Heteroskedastic BART

Ed George, Rob McCulloch, Matt Pratola

1. The Goal
2. Heteroskedastic BART
3. The Prior
4. Prior Choice
5. MCMC
6. Simulated 1-d Example
7. Real 1-d Example
8. Real 4-d Example
9. Real 90-d Example
10. Concluding Remarks

1. The Goal

Ensemble models combine many fits together to get on overall prediction.

Ensemble modeling is a very powerful methodology.

(as a practical matter, Random Forests and Boosting trees)

They allow us to search for high dimensional, complex relationships with relatively little bother.

(nonlinearity, interactions)

However they are focused on simple point predictions !!

$$E(Y \mid X = x)$$

There is a need for tools for looking for heteroskedasticity flexibly.

$$Var(Y \mid X = x)$$

2. Heteroskedastic BART

Our model is:

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

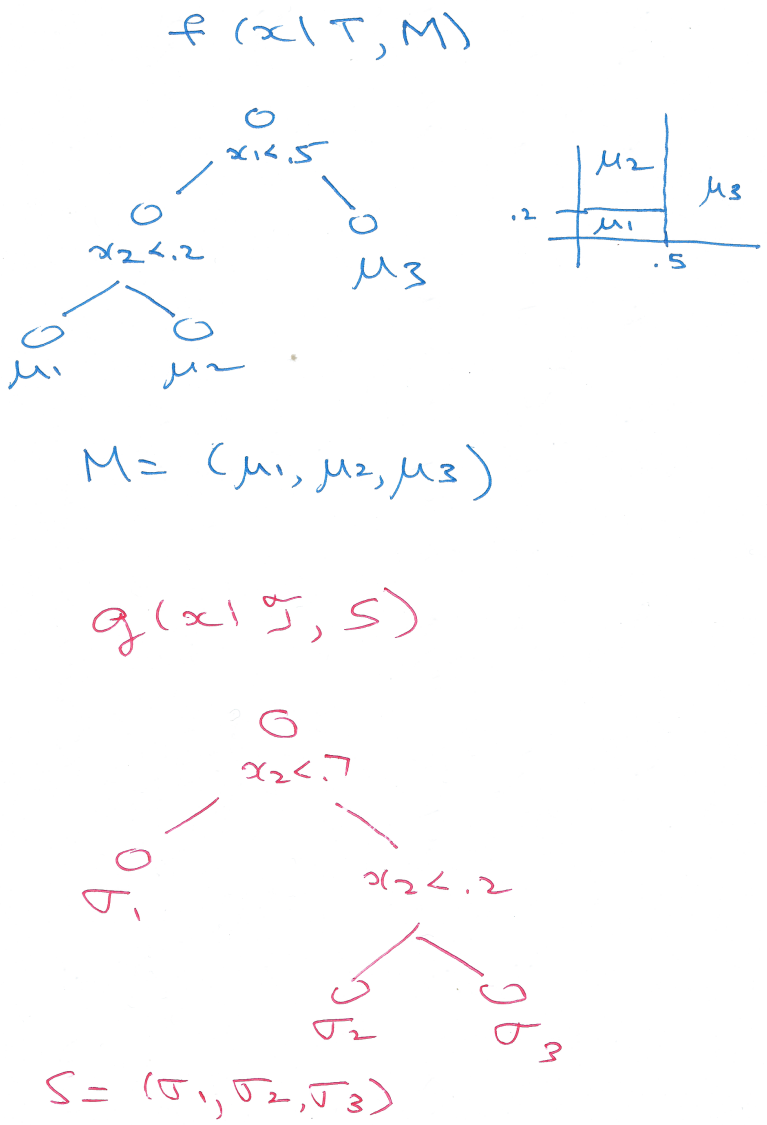
$$f(x) = \sum_{i=1}^{m} f(x \mid T_i, M_i)$$

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

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.

a tree for a mean

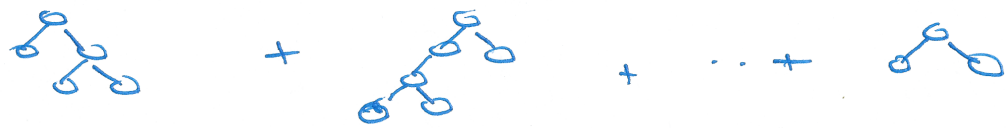


and a tree for a standard deviation

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

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

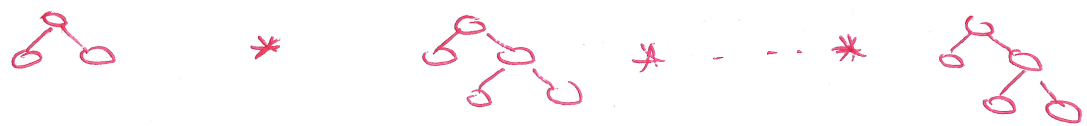
$$= \quad + \quad + \cdots +$$



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

$$g(x) = g(x \mid \sigma_1, S_1) * g(x \mid \sigma_2, S_2) * \cdots * g(x \mid \sigma_k, S_k)$$

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



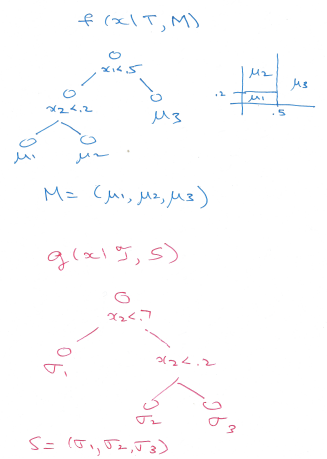
$$= \sigma_1 * \sigma_2 * \cdots * \sigma_k$$

3. The Prior

$$\pi((T_1, M_1), \ldots, (T_m, M_m), (\mathcal{T}_1, S_1), \ldots, (\mathcal{T}_k, S_k))$$

$$= \prod_{i=1}^{m} \pi(T_i, M_i) \prod_{i=1}^{k} \pi(\mathcal{T}_i, S_i)$$

$$= \prod_{i=1}^{m} \pi(T_i)\, \pi(M_i \mid T_i) \prod_{i=1}^{k} \pi(\mathcal{T}_i)\, \pi(S_i \mid \mathcal{T}_i)$$



$\pi(T_i)$, push towards small trees, "uniform" on rules.

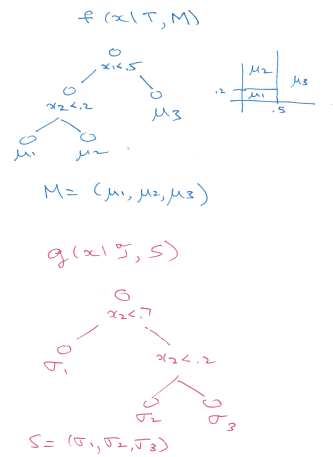
$\pi(M) = \prod_{i=1}^{B} \pi(\mu_i)$, $\mu_i \sim N(0, \tau^2)$, iid,

B is the number of bottom nodes.

$$\pi((T_1, M_1), \ldots, (T_m, M_m), (\mathcal{T}_1, S_1), \ldots, (\mathcal{T}_k, S_k))$$

$$= \prod_{i=1}^{m} \pi(T_i, M_i) \prod_{i=1}^{k} \pi(\mathcal{T}_i, S_i)$$

$$= \prod_{i=1}^{m} \pi(T_i)\, \pi(M_i \mid T_i) \prod_{i=1}^{k} \pi(\mathcal{T}_i)\, \pi(S_i \mid \mathcal{T}_i)$$



$\pi(\mathcal{T}_i)$ same as for T_i.

$\pi(S) = \prod_{i=1}^{B} \pi(\sigma_i)$, $\sigma_i \sim \frac{\nu\lambda}{\chi^2_\nu}$, iid.

4. Prior Choice

Key to Additive Trees - the Prior:

Use $\mu_i \sim N(0, \tau^2)$, iid, then, before we see the data,

$$f(x) = \sum_{i=1}^{m} \mu_i \sim N(0, m\,\tau^2)$$

Given m, relatively easy to think of a good choice for τ.

Or, at least, choose a sensible range of values to assess using cross-validation.

Key to Multiplicative Trees - the Prior:

Use $\sigma_i^2 \sim \frac{\nu\lambda}{\chi^2_\nu}$, iid, then

$$g(x) = \prod_{i=1}^{k} \sigma_i \sim ??$$

Given k, still fairly easy to think think about the prior.

For example, you can compute the mean or easily simulate.

Or, at least, choose a sensible range of values to assess using cross-validation.

Of course, the prior matters.

Choosing the prior, an example:

Suppose we were using a single error variance $\epsilon_i \sim N(0, \sigma^2)$, and we had an inverted chi-squared prior for the error variance that we were happy with:

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

This gives an expected value:

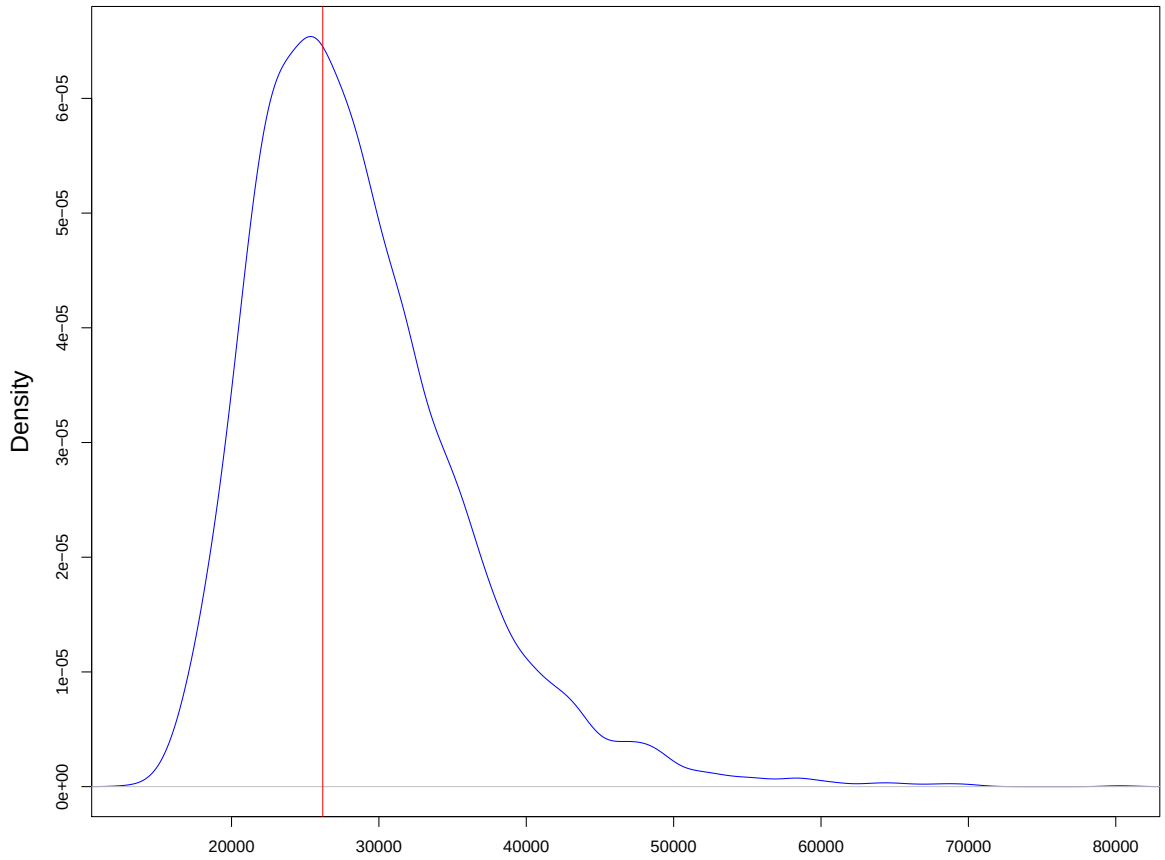
$$E[\sigma^2] = \frac{\nu\lambda}{\nu - 2}$$

In the cars data
(y is price of a used car, more later)
the sample standard deviation is
$s_y = 26190.17$.

We set
$\nu = 10$ and $\lambda = s_y^2$.

Get this prior for σ.

Red line at s_y.



Now suppose we want to choose ν' and λ' for our ensemble prior.

$$g(x) = \prod_{j=1}^{k} \sigma_j^2$$

where the σ_j^2 are mutually independent $\sim (\nu'\,\lambda')/\chi^2_{\nu'}$.

