The Remarkable Flexibility of BART

Rob McCulloch
Arizona State University

Collaborations with:
Hugh Chipman (Acadia), Ed George (Wharton)
Matt Pratola (Ohio State), Tom Shively (UT Austin)

University of Arizona
May 24, 2018

Part I. BART (Bayesian Additive Regression Trees)

Data: n observations of y and $x = (x_1, ..., x_p)$

Suppose: $Y = f(x) + \epsilon, \quad \epsilon$ symmetric with mean 0

Bayesian Ensemble Idea: Approximate unknown $f(x)$ by the form

$$f(x) = g(x; \theta_1) + g(x; \theta_2) + ... + g(x; \theta_m)$$

$$\theta_1, \theta_2, \ldots, \theta_m \quad \text{iid} \sim \pi(\theta)$$

and use the posterior of f given y for inference.

BART is obtained when each $g(x; \theta_j)$ is a regression tree.

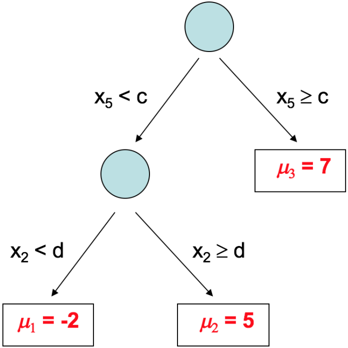
Key calibration: Using y, set $\pi(\theta)$ so that $\text{Var}(f) \approx \text{Var}(y)$.

Beginning with a Single Tree Model

Let T denote the tree structure including the decision rules

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

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



A single tree model:

$$Y = g(x; T, M) + \sigma z, \quad z \sim N(0,1)$$

Bayesian CART: Just add a prior $\pi(M, T)$

Bayesian CART Model Search
(Chipman, George, McCulloch 1998)

$$\pi(M, T) = \pi(M \mid T)\pi(T)$$

$\pi(M \mid T) : (\mu_1, \mu_2, \ldots, \mu_b)' \sim N_b(0, \tau^2 I)$

$\pi(T)$: Stochastic process to generate tree skeleton plus uniform prior on splitting variables and splitting rules.

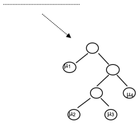
Closed form for $\pi(T \mid y)$ facilitates MCMC stochastic search for promising trees.

Moving on to BART

Bayesian Additive Regression Trees
(Chipman, George, McCulloch 2010)

The BART ensemble model

$$Y = g(x; T_1, M_1) + g(x; T_2, M_2) + \ldots + g(x; T_m, M_m) + \sigma z, \quad z \sim N(0, 1)$$



Each (T_i, M_i) identifies a single tree.

$E(Y \mid x, T_1, M_1, \ldots, T_m, M_m)$ is the sum of m bottom node μ's, one from each tree.

Number of trees m can be much larger than sample size n.

$g(x; T_1, M_1), g(x; T_2, M_2), ..., g(x; T_m, M_m)$ is a highly redundant "over-complete basis" with many many parameters.

Complete the Model with a Regularization Prior

$$\pi(\theta) = \pi((T_1, M_1), (T_2, M_2), \ldots, (T_m, M_m), \sigma)$$

π applies the Bayesian CART prior to each (T_j, M_j) independently so that:

- Each T small.
- Each μ small.
- σ will be compatible with the observed variation of y.

The observed variation of y is used to guide the choice of the hyperparameters for the μ and σ priors.

π is a "boosting prior" as it keeps the contribution of each $g(x; T_i, M_i)$ small, explaining only a small portion of the fit.

Connections to Other Modeling Ideas

$$Y = g(x;T_1,M_1) + \ldots + g(x;T_m,M_m) + \sigma z$$

plus

$$\pi((T_1,M_1),\ldots.(T_m,M_m),\sigma)$$

Bayesian Nonparametrics:

- Lots of parameters (to make model flexible)
- A strong prior to shrink towards simple structure (regularization, boosting)
- BART shrinks towards additive models with some interaction

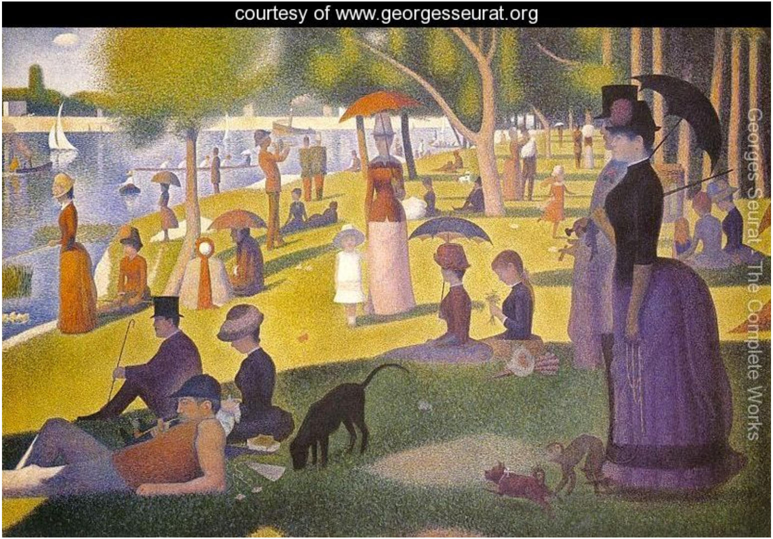
Dynamic Random Basis Elements:

- $g(x;T_1,M_1),\ldots,g(x;T_m,M_m)$ are dimensionally adaptive

Gradient Boosting:

- Fit becomes the cumulative effort of many weak learners

Build up the fit, by adding up tiny bits of fit ..



courtesy of www.georgesseurat.org

A Sketch of the BART MCMC Algorithm

$$Y = g(x;T_1,M_1) + \ldots + g(x;T_m,M_m) + \sigma z$$

plus

$$\pi((T_1,M_1),\ldots.(T_m,M_m),\sigma)$$

First, it is a "simple" Gibbs sampler:

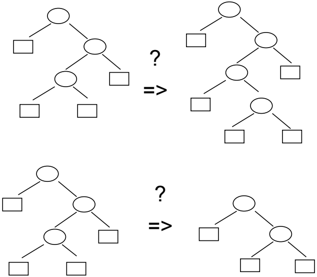
$$\begin{aligned} (T_i, M_i) \quad &| \quad \text{all other } (T_j, M_j), \text{ and } \sigma) \\ \sigma \quad &| \quad (T_1, M_1, \ldots, \ldots, T_m, M_m) \end{aligned}$$

To draw (T_i, M_i) above, subtract the contributions of the other trees from both sides to get a simple one-tree model.

We integrate out M to draw T and then draw $M \mid T$.

For the draw of T we use a Metropolis-Hastings within Gibbs step.

Our proposal moves around tree space by proposing local modifications such as the "birth-death" step:



propose a more complex tree

propose a simpler tree

... as the MCMC runs, each tree in the sum will grow and shrink, swapping fit amongst them ....

Using the MCMC Output to Draw Inference

Each iteration d results in a draw from the posterior of f

$$\hat{f}_d(\cdot) = g(\cdot; T_1, M_1) + \cdots + g(\cdot; T_m, M_m)$$

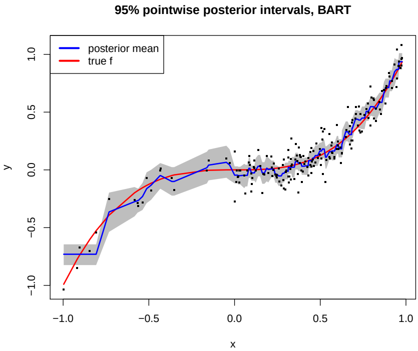
To estimate $f(x)$ we simply average the $\hat{f}_d(\cdot)$ draws at x

Posterior uncertainty is captured by variation of the $\hat{f}_d(x)$
eg, 95% HPD region estimated by middle 95% of values

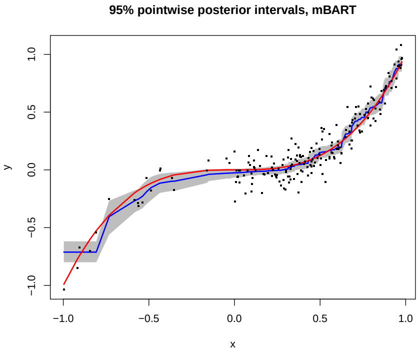
Can do the same with functionals of f.

Automatic Uncertainty Quantification

A simple simulated 1-dimensional example



95% pointwise posterior intervals, BART



95% pointwise posterior intervals, mBART

Note: mBART on the right plot still to be discussed

Example: Friedman's Simulated Data

$$Y = f(x) + \epsilon, \quad \epsilon \sim N(0,1)$$

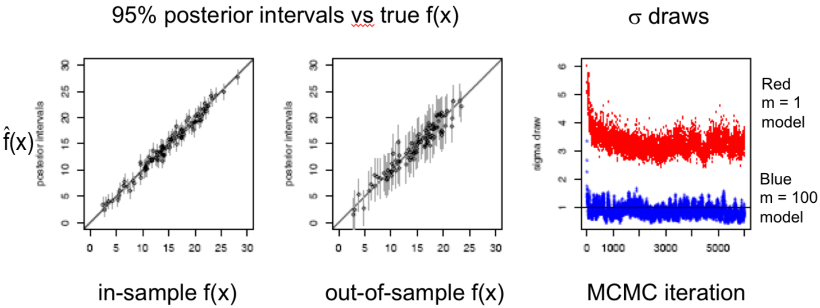
where

$$f(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - .5)2 + 10x_4 + 5x_5 + 0x_6 + \cdots + 0x_{10}$$

- x_i's iid $\sim$ *Uniform*$(0,1)$
- Only the first 5 x_i's matter!
- Friedman (1991) used $n = 100$ observations from this model to illustrate the potential of MARS
- BART handily outperforms competitors including random forests, neural nets and gradient boosting on this example.

