



1-2000


## Hierarchical Priors for Bayesian CART Shrinkage

Hugh A. Chipman

Edward I. George  
*University of Pennsylvania*

Robert E. McCulloch

Follow this and additional works at: [https://repository.upenn.edu/statistics\\_papers](https://repository.upenn.edu/statistics_papers)

 Part of the [Other Physical Sciences and Mathematics Commons](#), and the [Other Statistics and Probability Commons](#)

---

### Recommended Citation

Chipman, H. A., George, E. I., & McCulloch, R. E. (2000). Hierarchical Priors for Bayesian CART Shrinkage. *Statistics and Computing*, 10 (1), 17-24. <http://dx.doi.org/10.1023/A:1008980332240>

This paper is posted at ScholarlyCommons. [https://repository.upenn.edu/statistics\\_papers/521](https://repository.upenn.edu/statistics_papers/521)  
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).

---

## Hierarchical Priors for Bayesian CART Shrinkage

### Abstract

The Bayesian CART (classification and regression tree) approach proposed by Chipman, George and McCulloch (1998) entails putting a prior distribution on the set of all CART models and then using stochastic search to select a model. The main thrust of this paper is to propose a new class of hierarchical priors which enhance the potential of this Bayesian approach. These priors indicate a preference for smooth local mean structure, resulting in tree models which shrink predictions from adjacent terminal node towards each other. Past methods for tree shrinkage have searched for trees without shrinking, and applied shrinkage to the identified tree only after the search. By using hierarchical priors in the stochastic search, the proposed method searches for shrunk trees that fit well and improves the tree through shrinkage of predictions.

### Keywords

binary trees, tree shrinkage, Markov chain Monte Carlo, model selection, stochastic search, mixture models

### Disciplines

Other Physical Sciences and Mathematics | Other Statistics and Probability

# Hierarchical Priors for Bayesian CART Shrinkage

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

Working Paper 98-03

March 1998

Department of Statistics and Actuarial Science  
University of Waterloo

## ABSTRACT

The Bayesian CART (classification and regression tree) approach proposed by Chipman, George and McCulloch (1998) entails putting a prior distribution on the set of all CART models and then using stochastic search to select a model. The main thrust of this paper is to propose a new class of hierarchical priors which enhance the potential of this Bayesian approach. These priors indicate a preference for smooth local mean structure, resulting in tree models which shrink predictions from adjacent terminal node towards each other. Past methods for tree shrinkage have searched for trees without shrinking, and applied shrinkage to the identified tree only after the search. By using hierarchical priors in the stochastic search, the proposed method searches for shrunk trees that fit well and afterwards improves the tree through shrinkage of predictions.

Keywords: binary trees, tree shrinkage, Markov chain Monte Carlo, model selection, stochastic search, mixture models.

---

Hugh Chipman is Assistant Professor of Statistics, Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, ON N2L 3G1, hachipman@uwaterloo.ca. Edward I. George is the Ed and Molly Smith Chair in Business Administration and Professor of Statistics, Department of MSIS, University of Texas, Austin, TX 78712-1175, egeorge@mail.utexas.edu, and Robert E. McCulloch is Professor of Statistics, Graduate School of Business, University of Chicago, IL 60637, rem@gsb.uchicago.edu. This work was supported by NSF grant DMS 94.04408, Texas ARP grant 003658130, and research funding from the Faculty of Mathematics at the University of Waterloo and the Graduate Schools of Business at the University of Chicago and the University of Texas at Austin.

# 1 Introduction

Consider the setup where we observe  $y$  and a vector  $x = (x_1, \dots, x_p)$  of potential predictors of  $y$ . CART (classification and regression tree) models use binary trees to recursively partition the predictor space into regions within which the distribution of  $y$  is homogeneous. In contrast to standard approaches (e.g. Breiman, Friedman, Olshen and Stone (1984), Quinlan (1986), and Clark and Pregibon (1992)) which use greedy algorithms to select partitioning trees, Chipman, George and McCulloch (1998) (hereafter denoted CGM) and Denison, Mallick, and Smith (1998) proposed a Bayesian approach which puts a prior distribution on the set of all CART models and then uses stochastic search to select a model. The main thrust of this paper is to propose a new class of hierarchical priors which enhance the potential of this Bayesian approach. These priors result in shrunk trees, in which the posterior means in terminal nodes are shrunk towards each other. Shrinkage is greatest between terminal nodes that share many of the same common parents (i.e. are close to each other). An important difference from past work in tree shrinkage (Hastie and Pregibon 1990, Leblanc and Tibshirani 1996) is that shrinkage is integrated into the tree search, rather than applied to a tree after it has been found by tree search algorithms that do not take shrinkage estimation into account.