Then,

$$E(g(x)) = \prod_{j=1}^{k} E(\sigma_j^2) = \left[\frac{\nu'\lambda'}{(\nu'-2)}\right]^k$$

If we want the expected values to match up we can use:

$$\lambda' = \lambda^{1/k}$$

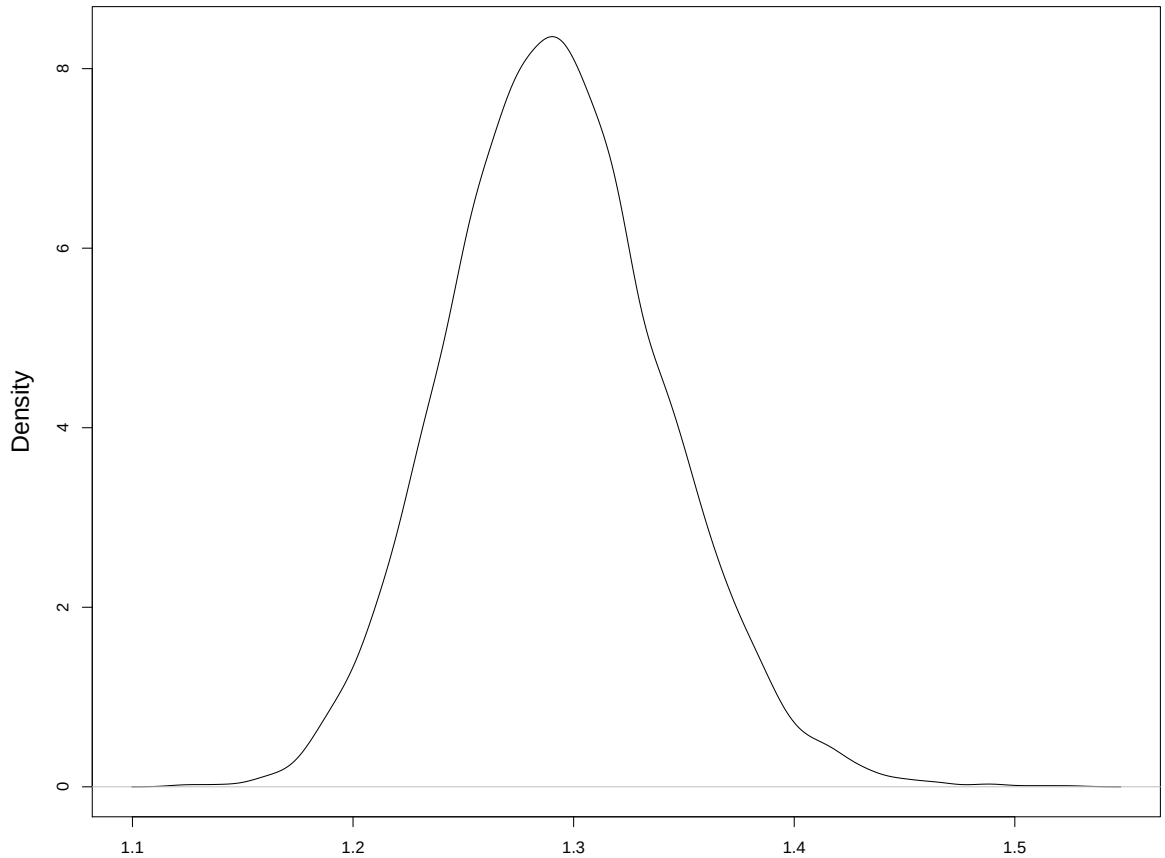
$$\nu' = \frac{2\nu^{1/k}}{\nu^{1/k} - (\nu - 2)^{1/k}}$$

Let's use $k = 40$ (forty multiplicative trees).

In our example this gives:

$\lambda' = 1.663056$

$\nu' = 360$

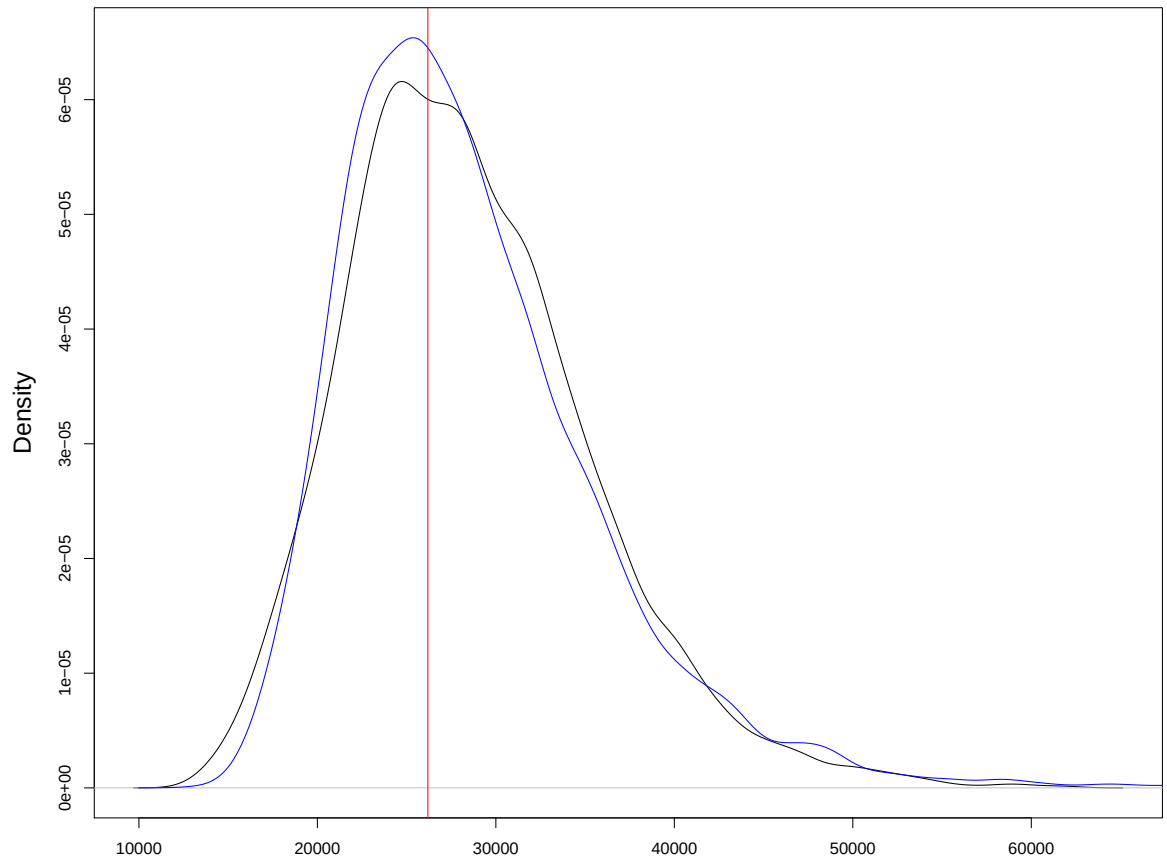


What happens when we multiply up $k = 40$ of these ??!!

blue: the $(\nu, \lambda) = (10, s_y^2)$ prior.

black: the product of 40 $(\nu, \lambda) = (360, (s_y^2)^{1/k})$ priors.

red: vertical at s_y.



looks amazingly good !!!

5. MCMC

Simple.

Because our prior on the bottom node σ values is conditionally conjugate, we can draw the same way we do for the μ bottom node parameters.

(1)

$$(\mathcal{T}_i, S_i) \mid \circ$$

Draw the tree/stan devs one at a time given everything else. Subtract off the mean fit and then divide by all the multiplicative trees but the i^{th}.

(2)

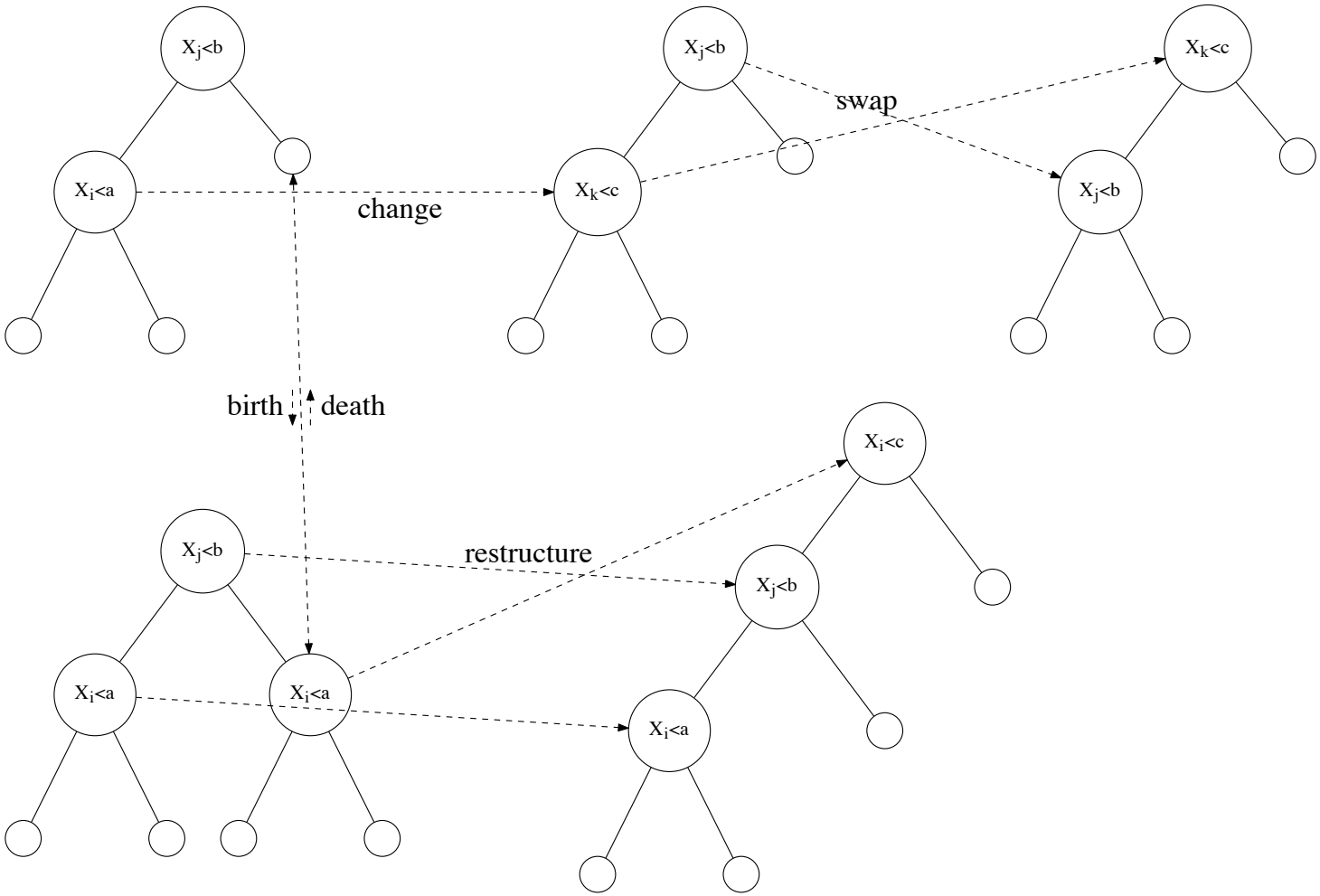
$$(\mathcal{T}_i, S_i) \mid \circ \sim \mathcal{T}_i \mid \circ,\ S_i \mid \mathcal{T}_i, \circ$$

(3)

