Fully Nonparametric Bayesian Additive Regression Trees

Ed George, Prakash Laud, Brent Logan, Robert McCulloch, Rodney Sparapani

Ed: Wharton, U Penn Prakash, Brent, Rodney: Medical College of Wisconsin Rob: Arizona State

1. The BART Model and Prior
2. Fully Nonparametric BART
3. Simulated Examples
4. Real Data
5. More on DPM

1. The BART Model and Prior

BART:

Bayesian Additive Regression Trees

Chipman, George, and McCulloch.

Regression Trees:

First, we review regression trees to set the notation for BART.

Note however, that even in the simple regression tree case, our Bayesian approach is very different from the usual CART type approach.

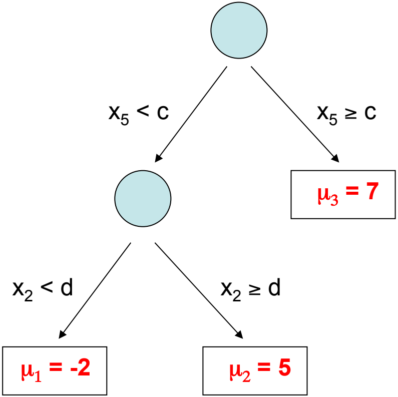
The model with have *parameters* and corresponding *priors*.

Regression Tree:

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.

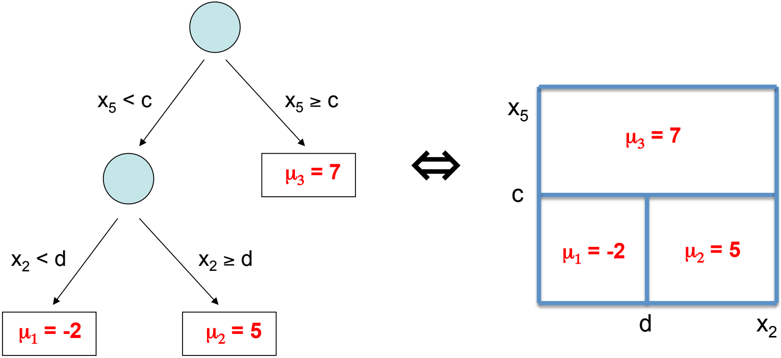
Let $g(x; \theta)$, $\theta = (T, M)$ be a regression tree function that assigns a μ value to x.



A single tree model:

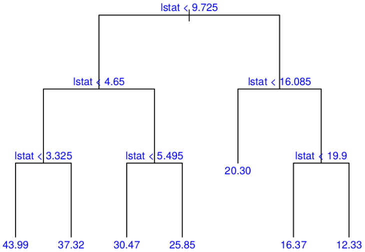
$$y = g(x; \theta) + \epsilon.$$

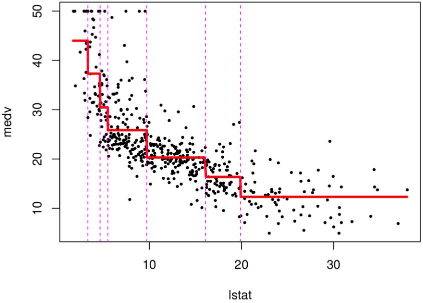
A coordinate view of $g(x;\theta)$



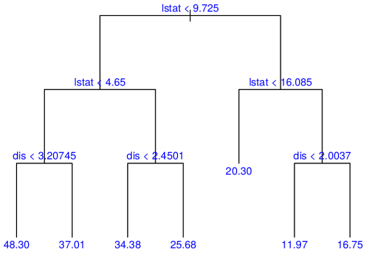
Easy to see that $g(x;\theta)$ is just a step function.

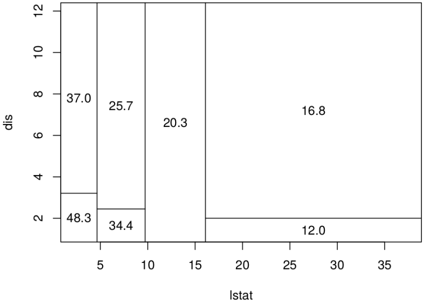
Here is an example of a simple tree with one x fit using standard CART methdology.



