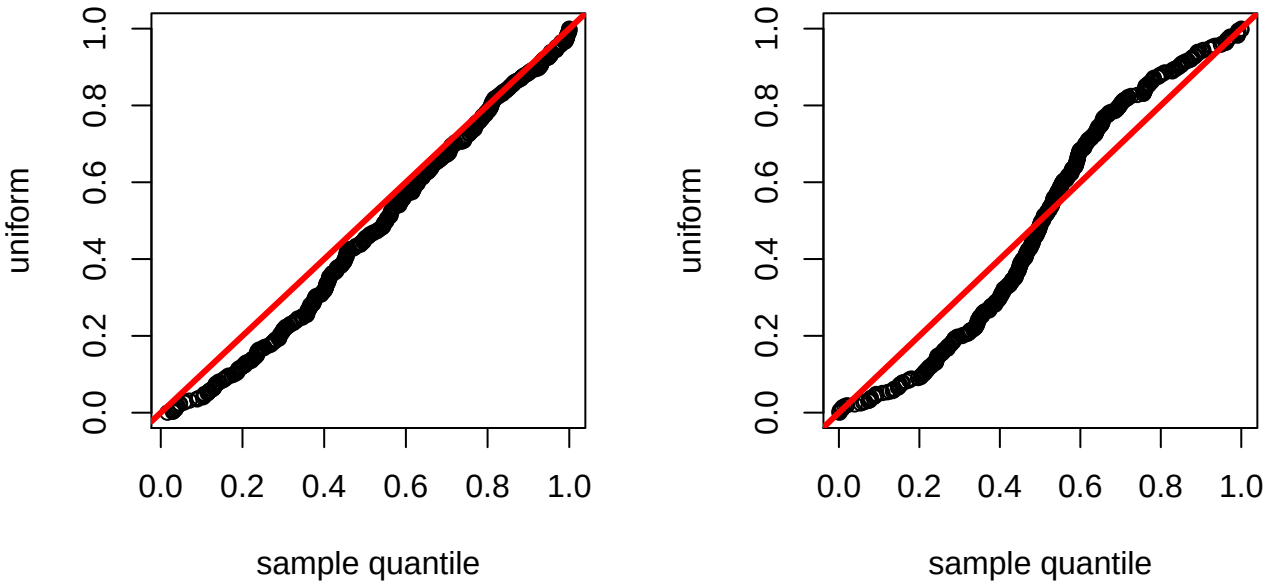

Figure 13: Used cars example. Predictive qq-plot of posterior draws for `price` calibrated to the uniform distribution for the heteroscedastic model with $\kappa = 10$ (left pane) and the homoscedastic model with $\kappa = 2$ (right pane).

as an empirical graphical validation for the high level of prior smoothing suggested by the cross-validated tuning of κ. The qq-plot for the heteroscedastic model is much closer to a straight line and dramatically better than the qq-plot for the homoscedastic version. The H-evidence plot in Figure 14 serves as an additional check for evidence of heteroscedasticity: clearly, the posterior 90% credible intervals do not cover the estimate of standard deviation from the homoscedastic model for a majority of the data. Even though there is considerable uncertainty about $s(\mathbf{x})$ at the higher levels, we have strong evidence that $s(\mathbf{x}) > 10,000$ at the the higher levels and $s(\mathbf{x}) < 2,500$ at the lower levels. These differences are *practically* significant.

Finally, a large benefit of the proposed model is the ability to perform inference on both the mean *and* variance of the observed process. A standard measure of variable activity in BART is to calculate the percentage of internal tree node splits on each variable across all the trees. With heteroscedastic BART, we can obtain this information for both the mean and standard deviation $s(\mathbf{x})$ as shown in Figure 15. These figures summarize some interesting findings. The majority of tree splits for the mean model occur on the first two predictors in the dataset, `mileage` and `year`. These two variables are also important in tree splits for the variance model as shown in the right pane. As we suspected earlier, this includes the