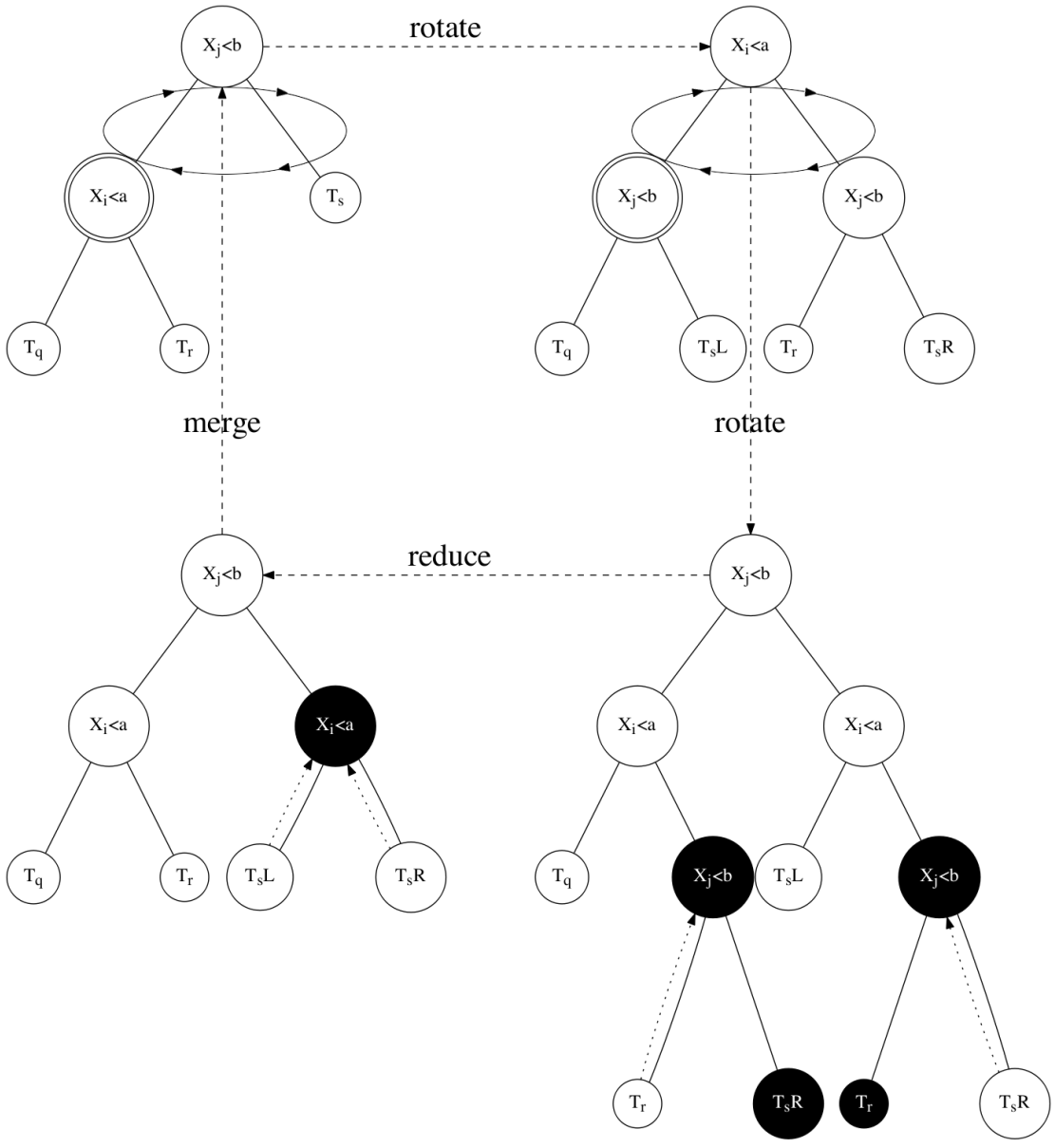
Draw $\mathcal{T}_i \mid \circ$ by integrating out the S_i which is straightforward given the conditionally conjugate prior.

Note:

The draw of $T_i \mid \circ$ or $\mathcal{T}_i \mid \circ$ involves Metropolis moves in tree space.

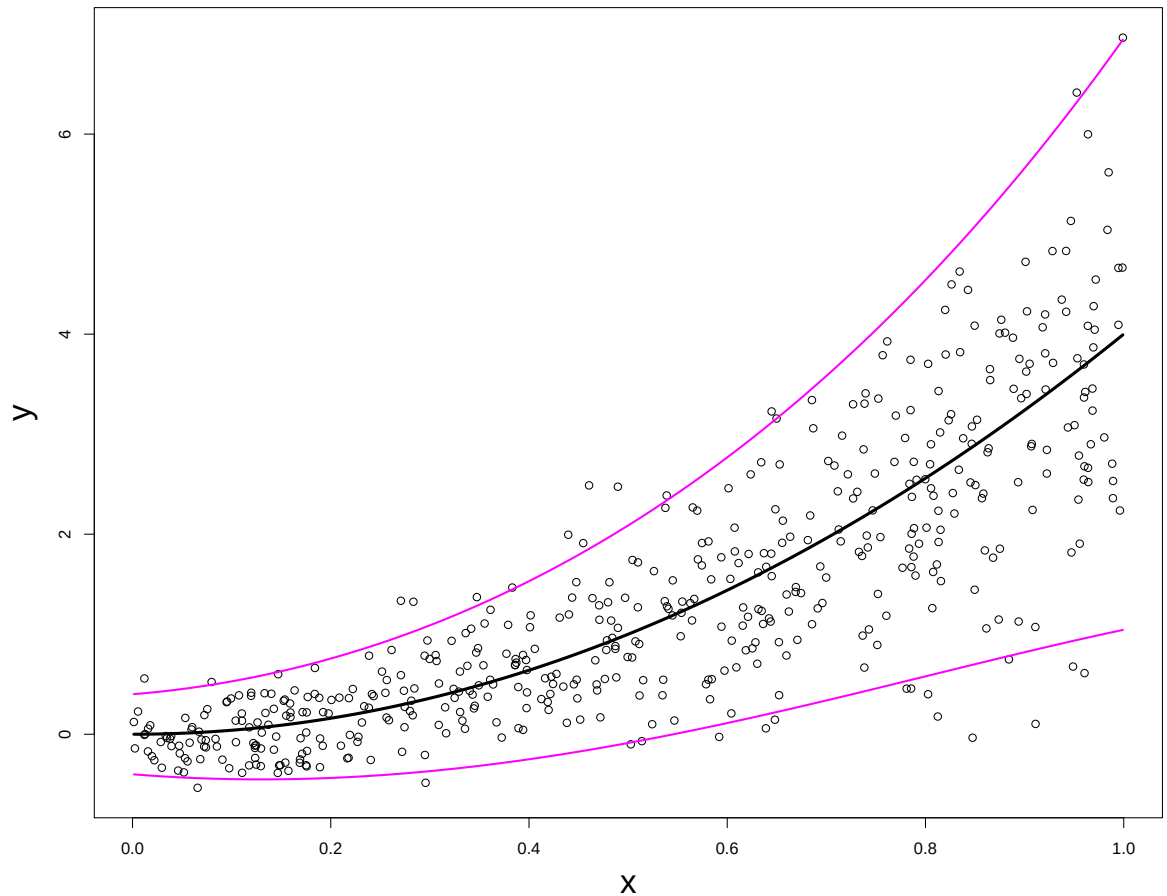


Matt Pratola has developed powerful moves for exploring the tree space:

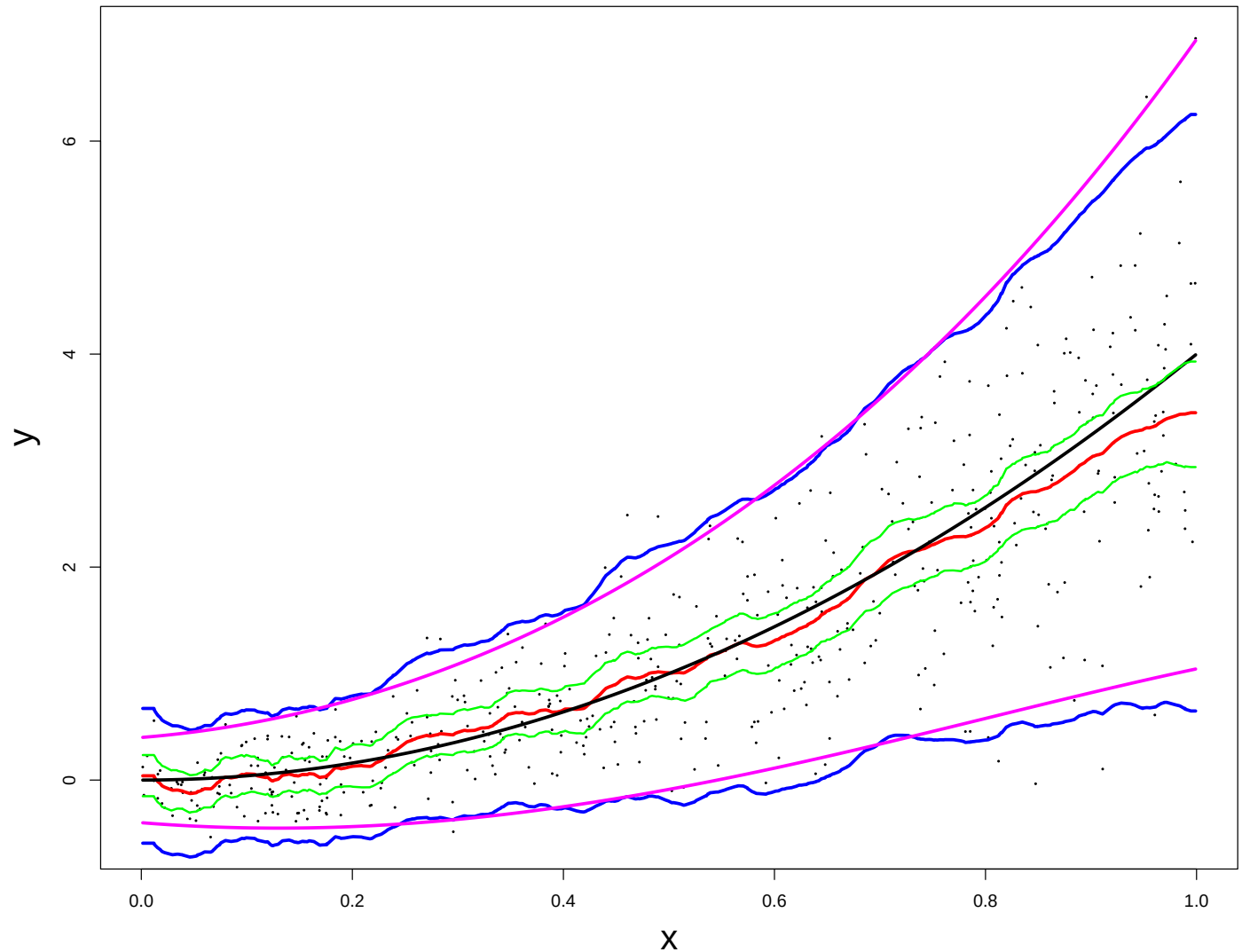


6. Simulated 1-d Example

simulated data with 1 x, black is true $f(x)$, magenta is $f(x) \pm 2\, g(x)$.