In Section 2, we review the Bayesian CART approach proposed by CGM. We propose a new class of hierarchical priors in Section 3, and discuss hyperparameter selection in Section 4. In Section 5 we present two simulated examples which illustrate the manner in which prior parameters influence model search and shrinkage, and make comparisons with the tree shrinkage methods of Hastie and Pregibon (1990).

## 2 Bayesian CART

We begin our discussion of prior selection for CART models with a description of the model space. A CART model has two main components: a binary tree  $T$  with  $b$  terminal nodes, and a parameter  $\Theta = (\theta_1, \dots, \theta_b)$  which associates the parameter value  $\theta_i$  with the  $i^{\text{th}}$  terminal node. The tree  $T$  assigns each value of  $y$  to a distinct terminal node. Starting at the root node of  $T$ , this is done by successively assigning  $y$  to left or right child nodes according to prechosen splitting rules of the form  $\{x \in A\}$  or  $\{x \notin A\}$ . For a given tree  $T$ , we indicate the assignment of  $y$  to a terminal node by letting  $y_{ij}$  denote the  $j^{\text{th}}$  observation of  $y$  assigned to the  $i^{\text{th}}$  terminal node,  $i = 1, 2, \dots, b$ ,  $j = 1, 2, \dots, n_i$ .

The parameter value  $\theta_i$  at the  $i^{\text{th}}$  terminal node then determines the distribution  $f(y_{ij}|\theta_i)$  of  $y_{ij}$  where  $f$  is a parametric family indexed by  $\theta_i$ . For example, CGM consider the CART model for which  $f(y_{ij}|\theta_i)$  is normal with  $\theta_i = (\mu_i, \sigma)$ . Assuming that, conditionally on  $(\Theta, T)$ , all  $y_{ij}$  values are independent, this model can be expressed as

$$y_{i1}, \dots, y_{in_i} | \theta_i \text{ iid} \sim N(\mu_i, \sigma^2), \quad i = 1, \dots, b. \quad (1)$$

A CART model is called a regression tree model or a classification tree model according to whether  $y_{ij}$  is quantitative or qualitative, respectively. In this paper, we will focus exclusively on regression tree models of the form (1) and will report elsewhere on hierarchical priors for other CART models.

Since a CART model is identified by  $(\Theta, T)$ , a Bayesian analysis of the problem proceeds by specifying a prior probability distribution  $p(\Theta, T) = p(\Theta | T)p(T)$ . For CART models in general, CGM propose specification of  $p(T)$  by a tree-generating stochastic process which “grows” trees from a single root tree by randomly “splitting” terminal nodes. In particular, they recommend using a prior  $p(T)$  which is implicitly determined by the following recursively defined process.

1. Begin by setting  $T$  to be the trivial tree consisting of a single root (and terminal) node denoted  $\eta$ .
2. Split the terminal node  $\eta$  with probability  $p_\eta = \alpha(1 + d_\eta)^{-\beta}$  where  $d_\eta$  is the depth of the node  $\eta$ , and  $\alpha \in (0, 1)$  and  $\beta \geq 0$  are prechosen control parameters.
3. If the node splits, randomly assign it a splitting rule as follows: First choose  $x_i$  uniformly from the set of available predictors. If  $x_i$  is quantitative, assign a splitting rule of the form  $\{x_i \leq s\}$  vs  $\{x_i > s\}$  where  $s$  is chosen uniformly from the available observed values of  $x_i$ . If  $x_i$  is qualitative, assign a splitting rule of the form  $\{x_i \in C\}$  vs  $\{x_i \notin C\}$  where  $C$  is chosen uniformly from the set of subsets of available categories of  $x_i$ . Next assign left and right children nodes to the split node, and apply steps 2 and 3 to the newly created tree with  $\eta$  equal to the new left and the right children (if nontrivial splitting rules are available).

Turning to  $p(\Theta | T)$ , CGM recommend the standard conjugate form for the normal regression tree model (1), namely

$$\mu_1, \dots, \mu_b | \sigma, T \text{ iid} \sim N(\bar{\mu}, \sigma^2/a) \quad (2)$$

and

$$\sigma^2 | T \sim \text{IG}(\nu/2, \nu\lambda/2) \quad (\Leftrightarrow \nu\lambda/\sigma^2 \sim \chi_\nu^2). \quad (3)$$

Note that the use of a conjugate form allows for the analytical simplification

$$\begin{aligned} p(Y | X, T) &= \int p(Y | X, \Theta, T)p(\Theta | T)d\Theta \\ &= \frac{c a^{b/2}}{\prod_{i=1}^b (n_i + a)^{1/2}} \left( \sum_{i=1}^b (s_i + t_i) + \nu\lambda \right)^{-(n+\nu)/2} \end{aligned} \quad (4)$$

where  $c$  is a constant which does not depend on  $T$ ,  $s_i$  is  $(n_i - 1)$  times the sample variance of the  $Y_i$  values,  $t_i = \frac{n_i a}{n_i + a} (\bar{y}_i - \bar{\mu})^2$ , and  $\bar{y}_i$  is the average value in  $Y_i$ . This simplification is particularly useful because it substantially speeds up the stochastic search described below.

