Heteroscedastic BART Using Multiplicative Regression Trees

M. T. Pratola , H. A. Chipman , E. I. George , and R. E. McCulloch

1. HeterBART

BART, *Bayesian Additive Regression Trees*
(Chipman, George, and McCulloch (2010))

BART flexibly fits the conditional mean of a response.

HeterBART flexibly fits the conditional mean and the conditional variance.

BART, *Bayesian Additive Regression Trees*

fits the basic model:

$$Y_i = f(x_i) + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2)$$

by,

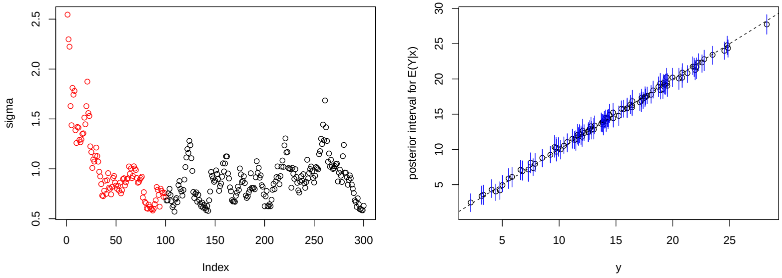
- expressing f as a sum of regression tree.
- putting prior information on each regression tree so that each tree makes a small contribution to the overall fit.
- putting prior information on each regression tree and σ to "regularize" the model so that it does not overfit.
- Drawing from the posterior of the trees, and hence f, using an effective MCMC.

```
##simulate data (example from Friedman MARS paper)
f = function(x){
10*sin(pi*x[,1]*x[,2]) + 20*(x[,3]-.5)^2+10*x[,4]+5*x[,5]
}

sigma = 1.0  #y = f(x) + sigma*z , z~N(0,1)
n = 100      #number of observations
set.seed(99)
x=matrix(runif(n*10),n,10) #10 variables, only first 5 matter
Ey = f(x)
y=Ey+sigma*rnorm(n)

set.seed(99)
bartFit = bart(x,y,ndpost=200) #default is ndpost=1000, this is to run example
plot(bartFit) # plot bart fit
```

At left: MCMC draws of σ_d, where d indexes MCMC draws.



At right:

For each training observation (x_i, y_i) plot the mean of $f_d(x_i)$ (the dot) and a solid blue line from the 2.5% quantile to the 97.5% quantils of $f_d(x_i)$.

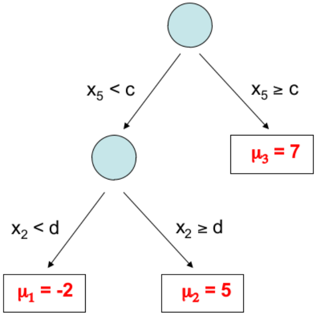
BART gives fit and uncertainty!!!

A single tree model:

Let T denote the tree structure including the decision rules.

Let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denote the set of bottom node μ's.

Let $f(x; T, M)$ be a regression tree function that assigns a μ value to x.



A single tree model:

$$y = f(x; T, M) + \epsilon.$$

The BART model:

$$Y = f(x) + \sigma Z$$

and

$f(x)$ is represented as the sum of *many* single tree models:

$$f(x) = \sum_{i=1}^{m} f(x \mid T_i, M_i)$$

where each (T_i, M_i) represents a single tree model.

m is hundreds , thousands.

The HeterBART model:

$$Y = f(x) + g(x)\, Z$$

$$f(x) = \sum_{i=1}^{m} f(x \mid T_i, M_i)$$

$$g(x) = \prod_{i=1}^{k} g(x \mid \mathcal{T}_i, S_i)$$

Each (T_i, M_i) gives a tree model for a mean.

Each $(\mathcal{T}_i, S_i)$ gives a tree model for a standard deviation.

$Z \sim N(0, 1)$.

$$Y = f(x) + g(x)\,Z$$

$$f(x) = f(x \mid T_1, M_1) + f(x \mid T_2, M_2) + \cdots + f(x \mid T_m, M_m)$$

$=$  $+$  $+ \cdots +$ 

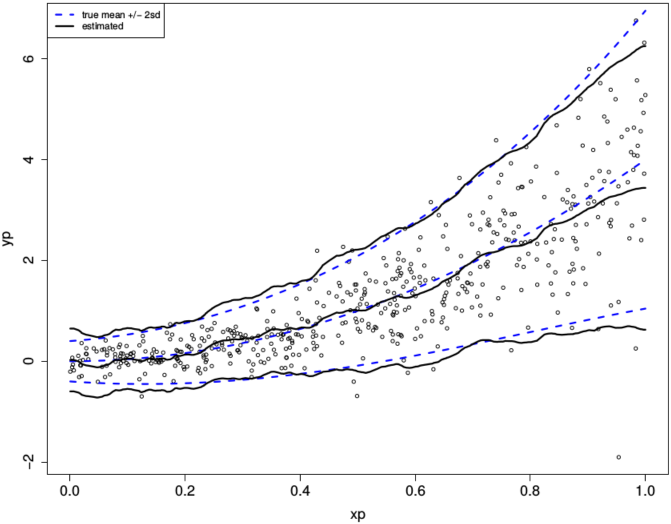
$$= \mu_1 + \mu_2 + \cdots + \mu_m$$

$$g(x) = g(x \mid \tilde{T}_1, S_1) * g(x \mid \tilde{T}_2, S_2) * \cdots * g(x \mid \tilde{T}_K, S_K)$$