Applying BART to the Friedman Data

With $n = 100$ observations and $m = 100$ trees



Detecting Low Dimensional Structure in High Dimensional Data

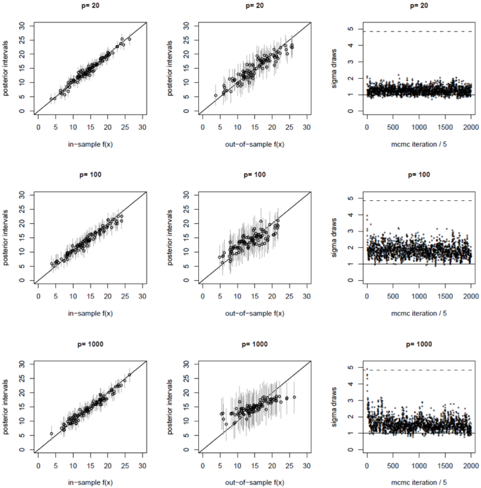
Added many useless x's to Friedman's example

With only 100 observations on y and *1000 x's, BART yielded "reasonable" results !!!!*

In-sample post int vs f(x)

Out-of-sample post int vs f(x)

σ draws



20 x's

100 x's

1000 x's

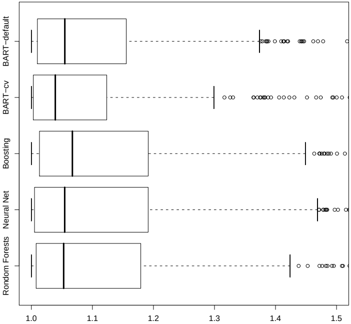
Out of Sample Prediction

Predictive comparisons on 42 data sets.

Data from Kim, Loh, Shih and Chaudhuri (2006) (thanks Wei-Yin Loh!)

- $p = 3$ to 65, $n = 100$ to 7,000.
- for each data set 20 random splits into 5/6 train and 1/6 test
- use 5-fold cross-validation on train to pick hyperparameters (except BART-default!)
- gives 20*42 = 840 **out-of-sample predictions**, for each prediction, divide rmse of different methods by the smallest

+ each boxplots represents 840 predictions for a method
+ 1.2 means you are 20% worse than the best
+ BART-cv best
+ BART-default (use default prior) does amazingly well!!

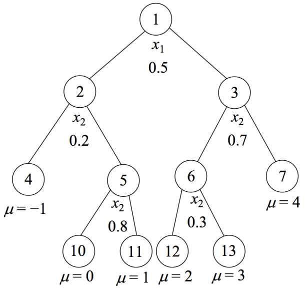


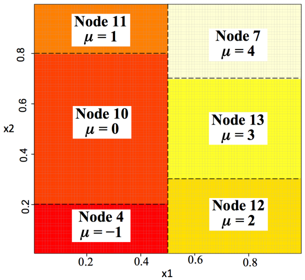
Part II. mBART - Monotone BART

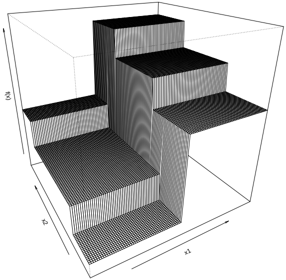
Multidimensional Monotone BART
(Chipman, George, McCulloch, Shively 2018)

Idea:
Approximate multivariate monotone functions by the sum of many single tree models, each of which is monotonic.

An Example of a Monotonic Tree

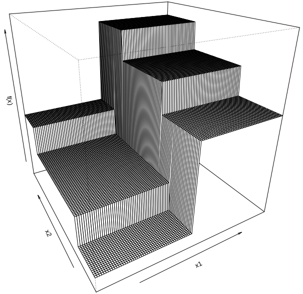






Three different views of a bivariate monotonic tree.

What makes this single tree monotonic?

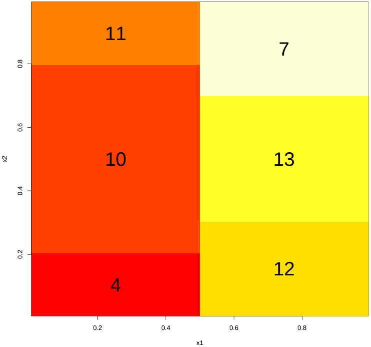


A function g is said to be *monotonic* in x_i if for any $\delta > 0$,

$$
\begin{aligned}
& g(x_1, x_2, \ldots, x_i + \delta, x_{i+1}, \ldots, x_k; T, M) \\
& \quad \geq g(x_1, x_2, \ldots, x_i, x_{i+1}, \ldots, x_k; T, M).
\end{aligned}
$$

For simplicity and wlog, let's restrict attention to monotone nondecreasing functions.

To implement this monotonicity in "tree language" we simply constrain the mean level of a node to be greater than those of it below neighbors and less than those of its above neighbors.



- node 7 is disjoint from node 4.
- node 10 is a below neighbor of node 13.
- node 7 is an above neighbor of node 13.

The mean level of node 13 must be greater than those of 10 and 12 and less than that of node 7.

The mBART Prior

Recall the BART parameter

$$\theta = ((T_1, M_1), (T_2, M_2), \ldots, (T_m, M_m), \sigma)$$

Let $S = \{\theta : \textit{each tree is monotonic in a desired subset of } x_i's\}$

To impose the monotonicity we simply truncate the BART prior $\pi(\theta)$ to the set S

$$\pi^*(\theta) \propto \pi(\theta)\, I_S(\theta)$$

A New BART MCMC "Christmas Tree" Algorithm

$$\pi((T_1, M_1), (T_2, M_2), \ldots, (T_m, M_m), \sigma \mid y))$$

Bayesian backfitting again: Iteratively sample each (T_j, M_j) given (y, σ) and other (T_j, M_j)'s

Each $(T^0, M^0) \to (T^1, M^1)$ update is sampled as follows:

- Denote move as $(T^0, M^0_{Common}, M^0_{Old}) \to (T^1, M^0_{Common}, M^1_{New})$
- Propose T^* via birth, death, etc.
- If M-H with $\pi(T, M \mid y)$ accepts (T^*, M^0_{Common})
    - Set $(T^1, M^1_{Common}) = (T^*, M^0_{Common})$
    - Sample M^1_{New} from $\pi(M_{New} \mid T^1, M^1_{Common}, y)$

Only $M^0_{Old} \to M^1_{New}$ needs to be updated.