Finally, stochastic search for high posterior models under the above setup is performed by using the following Metropolis-Hastings algorithm which simulates a Markov chain  $T^0, T^1, T^2, \dots$  with limiting distribution  $p(T | Y, X)$ . Starting with an initial tree  $T^0$ , iteratively simulate the transitions from  $T^i$  to  $T^{i+1}$  by the two steps:

1. Generate a candidate value  $T^*$  with probability distribution  $q(T^i, T^*)$ .

2. Set  $T^{i+1} = T^*$  with probability

$$\alpha(T^i, T^*) = \min \left\{ \frac{q(T^*, T^i) p(Y | X, T^*) p(T^*)}{q(T^i, T^*) p(Y | X, T^i) p(T^i)}, 1 \right\}. \quad (5)$$

Otherwise, set  $T^{i+1} = T^i$ .

In (5),  $p(Y | X, T)$  is obtained from (4), and  $q(T, T^*)$  is the kernel which generates  $T^*$  from  $T$  by randomly choosing among four steps:

- GROW: Randomly pick a terminal node. Split it into two new ones by randomly assigning it a splitting rule as in step 3 of the prior.
- PRUNE: Randomly pick a parent of two terminal nodes and turn it into a terminal node by collapsing the nodes below it.
- CHANGE: Randomly pick an internal node, and randomly reassign it a splitting rule as in step 3 of the prior.
- SWAP: Randomly pick a parent-child pair which are both internal nodes. Swap their splitting rules unless the other other child has the identical rule. In that case, swap the splitting rule of the parent with that of both children.

CGM note that chains simulated by this algorithm tend to quickly gravitate towards a region where  $P(T | Y, X)$  is large, and then stabilize, moving locally in that region for a long time. Evidently, this is a consequence of a proposal distribution which makes local moves over a sharply peaked multimodal posterior. To avoid wasting time waiting for mode to mode moves, CGM recommend search with continual restarts of the algorithm, saving the most promising trees from each run.

### 3 A Hierarchical Prior for Regression Trees

Although it is easy to describe and implement, the simple independence prior (2) for the terminal node means may not provide enough structure. For example, the independence choice makes the prior on larger sets (large  $b$ ) of  $\theta$  values much more diffuse than the prior on smaller sets (small  $b$ ). This builds into our posterior calculation a preference for smaller trees beyond that expressed in our prior for  $T$ . Also, there is a natural intuition that may lead us to believe that there should be prior dependence in the  $\theta$  values. We may feel that a pair of  $\theta$  values that correspond to regions which are nearby in the predictor space should be more similar than a pair corresponding to regions which are far apart. Put another way, we may want to incorporate local smoothness of the model surface through our prior. This idea of local similarity is developed in a non-Bayesian framework by Hastie and Pregibon (1990) (also Clark and Pregibon 1992), and LeBlanc and Tibshirani (1996).

We now suggest how the tree structure of a CART model provides a natural way to model such prior dependence for normal regression tree model (1). The basic idea is to consider the bottom node means  $\mu_1, \dots, \mu_b$  as arising from a hierarchical Bayesian model based on the tree. To specify this model, we use the following notation which is illustrated

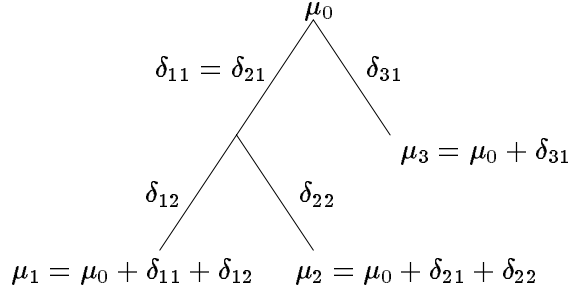


Figure 1: Hierarchical model with mean shifts  $\delta_{ij}$  and terminal node means  $\mu_i$ .

in Figure 1. For the end node with mean  $\mu_i$ , let  $\delta_{i1}, \dots, \delta_{id(i)}$  be sequence of real-valued mean shifts such that

$$\mu_i = \mu_0 + \sum_{j=1}^{d(i)} \delta_{ij}. \quad (6)$$

The idea is that  $\delta_{ij}$  represents the additive contribution of the depth  $j$  node on the tree path leading to  $\mu_i$ . Note that the depth of the final node leading to  $\mu_i$  is  $d(i)$ . Because of the binary tree structure leading to the bottom nodes, many of the mean shift values  $\delta_{ij}$  will be identical. Indeed,  $\delta_{ij} = \delta_{i'j}$  whenever the paths leading to means  $\mu_i$  and  $\mu_{i'}$  share a node at depth  $j$ .

