Supplement to Heteroscedastic BART Via Multiplicative Regression Trees

M. T. Pratola[*], H. A. Chipman[†], E. I. George[‡], and R. E. McCulloch[§]

August 27, 2019

---

[*] Department of Statistics, The Ohio State University, 1958 Neil Avenue, 404 Cockins Hall, Columbus, OH 43210-1247 (mpratola@stat.osu.edu).

[†] Department of Statistics, Acadia University.

[‡] Department of Statistics, The Wharton School, University of Pennsylvania.

[§] School of Mathematical and Statistical Sciences, Arizona State University.

1 Simulated Example

Here we provide additional plots for the simulated dataset example discussed in the paper.

Figure 1 informally examines the performance of the MCMC chain by displaying sequences of post burn-in draws for certain marginals. The top panel displays the draws of σ from the homoscedastic BART model $Y = f(x) + \sigma Z$. For BART, this plot is a simple way to get a feeling for the performance of the MCMC. We can see that the draws vary about a fixed level with an appreciable but moderate level of autocorrelation. For HBART, there is no simple summary of the overall error level comparable to the homoscedastic σ. The middle panel plots draws of $s(x)$ for $x = .12, .42, .63, .79, .91$. The bottom panel plots the draws of $\bar{s} = \frac{1}{n}\sum s(x_i)$, the average s value for each MCMC draw of s. Here the x_i are from the test data. In all plots, the MCMC appears to be reasonably well behaved.

2 Cars Example

Here we provide additional plots for the cars example discussed in the paper.

The relationship between active categorical predictors and the response variable `price` and continuous predictors `mileage` and `year` are summarized in Figure 2. Note that the categorical predictor `color` does not appear in this figure as it has little marginal effect on the response.

A plot of the log-transformed `Price` versus the continuous predictors is shown in Figure 3.

3 Million Songs Example

Here we provide additional plots for the million songs example discussed in the paper.

First, both the BART and HBART models underwent 10k iterations of burn-in, after which 5k iterations were saved as samples from the posterior. Plots of σ and the average of $s(x)$ are shown for the BART and HBART posterior samples in Figure 4. The posterior plots suggest both models are reasonably converged. Figure 4 also demonstrates the wide difference between the models: BART's average estimate of σ around 8.8 is about 11% large than the average of the $s(x)$ posterior samples at around 7.96. Meanwhile, the range of posterior samples in HBART for $s(x)$ varies from 1.04 to 51.9, suggesting that HBART provides much more confident predictions for some aspects of the dataset while accounting for great uncertainty in modeling other aspects of the dataset.

The H-evidence plot, shown in Figure 5, clearly confirms the presence of heteroscedasticity for the million songs dataset.

4 Fishery and Alcohol Examples

In this section we very briefly present results from two more examples. In the first example the dependent variable y is the daily catch of fishing boats in the Grand Bank fishing grounds (Fernandez et al., 2002). The explanatory $\mathbf{x}$ variables capture time, location, and characteristics of the boat. After the creation of dummies for categorical variables, the dimension of $\mathbf{x}$ is 25. In the second example, the dependent variable y is the number of alcoholic beverages consumed in the last two weeks (Kenkel and Terza, 2001). The explanatory $\mathbf{x}$ 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 $\mathbf{x}$ is 35.

In both of the examples the response is constrained to be positive and there is a set of observations with $y = 0$ so that there is a clear sense in which our model $Y = f(\mathbf{x}) + s(\mathbf{x})Z$ does not account for these features of Y. In both previous papers, careful modeling was done to capture the special nature of the dependent variable. Our interest here is to see how well our model can capture the data given our flexible representations of f and s in the presence of a clear mispecification.

