Heteroscedastic BART Using Multiplicative Regression Trees

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

1. HeterBART
2. Simulated Example
3. Cars Example
4. Fish and Alcohol Examples
5. Prior and MCMC
6. Conclusion

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:

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

by,

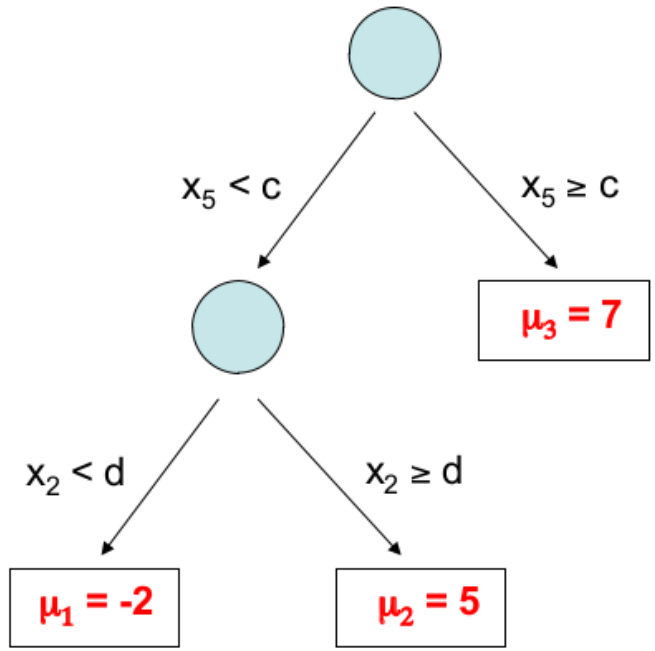
- expressing f as a sum of regression trees (*ensemble modeling*).
- putting prior information on each regression tree so that each tree makes a small contribution to the overall fit (*each tree is a weak learner - as in boosting*).
- putting prior information on each regression tree and σ so that it does not overfit (*use the prior to regularize the fit*).
- Drawing from the posterior of the trees, and hence f, using an effective MCMC.

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:

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

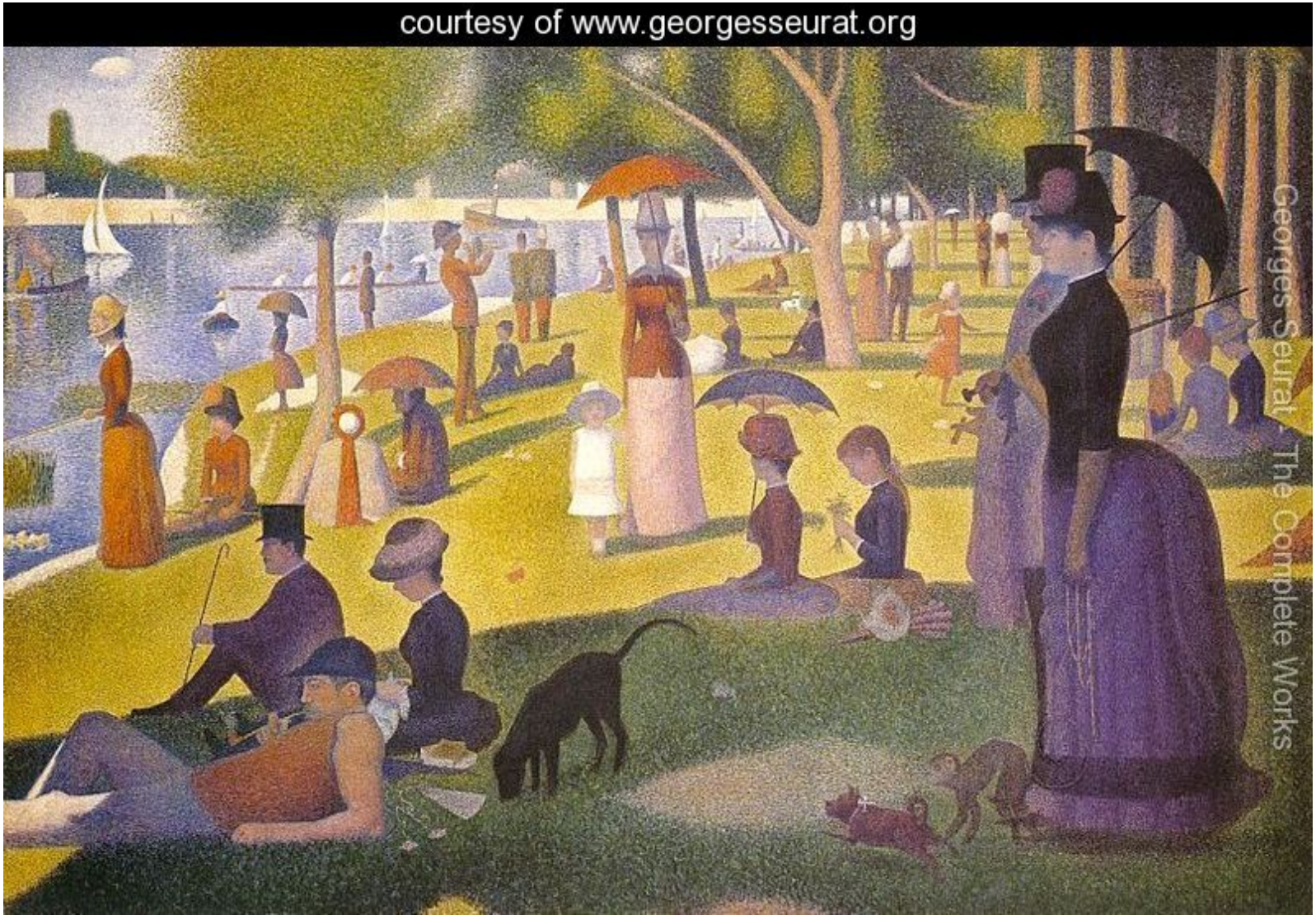
$$f(x) = \sum_{i=1}^{m} f(x; T_i, M_i)$$

where,

- each $f(x; T_i, M_i)$ represents a single tree model.
- m is hundreds, thousands.
- prior pushes each T_i to be a small tree.
- prior shrinks all the μ in all the M_i towards 0.

Note that each T is inferred so that the size of each tree and hence the number of μ is not fixed.

Make the overall fit, the sum of little dabs of fit !!