Under the normal mean-shift model (1), a conjugate prior form for this hierarchical model is obtained by putting a zero-mean, normal prior on each of the mean shifts, namely

$$\delta_{ij} \mid \sigma^2, T \sim N(0, \sigma^2 v_{ij}), \quad (7)$$

and assuming that for all  $i, j, i', j'$ ,  $\delta_{ij}$  and  $\delta_{i'j'}$  are independent unless  $\delta_{ij} = \delta_{i'j}$ . The grand mean is also treated as normal

$$\mu_0 \mid \sigma^2, T \sim N(\bar{\mu}, \sigma^2 v_0), \quad (8)$$

independently of the  $\delta_{ij}$ 's. This mean shift prior structure induces a multivariate normal prior on the bottom node means, namely

$$\mu \equiv (\mu_1, \dots, \mu_b)' \mid \sigma^2, T \sim N_b(\bar{\mu} \mathbf{1}, \sigma^2 \Sigma_T). \quad (9)$$

The  $ii$ th diagonal element of  $\Sigma_T$  is  $v_0 + \sum_{j=1}^{d(i)} v_{ij}$ . The  $ii'$ th off-diagonal element of  $\Sigma_T$  is  $v_0 + \sum_{j=1}^{d(i,i')} v_{ij}$ , where  $d(i, i')$  is the largest value of  $j$  for which  $\delta_{ij} = \delta_{i'j}$ . To complete the prior specification, the bottom node variance prior is assumed inverse gamma

$$\sigma^2 \mid T \sim \text{IG}(\nu/2, \nu\lambda/2) \quad (10)$$

as in (3), independently of all the other parameters.

Analytical elimination of  $\mu$  and  $\sigma$  from

$$p(\mu, \sigma, T \mid Y) \propto p(Y \mid \mu, \sigma, T) p(\mu \mid \sigma, T) p(\sigma \mid T) p(T) \quad (11)$$

for this hierarchical model is feasible. Integrating out  $\mu$  yields

$$p(\sigma, T | Y) \propto \sigma^{-(n+\nu+1)} |\Sigma_T|^{-1/2} |N_n + \Sigma_T^{-1}|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} (\nu\lambda + S_y^2) \right\} p(T), \quad (12)$$

where

$$S_y^2 = Y'Y + \bar{\mu}^2 1' \Sigma_T^{-1} 1 - (N_n \bar{y} + \Sigma_T^{-1} 1 \bar{\mu})' (N_n + \Sigma_T^{-1})^{-1} (N_n \bar{y} + \Sigma_T^{-1} 1 \bar{\mu}), \quad (13)$$

$\bar{y} = (\bar{y}_1, \dots, \bar{y}_b)$  and  $N_n$  is the diagonal matrix with  $i$ th diagonal element  $n_i$ . Finally, integrating out  $\sigma$  from (12) yields

$$p(Y | X, T) = c |\Sigma_T|^{-1/2} |N_n + \Sigma_T^{-1}|^{-1/2} (\nu\lambda + S_y^2)^{-(n+\nu)/2}, \quad (14)$$

where  $c$  is a constant which does not depend on  $T$ .

Stochastic search of the posterior under this hierarchical prior can now be carried out using the same Metropolis-Hasting algorithm discussed in Section 2, with  $p(Y | X, T)$  from (14) inserted into  $\alpha$  in (5). Because shrinkage of terminal node means is incorporated in (14), the stochastic search will prefer trees with shrunk means that fit well. Note also that once a particular tree  $T$  is selected, the hierarchical Bayes model can be used to directly obtain easily computable shrinkage estimates

$$E(Y | X, T) = (N_n + \Sigma_T^{-1})^{-1} (N_n \bar{y} + \Sigma_T^{-1} 1 \bar{\mu}). \quad (15)$$

Note also that this posterior mean figures prominently in the stochastic search through the posterior evaluation of  $T$  via (13) and (14).

## 4 Hyperparameter Selection

We here consider and recommend choices for the prior hyperparameters for hierarchical Bayesian CART. This entails choosing  $\alpha$  and  $\beta$  for step 2 of the tree generating prior discussed in Section 2,  $\nu$  and  $\lambda$  for the prior (10) on  $\sigma^2$ ,  $\bar{\mu}$  and  $v_0$  for the prior (8) on  $\mu_0$ , and the  $v_{ij}$ 's for the priors (7) on the  $\delta_{ij}$ 's. Because the overall prior complexity can make subjective choices difficult, we recommend some automatic choices based on the observed  $Y$ .