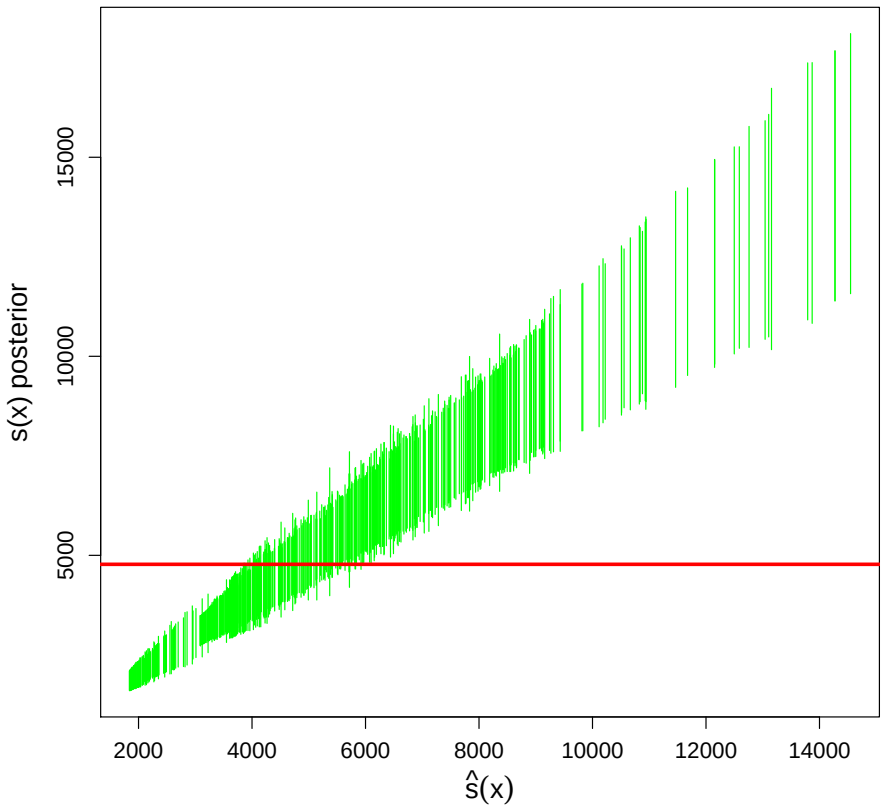

Figure 14: Used cars example. H-evidence plot. 90% posterior credible intervals for $s(\mathbf{x})$ for the heteroscedastic BART model versus observation index sorted by level of the posterior mean standard deviation, $\hat{s}(\mathbf{x})$. The solid horizontal line shows the estimate of σ from the homoscedastic BART model for reference.

`year` variable which agrees with our exploratory plot in Figure 11 where a large jump in the spread of the response variable seems to occur from 2007 onwards. `mileage` is also an important variable to model the spread of the data, which is not surprising since one would expect cars with similar characteristics with mileage near 0 may have well determined values while as the mileage increases the determination of value becomes more complex. Interestingly, a third variable seems strongly important for the variance trees while not particularly important for the mean trees: `trim.other`. As shown in Figure 11, cars with `trim.other` unusually span the entire range of years in the dataset, so it seems sensible that it may be an important variable in modeling the spread of the data.