The HeterBART model:

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

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

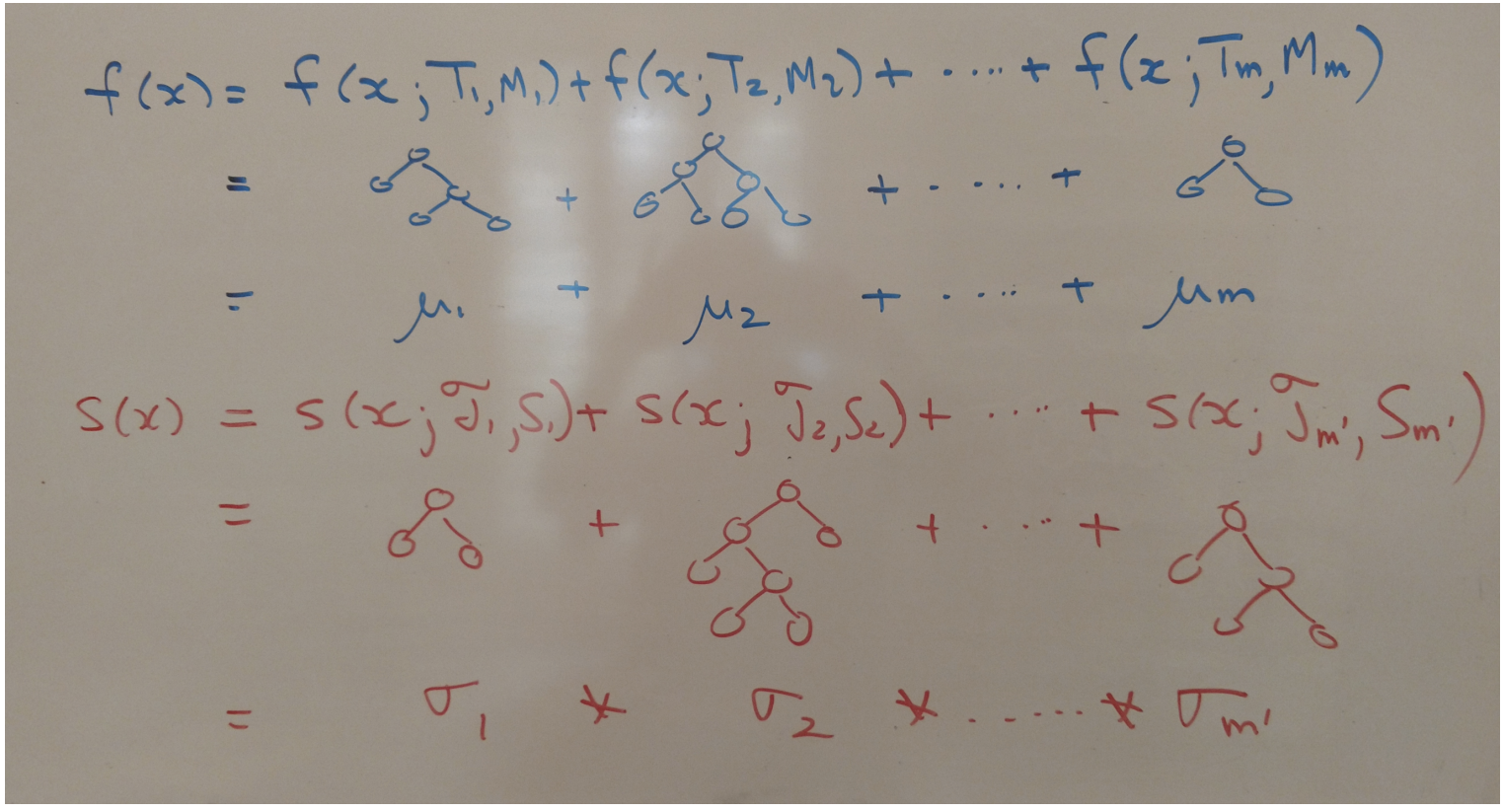
$$s(x) = \prod_{i=1}^{m'} s(x; \mathcal{T}_i, S_i)$$

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.

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

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


f(x) = f(x; T1, M1) + f(x; T2, M2) + ··· + f(x; Tm, Mm)
= + + ··· +
= μ1 + μ2 + ··· + μm
S(x) = S(x; σ J1, S1) + S(x; σ J2, S2) + ··· + S(x; σ Jm', Sm')
= + + ··· +
= σ1 * σ2 * ····· * σm'

At each MCMC iteration we have draws of all the

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

and

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

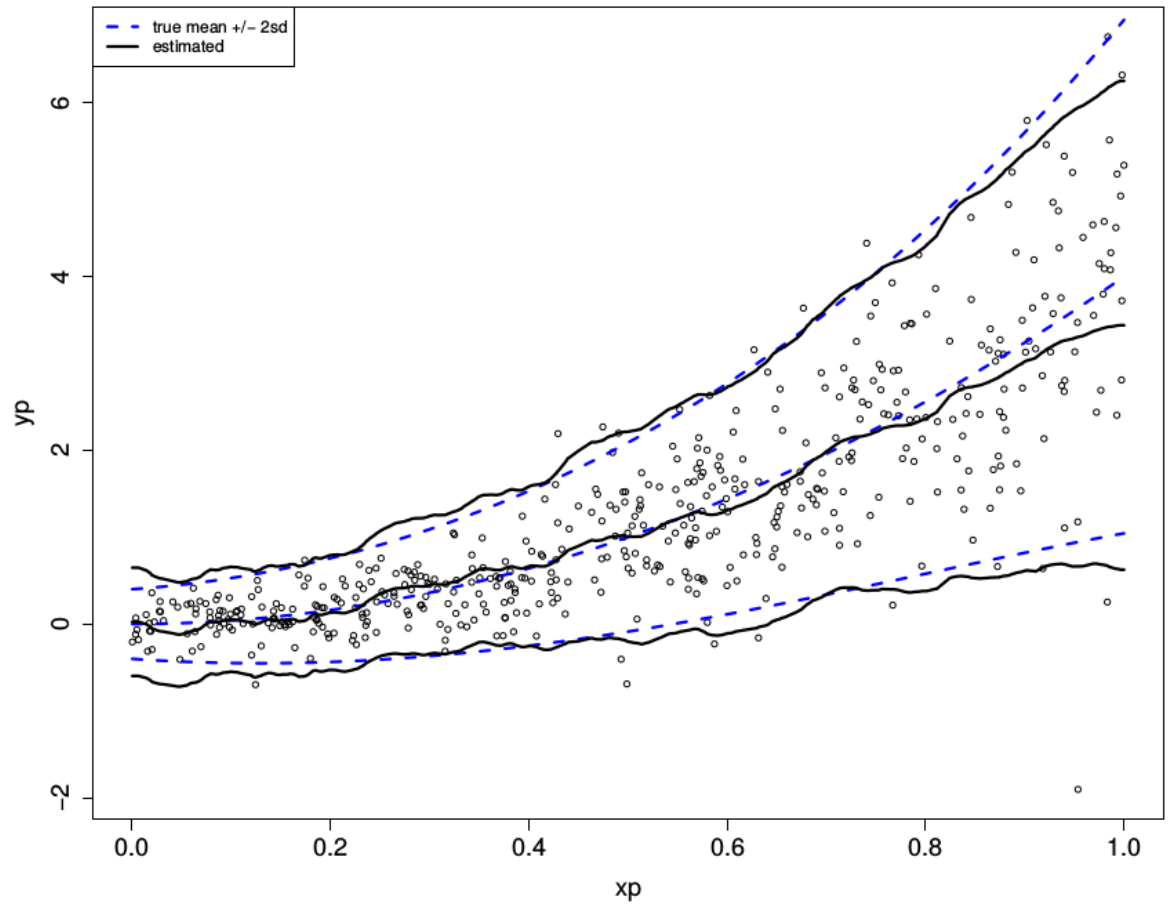
At MCMC iteration d we have a draw f_d of the function f and a draw s_d of the function s.