Beginning with  $\alpha$  and  $\beta$  for the tree prior, these hyperparameters determine the splitting probability  $p_\eta = \alpha(1 + d_\eta)^{-\beta}$ , which in turn controls the size and shape of the generated trees. For example, larger  $\alpha$  puts larger probability on larger trees, and larger  $\beta$  puts larger probability on “bushy” trees. CGM recommend selecting  $\alpha$  and  $\beta$  on the basis of sample characteristics under various choices. For example, when the number of potential predictors  $p$  is large, the choices  $(\alpha, \beta) = (.95, .5), (.95, 1), (.95, 1.5)$  yield mean tree sizes of about 7.0, 3.7 and 2.9 respectively.

We next consider the choice of  $\nu$  and  $\lambda$  for the prior (10) on the residual variance  $\sigma^2$ . Because the normal regression tree model (1) explains the variation of  $Y$ ,  $\sigma$  is likely to be smaller than the sample standard deviation of  $Y$ , say  $s^*$ , and larger than a pooled standard deviation estimate, say  $s_*$ , such as might be obtained from a deliberate overfitting of the data by a greedy algorithm. Using these values as guides,  $\nu$  and  $\lambda$  could then be chosen so



that the prior for  $\sigma$  assigns substantial probability to the interval  $(s_*, s^*)$ . Alternatively, if a (possibly very rough) estimate  $\hat{\sigma}$  were available, one might simply set  $\lambda = \hat{\sigma}^2$  and  $\nu = 5$  (or some small number). Note that  $\nu = 5$  corresponds to a prior on  $\sigma^2$  with 1<sup>st</sup> percentile  $0.33\lambda$  and 99<sup>th</sup> percentile  $9.02\lambda$ .

For the prior (8) on the grand mean  $\mu_0$ , a reasonable automatic choice is to center it at  $\bar{\mu} = \bar{y}$ , the sample mean. The variance of this prior,  $\sigma^2 v_0$ , should be large enough to allow  $\mu_0$  to take any value in the observed range of the responses with at least modest probability. If  $\Delta Y$  is the range of the response, this is essentially obtained when  $\pm 2\sqrt{\sigma^2 v_0} = \Delta Y$ . Replacing the unknown  $\sigma^2$  by its prior expectation  $\lambda$ , yields the automatic choice

$$v_0 = \Delta Y^2 / 4\lambda.$$

Finally, an automatic choice of the  $v_{ij}$  values for the mean shift priors (7) is facilitated by imposing the constraint that all mean shifts  $\delta_{ij}$  have the same variance, so that  $v_{ij} \equiv v_1$ . Conditionally on the grand mean  $\mu_0$ , a terminal node mean  $\mu_i$  of depth  $d_i$  will have variance  $d_i \sigma^2 v_1$ . Assuming that the grand mean is near the center of the data, it seems reasonable to require that  $P(|\mu_i - \mu_0| > \Delta Y/2)$  be small. This is then roughly obtained when

$$3\sqrt{d_i \sigma^2 v_1} = \Delta Y/2.$$

If a (possibly very rough) estimate of average depth  $\bar{d}$  were available, a choice for  $v_{ij} \equiv v_1$  is obtained as

$$v_1 = \frac{\Delta Y^2}{36\bar{d}\lambda}.$$

## 5 Two Simulated Examples

### 5.1 A One Dimensional Smooth Function

In this section, we use a continuous response model with one predictor to illustrate the effect of various prior settings. Data are simulated as

$$y_i = 8 \frac{e^{-20+5x_i}}{1 + e^{-20+5x_i}} + 5 \frac{e^{-20+2x_i}}{1 + e^{-20+2x_i}} + 2\epsilon_i, \quad i = 1, \dots, 200,$$

where  $\epsilon_i$  are independent standard normal variates, and the 200  $x$  values are drawn from a continuous uniform distribution on  $(0,15)$ . All  $y_i$  values are centered about 0 (by subtraction of the mean  $\bar{y}$ ) before analysis.

A factorial experiment in three variables (tree size prior, hierarchical/independence prior, degree of shrinkage in hierarchical case) was conducted. The six combinations are given in Table 1. Cases 1 and 2 represent mild hierarchical shrinkage, 3 and 4 more substantial hierarchical shrinkage, and 5 and 6 shrinkage under independence priors. Each of these three groups has two elements corresponding to priors for large trees ( $\beta = 0.1$ ) and small trees ( $\beta = 1.0$ ).

For each of the six cases, the stochastic search algorithm was run with 100 restarts and 2000 steps per start. The model with the largest log likelihood was recorded. and the corresponding fitted values are given in Figure 2. Each row corresponds to one of the six settings mentioned above. The two columns present the posterior mean (left column) and

number	mean prior	tree prior ( $\beta$ )	shrinkage ( $v_1$ )
1	Hierarchical	0.1	0.50
2	Hierarchical	1.0	0.50
3	Hierarchical	0.1	0.05
4	Hierarchical	1.0	0.05
5	Independence	0.1	—
6	Independence	1.0	—

Table 1: Settings for the six different runs of Bayesian CART. For the independence case, there is no shrinkage parameter  $v_1$ .