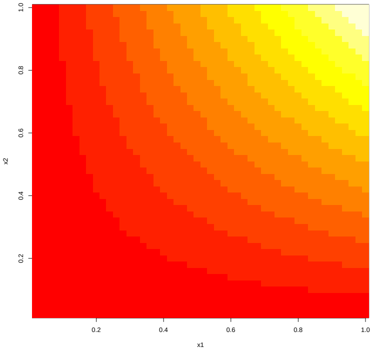
Works for both BART and monotone BART.

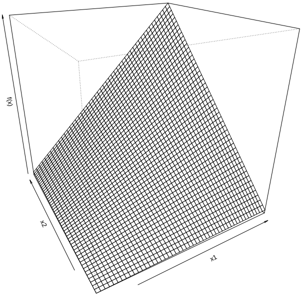
Example: Product of two x's

Let's consider a very simple simulated monotone example:

$$Y = x_1\, x_2 + \epsilon, \quad x_i \sim \text{Uniform}(0, 1).$$

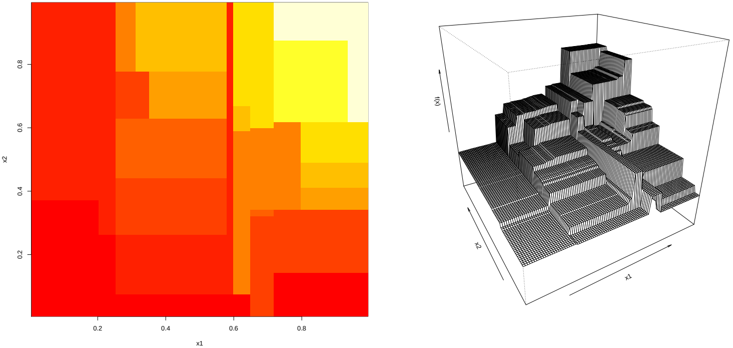
Here is the plot of the true function $f(x_1, x_2) = x_1\, x_2$





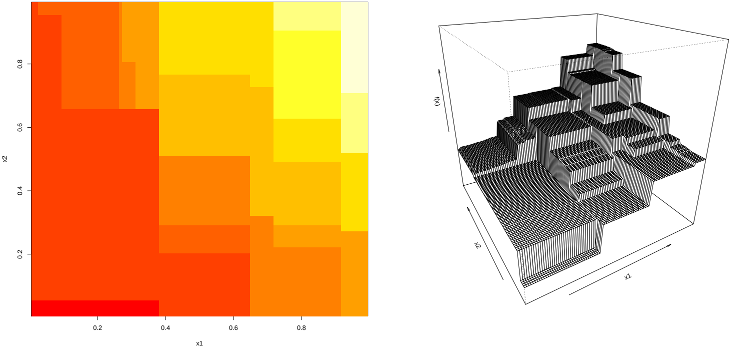
First we try a single (just one tree), unconstrained tree model.

Here is the graph of the fit.



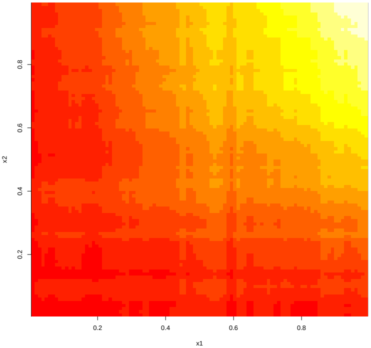
The fit is not terrible, but there are some aspects of the fit which violate monotonicity.

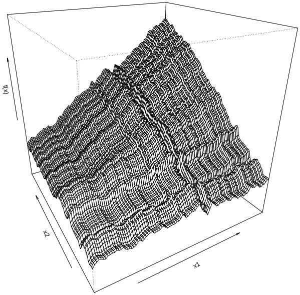
Here is the graph of the fit with the monotone constraint:



We see that our fit is monotonic, and more representative of the true f.

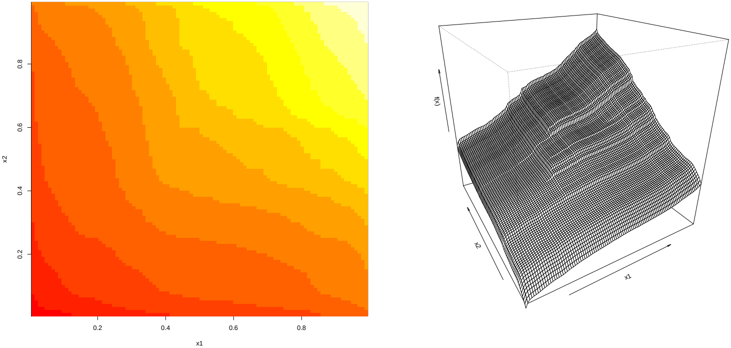
Here is the unconstrained BART fit:





Much better (of course) but not monotone!