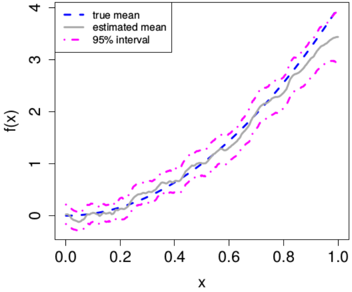
$=$  $*$  $* \cdots *$ 

$$= \sigma_1 * \sigma_2 * \cdots * \sigma_K$$

2. Simulated Example

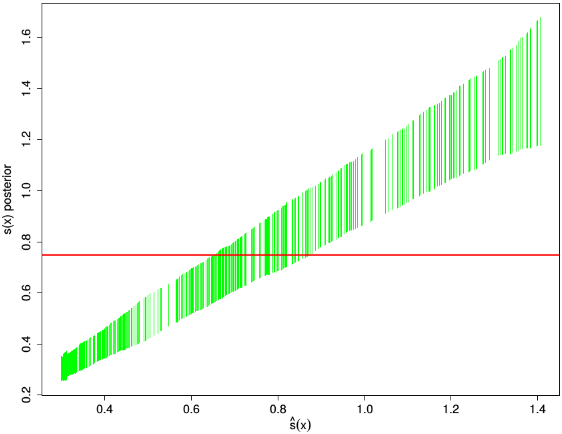


Left: Inference for f.



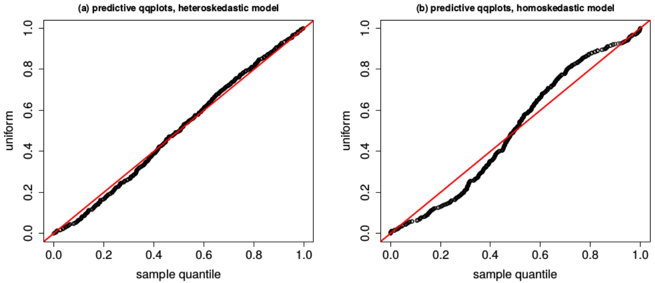
Right: Inference for g.

Given $\{x_i\}$, sort by $\hat{s}(x_i)$ then plot 95% quantile intervals for $s(x_i)$ vs $\hat{s}(x_i)$.



Given training or test (x_i, y_i):

- for each f_d g_d draw, let $\tilde{y}_{id} = f_d(x_i) + g_d(x_i) z_d$, z standard normal.
- for each i compute the percentile of y_i in the draws $\tilde{y}_{id}$.



If the model is right, the percentiles should look like draws from the uniform.

Compare to the uniform using qqplots.

Usually, for numeric responses we check our out-of-sample predictions using RMSE.

That just checks the point prediction.

Our Bayesian model give us a full predictive distribution for

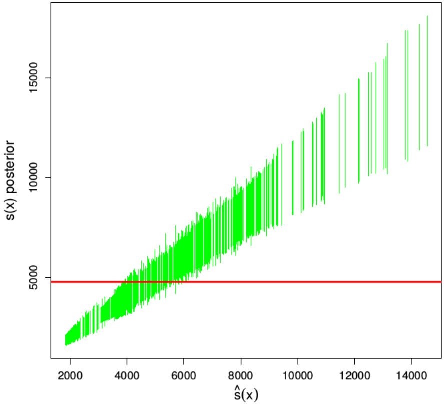
$$Y|x$$

The qqplots allows us to asses the full distributional fit, rather than just the point prediction.

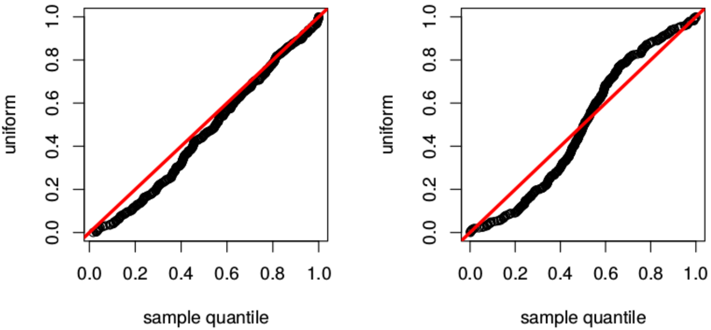
3. Cars Example

Real example, with 15 predictor variables.

So we are "nonparametrically" estimating two functions of 15 variables.



s(x) posterior
5000
10000
15000
2000
4000
6000
8000
10000
12000
14000
ŝ(x)


uniform
0.0
0.2
0.4
0.6
0.8
1.0
sample quantile
uniform
0.0
0.2
0.4
0.6
0.8
1.0
sample quantile