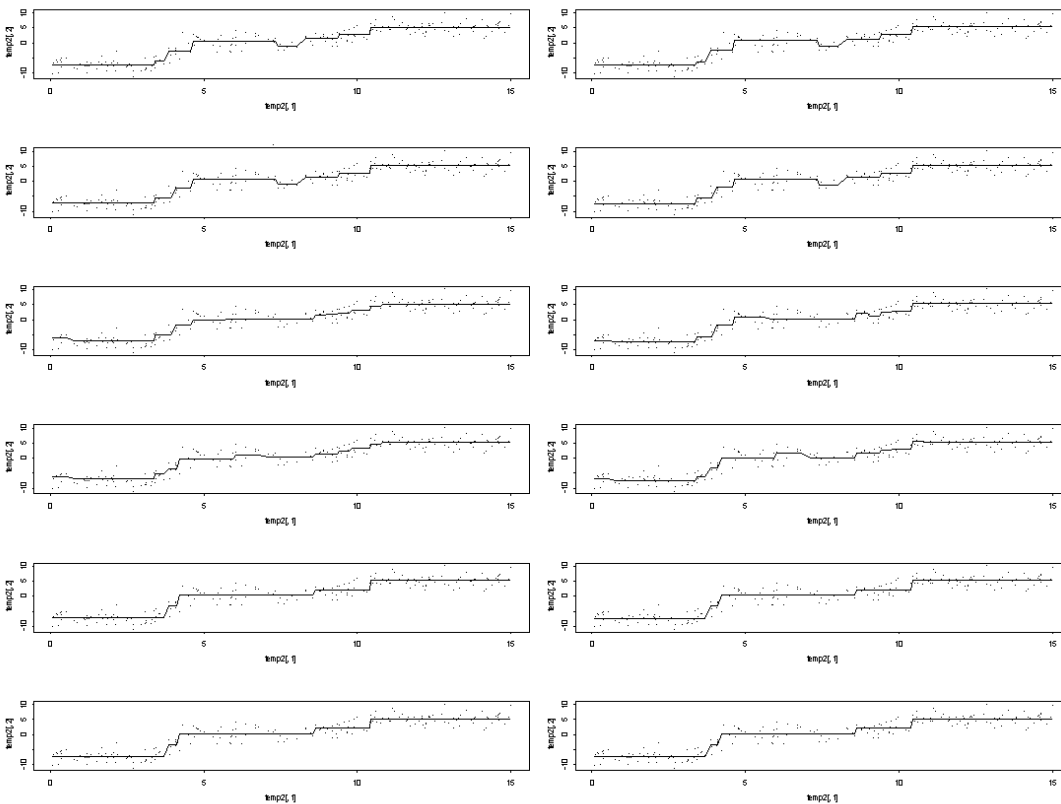


Figure 2: Posterior means (left column) and raw means (right column) resulting from different tree searches, one dimensional example. See text and Table 1 for description of the six different searches.

the “raw” mean (right column), which is the sample mean of the  $y_i$  falling in each partition. We make the following observations (indexes 1-6 refer to rows 1-6 of Figure 2 and Table 1):

1. The effect of the tree prior (1 vs. 2, 3 vs. 4, 5 vs. 6) is reasonably small. Whether a prior on small or large trees was used, roughly the same trees were found (for fixed values of other prior parameters).
2. In comparison to independence priors (5,6), hierarchical priors (1-4) lead the search to identify trees with more terminal nodes and more steps in regions of large curvature.
3. Within the hierarchical priors (1-4), an increase in shrinkage (3,4) produces trees with more terminal nodes.
4. In comparison with raw means (right column), hierarchical shrinkage produces estimates less affected by random noise in the data. For example in the right column of rows 3 and 4, we see that the raw means fluctuate in regions where the true mean is stable. The corresponding regions in the left panel (posterior means under hierarchical shrinkage) are more stable.

## 5.2 A Two Dimensional Smooth Function

For our second example, data will be generated as

$$y = x_{i1}x_{i2} + 1.5\epsilon_i = \mu_i + 1.5\epsilon_i, \quad i = 1, \dots, 400, \quad (16)$$

where  $\epsilon$  is standard normal. The  $(x_1, x_2)$  pairs are on a 20 by 20 grid ranging from -2 to +2 in each variable. 60 realizations of this data were simulated.

The greedy CART algorithm (as implemented in S-Plus) is used to fit a maximal tree (i.e., one in which splitting terminates only when observations are identical or a node contains a single observation). 10-fold cross-validation is then used to determine the degree of shrinkage. A pruned tree is also fit by 10-fold cross-validation. The accuracy of fitted values  $\hat{\mu}_i$  is measured via root mean squared error relative to the known mean  $\mu_i$ :

$$\text{RMSE} = \frac{1}{400} \sum_{i=1}^{400} (\hat{\mu}_i - \mu_i)^2$$

The Bayesian CART algorithm was used to estimate the function with four different prior settings. Three hierarchical shrinkage priors were used with varying degrees of shrinkage on the terminal node means. The fourth prior was an independence prior in which the terminal node means were assumed independent of each other.