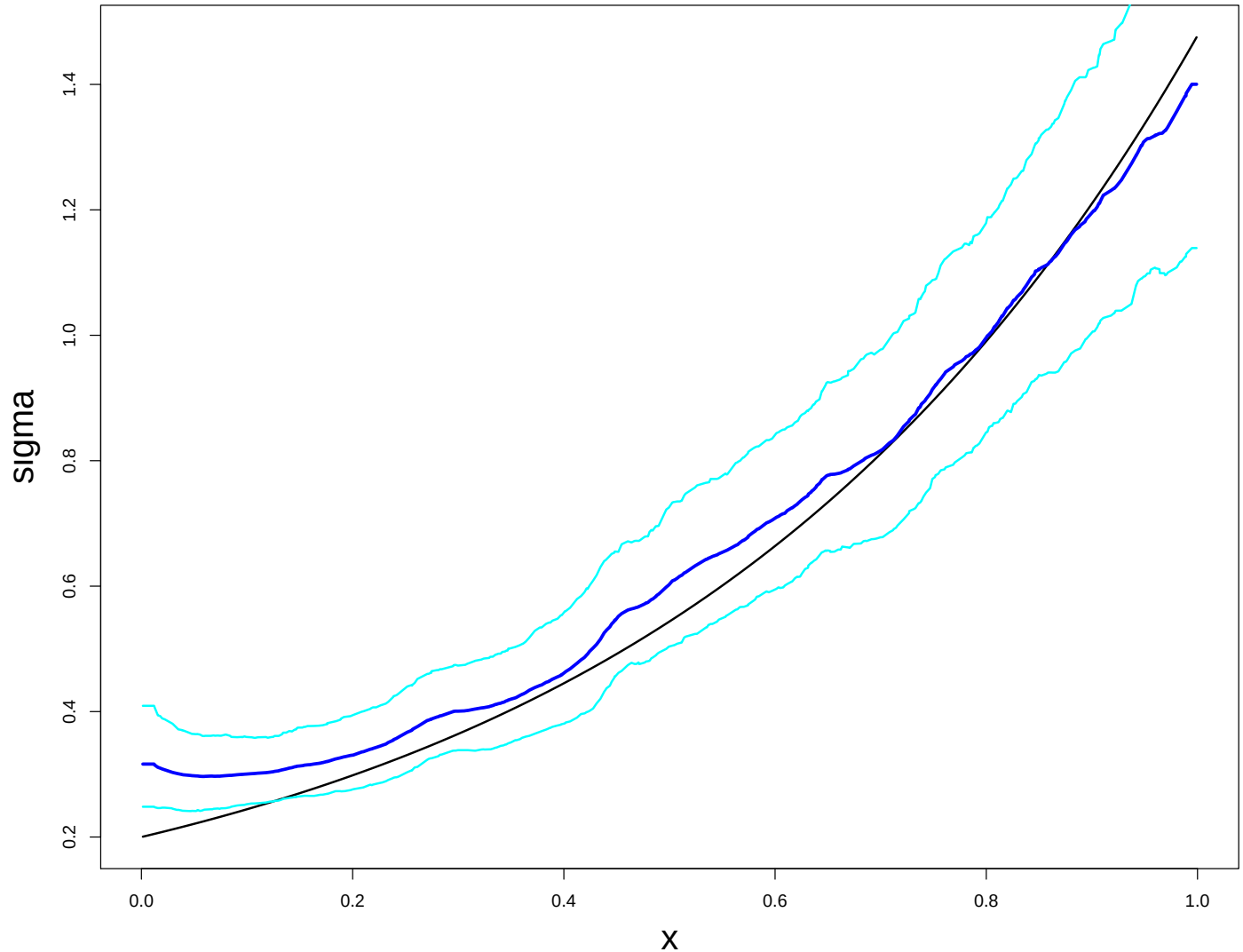


red: $\hat{f}(x)$ (posterior mean),
green: pointwise 90% posterior intervals for $f(x)$,
blue: $\hat{f}(x) \pm 2\,\hat{g}(x)$.



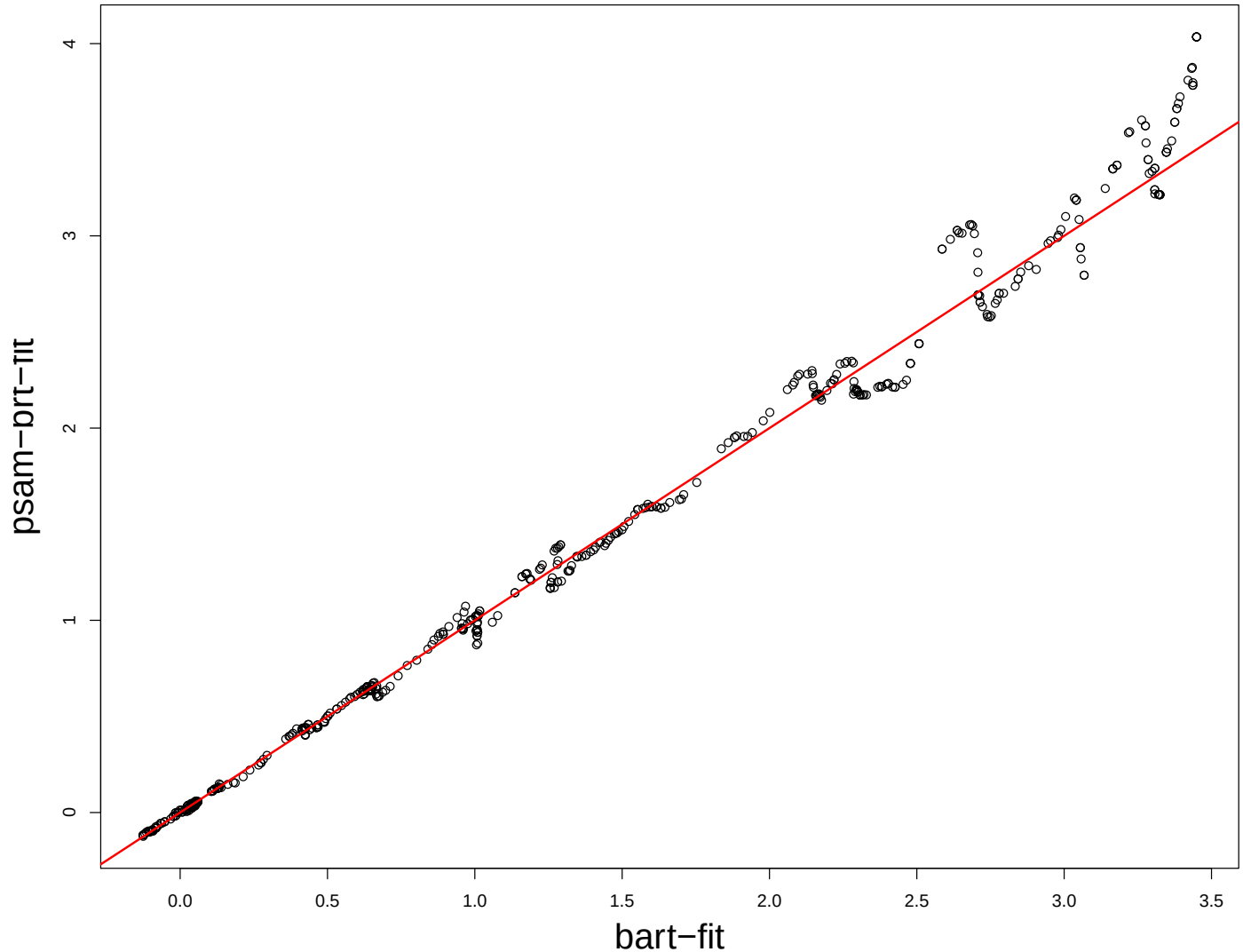
blue: $\hat{g}(x)$.

cyan: pointwise 90% posterior intervals for $g(x)$



bart (constant σ) fit $\hat{f}(x)$ vs.

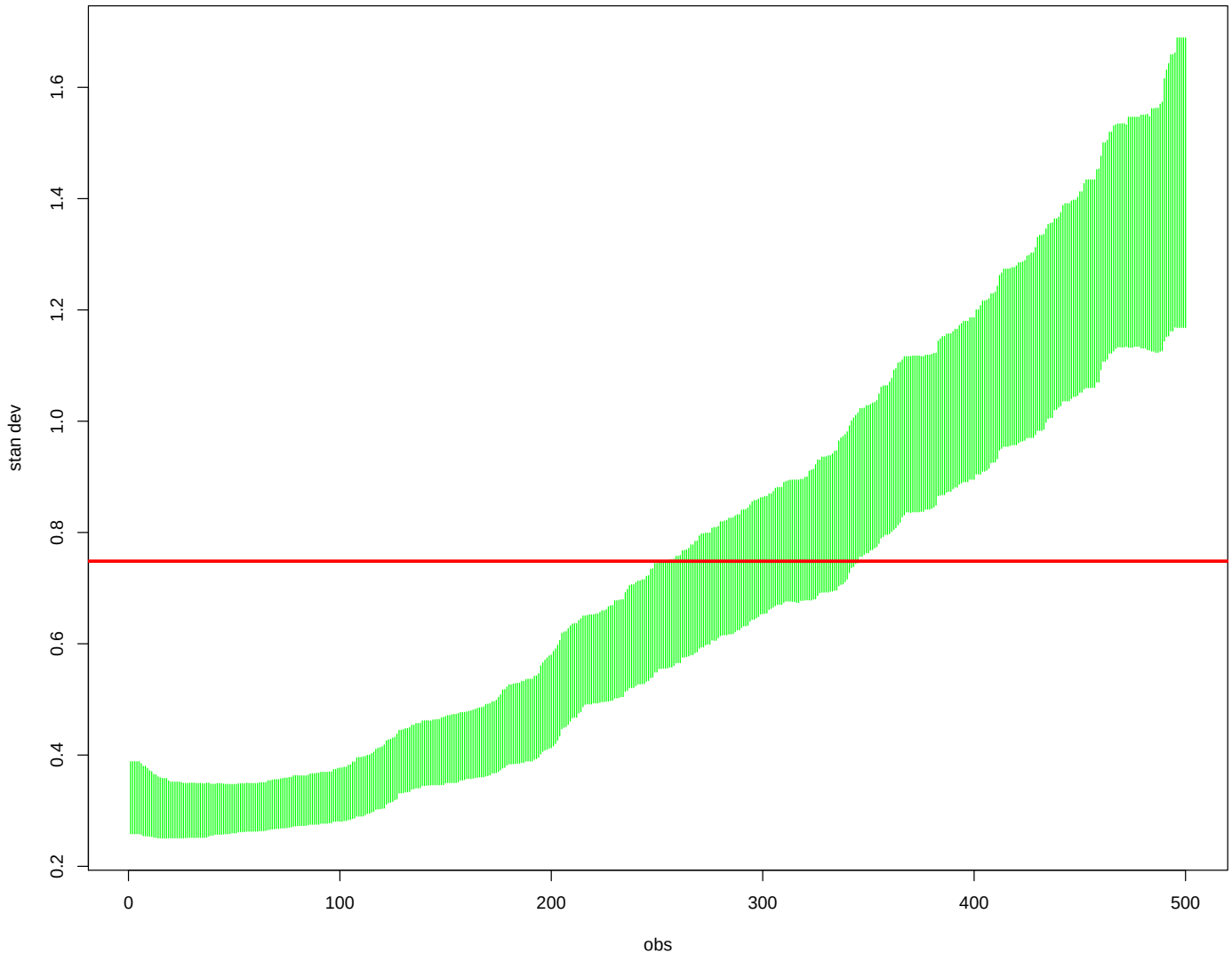
"psam" (product of stan devs, addition of means) fit.



Will estimates for $f(x)$ be different enough in practice to make point estimates better ?

red: bart estimate of σ.

green: point wise 90% intervals for $g(x)$, ordered by $\hat{g}(x)$.

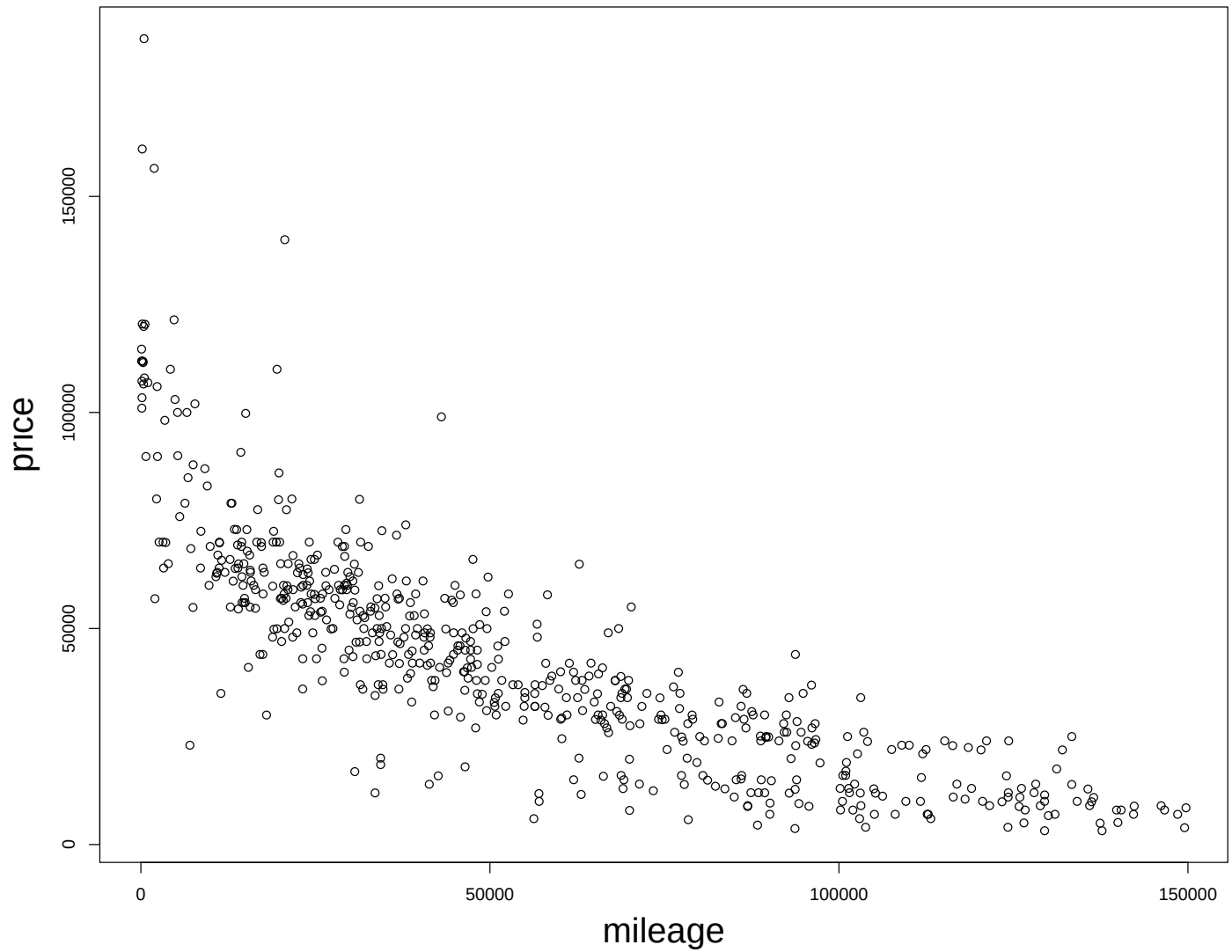


Clear evidence of hetero.