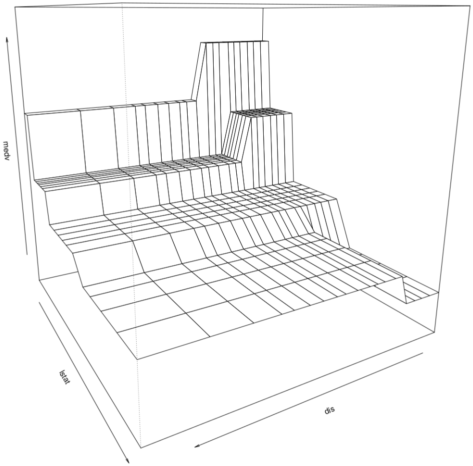


Here is an example with 2 x variables.



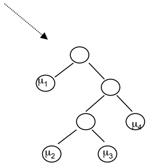


And here is the corresponding function (our g).



The BART Model

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



$m = 200, 1000, \ldots, \text{big}, \ldots$.

$f(x \mid \cdot)$ is the sum of all the corresponding μ's at each bottom node.

Such a model combines additive and interaction effects.

All parameters but σ are unidentified !!!!

...the connection to Boosting is obvious...

But,..

Rather than simply adding in fit in an iterative scheme, we will explicitly specify a prior on the model which directly impacts the performance.

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 wants:

- Each T small.
- Each μ small.
- "nice" σ (smaller than least squares estimate).

We refer to π as a regularization prior because it restrains the overall fit.

In addition, it keeps the contribution of each $g(x; T_i, M_i)$ model component small.

Prior on T

We specify a process we can use to draw a tree from the prior.

The probability a current bottom node, at depth d, gives birth to a left and right child is

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

The usual BART defaults are

$$\alpha = \text{“base”} = .95, \;\; \beta = \text{“power”} = 2.$$

This makes non-null but small trees likely.

```
nbottom
    1      2      3      4      5
  0.05   0.55 0.28  0.09  0.03
```

Splitting variables and cutpoints are drawn uniformly from the set of “available” ones.

Prior on M

Let θ denote all the parameters.

$$f(x \mid \theta) = \mu_1 + \mu_2 + \cdots \mu_m.$$

where μ_i, is the μ in the bottom node x falls to in the i^{th} tree.

Let $\mu_i \sim N(0, \tau^2)$, iid.

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

In practice we often, unabashadly, use the data by first centering and then choosing τ so that

$$f(x \mid \theta) \in (y_{min}, y_{max}), \text{ with high probability.}$$

This gives:

$$\tau \propto \frac{1}{\sqrt{m}}.$$

Prior on σ

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

Default: $\nu = 3$.

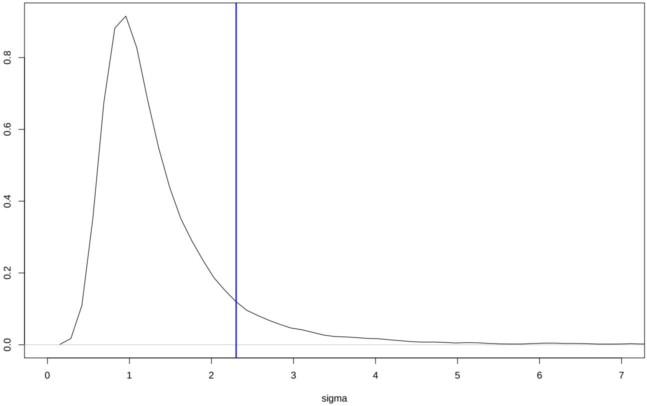
λ:

Get a reasonable estimate of $\hat{\sigma}$ of sigma then choose λ to put $\hat{\sigma}$ at a specified quantile of the σ prior.