Figures 6 and 7 present the results for the fish data using the same displays we have employed in our previous examples. In Figure 6 we see very strong evidence of heteroscedasticity. Our product of trees representation of s enables the model to represent the data by being quite certain that for some $\mathbf{x}$ the error standard deviation should be small. In Figure 7 we see the (in-sample) qqplots. While the qqplot for the HBART model is not perfect, it is a dramatic improvement over the homoscedastic fit and may be sufficiently accurate for practical purposes. In Fernandez et al. (2002) a particular feature of the data (a lump of responses at zero) was noted and a model was developed to capture this feature. For numerical summaries, we find the RMSE is $4,139$ for HBART and $3,883$ for BART while the e-statistic is 3.19 for HBART and 14.42 for BART.

In the left panel of Figure 7 we have also plotted the qqplot obtained from the plug-in model $Y \sim N(\hat{f}(\mathbf{x}), \hat{s}(\mathbf{x})^2)$. This is represented by a dashed line. It is difficult to see because it coincides almost exactly with the qqplot plot obtained from the full predictive distribution.

Our feeling is that in many applications the representation $Y \sim N(\hat{f}(\mathbf{x}), \hat{s}(\mathbf{x})^2)$ may be adequate and has an appealing simplicity. Many users will be able to understand this output easily without knowledge of the representations of f and s.

Figures 8 and 9 give results for the Alcohol data again using the same format. In this example the inference suggests that the homoscedastic version is adequate and the (in-sample) qqplots are very similar. In this case, even without the heteroscedastic model the flexible f captures the patterns reasonably well, although the qqplots are not perfect. Here the RMSE is 1.338 for HBART and 1.339 for BART while the e-statistic is 2.50 for HBART and 2.26 for BART.

Acknowledgments

This research was partially supported by the US National Science Foundation grants DMS-1106862, 1106974 and 1107046, the STATMOS research network on Statistical Methods in Oceanic and Atmospheric Sciences. H. A. Chipman acknowledges support from the Natural Sciences and Engineering Research Council of Canada. E. I. George acknowledges support from NSF grant DMS-1406563.