The automatic choices of prior parameters discussed in Section 4 were used. In particular, the quantity  $\Delta Y$  was taken to be the distance between the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the  $y$ 's, and  $\hat{\sigma}$  was the residual standard error of the cross-validated shrunk greedy tree. Both these quantities were calculated for each of the 60 realizations of the data. The use of output from the greedy trees is appropriate for two reasons. First, it calibrates the Bayesian method so that it should perform in a similar fashion to the greedy cart for each realized data set. Second, in any real problem we would fit greedy trees as an initial exploratory step before using Bayesian CART. Parameters for tree size were fixed at  $\alpha = .95, \beta = 1.0$ ,

method	GS	GP	BS1	BS1raw	BS2	BS2raw	BS3	BS3raw	BI	BIraw
mean	0.776	0.775	0.633	0.673	0.612	0.666	0.606	0.634	0.717	0.731
std. err.	0.011	0.013	0.006	0.007	0.005	0.006	0.006	0.007	0.007	0.008

Table 2: Average root mean squared error over 60 simulations. The standard errors are for the averages. Methods are abbreviated as follows: GS=greedy shrunk, GP=greedy prune BS=Bayes shrunk, with 1 having least shrinkage and 3 the most, BI = Bayes independent, raw= $\mu$  calculated without shrinkage.

giving a prior expected number of nodes of roughly four. Although this prior is tight, our experience in the last example illustrates that the prior on tree size has minimal influence compared to the prior on the mean shifts  $\delta_{ij}$ .

For comparison, Bayesian CART was also run with an independence prior on the terminal node means. The prior on means was taken to be  $N(\bar{y}, 4\sigma^2/\Delta Y)$ , with  $\sigma = \hat{\sigma}$ . This choice is comparable to the hierarchical prior.

For Bayesian CART, 20 restarts and 4000 steps per start were used. Preliminary experimentation with shorter runs indicated that even very short runs (eg 2 restarts and 200 steps per start) produced trees with predictive error on par with shrunk greedy trees. Since increasing the computational time produced substantial gains in RMSE, we report results for these longer runs here.

Table 2 gives the average RMSE, and standard errors of these estimates for the 60 realizations. Each method was applied to the same 60 realizations of data from (16). All Bayesian CART estimates with shrinkage (BS1, BS2, BS3) have better mean performance than shrunk greedy trees. Examination of the raw fitted values (i.e estimates  $\hat{\mu}_i$  are constructed by sample means without shrinkage, but using the trees identified by the hierarchical stochastic search) reveals that while they do not perform as well as the shrunk estimates, they are better than the shrunk or pruned greedy trees. Evidently the hierarchical prior is guiding the search to identify better trees, in addition to producing better estimates via shrinkage. The effect of the hierarchical priors is also evident in comparison with the Bayesian independence tree, which has a larger error. It appears that visiting many trees does not necessarily mean that good trees will be found. The hierarchical prior guides the search to find better trees.

Boxplots of the RMSE values for each method across the 60 realizations are given in Figure 3. Values for the raw means are omitted. Figures 4 and 5 give one realization of the data and corresponding fitted values from selected methods. In the first figure, the tree identified by independence Bayes differs substantially from that found by hierarchical Bayes. The hierarchical tree seems to better capture the saddlepoint nature of the function.

In Figure 5, we see that the shrunk greedy tree has more residual noise than the hierarchical Bayes estimates. Our initial reaction was that the Bayes procedure was superior simply because it was finding smaller trees. We choose however to report these results, since they are based on automatic choices, and the use of  $\hat{\sigma}$  from the shrunk greedy tree as input to our Bayesian procedure means that they should be calibrated. If automatic choices can produce better fits with the hierarchical Bayes estimates, this is an distinct advantage. This is especially the case in more substantial (i.e. higher dimensional) problems, where the extent of over or under fitting may not be evident.