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

$$\hat{f}(x) = \frac{1}{D} \sum_{d=1}^{D} f_d(x)$$

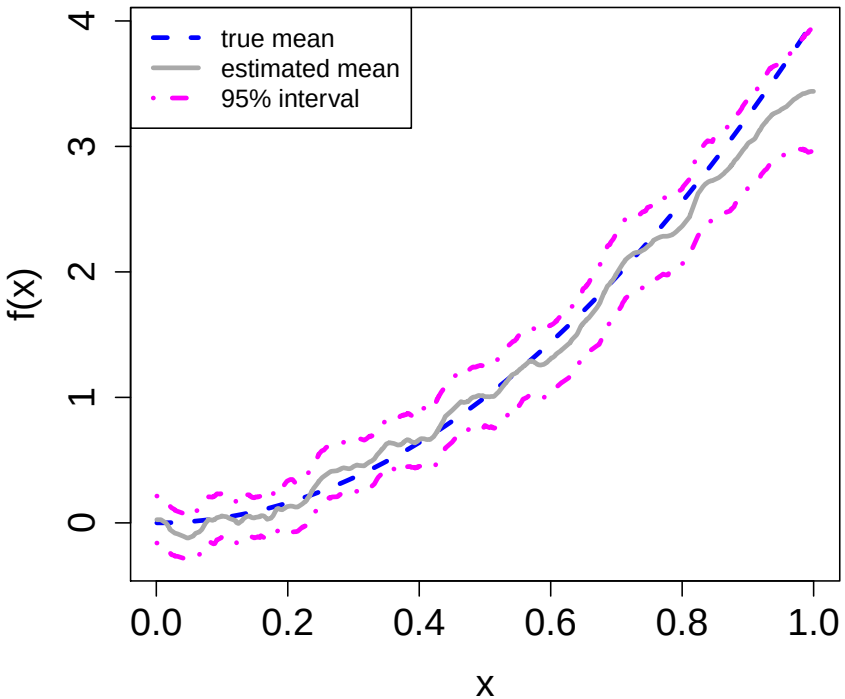
2. Simulated Example

$\hat{f}(x)$ and $\hat{f}(x) \pm 2\,\hat{s}(x)$ where hat is the posterior mean.

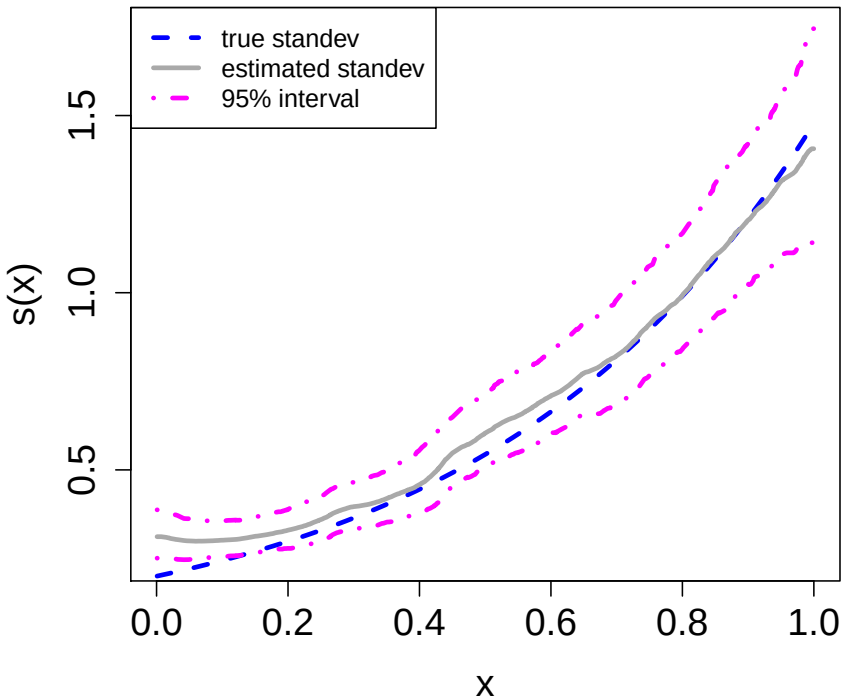


Pointwise intervals.

Inference for f.



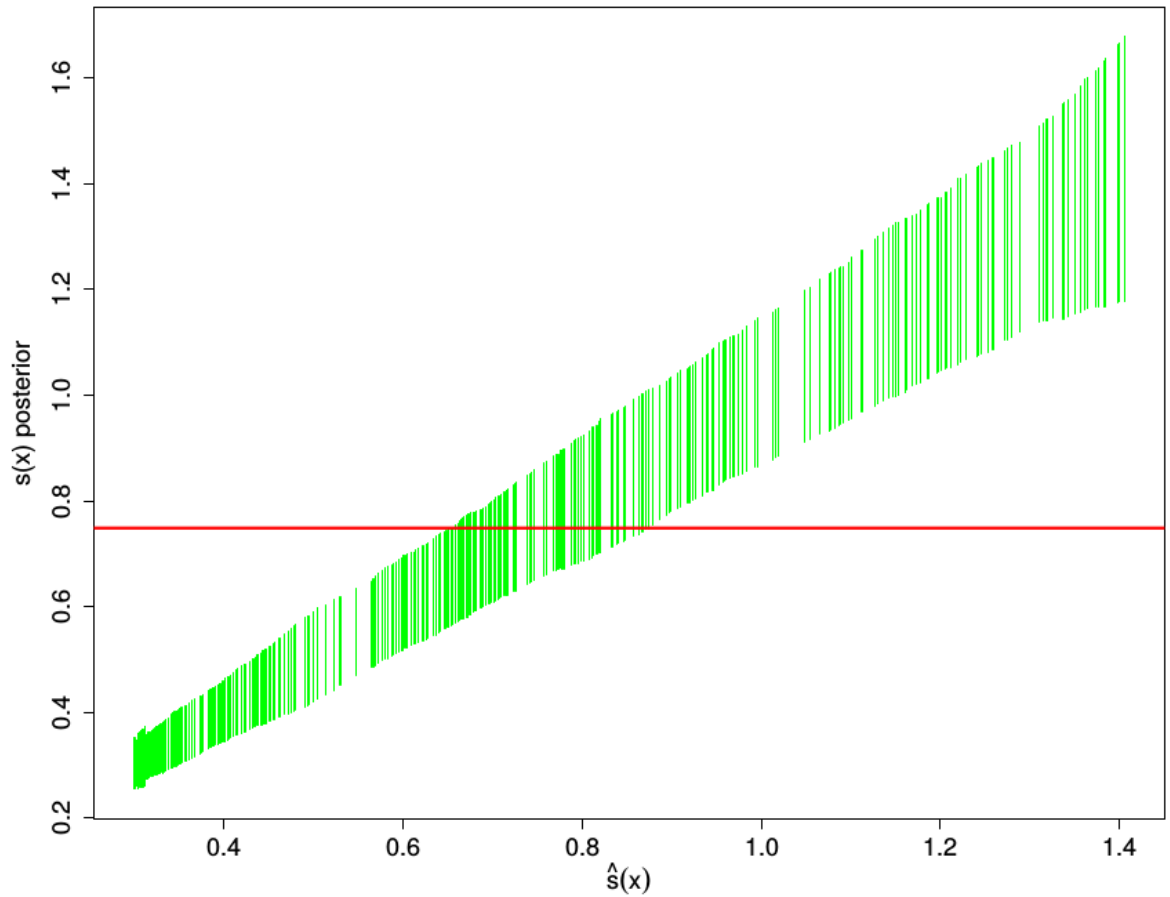
Inference for s.