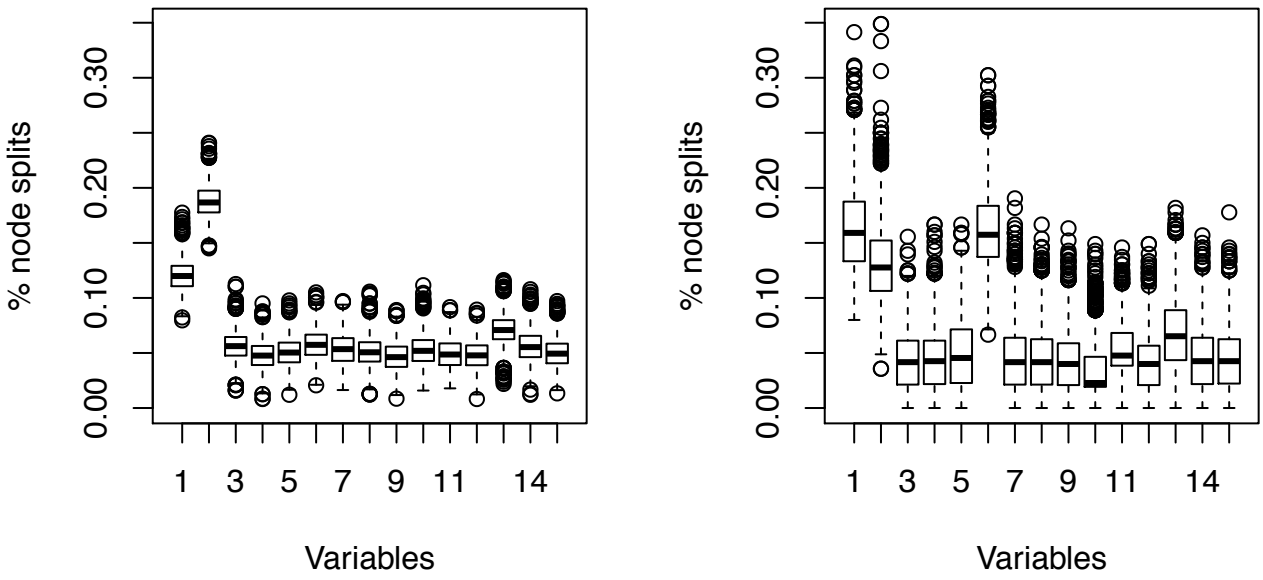

Figure 15: Used cars example. Posterior variable activity from the heteroscedastic BART model for the mean (left pane) and $s(\mathbf{x})$ (right pane).

4.3 Fishery and Alcohol Examples

In this section we very briefly present results from two more examples. In the first example the dependent variable y is the daily catch of fishing boats in the Grand Bank fishing grounds (Fernandez et al., 2002) The explanatory $\mathbf{x}$ variables capture time, location, and characteristics of the boat. After the creation of dummies for categorical variables, the dimension of $\mathbf{x}$ is 25. In the second example, the dependent variable y is the number of alcoholic beverages consumed in the last two weeks. (Kenkel and Terza, 2001) The explanatory $\mathbf{x}$ variables capture demographic and physical characteristics of the respondents as well as a key treatment variable indicating receipt of advice from a physician. After the creation of dummies for categorical variables, the dimension of $\mathbf{x}$ is 35.

In both of the examples the response is constrained to be positive and there is a set of observations with $y = 0$ so that there is a clear sense in which our model $Y = f(\mathbf{x}) + s(\mathbf{x})Z$ is inappropriate. In both previous papers, careful modeling was done to capture the special nature of the dependent variable. Our interest here is to see how well our model can capture the data given our flexible representations of f and s in the presence of a clear mispecification.

Figures 16 and 17 present the results for the fish data using the same displays we have employed in our previous examples. In Figure 16 we see very strong evidence of heteroscedasticity. Our product of trees representation of s enables the model to represent the data by being quite certain that for some $\mathbf{x}$ the error standard deviation should be small. Does this work? In Figure 17 we see the (in-sample) qqplots. While the qqplot for the heteroscedastic BART model is not perfect, it is a dramatic improvement over

the homoscedastic fit and may be sufficiently accurate for practical purposes.

In the left panel of Figure 17 we have also plotted the qqplot obtained from the plug-in model $Y \sim N(\hat{f}(\mathbf{x}), \hat{s}(\mathbf{x})^2)$. This is represented by a dashed line. It is difficult to see because it coincides almost exactly with the qqplot plot obtained from the full predictive distribution.

Our feeling is that in many applications the representation $Y \sim N(\hat{f}(\mathbf{x}), \hat{s}(\mathbf{x})^2)$ may be adequate and has an appealing simplicity. Many users will be able to understand this output easily without knowledge of the representations of f and s.

Figures 18 and 19 give results for the Alcohol data again using the same format. In this example the inference suggests that the homoscedastic version is adequate and the (in-sample) qqplots are very similar. In this case, even without the heteroscedastic model the flexible f captures the patterns reasonably well, although the qqplots are not perfect.