And, finally, the constrained BART fit:



Not Bad!

Same method works with any number of x's!

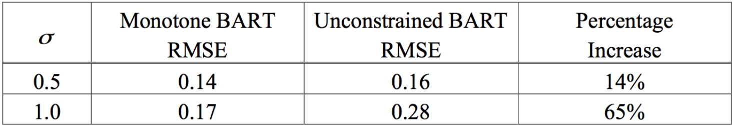
A 5-Dimensional Example

$$Y = x_1 \, x_2^2 + x_3 \, x_4^3 + x_5 + \epsilon,$$

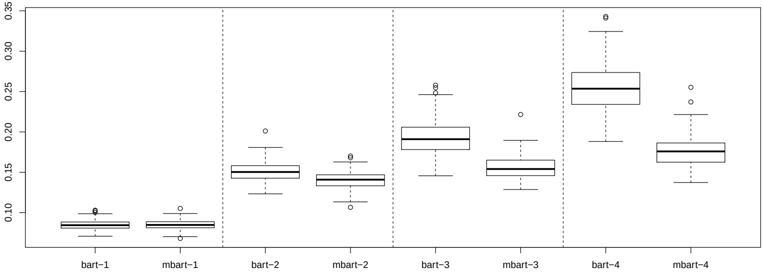
$$\epsilon \sim N(0, \sigma^2), \quad x_i \sim \text{Uniform}(0, 1).$$

We simulated 5,000 observations, with $\sigma = .1$.

RMSE improvement over unconstrained BART



σ	Monotone BART RMSE	Unconstrained BART RMSE	Percentage Increase
0.5	0.14	0.16	14%
1.0	0.17	0.28	65%



$$\sigma = 0.2, 0.5, 0.7, 1.0$$

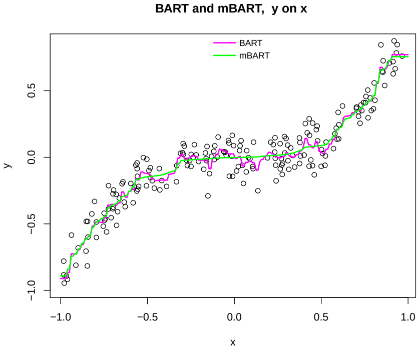
Discovering Monotonicity with mBART

Suppose we are not sure if a function is monotone.

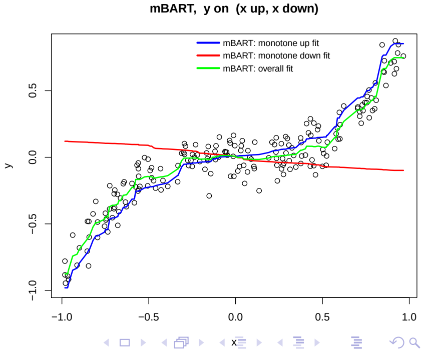
Good news! mBART can be deployed to estimate the monotone components of f.

Thus monotonicity can be discovered rather than imposed!

Example: Suppose $Y = x^3 + \epsilon$.

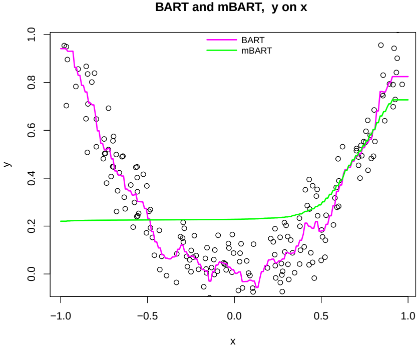


BART and mBART, y on x

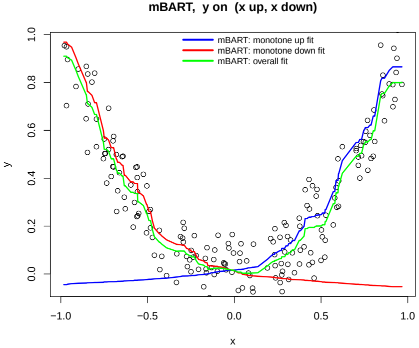


mBART, y on (x up, x down)

Example: Suppose $Y = x^2 + \epsilon$.

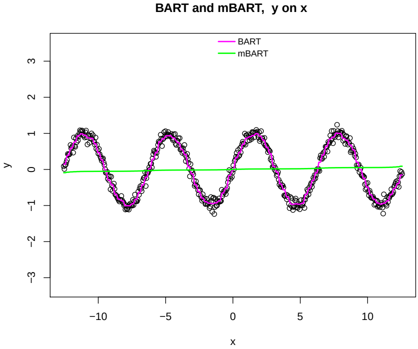


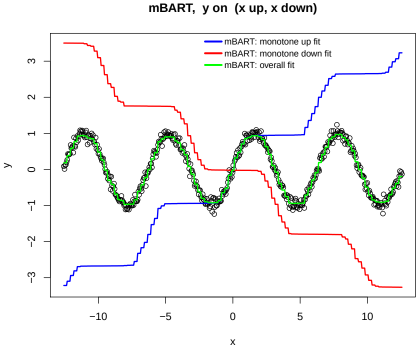
BART and mBART, y on x



mBART, y on (x up, x down)

Example: Suppose $Y = sin(x) + \epsilon$.