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

Our 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 $\hat{s}(x_i)$.



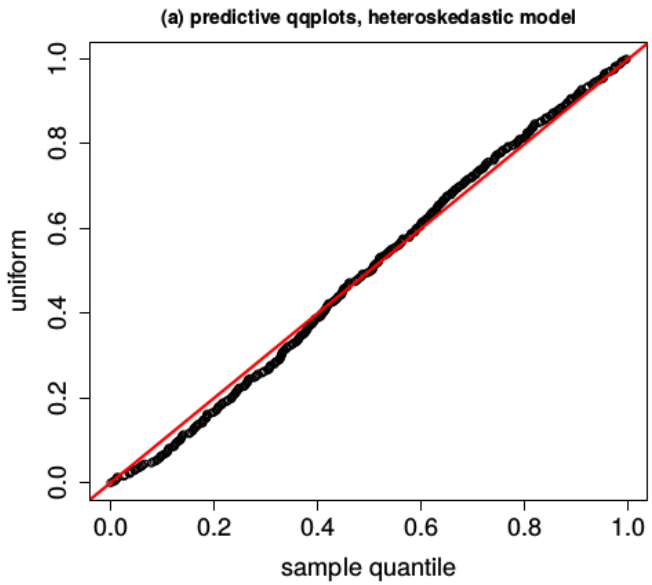
Given training or test (x_i, y_i):

- 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 percentile of y_i in the draws $\tilde{y}_{id}$.

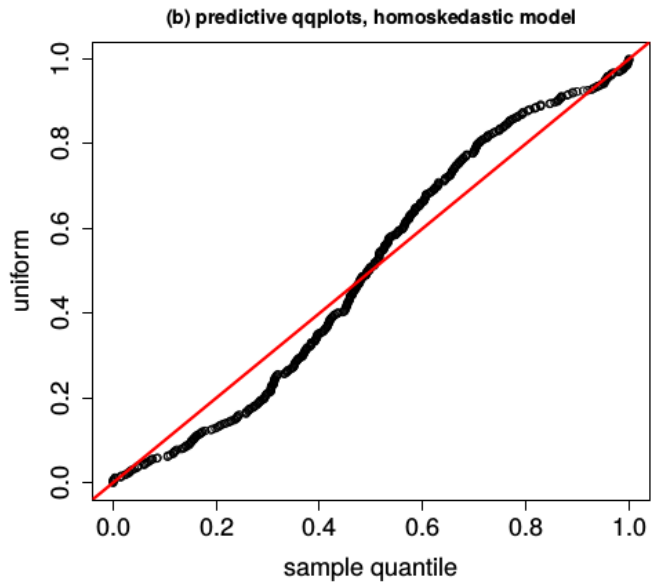
If the model is right, the percentiles should look like draws from the uniform.

Compare to the uniform using qqplots.

heterBART



BART



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

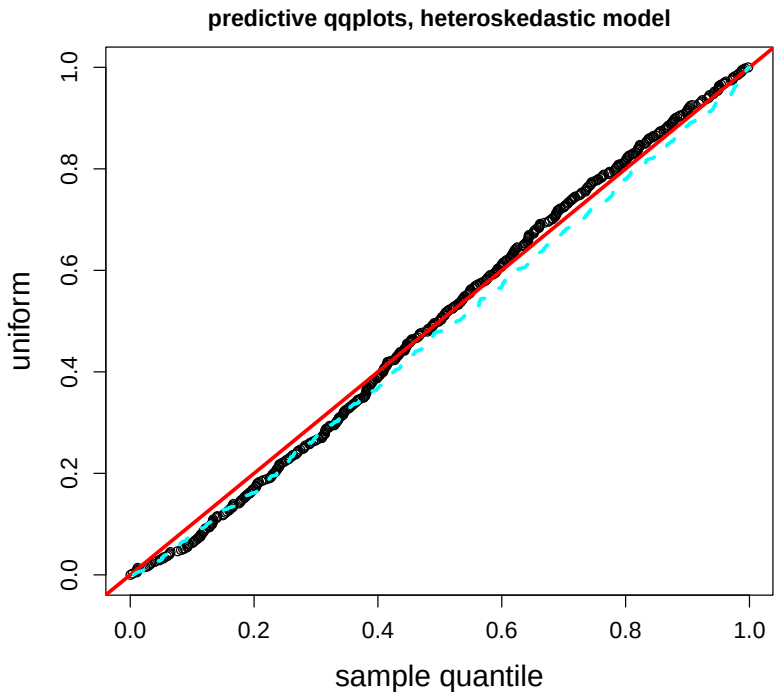
That just checks the point prediction.

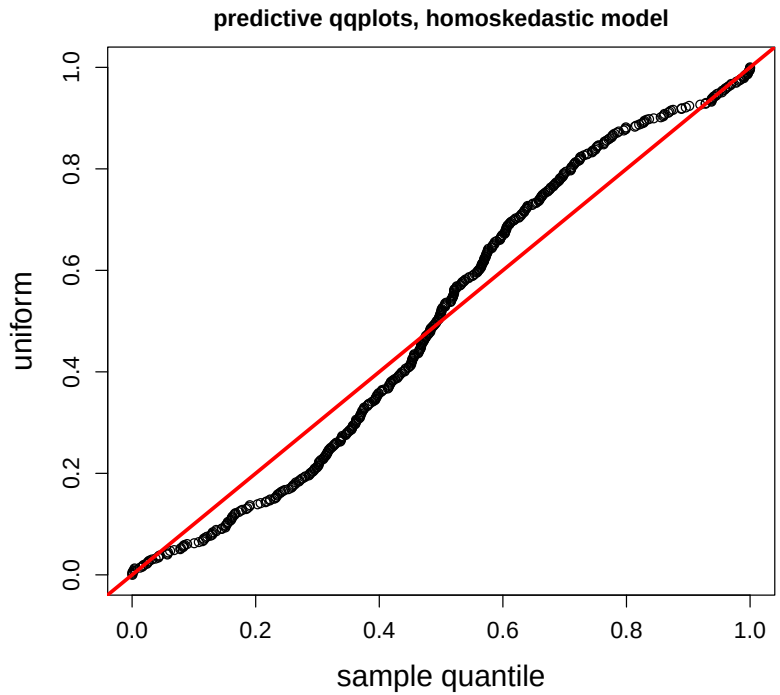
Our Bayesian model give us a full predictive distribution for

$$Y|x$$

The qqplots allow us to assess the full distributional fit, rather than just the point prediction.

Of course, the Bayesian predictive may be more spread out than the "true" $Y|x$ since it reflects our underlying uncertainty.