7. Real 1-d Example

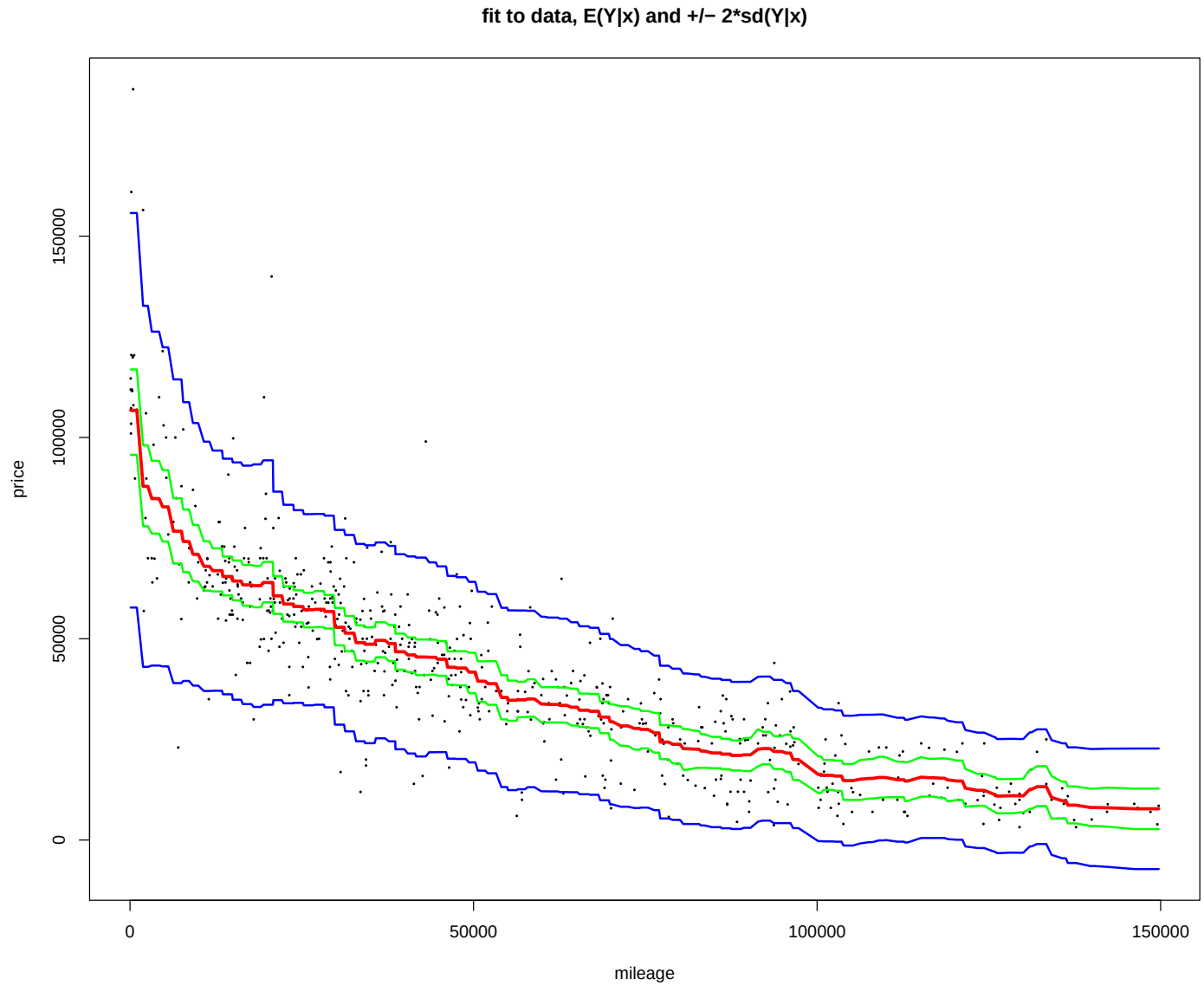
Each observation corresponds to the sale of a used car.

Y = price, x=mileage.

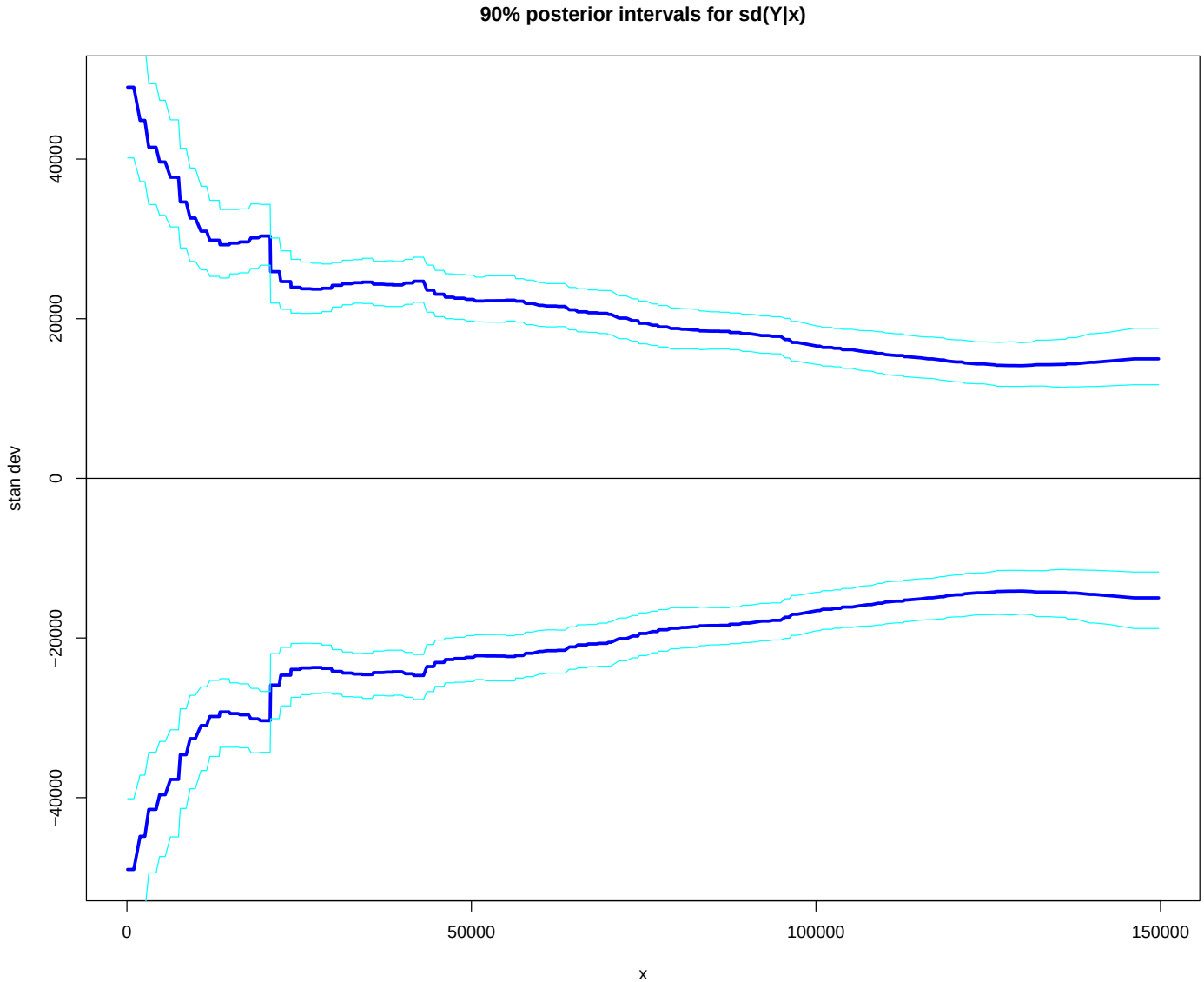


red: $\hat{f}(x)$.
blue: $\hat{f}(x) \pm 2\hat{g}(x)$.

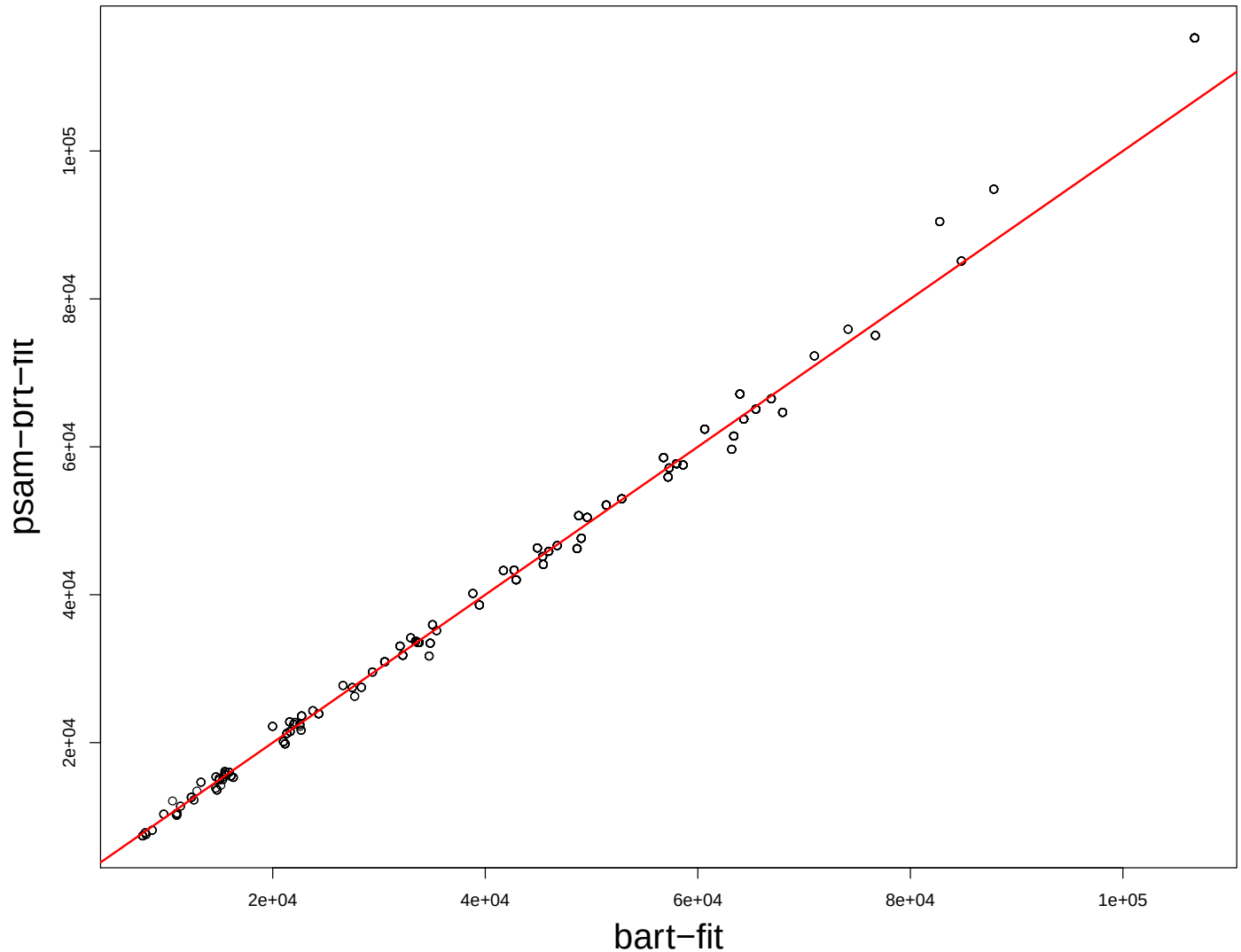
green: posterior intervals for f.