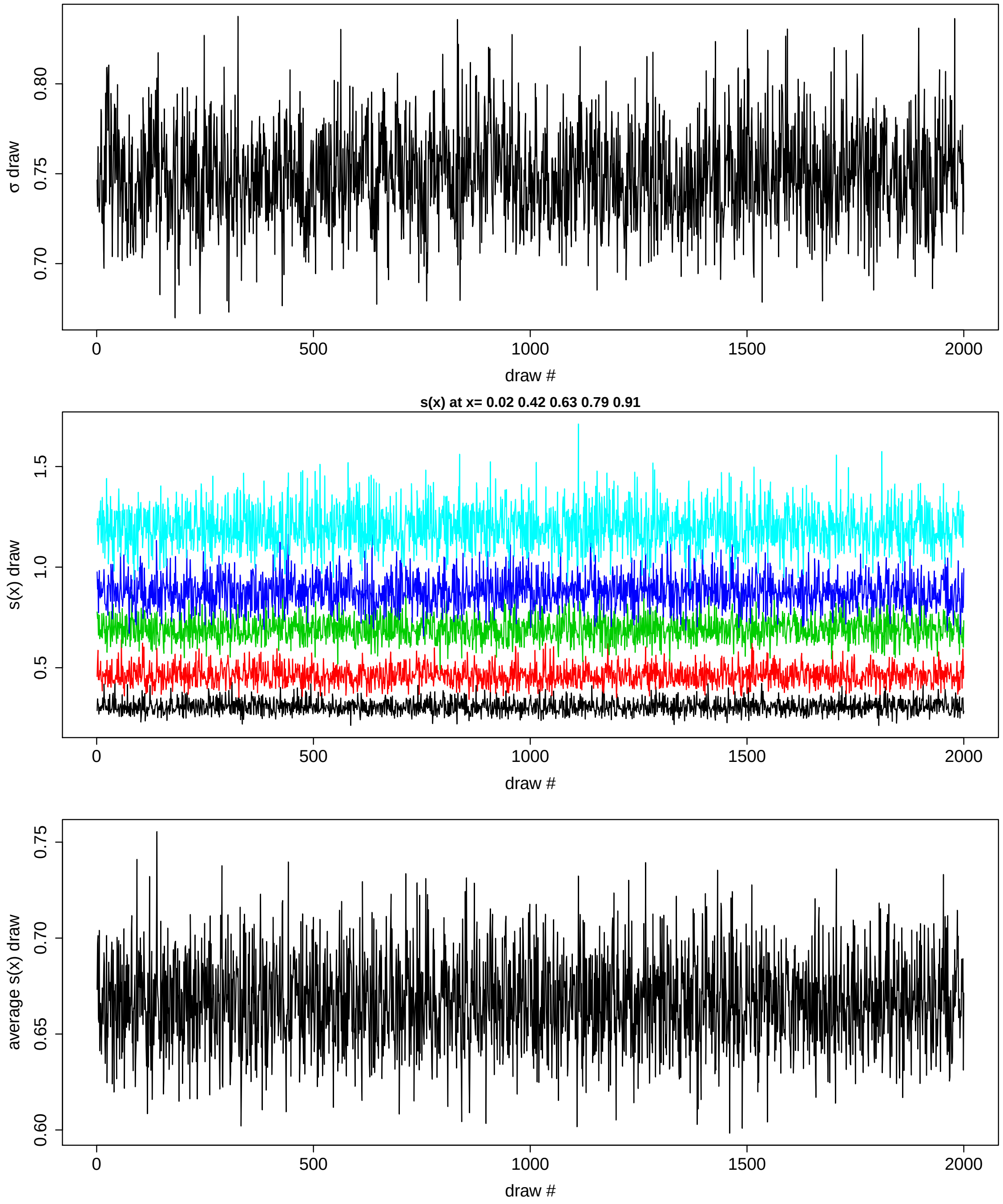

Figure 1: Simulated example. Top panel: MCMC draws of σ in homoscedastic BART. Middle panel: MCMC draws of $s(x)$ for five different x in HBART. Bottom panel: MCMC draws of $\bar{s}$ the average of $s(x_i)$ for each MCMC draw in HBART.