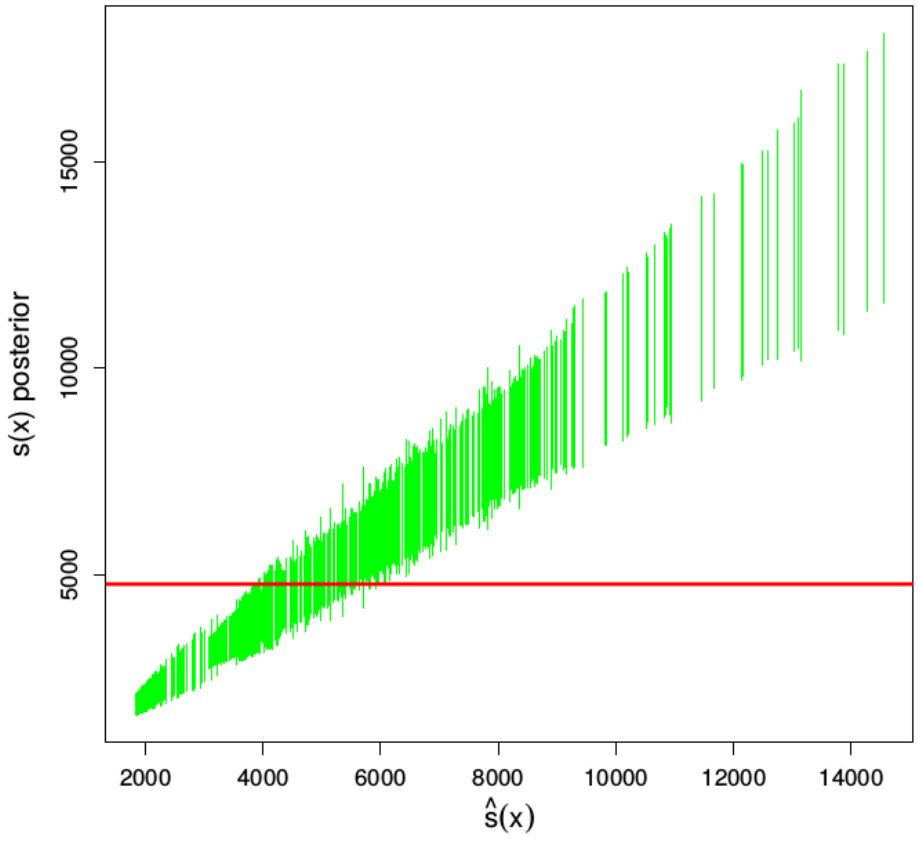
The magenta dashed line is the version we get by computing the percentile of y_i for $Y \sim N(\hat{f}(x_i), \hat{s}(x_i)^2)$, that is, we just plug in estimates rather than using the full predictive.

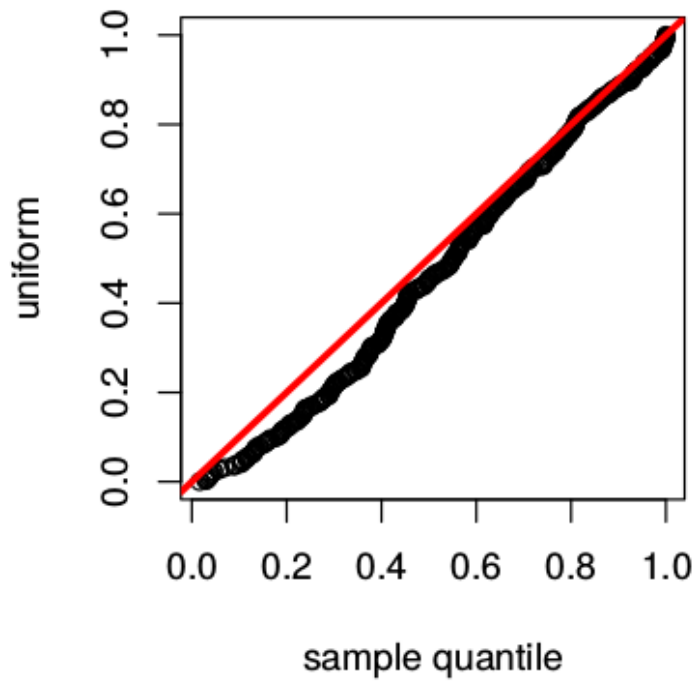
3. Cars Example

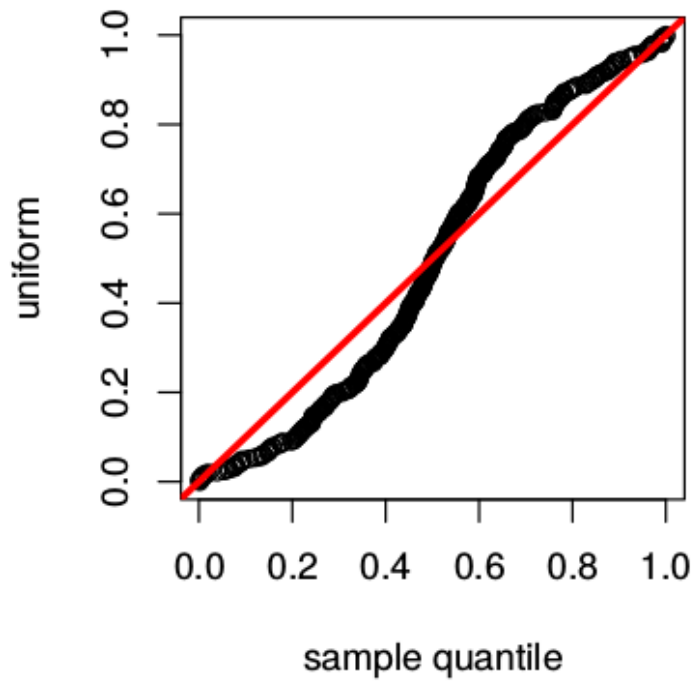
Real example, with 15 predictor variables.

Y is the price of a used car, x is characteristics of the car.

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)





not perfect, but pretty good !!.

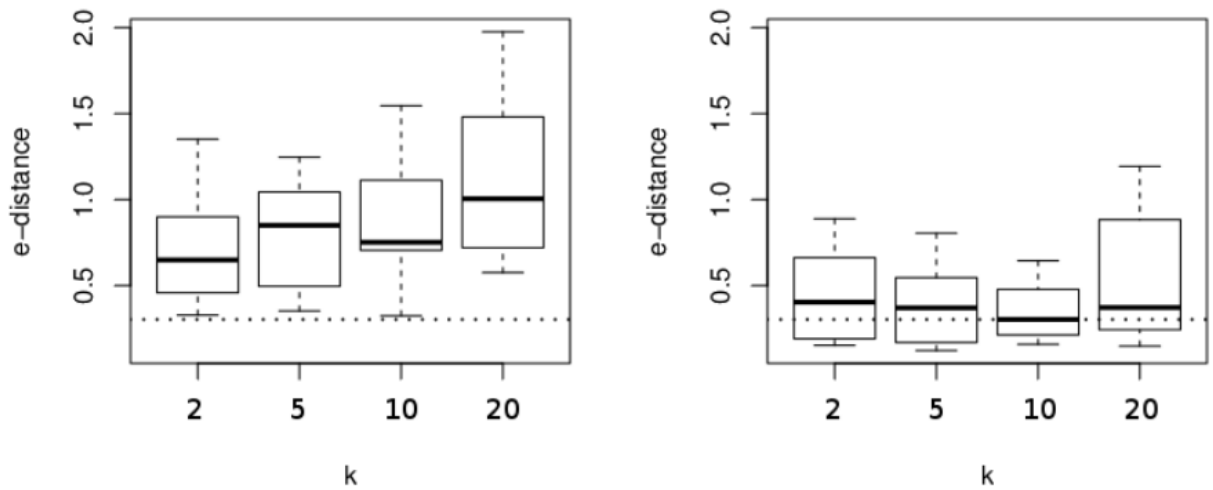
Cross Validation

k is a key prior parameter which determines how smooth the function f is.