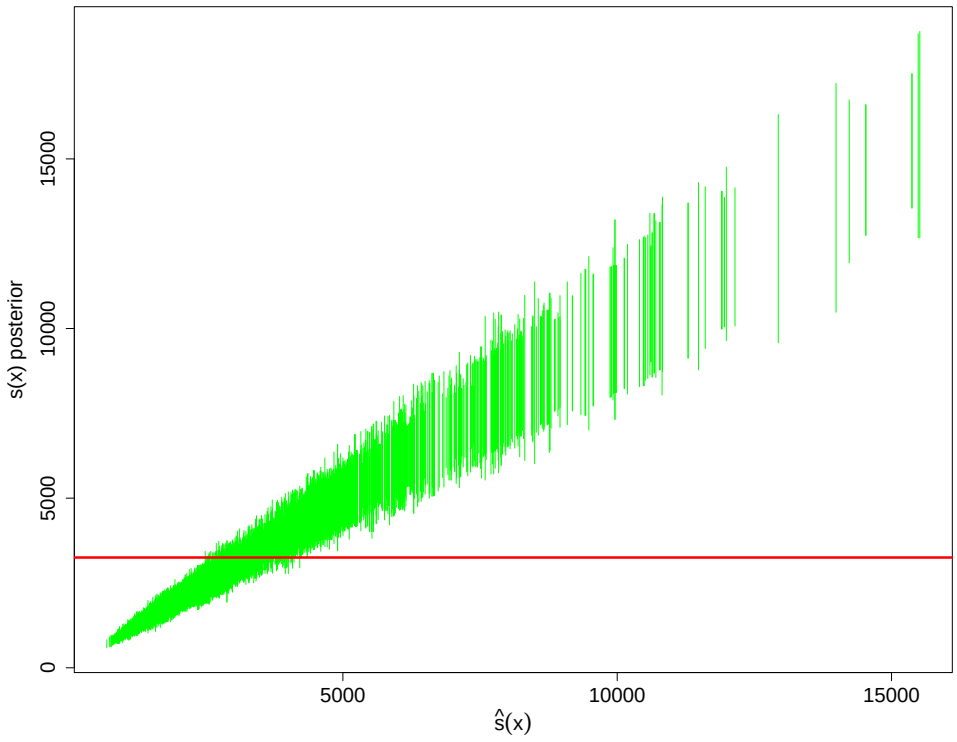

Figure 16: Fishery example. H-evidence plot. Posterior intervals for $s(\mathbf{x}_i)$ sorted by $\hat{s}(\mathbf{x}_i)$. The solid horizontal line is draw at the estimate of σ obtained from fitting homoscedastic BART.