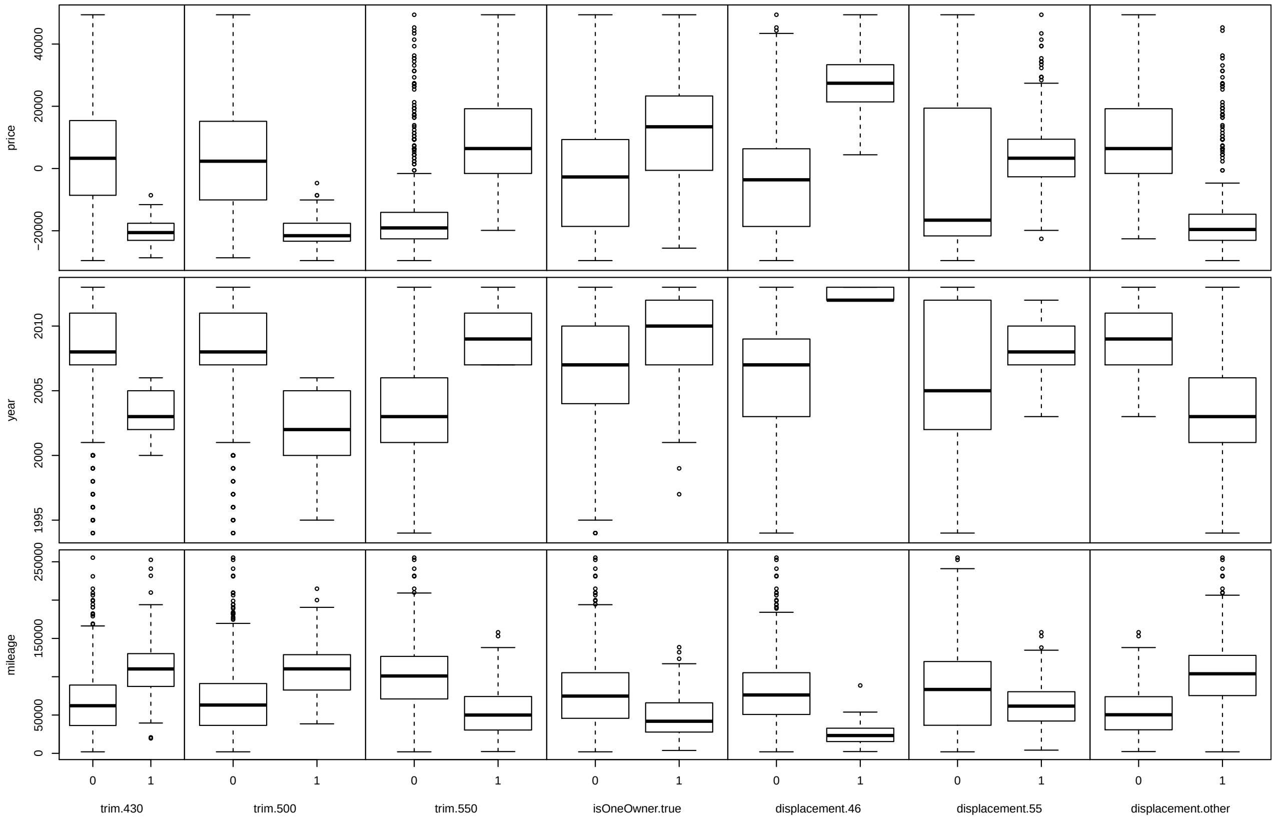

Figure 2: Used cars example. Summary of response variable `price` and continuous predictors `mileage` and `year` by the levels of the important categorical predictors `trim`, `isOneOwner` and `displacement`.

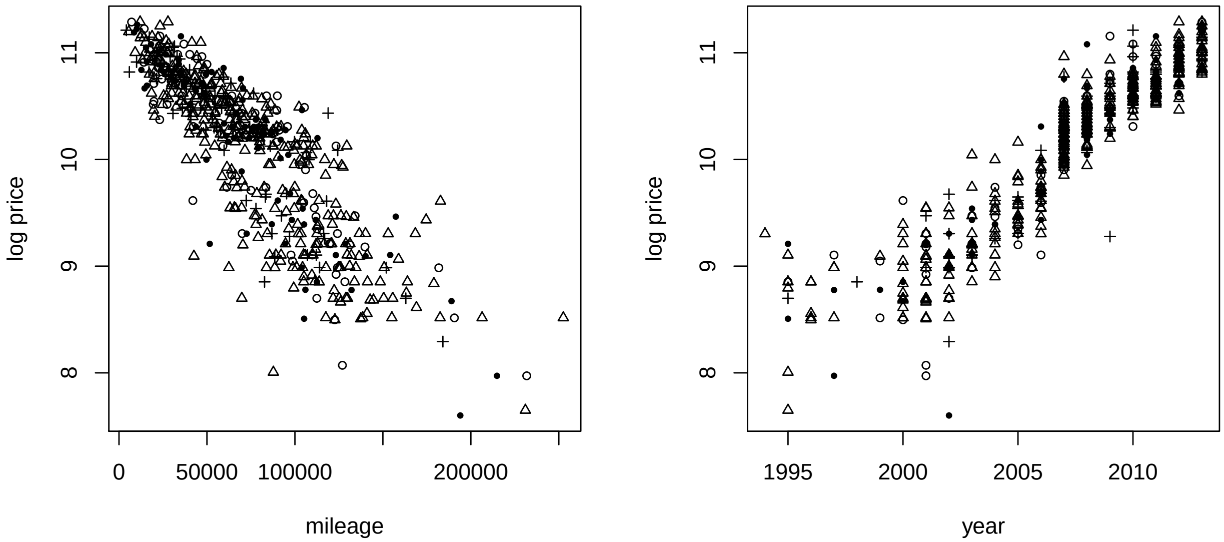

Figure 3: Used cars example. Summary of log(`price`) and other continuous variables coded by level of `trim`. `trim.430` shown by solid dots, `trim.500` by '+', `trim.550` by triangles and `trim.other` by 'o'.