Default: quantile $=$ .9

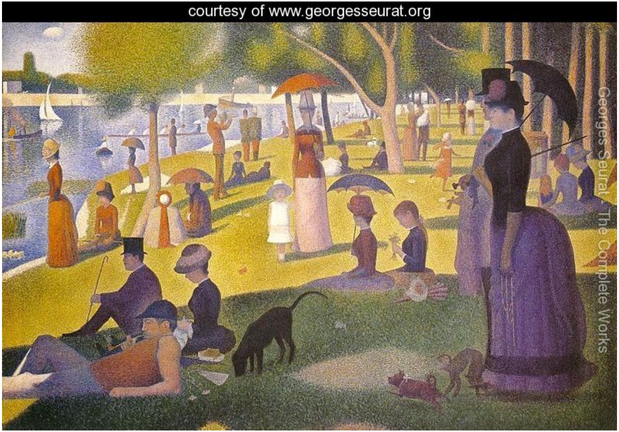
Default: if $p < n$, $\hat{\sigma}$ is the usual least squares estimate, else $sd(y)$.

Solid blue line at $\hat{\sigma}$.



Conjecture: Most "failures" of BART are due to this default.

Why does it work???


courtesy of www.georgesseurat.org
Georges Seurat, The Complete Works

Boosting: Freund and Schapire, Jerome Friedman

Note:

I really want to be able to pick a (data based) default prior so I can put out my R package and people can get good results without too much effort.

Contrast this with Deep Neural Nets, which are hard to fit.

But, you can pretty easily put choose a prior for $f(x)$ and σ!!!

Constrast this with Deep Neural Nets, where it is very hard to think about the prior.

2. Fully Nonparametric BART

BART

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

where f is a sum of trees.

- normal errors are embarrassing.
- prior on σ is flawed.
- normal errors may lead to influential observations and poorly calibrated predictive intervals.

Obvious Solution:

Use DPM (Dirichlet Process Mixtures) in the classic Escobar and West manner to model the errors “non parametrically”.

Tried this in the past with mixed success.

The DPM stuff is tricky.

...not all obvious that you can get away with flexible f and flexible errors !!!

The Goal: Goes in the R-package so people can use it with automatic priors and reliably get sensible results.

The MCW crowd (Prakash is a long-time nonparametric Bayesian) have a lot of experience with DPM.

Prakash has recent work on choosing priors for DPM:

Low Information Omnibus (LIO) Priors for Dirichlet Process Mixture Models
(Yushu Shi, Michael Martens, Anjishnu Banerjee, and Purushottam Laud)

Cautiously optimistic that we have a scheme that is close to working.

DPMBART

$$Y_i = f(x_i) + \mu_i + \sigma_i \, Z_i, \quad Z_i \sim N(0,1).$$

each observation gets to have its own (μ_i, σ_i).

But, the DPM machinery allows us to uncover a set of $(\mu_j^*, \sigma_j^*), \; j = 1, 2, \ldots, I$ such that each

$$\text{for each } i, \; (\mu_i, \sigma_i) = (\mu_j^*, \sigma_j^*), \;\; \text{for some } j.$$

In our real example, $n = 1,479$, $I \sim 100$.

Even though each observation can have it's own (μ_i, σ_i), *subsets of the obserations have the same* (μ, σ) *so that there is a relatively small number of unique values.*

Markov Chain Monte Carlo (MCMC):

$$\{\mu_i, \sigma_i\} \mid f, \quad f \mid \{\mu_i, \sigma_i\}$$

At each draw d we have

$$f^d, \{(\mu_i^d, \sigma_i^d)\},\ i = 1, 2, \ldots, n$$

where at each draw, many of the (μ, σ) pairs are repeats.

For example,

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

Connection to Mixture of Normals

At each draw d we have

$$f, \{(\mu_i, \sigma_i)\},\ i = 1, 2, \dots, n$$

Let

$$\{(\mu_j^*, \sigma_j^*)\},\ j = 1, 2, \dots, I$$

be the unique (μ, σ) pairs.