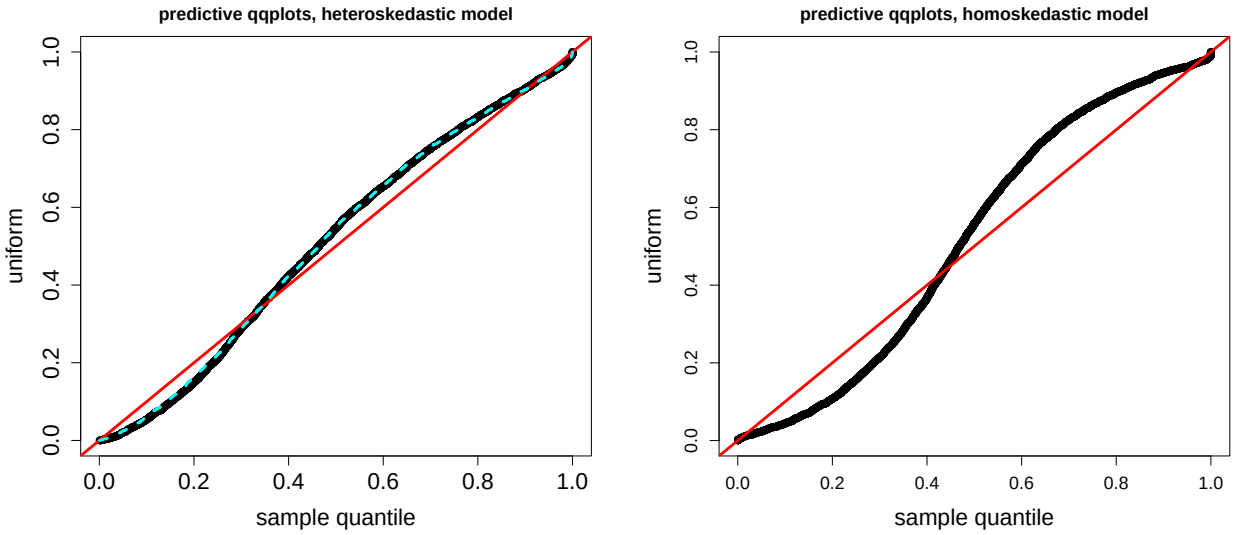

Figure 17: Fishery example. Predictive qq-plots. Left panel: heteroscedastic BART. Right panel: homoscedastic BART.

5 Conclusion

In this paper we proposed a Bayesian non-parametric regression model that allows for flexible inference for both the conditional mean and variance using ensembles of trees. Modeling the mean takes the sum-of-trees approach made popular in the BART methodology while we introduced a product-of-trees ensemble for modeling the variance of the observed process. Our specification of the product-of-trees model gives us the conditional conjugacy needed for a relatively simple MCMC algorithm. It also allows for an approach to prior specification which is no more complex to use than specifying priors in homoscedastic BART.

Our heteroscedastic model provides for a richer inference than standard BART. In the context of our examples we developed tools for visualizing and tuning the inference. The H-evidence plot allows us to quickly assess the evidence for predictor dependent variance. The predictive qq-plot allows us to assess the full predictive distributional fit to the data. While we often obtain good results from default prior specifications, it may be desirable to tune the inference by using cross-validation for prior choice. In our experience, calibrating the priors requires only cross-validating the hyperparameter 'κ' in the prior mean model, thereby offering a fairly low level of complexity in terms of using the approach in real applied analyses. This is fairly surprising given the amount of flexibility that can be afforded by the model when the data demands it. Rather than using mean square error to assess out-of-sample performance in cross-validation, we use the e-statistic to summarize the predictive qq-plot.

We demonstrated the proposed method on a simulated example, used cars sales data, a fisheries dataset and data on alcohol consumption. The simulated example, where the true mean and variance functions