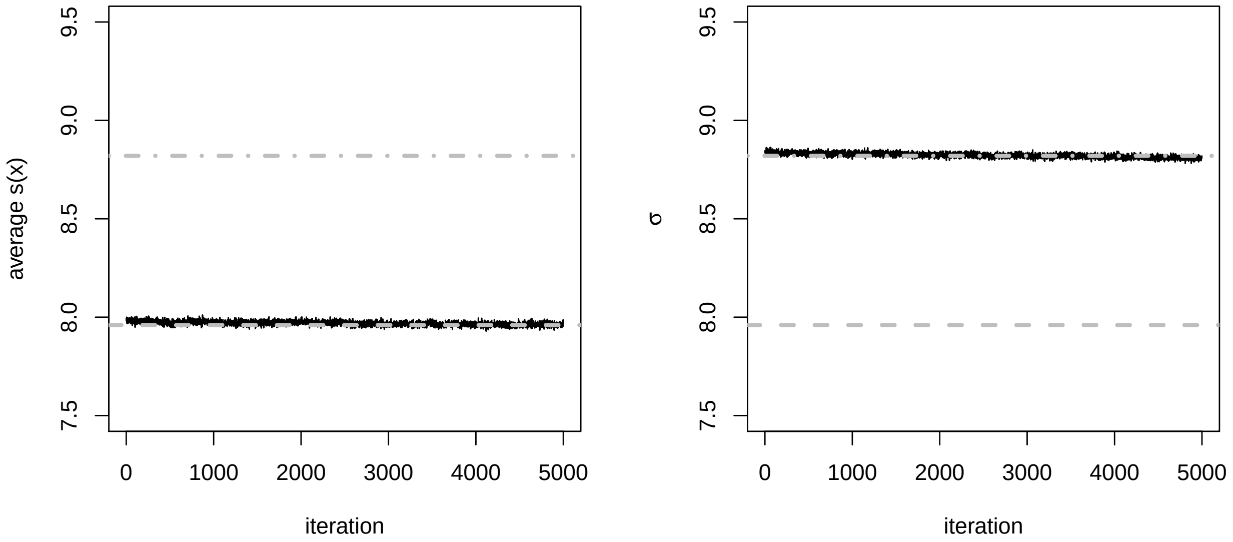

Figure 4: Million songs example. Posterior samples of the average (over x) of $s(x)$ from HBART (left panel) and posterior samples of σ from BART (right panel). The dashed line represents the average of the posterior HBART samples while the dashed-dotted line represents the average of the posterior BART samples.

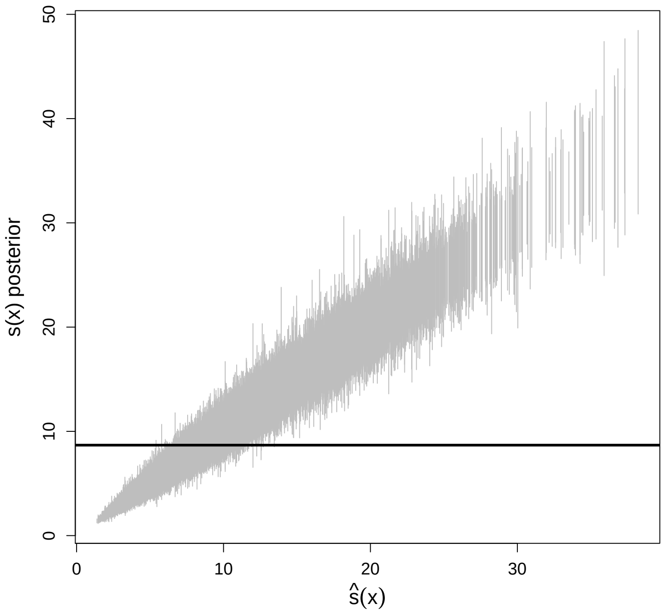

Figure 5: Million songs example. H-evidence plot. Posterior intervals for $s(x_i)$ sorted by $\hat{s}(x_i)$. The solid horizontal line is drawn at the estimate of σ obtained from fitting homoscedastic BART (the credible interval is too narrow to visualize on the scale of this plot).