Let

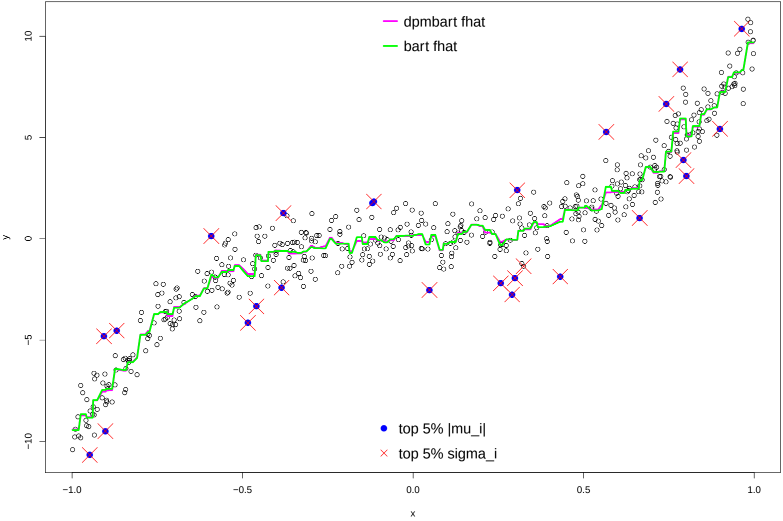
$$p_j = \frac{\#\left[(\mu_i, \sigma_i) = (\mu_j^*, \sigma_j^*)\right]}{n}$$

Then

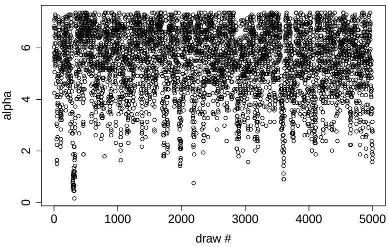
$$\epsilon \approx \sum_{j=1}^{I} p_j \, N(\mu_j^*, (\sigma_j^*)^2)$$

3. Simulated Examples

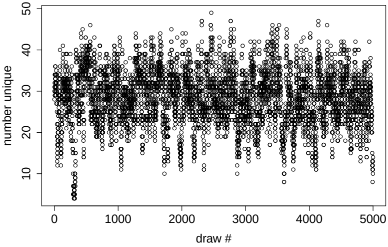
Simulated data with t_{20} (essentially normal) errors. $n = 500$.



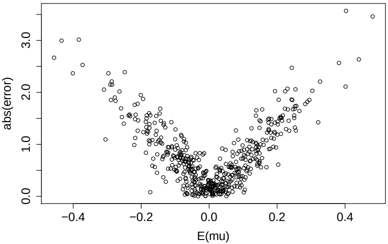
alpha draws



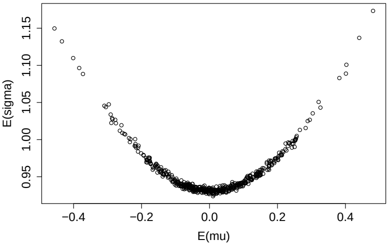
draws of number unique (mu,sigma)



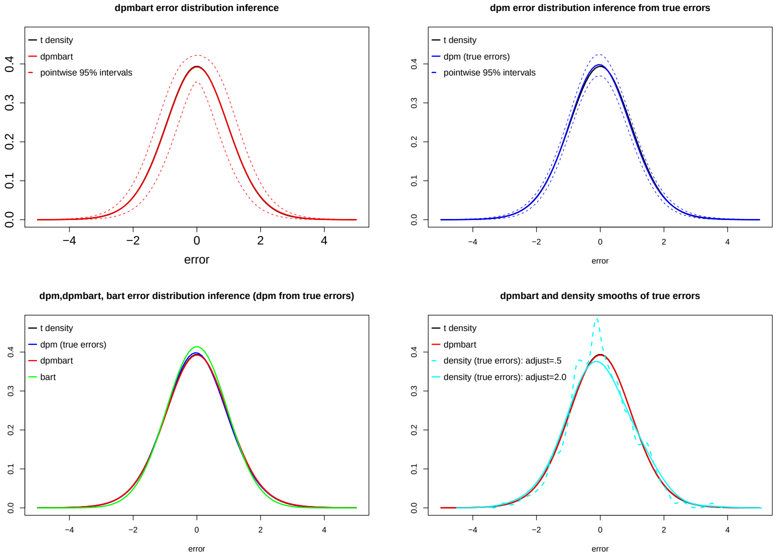
E(mu) vs. abs(y−fx)