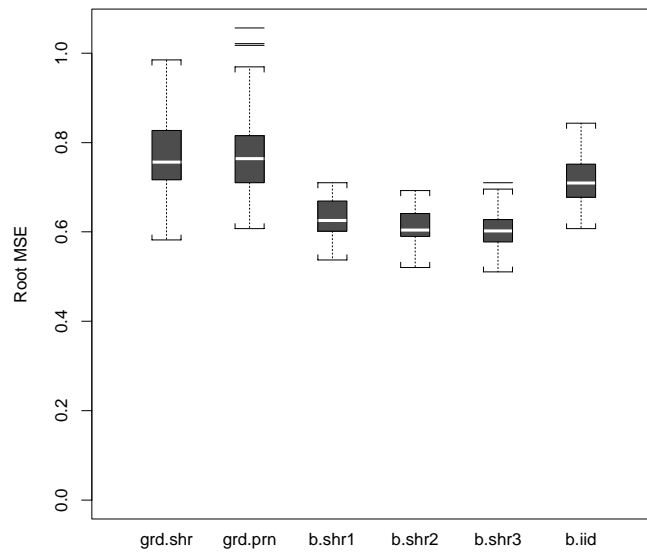


Figure 3: Boxplots of RMSE for 60 realized data sets. The six methods reported are shrunk greedy trees (grd.shr), pruned greedy trees (grd.prn), low, medium, and high shrinkage hierarchical Bayes trees (b.shr 1,2,3), and independence Bayes trees (b.iid).

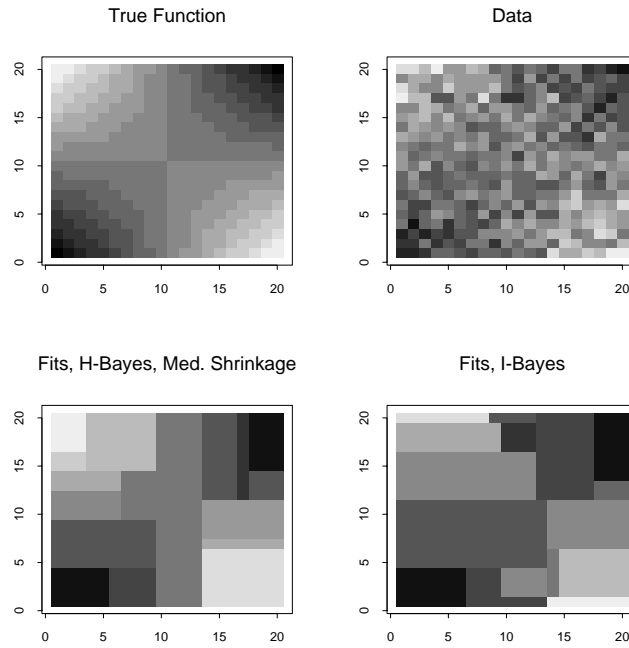


Figure 4: Comparison of true function and one realization of data with estimates found by hierarchical Bayes shrinkage and independence Bayes.

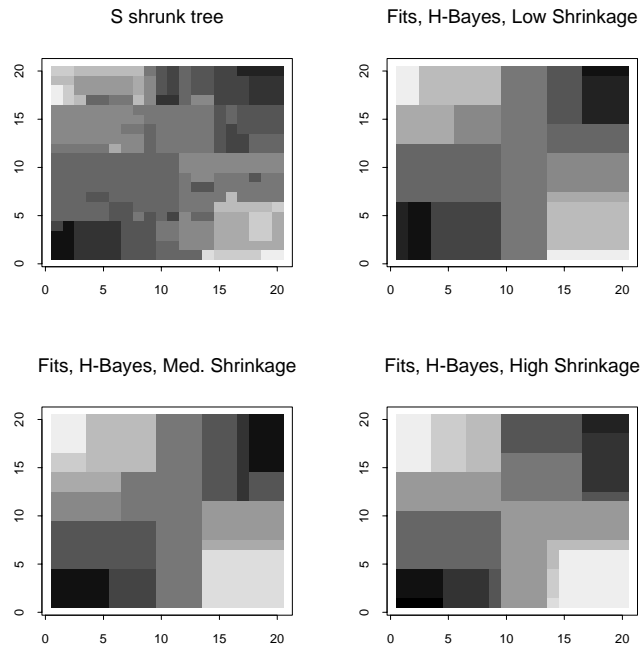


Figure 5: Comparison of estimates found by shrunk greedy tree and three hierarchical Bayes shrinkage methods (low, medium, high shrinkage).

## References

- Breiman, L., Friedman, J. Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, Wadsworth.
- Chipman, H., George, E.I. & McCulloch, R.E. (1998) “Bayesian CART Model Search (with discussion)”, *Journal of the American Statistical Association*, (in press).
- Clark, L., and Pregibon, D. (1992), “Tree-Based Models” in *Statistical models in S*, J. Chambers and T. Hastie, Eds., Wadsworth.
- Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998) “A Bayesian CART Algorithm”, *Biometrika*, to appear.
- Hastie, T., and Pregibon, L. (1990), “Shrinking Trees”, AT&T Bell Laboratories Technical Report.
- LeBlanc, M. and Tibshirani, R. (1996), “Monotone Shrinkage of Trees”, technical report, University of Toronto Department of Statistics.
- Quinlan, J. R. (1986) “Induction of Decision Trees”, *Machine Learning*, 1, 81–106.