Part III. HBART - Heteroscedastic BART

Heteroscedastic BART via Multiplicative Regression Trees
(Pratola, Chipman, George, McCulloch 2018)

- BART flexibly fits the conditional mean with a sum of trees.
- HBART flexibly fits the conditional mean with a sum of trees *and* the conditional variance with a product of trees.

The HBART model:

$$Y = f(x) + s(x)\, Z, \quad Z \sim N(0,1)$$

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

$$s^2(x) = \prod_{i=1}^{m'} h(x; \mathcal{T}_i, S_i)$$

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

Each $(\mathcal{T}_i, S_i)$ identifies a tree model for a variance component.

At each MCMC iteration we have draws of all the

$$(T_i, M_i), \quad i = 1, 2, \ldots, m$$

and

$$(\mathcal{T}_i, S_i), \quad i = 1, 2, \ldots, m'$$

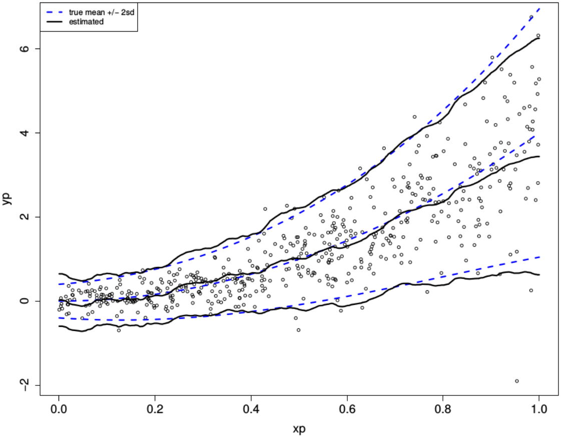
At MCMC iteration d we have a draw f_d of the function f and a draw s_d^2 of the function s^2.

So, for example, at any x, we could use

$$\hat{f}(x) = \frac{1}{D} \sum_{d=1}^{D} f_d(x), \qquad \hat{s}^2(x) = \frac{1}{D} \sum_{d=1}^{D} s_d^2(x)$$

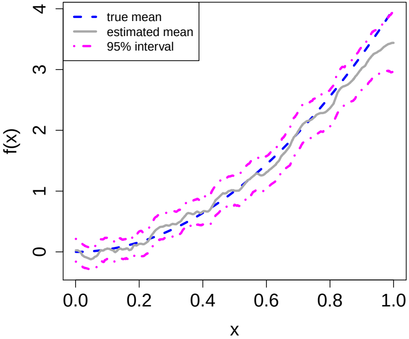
A Simple 1-Dimensional Simulated Example

$\hat{f}(x)$ and $\hat{f}(x) \pm 2\,\hat{s}(x)$

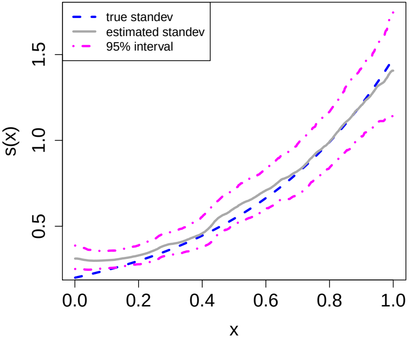


Pointwise intervals.

Inference for f.



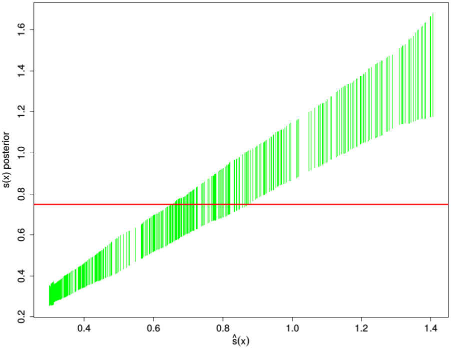
Inference for s.



The previous displays used the fact that x is one-dimensional.

But the next two displays can be used with a vector x of any dimension.

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



The horizontal line is the ordinary BART estimate of σ.

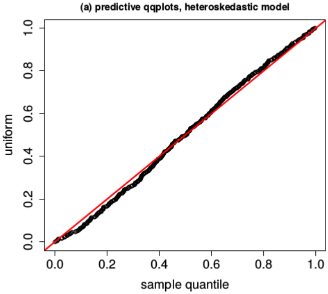
This "H-evidence" plot is useful for gauging the extent of heteroscedasticity.

Given training or test (x_i, y_i):

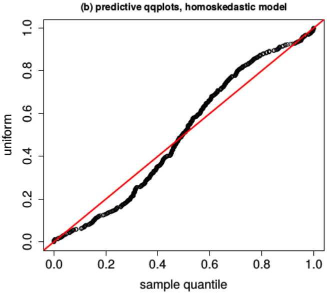
- for each (f_d, s_d) draw, let $\tilde{y}_{id} = f_d(x_i) + s_d(x_i) z_d$, z standard normal.
- for each i compute the quantile of y_i in the draws $\tilde{y}_{id}$.

If the model is right, the quantiles should look like draws from the uniform. This can be gauged with qq-plots.

HBART



BART



For numeric responses, we typically check out-of-sample predictions using RMSE.

However, that just checks the point prediction.

Our Bayesian model give us a full predictive distribution for

$$Y \mid x$$

With the qq-plots, we assess the full distributional fit, rather than just the point prediction.