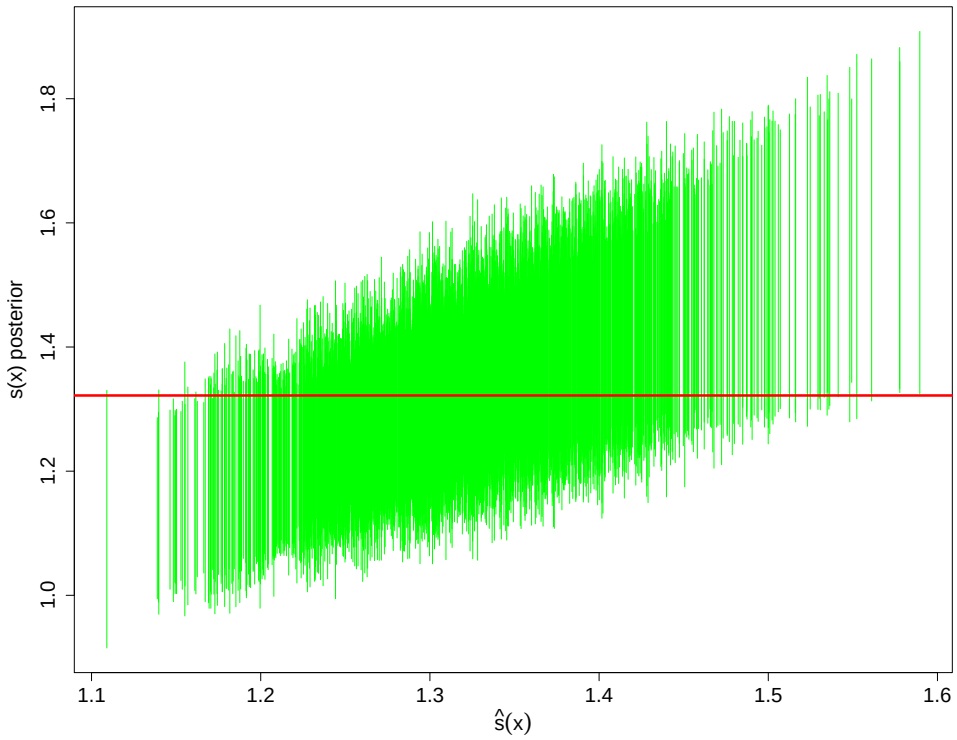

Figure 18: Alcohol example. H-evidence plot. Posterior intervals for $s(\mathbf{x}_i)$ sorted by $\hat{s}(\mathbf{x}_i)$. The solid horizontal line is draw at the estimate of σ obtained from fitting homoscedastic BART.

were known, resulted in very accurate estimation while relying on nothing more than the default priors. In the used cars data where a deeper analysis was performed, the model captures clear evidence of heteroscedasticity. Analysis of this data also revealed interesting differences in variable activity, where the trees for the mean had 2 highly active predictors while the trees for the variance had 3 highly active predictors. The ability to extract such inferential information for both the mean and variance components of the data may have important practical consequences.

In the cars data, the qq-plot (Figure 13) indicates that our model has done a very effective job of capturing the conditional distribution of sales price given the car characteristics. In the Fishery example, the fit from the heteroscedastic model is dramatically better than the homoscedastic fit (Figure 17) but not perfect. In the alcohol data example (Figures 19 and 18) we can easily infer that the heteroscedastic version does not improve the fit.

In the Fishery data example we also note (Figure 17) that the plug-in model $Y|\mathbf{x} \sim N(\hat{f}(\mathbf{x}), \hat{s}(\mathbf{x})^2)$ does as well as the full predictive and may be an adequate approximation to the conditional distribution. In Fernandez et al. (2002) a particular feature of the data (a lump of responses at zero) was noted and a

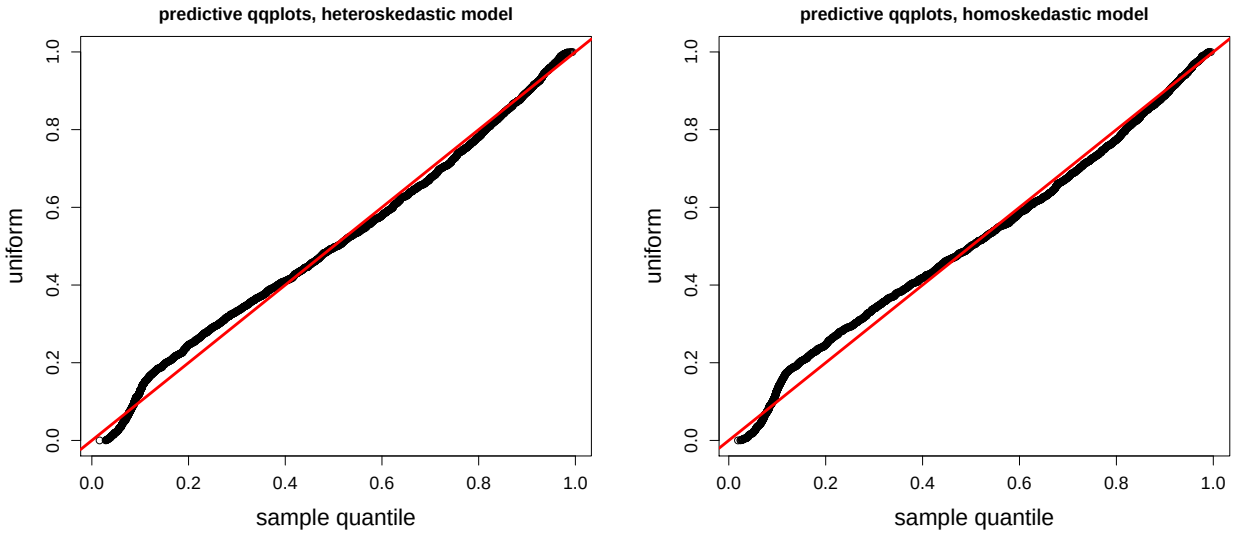

Figure 19: Alcohol example. Predictive qq-plots. Left panel: heteroscedastic BART. Right panel: homoscedastic BART.

model was developed to capture this feature. Our model captures much of the pattern in the data and does so in a way that is both automatic and simply interpretable.