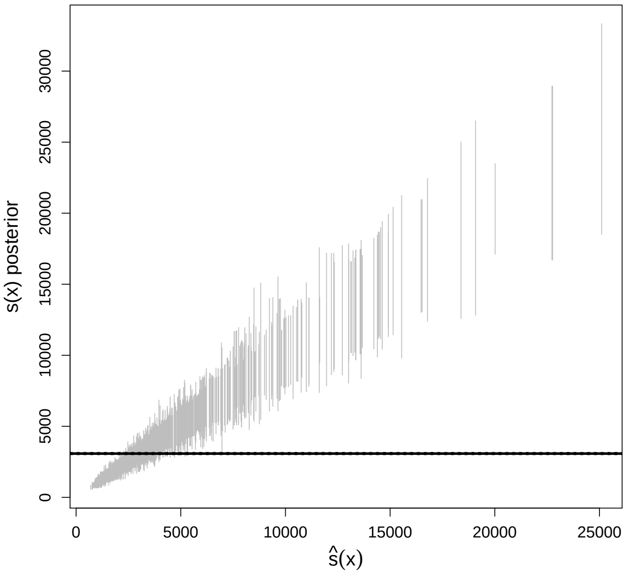

Figure 6: Fishery example. H-evidence plot. Posterior intervals for $s(\mathbf{x}_i)$ sorted by $\hat{s}(\mathbf{x}_i)$. The solid horizontal line is drawn at the estimate of σ obtained from fitting BART.

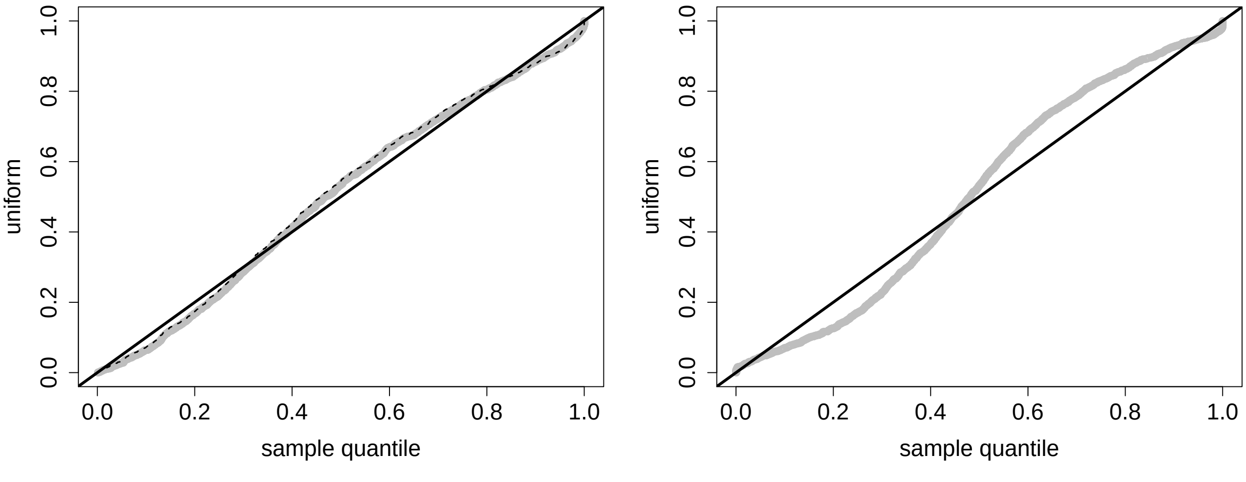

Figure 7: Fishery example. Predictive qq-plots. Left panel: HBART. Right panel: BART.