Used Car Prices Example

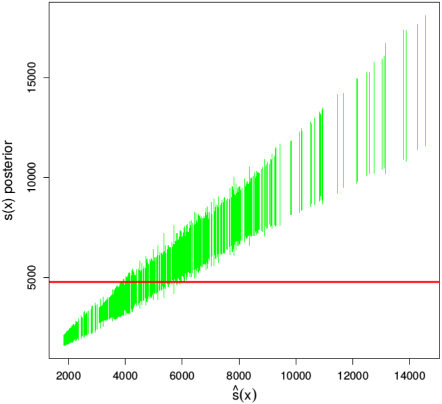
Real example, with 15 predictor variables.

y is the price of a used car

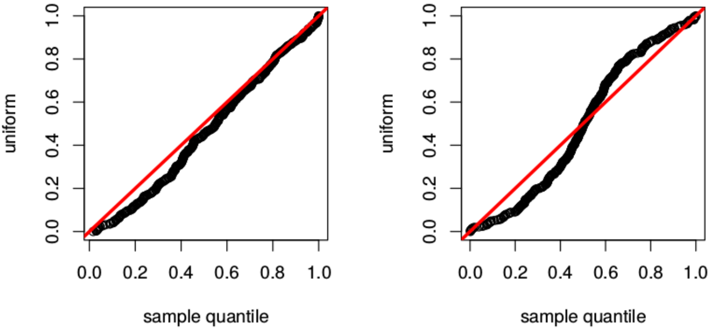
x is characteristics of the car: mileage, year, trim, color, etc.

So we are “nonparametrically” estimating both the mean and the variance of the price, each as a function of 15 variables.

The H-evidence plot reveals substantial heteroscedasticity



The qq-plots show a clear improvement over BART



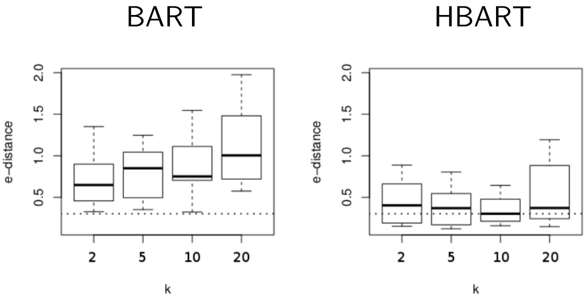
Not perfect, but pretty good !!.

Hyperparameter Selection via e-distance Cross Validation

K is a key prior hyperparameter which determines how smooth the function f is.

Rather than using RMSE cross-validation for choosing K, we use the e-distance measure between the empirical quantiles and the uniform quantiles.

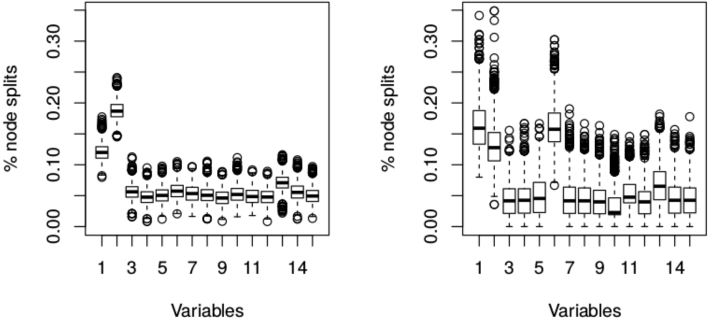
Each boxplot tells us how good the qq-plot looks on a bunch of randomly chosen test data sets.


BART

HBART

Variable selection for both $f(x)$ and $s(x)$:

f at left.
s at right.



$s(x)$ uses `trim.other` in addition to `mileage` and `year`.

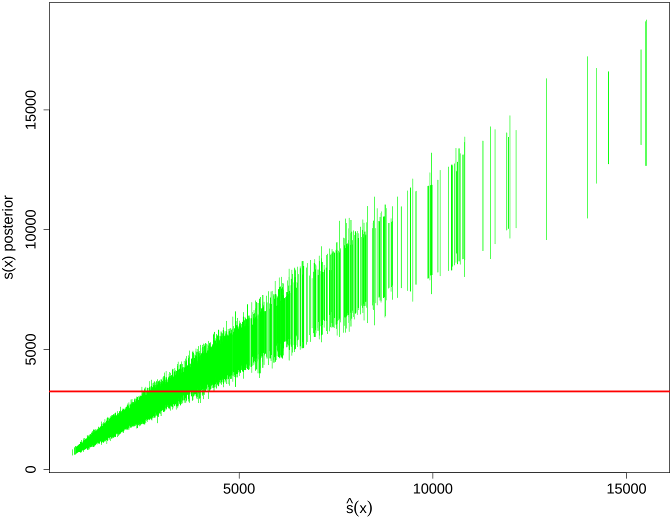
Fish and Alcohol Examples

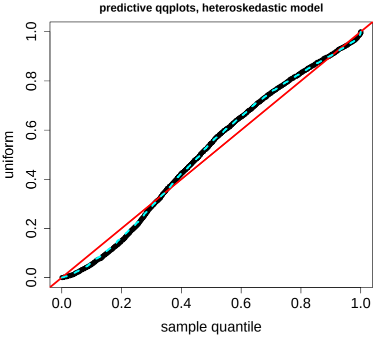
Fish

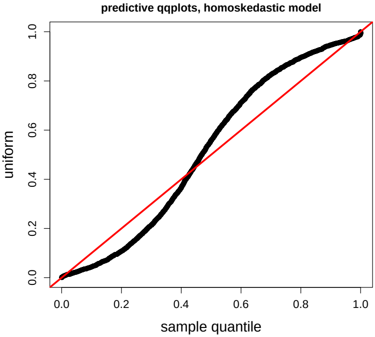
The dependent variable y is the daily catch of fishing boats in the Grand Bank fishing grounds (Fernandez et al., 2002).