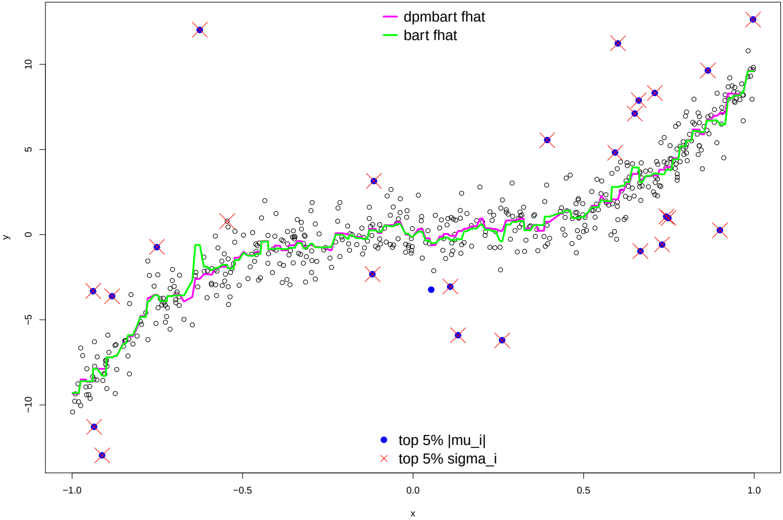
E(mu) vs E(sigma)

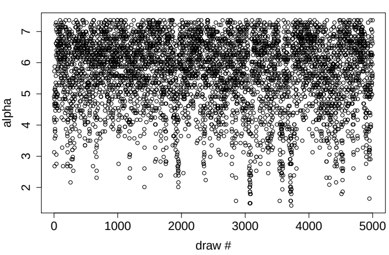


Inference for the error distribution:

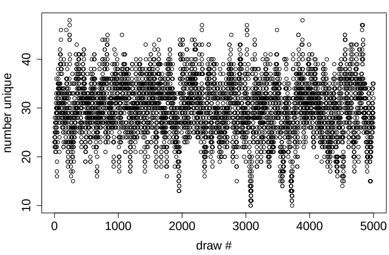


Simulated data with t_3 errors.

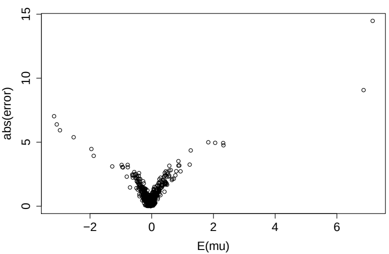




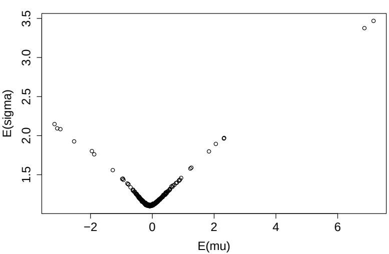
alpha draws



draws of number unique (mu,sigma)