The homoscedastic BART model $Y = f(\mathbf{x}) + \sigma Z,\ Z \sim N(0,1)$ is clearly very restrictive in modeling the error term. Our heteroscedastic ensemble model $Y = f(\mathbf{x}) + s(\mathbf{x}) Z,\ Z \sim N(0,1)$, moves away from the BART model but still assumes an additive error structure with conditionally Normal errors. While we are currently working on ways to relax these assumptions, there are a great many ways in which the model may be elaborated. Our feeling is that, in many applications, by focusing on the first and second moments, the model presented in this paper will capture much of the structure to be found in the data in a relatively simple way.

Acknowledgments

This research was partially supported by the US National Science Foundation grants DMS-1106862, 1106974 and 1107046, the STATMOS research network on Statistical Methods in Oceanic and Atmospheric Sciences. E. I. George acknowledges support from NSF grant DMS-1406563. H. Chipman acknowledges support from the Natural Sciences and Engineering Research Council of Canada.

References

Allen, G., Grosenick, L., and Taylor, J. (2013). "A Generalized Least-Square Matrix Decomposition." *Journal of the American Statistical Association*, 109, 145–159.

Bleich, J. and Kapelner, A. (2014). "Bayesian Additive Regression Trees With Parametric Models of Heteroskedasticity." *arXiv:1402.5397v1*, 1–20.

Box, G. E. and Cox, D. R. (1964). "An analysis of transformations." *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

Breiman, L. (2001). "Random Forests." *Machine Learning*, 45, 5–32.

Carroll, R. J. (1988). *Transformation and Weighting in Regression*. Chapman and Hall.

Chipman, H., George, E., and McCulloch, R. (1998). "Bayesian CART Model Search." *Journal of the American Statistical Association*, 93, 443, 935–960.

— (2002). "Bayesian Treed Models." *Machine Learning*, 48, 299–320.

— (2010). "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, 4, 1, 266–298.

Cook, R. D. (2007). "Fisher Lecture: Dimension Reduction in Regresssion." *Statistical Science*, 22, 1–26.

Daye, Z. J., Chen, J., and Li, H. (2012). "High-Dimensional Heteroscedastic Regression with an Application to eQTL Data Analysis." *Biometrics*, 68, 316–326.

Denison, D., Mallick, B., and Smith, A. (1998). "A Bayesian CART Algorithm." *Biometrika*, 85, 2, 363–377.

Fernandez, C., Ley, E., and Steel, M. (2002). "Bayesian Modelling of Catch in a North-West Atlantic Fishery." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 51, 3, 257–280.

Freidman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 28(5), 1189–1232.

Freund, Y. and Schapire, R. E. (1997). "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of Computer and System Sciences*, 55(1), 119–139.

Gramacy, R. and Lee, H. (2008). "Bayesian Treed Gaussian Process Models with an Application to Computer Modeling." *Journal of the American Statistical Association*, 103, 483, 1119–1130.

Kenkel, D. and Terza, J. (2001). "The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect." *Journal of Applied Econometrics*, 16, 165–184.

Pratola, M. T. (2016). "Efficient Metropolis-Hastings Proposal Mechanisms for Bayesian Regression Tree Models." *Bayesian Analysis*, 11, 885–911.

Sang, H. and Huang, J. (2012). "A full scale approximation of covariance functions for large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 111–132.

Székely, G. J. and Rizzo, M. L. (2004). "Testing for equal distributions in high dimension." *InterStat*, 5, 1–6.

Taddy, M., Gramacy, R., and Polson, N. (2011). "Dynamic Trees for Learning and Design." *Journal of the American Statistical Association*, 106, 493, 109–123.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society B*, 58, 267–288.

Yeo, I. K. and Johnson, R. A. (2000). "A new family of power transformations to improve normality or symmetry." *Biometrika*, 87, 954–959.