The 25 explanatory x variables capture time, location, and characteristics of the boat.

The H-evidence plot reveals substantial heteroscedasticity







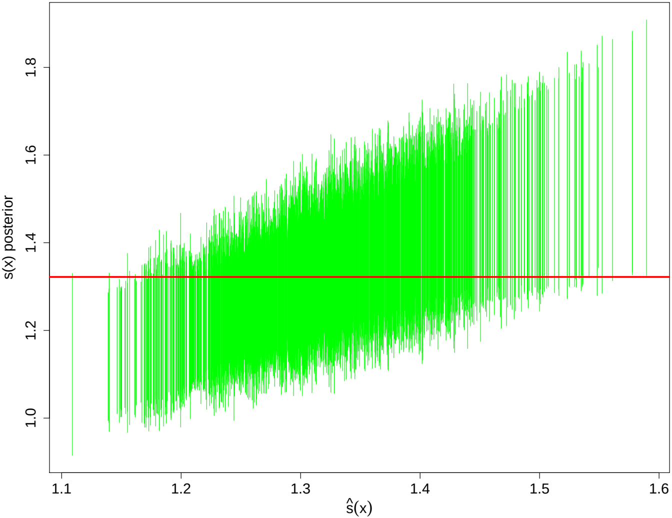
Even though we know $y \geq 0$, *simple HBART is not too bad!!*

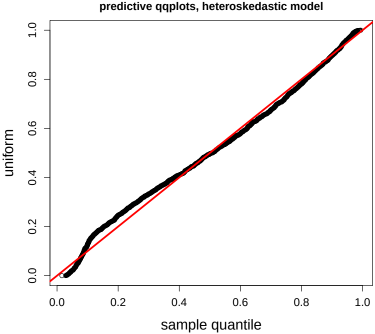
Alcohol

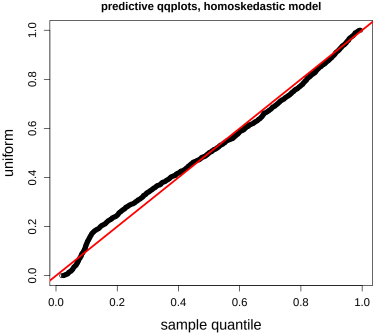
The dependent variable y is the number of alcoholic beverages consumed in the last two weeks. (Kenkel and Terza, 2001).

The 35 explanatory x variables capture demographic and physical characteristics of the respondents as well as a key treatment variable indicating receipt of advice from a physician.

The H-evidence plot does not show heteroscedasticity







Even though we know $y \geq 0$*, ordinary BART is not too bad!!*

Concluding Remarks

- Despite its many compelling successes in practice, theoretical frequentist support for BART is only now just beginning to appear.
- In particular, Rockova and van der Pas (2017) *Posterior Concentration for Bayesian Regression Trees and Their Ensembles* recently obtained the first theoretical results for Bayesian CART and BART, showing near-minimax posterior concentration when $p > n$ for classes of Holder continuous functions.
- Software for BART, mBART and HBART is available at http://www.rob-mcculloch.org/ and on CRAN.

Thank You!

The Prior for heterBART

Key to BART is the simple prior on the bottom node μ parameters.

In the prior, they are iid with

$$\mu \sim N(0, \tau^2).$$

$$f(x) = \sum_{i=1}^{m} \mu_i$$

so that,

$$f(x) \sim N(0, m\tau^2).$$

This makes the prior choice simple and greatly simplifies the MCMC since at a key point we have conditional conjugacy which allows us to integrate out the μ's in a tree analytically.

For the S_i (standard deviations in the bottom nodes of the $\mathcal{T}_i$) we use:

$$\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}, \quad iid.$$

Then

$$s(x) = \prod_i \sigma_i$$

This prior is not as simple as the μ one but by a simple moment-matching strategy, we have a good heuristic for the choice of ν and λ.

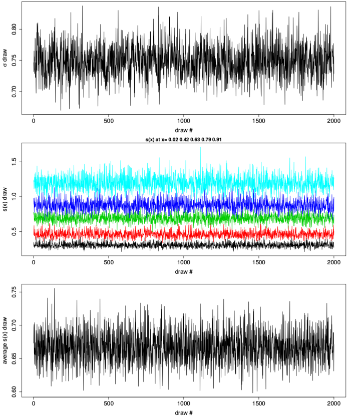
And, we use the same priors for T and $\mathcal{T}$.

And, the simplicity of the BART MCMC is maintained!!

Note:

The MCMC is actually pretty complex as it used Pratola's enhanced MCMC on a single tree which can work better than the original BART proposal moves.

Seems to work pretty good!!



Top: draws of σ in BART.

Midddle: draws of $s(x_i)$ for 5 i in heterBART.

Bottom: draws of average $s(x_i)$ in heterBART.