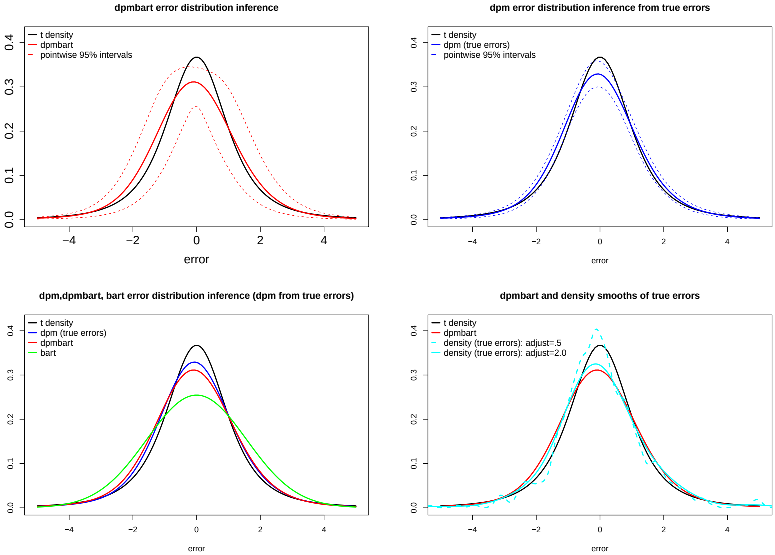


E(mu) vs. abs(y−fx)



E(mu) vs E(sigma)

Inference for the error distribution:

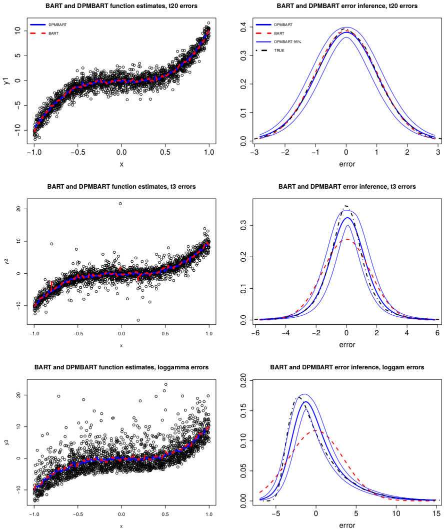


Three basic examples: t20, t3, skewed.

If the error is close to normal, then dpmbart is close to bart.

If the error in non-normal, dpmbart is much closer to the truth, but shrunk a bit towards bart.

In these examples, $\hat{f}$ for dpmbart and bart are pretty much the same but with lower signal/sample sizes this does not have to be the case.



BART and DPMBART function estimates, t20 errors

BART and DPMBART error inference, t20 errors

BART and DPMBART function estimates, t3 errors

BART and DPMBART error inference, t3 errors

BART and DPMBART function estimates, loggamma errors

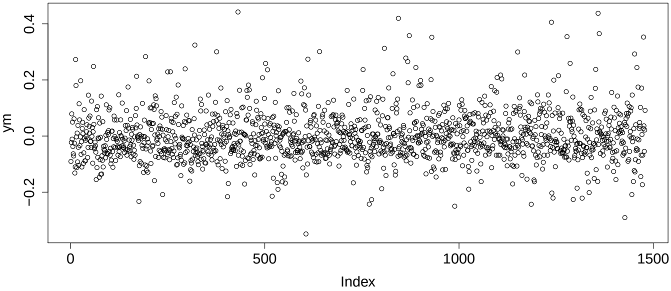
BART and DPMBART error inference, loggam errors

4. Real Data

Using one month of a much larger data set I am working on.

y: return on cross-section of firms
x: things about the firm measured the previous month.

y:



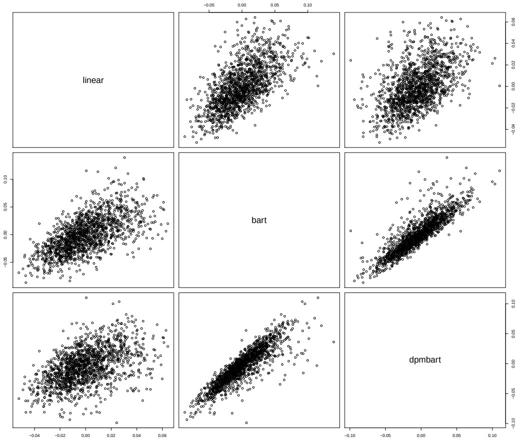
Multiple regression results:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.0309918  0.0115973   2.672 0.007616 **
r1_1              -0.0384859  0.0077977  -4.936  8.9e-07 ***
r12_2             -0.0326876  0.0077786  -4.202  2.8e-05 ***
idiosyncraticvol   0.0068535  0.0098193   0.698 0.485311
seasonality       -0.0118890  0.0076687  -1.550 0.121277
industrymom        0.0006992  0.0081369   0.086 0.931536
ln_turn            0.0311180  0.0085812   3.626 0.000297 ***
me                -0.0271681  0.0093472  -2.907 0.003709 **
an_cbprofitability 0.0096240  0.0077403   1.243 0.213935
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.08476 on 1470 degrees of freedom
Multiple R-squared:  0.05543,   Adjusted R-squared:  0.05029
F-statistic: 10.78 on 8 and 1470 DF,  p-value: 7.626e-15
```

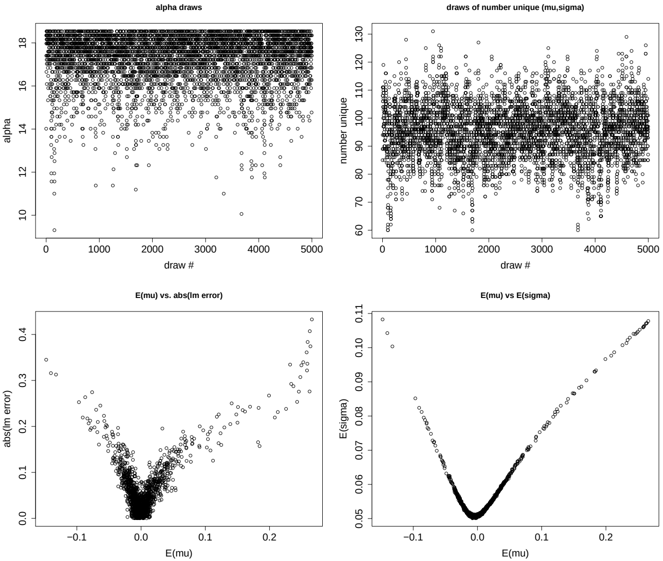
It’s like looking for a needle in a haystack !!!

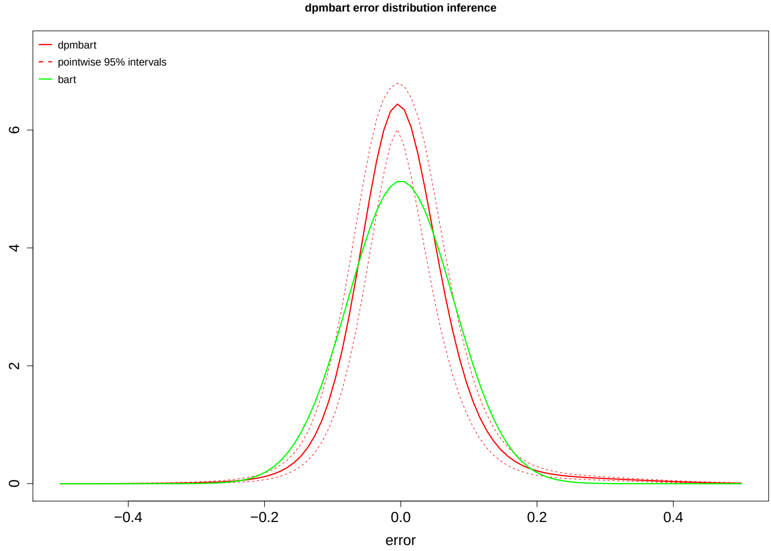
Compare the $\hat{f}$: linear, bart, dpmbart:



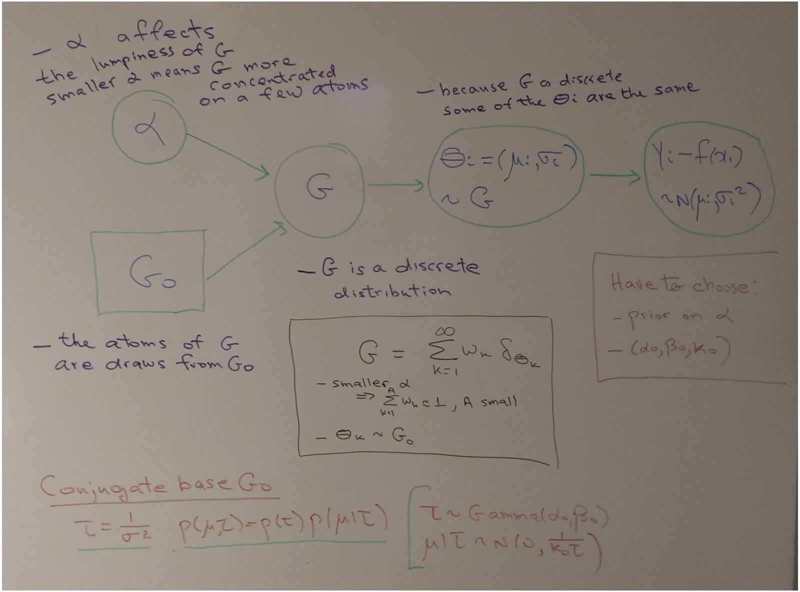
dpmbart a little different from bart because it is not pulled around by the outliers ???

Note: the "errors" are now the errors from the multiple regression since we don't have the $y - f(x)$ we had for the simulated data.




dpmbart error distribution inference
dpmbart
pointwise 95% intervals
bart
0
2
4
6
−0.4
−0.2
0.0
0.2
0.4
error

5. More on DPM



– α affects the lumpiness of G
smaller α means G more concentrated on a few atoms

– because G is discrete
some of the Θ_i are the same

– G is a discrete distribution

– the atoms of G are draws from G_0

$$G = \sum_{k=1}^{\infty} w_k \delta_{\Theta_k}$$

– smaller α
$\Rightarrow \sum_{k=1}^{A} w_k \approx 1$, A small

– $\Theta_k \sim G_0$

Have to choose:

– prior on α

– (α_0, β_0, k_0)

Conjugate base G_0