Rather than using RMSE to use cross-validation to choose the prior we use the e-distance measure of how good the qq-plot is.

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

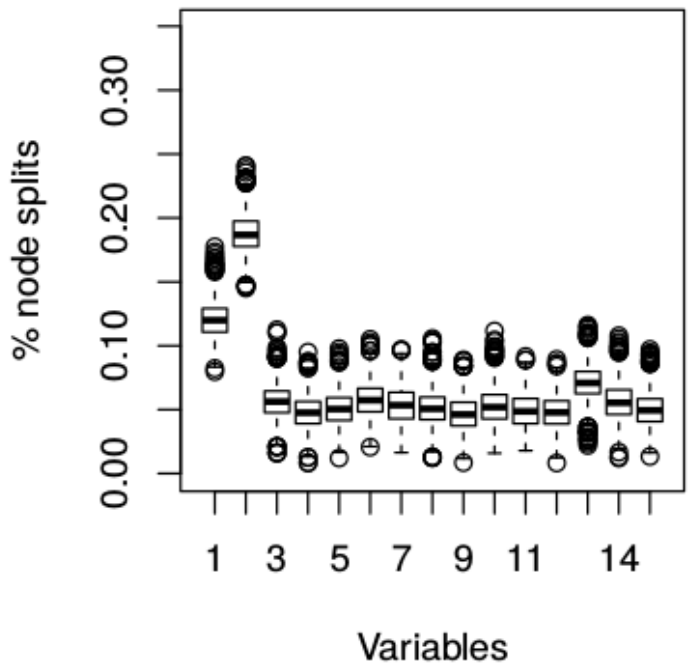
BART. heterBART.

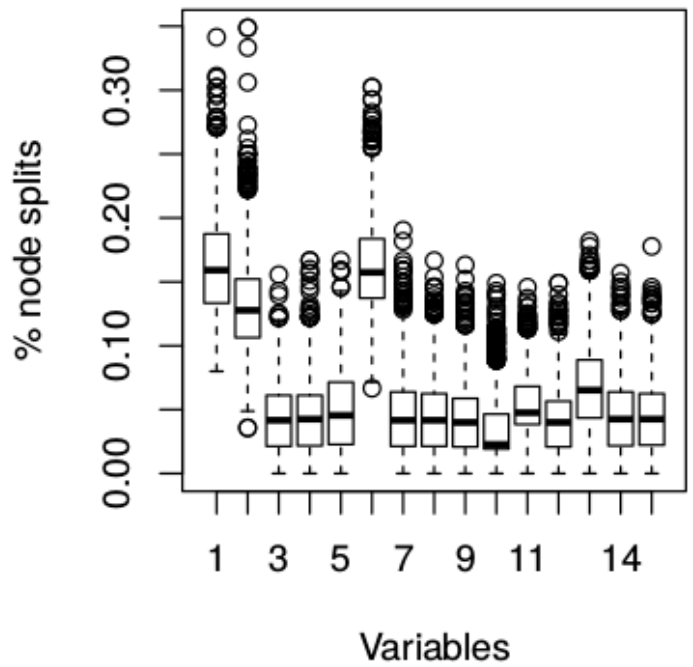


After we account for the hetero, we can fit a smoother f.

Variable selection:

f at left.
s at right.