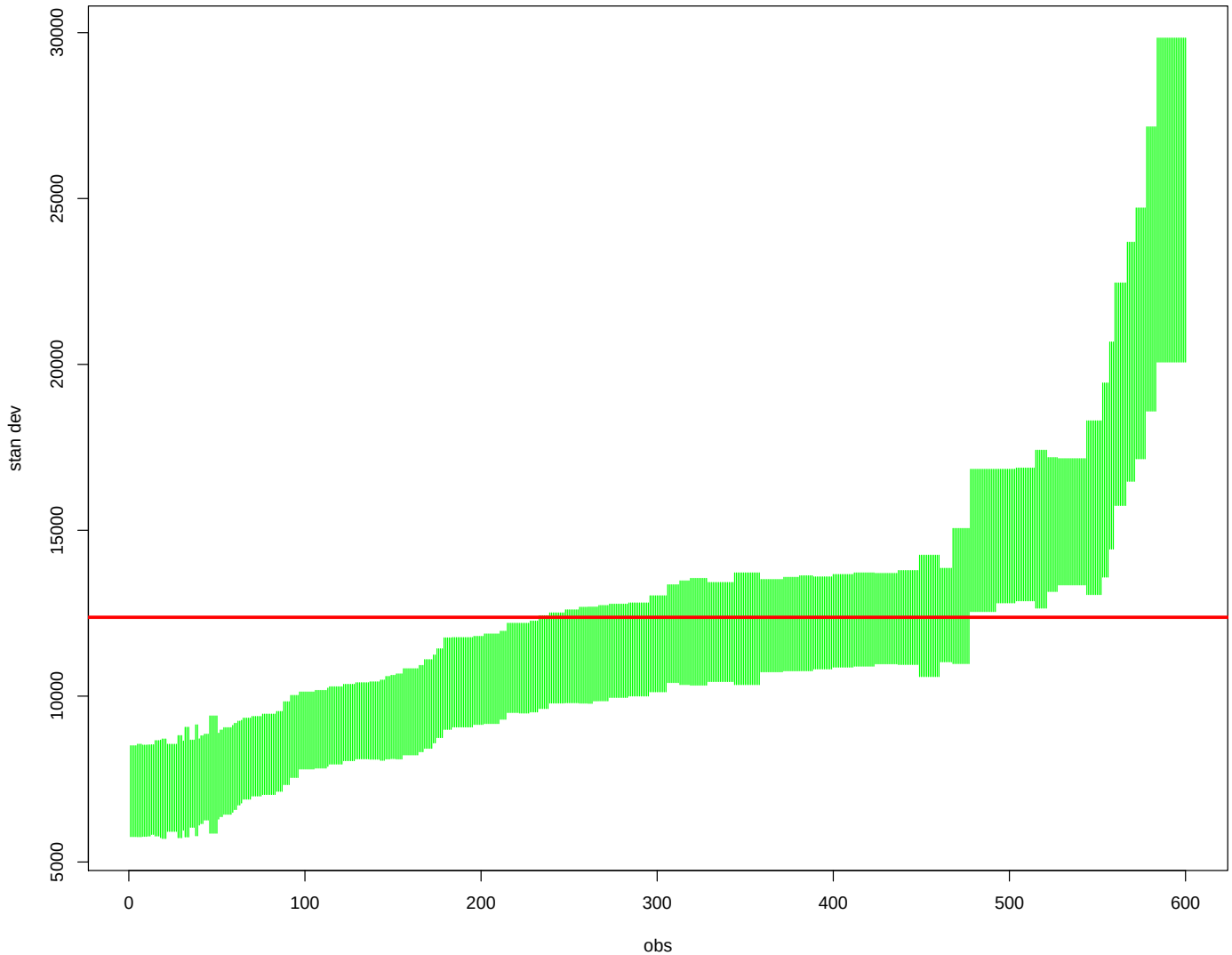
$\pm 2\,\hat{g}(x)$ and intervals.



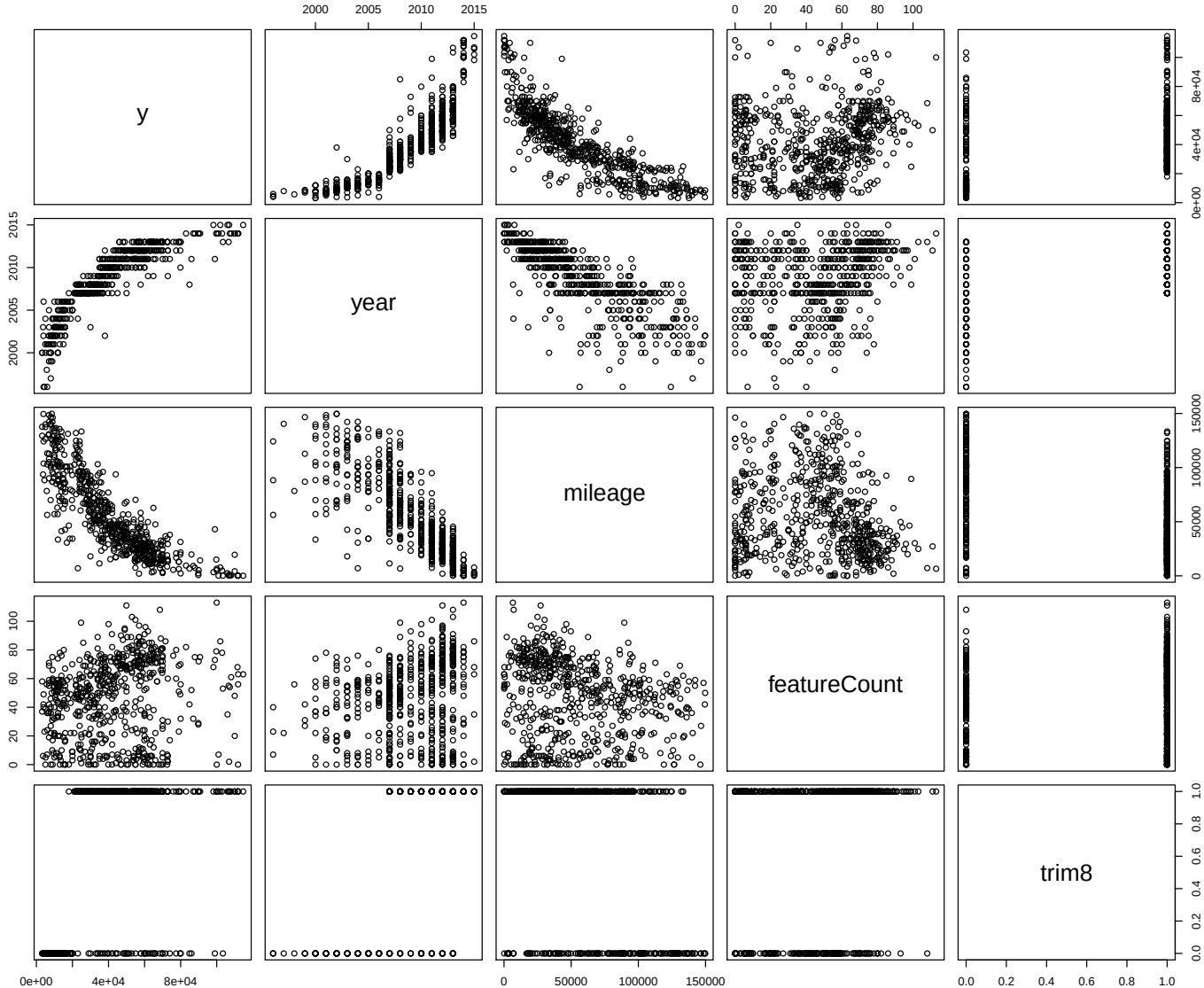
Fit for f very similar to bart.



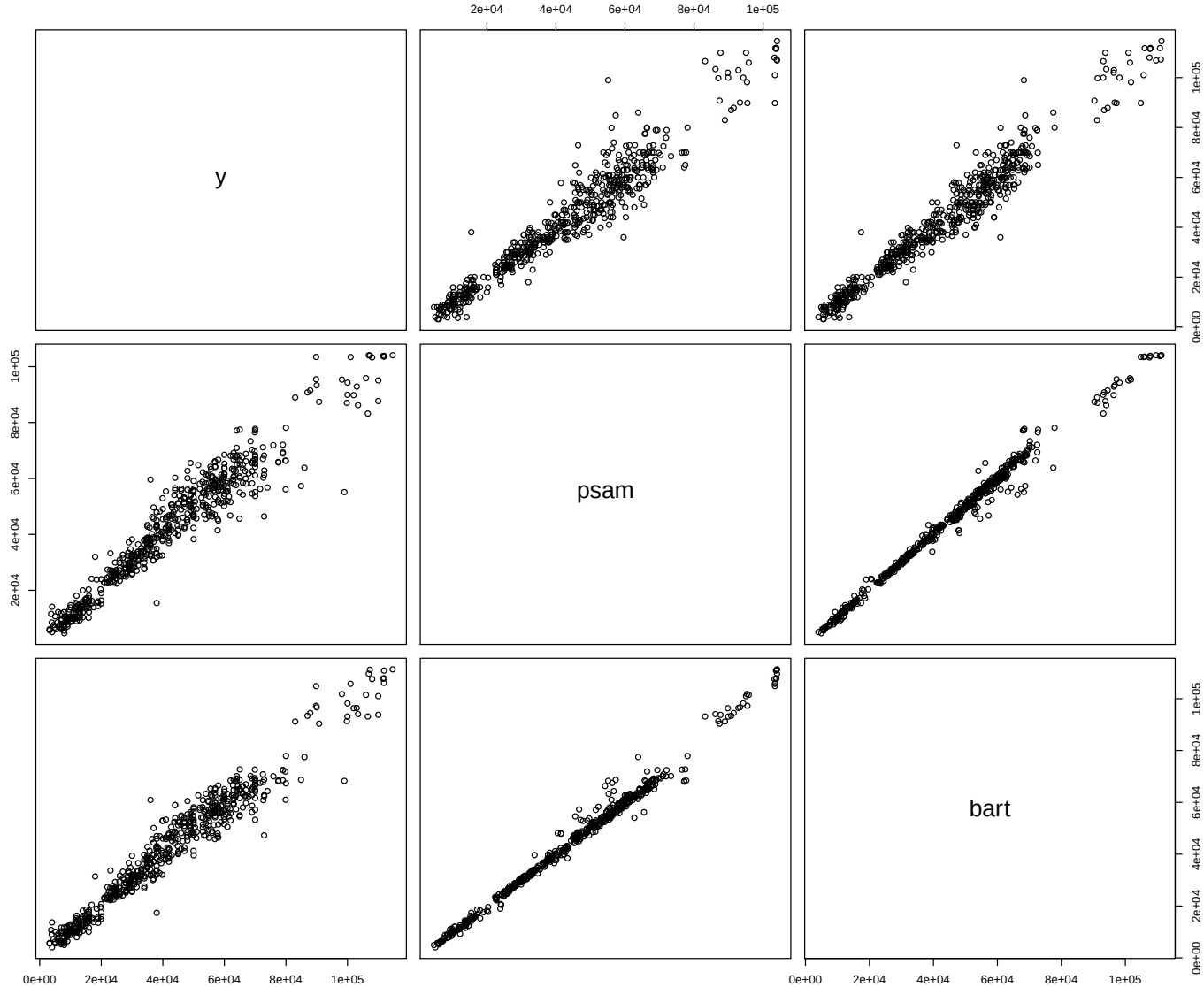
Clear evidence of heter.