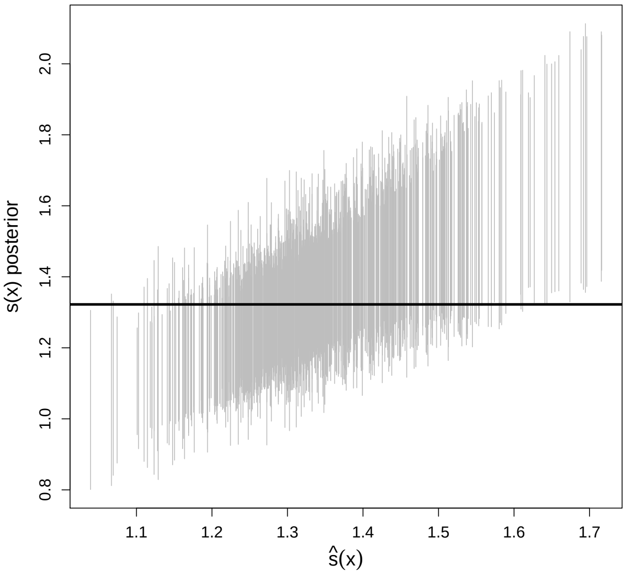

Figure 8: Alcohol example. H-evidence plot. Posterior intervals for $s(\mathbf{x}_i)$ sorted by $\hat{s}(\mathbf{x}_i)$. The solid horizontal line is draw at the estimate of σ obtained from fitting BART.

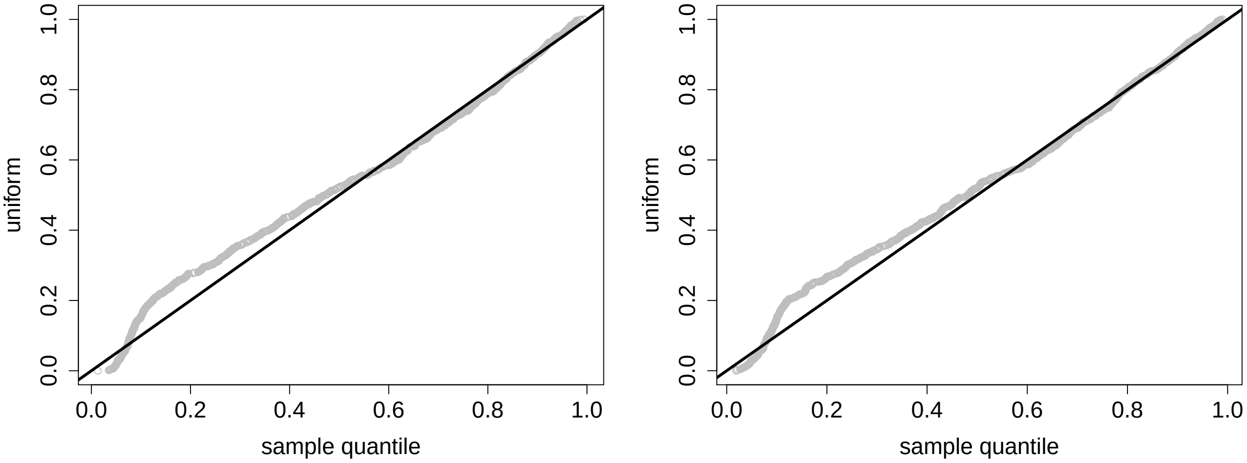

Figure 9: Alcohol example. Predictive qq-plots. Left panel: HBART. Right panel: BART.

References

Fernandez, C., Ley, E., and Steel, M. (2002). “Bayesian Modelling of Catch in a North-West Atlantic Fishery.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 51, 3, 257–280.

Kenkel, D. and Terza, J. (2001). “The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect.” *Journal of Applied Econometrics*, 16, 165–184.