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

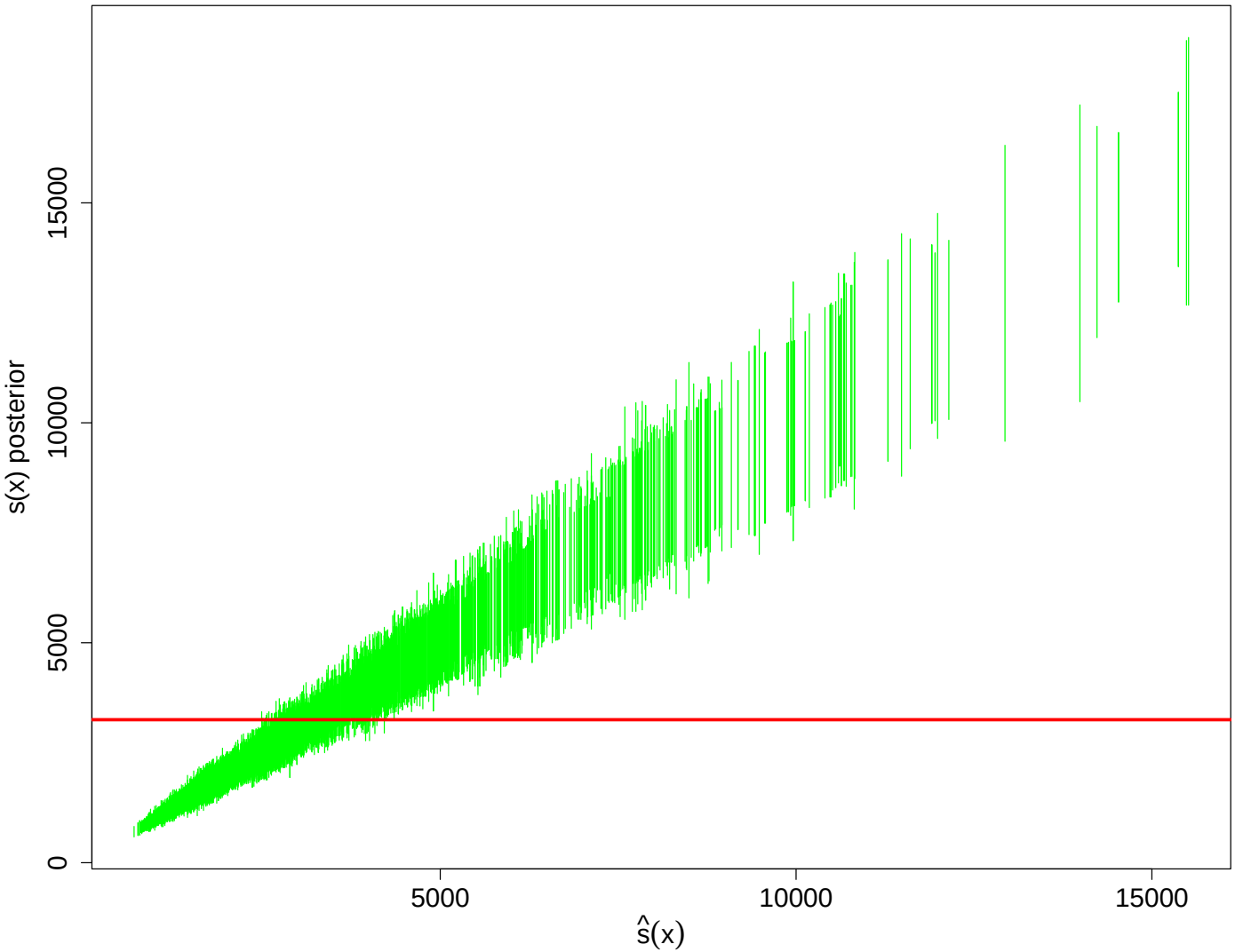
4. Fish and Alcohol Examples

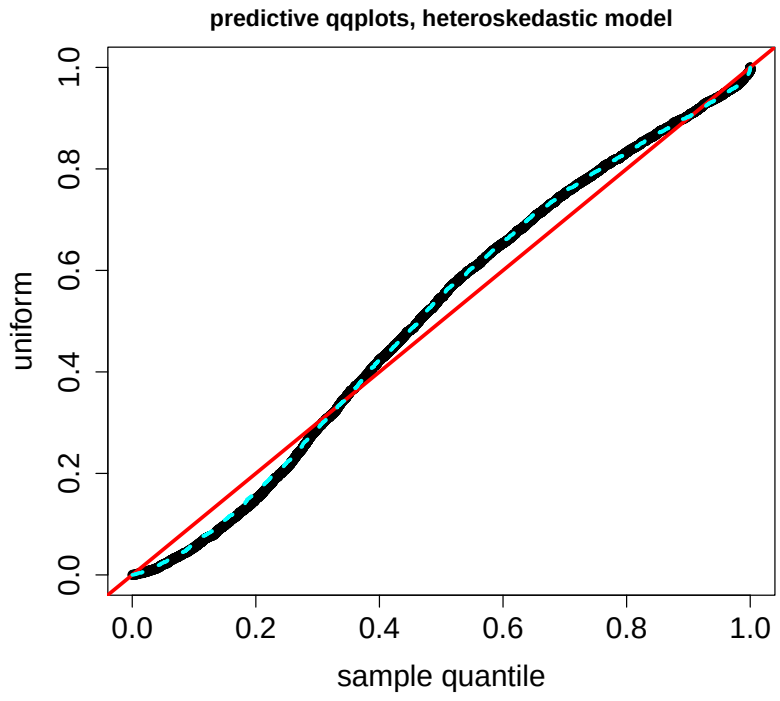
Fish

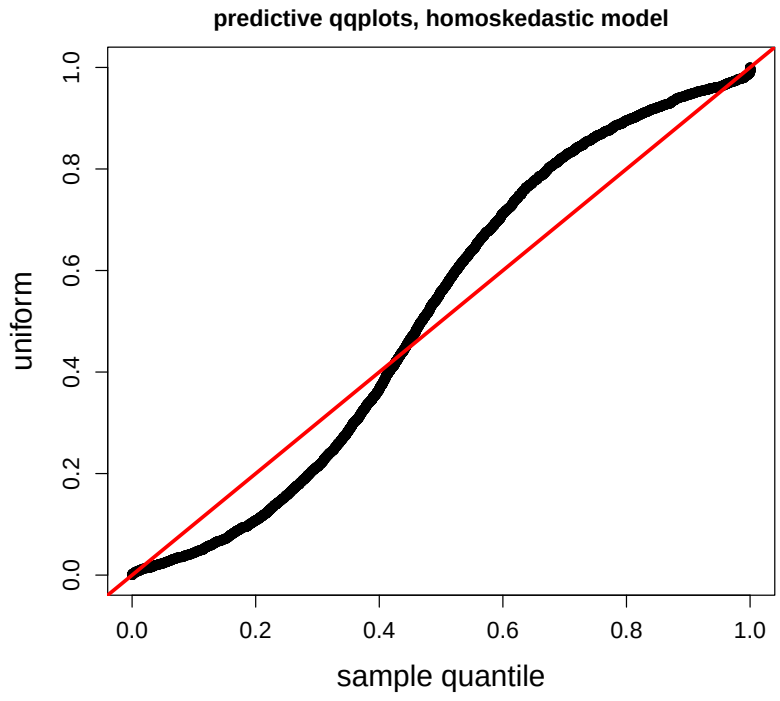
The dependent variable y is the daily catch of fishing boats in the Grand Bank fishing grounds (Fernandez et al., 2002).

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

After the creation of dummies for categorical variables, the dimension of x is 25.


s(x) posterior
0
5000
10000
15000
5000
10000
15000
ŝ(x)





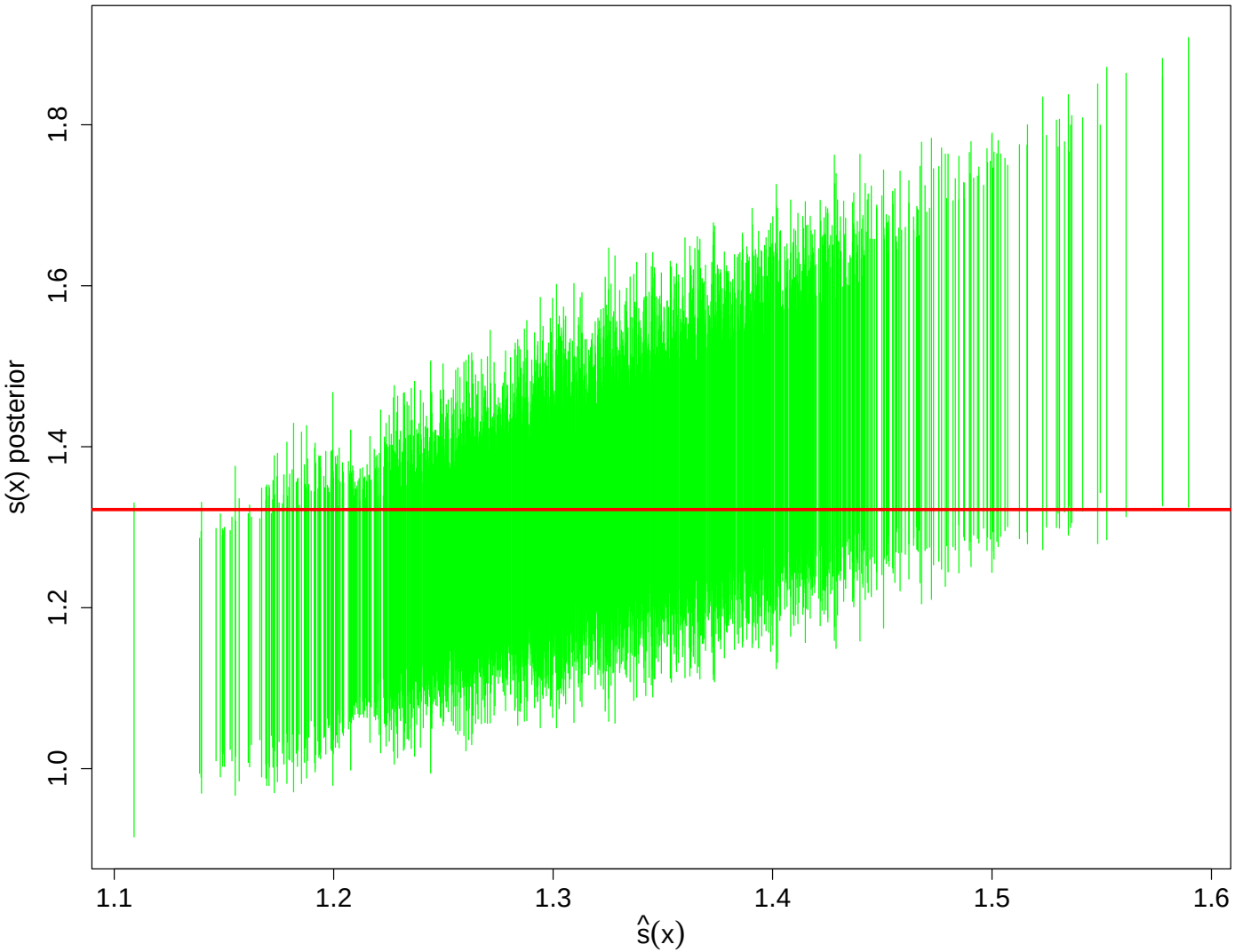
Even though we know $y \geq 0$, *simple heterBART is not too bad!!*