8. Real 4-d Example

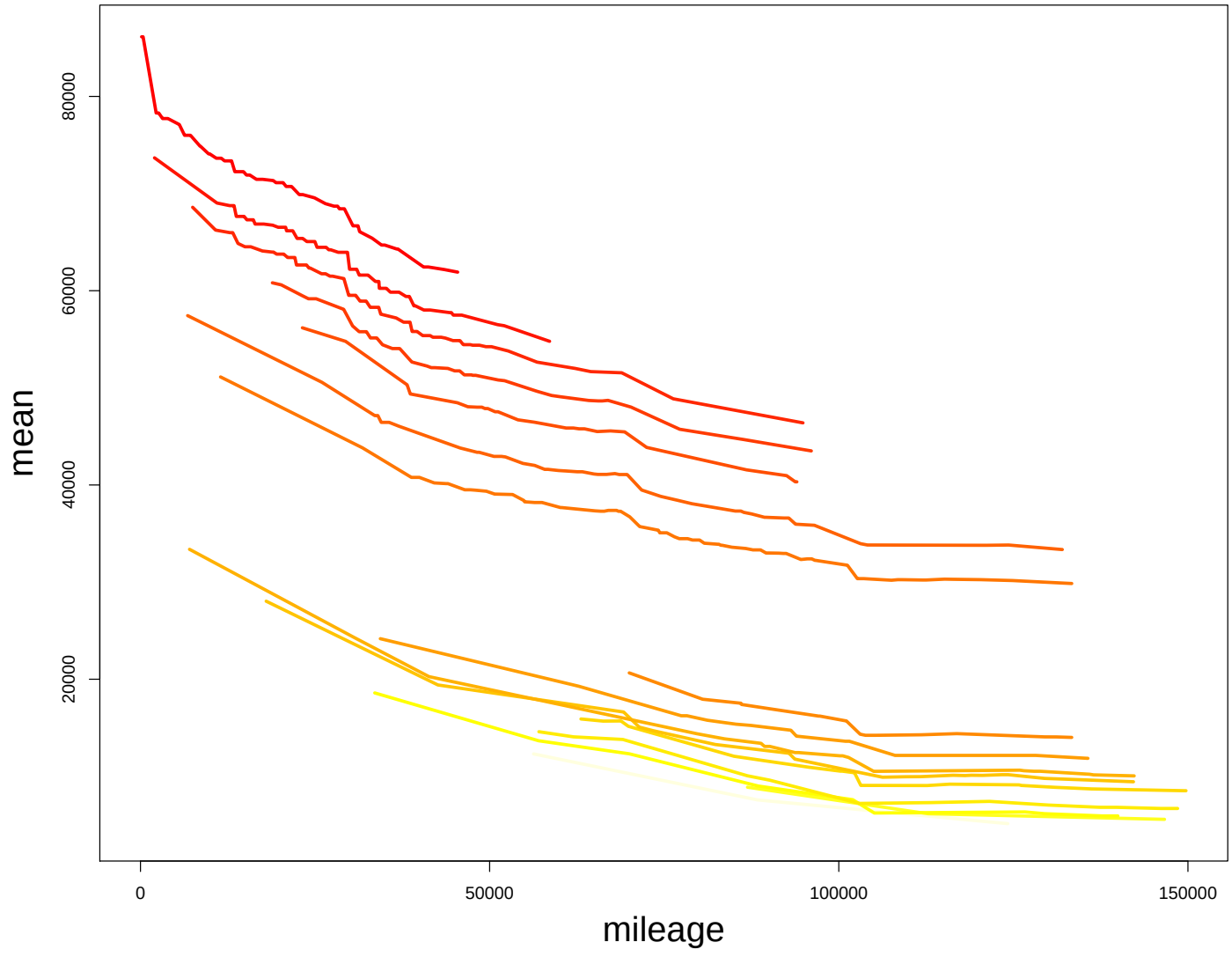


$\hat{f}$ very similar to bart.

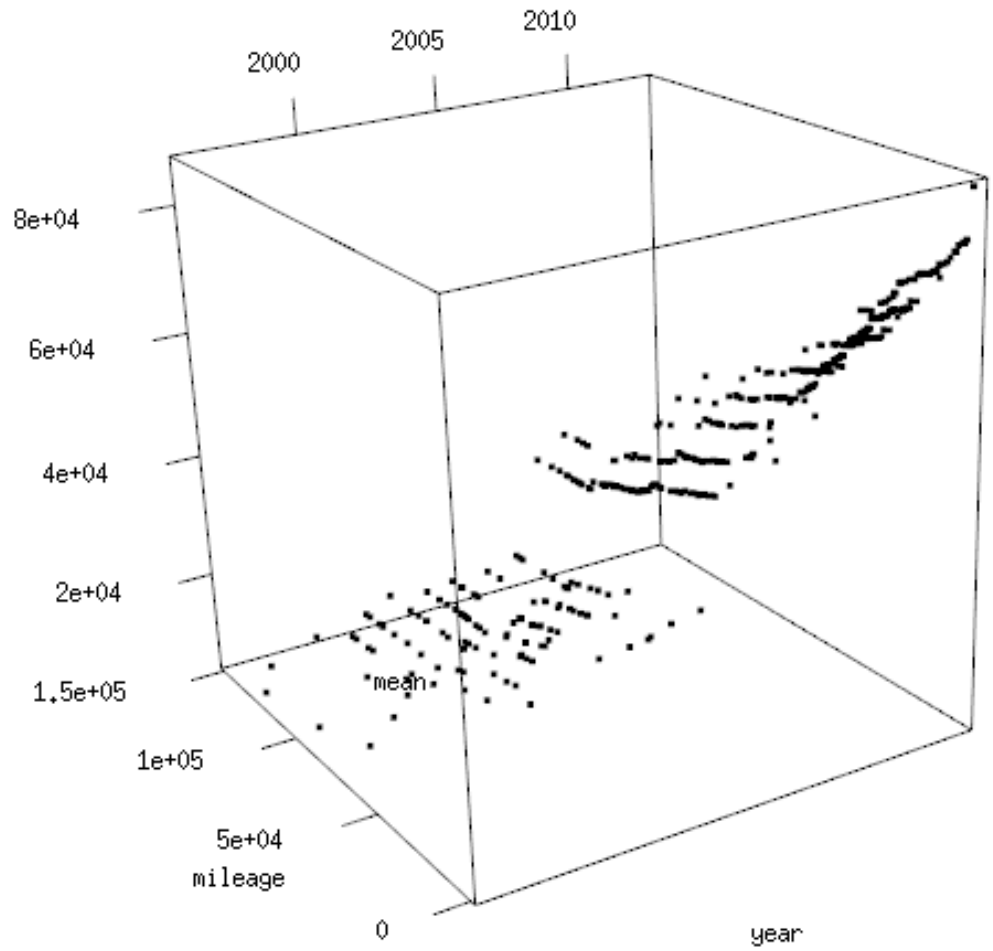


To visualize fixed `featureCount` and `trim8` and then used sample `year` and `mileage`.

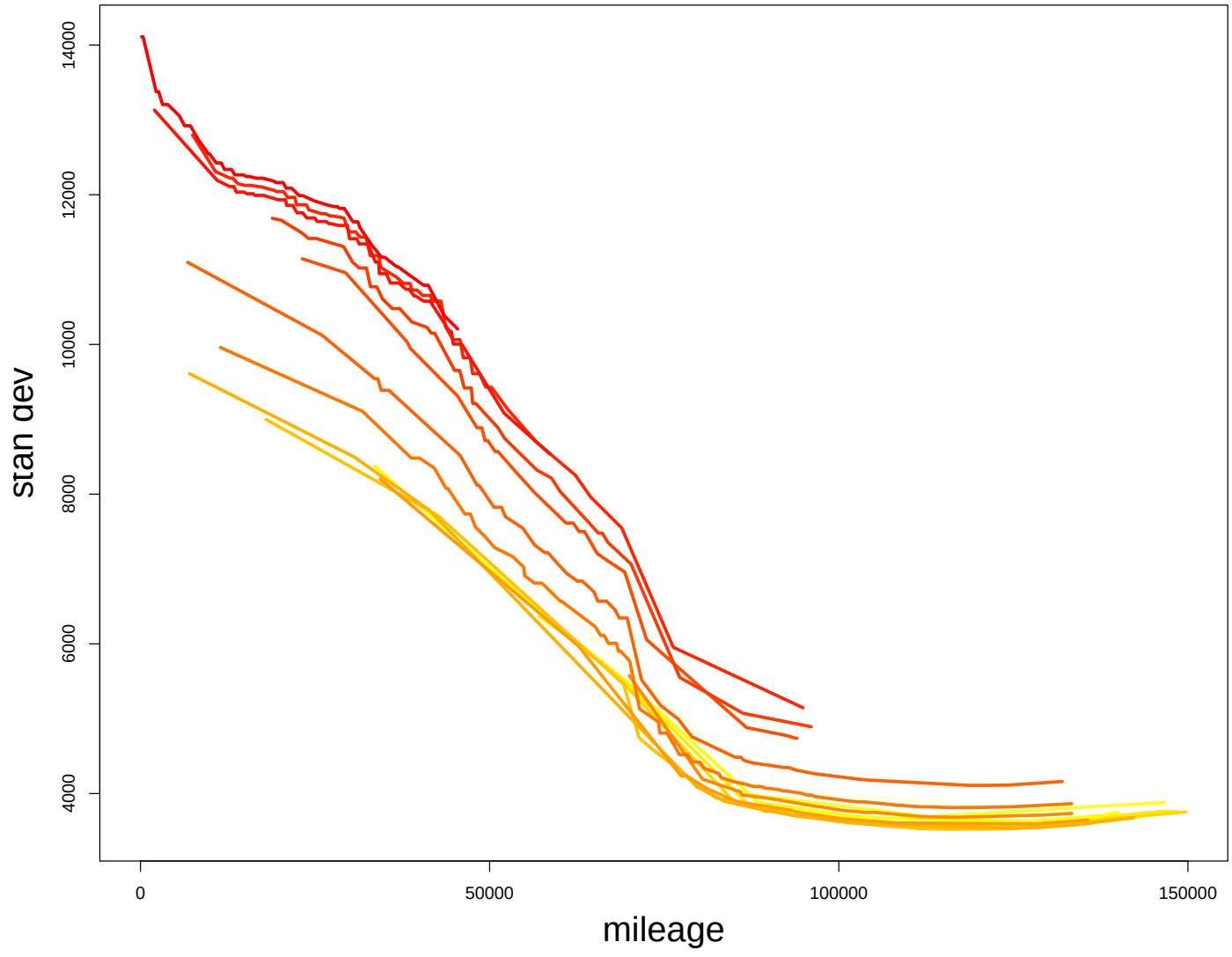
$\hat{f}(x)$ vs mileage for different years.



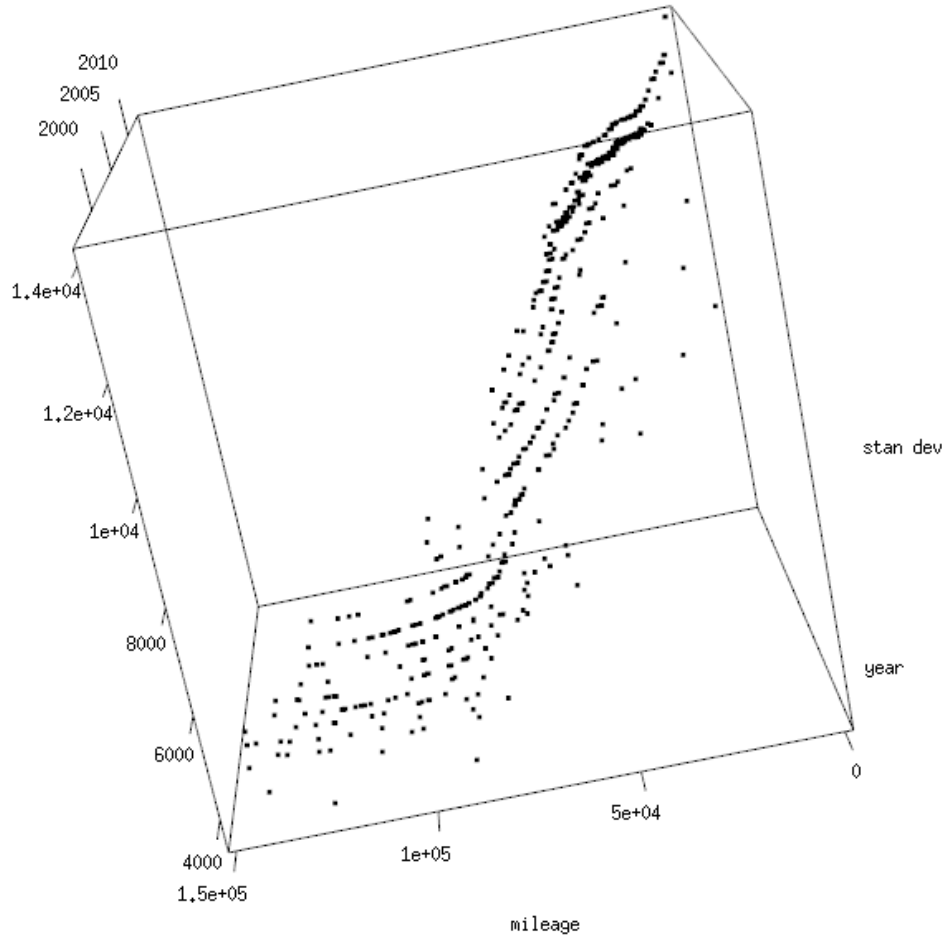
$\hat{f}(x)$ vs mileage and years.