$\tau = \frac{1}{\sigma^2}$ $\quad p(\mu, \tau) = p(\tau)\, p(\mu \mid \tau)$

$$\left[\begin{array}{l} \tau \sim \text{Gamma}(\alpha_0, \beta_0) \\ \mu \mid \tau \sim N\left(0, \frac{1}{k_0 \tau}\right) \end{array}\right.$$

Prior on α:

Used construction of Conley, Hanson, McCulloch, and Rossi.

- discrete distribution for α.
- you get to pick (I_{min}, I_{max}) range for number of unique θ values.
  Default was $I_{min} = 1$, $I_{max} \approx .1n$.
- In our examples, draws of α bumped up against upper limit.
  This could be good in that we want the prior conservative.

(μ, τ):

τ:

For $\tau = 1/\sigma^2$ we used an approach similar to the BART default, but we tighten up up a bit.

$$\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}, \;\; \nu = 2\alpha_o, \lambda = \beta_o/\alpha_o.$$

- bart: $\nu = 3$, dpmbart: $\nu = 10$.
- bart: choose λ to put $\hat{\sigma}$ at quantile $= .9$, dpmbart: quantile $= .95$.

The bart default gets $\hat{\sigma}$ from the multiple regression.

μ:

$$\mu \sim \frac{\sqrt{\lambda}}{\sqrt{k_o}}\, t_\nu .$$

let e_i be the residuals from the multiple regression.

Let k_s be scaling for the μ marginal.

Let k_o solve:

$$max|e_i| = k_s \frac{\sqrt{\lambda}}{\sqrt{k_o}}.$$

Default: $k_s = 10$.

Comments:

- You can't be too diffuse on the base measure.
- Would prefer not to extend the hierarchy and put priors on the base hyperparameters (a common practice).
- BART default depends on the standard deviation of the regression residuals, DPMBART depends on the sd of the resids *and* the overall scale of the resids.
- $k_s = 10$ may seem large, you don't have to cover the residual range, as μ get's bigger, σ gets bigger and you can't be too spread out.
- We would be happy to keep the dpm prior somewhat conservative in that we nail the normal error case but miss slightly on the non-normal cases: *DO NO HARM*.