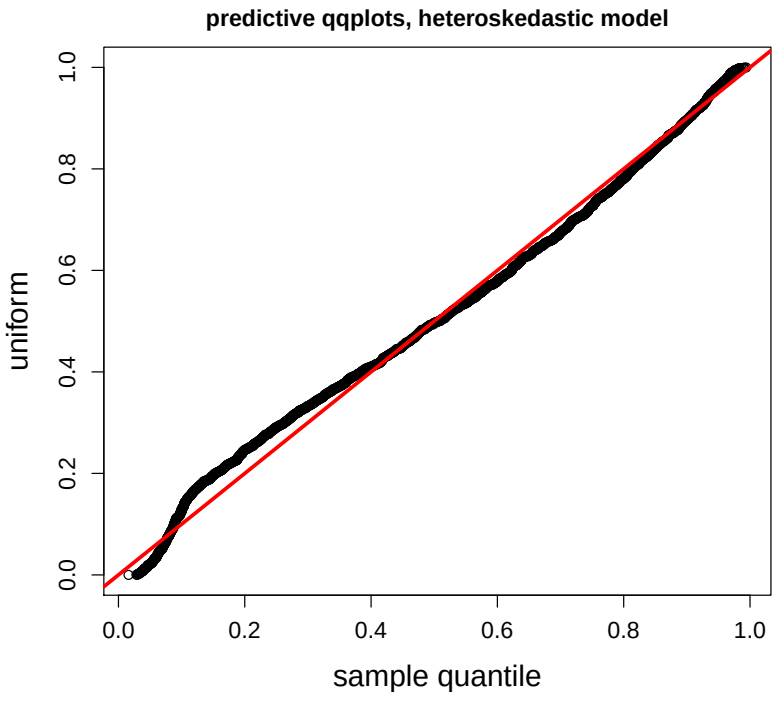
Alcohol

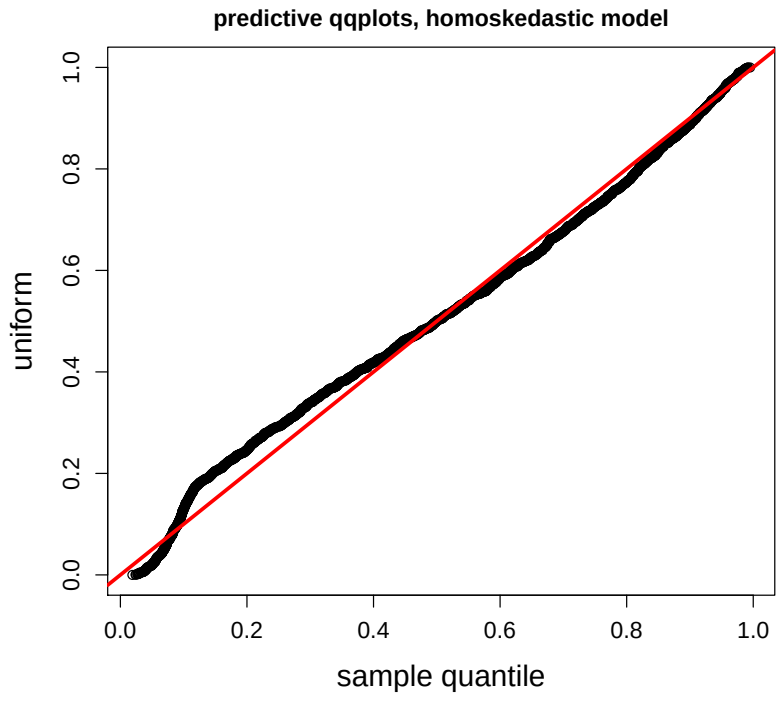
The dependent variable y is the number of alcoholic beverages consumed in the last two weeks. (Kenkel and Terza, 2001).

The 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.

After the creation of dummies for categorical variables, the dimension of x is 35.


s(x) posterior
1.0
1.2
1.4
1.6
1.8
1.1
1.2
1.3
1.4
1.5
1.6
ŝ(x)





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

5. Prior and MCMC

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, the simplicity of the BART MCMC is maintained!!
And, we use the same priors for T and $\mathcal{T}$.

Note:

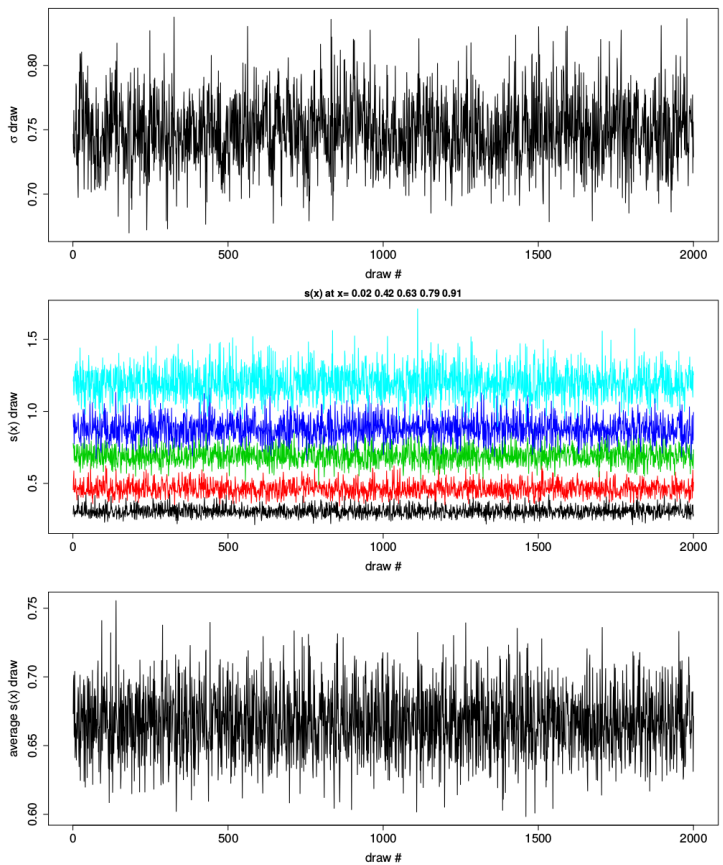
The MCMC is actually pretty complex as it used Pratola's enhanced MCMC on a single tree which can work better than the orginal Chipman, George, McCulloch moves.

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.



6. Conclusion

Adding in a fit the variance seems like a nice enhancement to ensemble modeling.

In some applications, a point prediction is all you want but sometimes you want the plus and minus!!

We still have normal errors and we working on this, but you have to be careful. You can't make things *too* flexible.

And there is something to be said for

$$Y|x \sim N(f(x), s(x)^2).$$