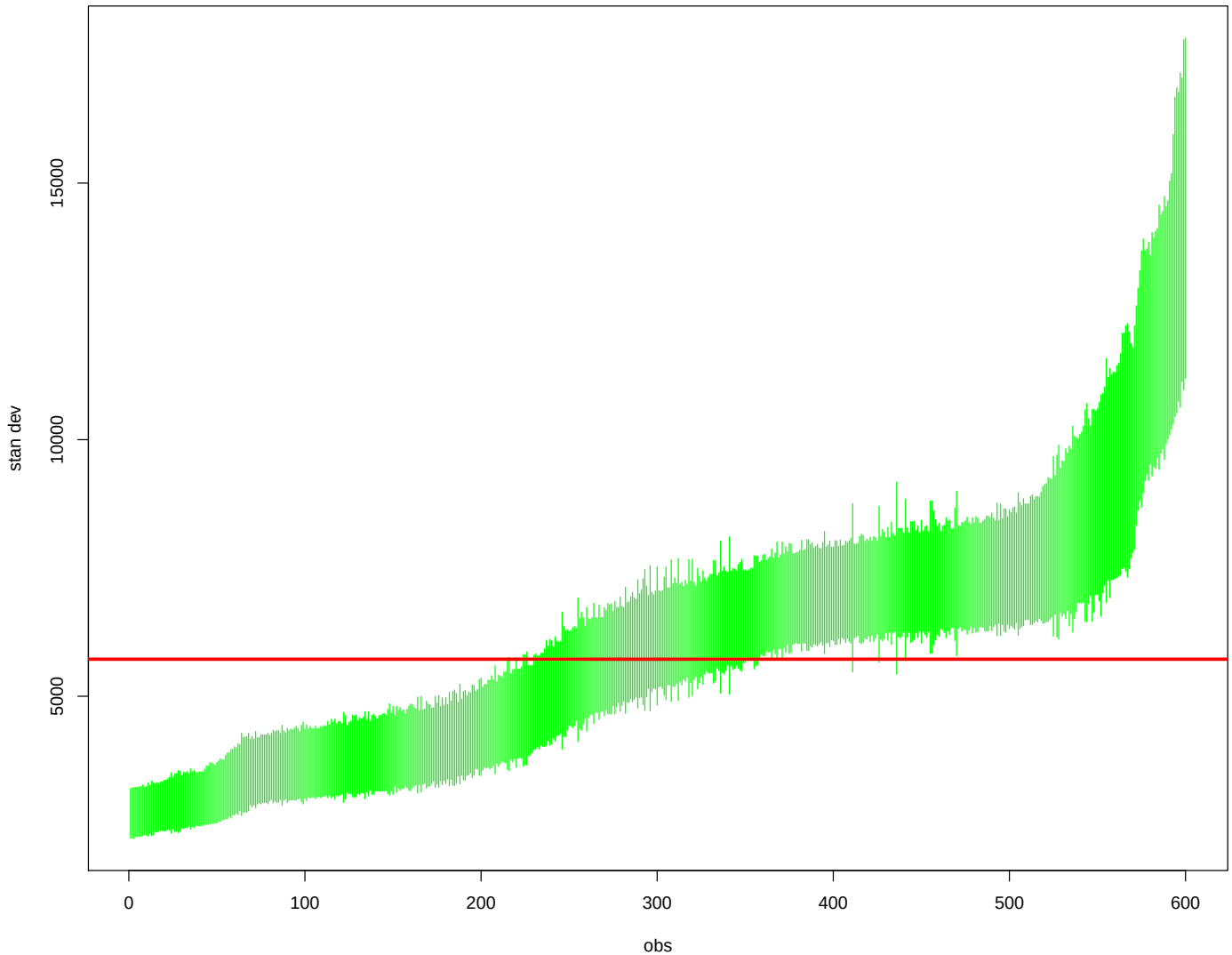
$\hat{g}(x)$ vs mileage for different years.



$\hat{g}(x)$ vs mileage and years.



Evidence of hetero.

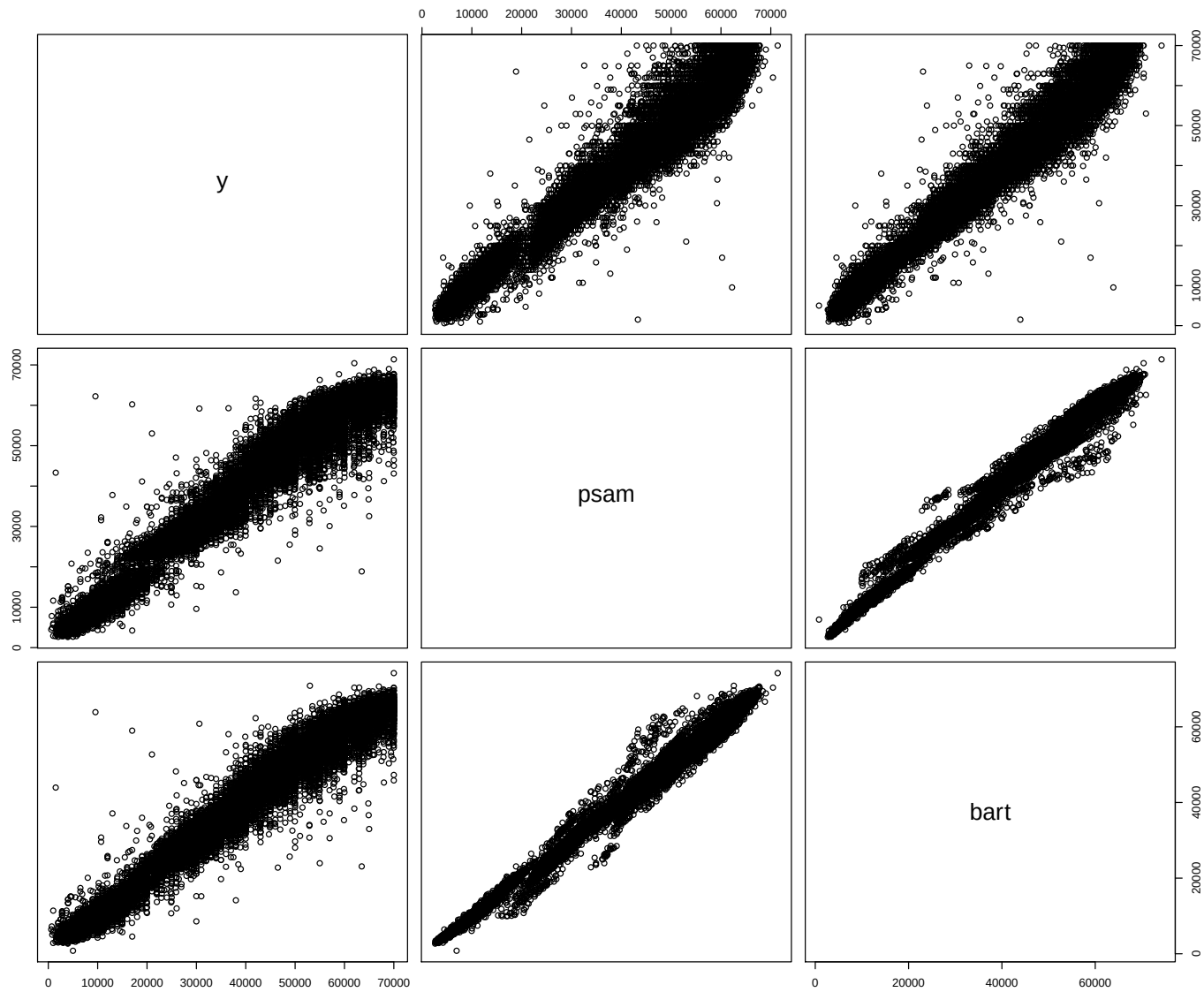


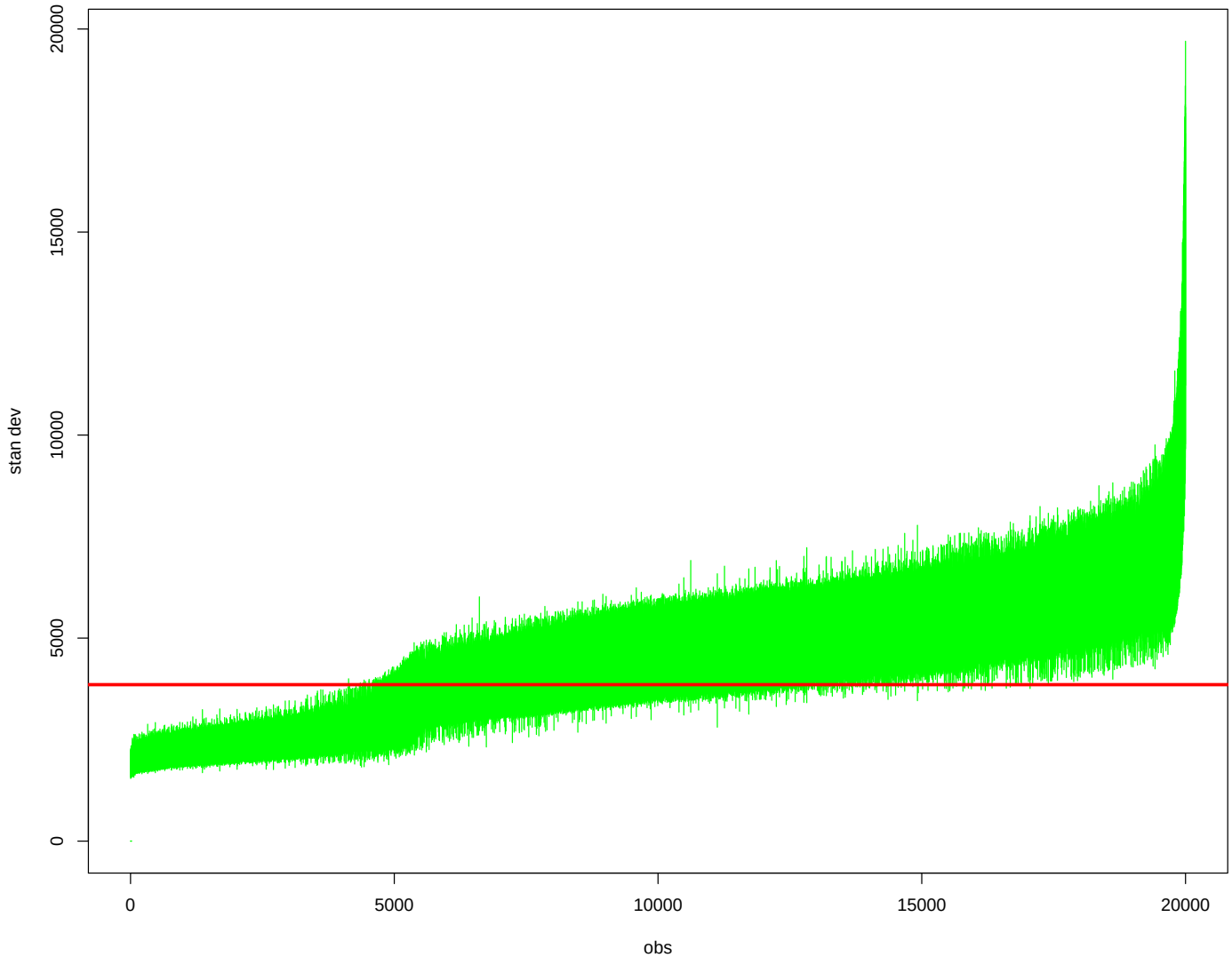
9. Real 90-d Example

n=20,000.

p=90.

Car data before variable selection and random subsampling.


y
psam
bart
0
10000
20000
30000
40000
50000
60000
70000



outliers + smaller variance for cheap old cars?

10. Concluding Remarks

The basic BART ideas lend themselves to the development of a rich class on flexible Bayesian models.

See rbart.pdf vignette for heterobart.

Not too complicated, and much richer than a lot of ML tools !!!!!!!!