Tree-Based Predictive Models for Noisy Input Data

Kevin McCoy

Rice University
Department of Statistics

Conference of Texas Statisticians
April 10 – 11, 2026

Decision Trees

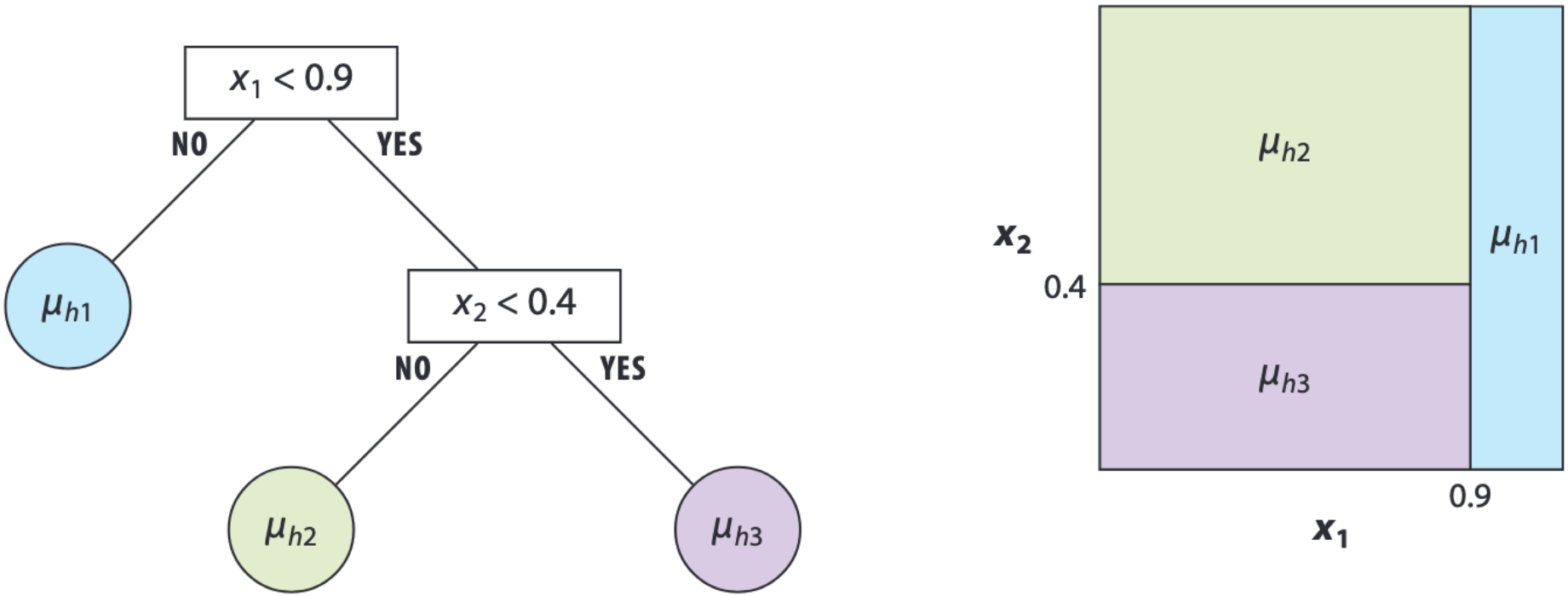


Figure: An example of a decision tree model, which can be visualized as a binary tree or a rectangular partition of the data space. Figure borrowed from [Hill et al., 2020].

[Breiman et al., 1984]

A Bayesian Ensemble of Trees

Bayesian Additive Regression Trees (BART) are a Bayesian nonparametric sum-of-trees model introduced by [Chipman et al., 2010]. We assume the true underlying function is:

$$Y = f(x) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Classic Regression

$$\mathbb{E}[y_i|x_i] \approx x_i^\top \beta$$

models conditional mean as linear combination of features

BART

$$\mathbb{E}[y_i|x_i] \approx \sum_{h=1}^{m} g(x_i; T_h, M_h)$$

models conditional mean as sum of decision trees

Measurement Error

Definition

Measurement error refers to the idea that a particular variable in a statistical model cannot be quantified with precision. Instead, only a rough surrogate for that value can be obtained.

- Measurement error may refer to random noise, sampling error, or variation associated with the measuring process itself [Yi et al., 2021].
- Incredibly common in fields such as epidemiology, environmental science, economics, and survey sampling.
- Many statistical models allow for uncertainty in the outcome variable, but typically assume that the predictors are fixed values observed without noise.

Linear Regression Example

- Unobserved "True" Data: (X_i, y_i)
- Observed Data: (X_i^*, y_i^*)
  - $y_i^* = \beta_0 + \beta_x X_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
  - $X_i^* = X_i + e_i, \qquad e_i \sim \mathcal{N}(0, \sigma_e^2)$

Linear Regression Example

- Unobserved "True" Data: (X_i, y_i)
- Observed Data: (X_i^*, y_i^*)
  - $y_i^* = \beta_0 + \beta_x X_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
  - $X_i^* = X_i + e_i, \qquad e_i \sim \mathcal{N}(0, \sigma_e^2)$

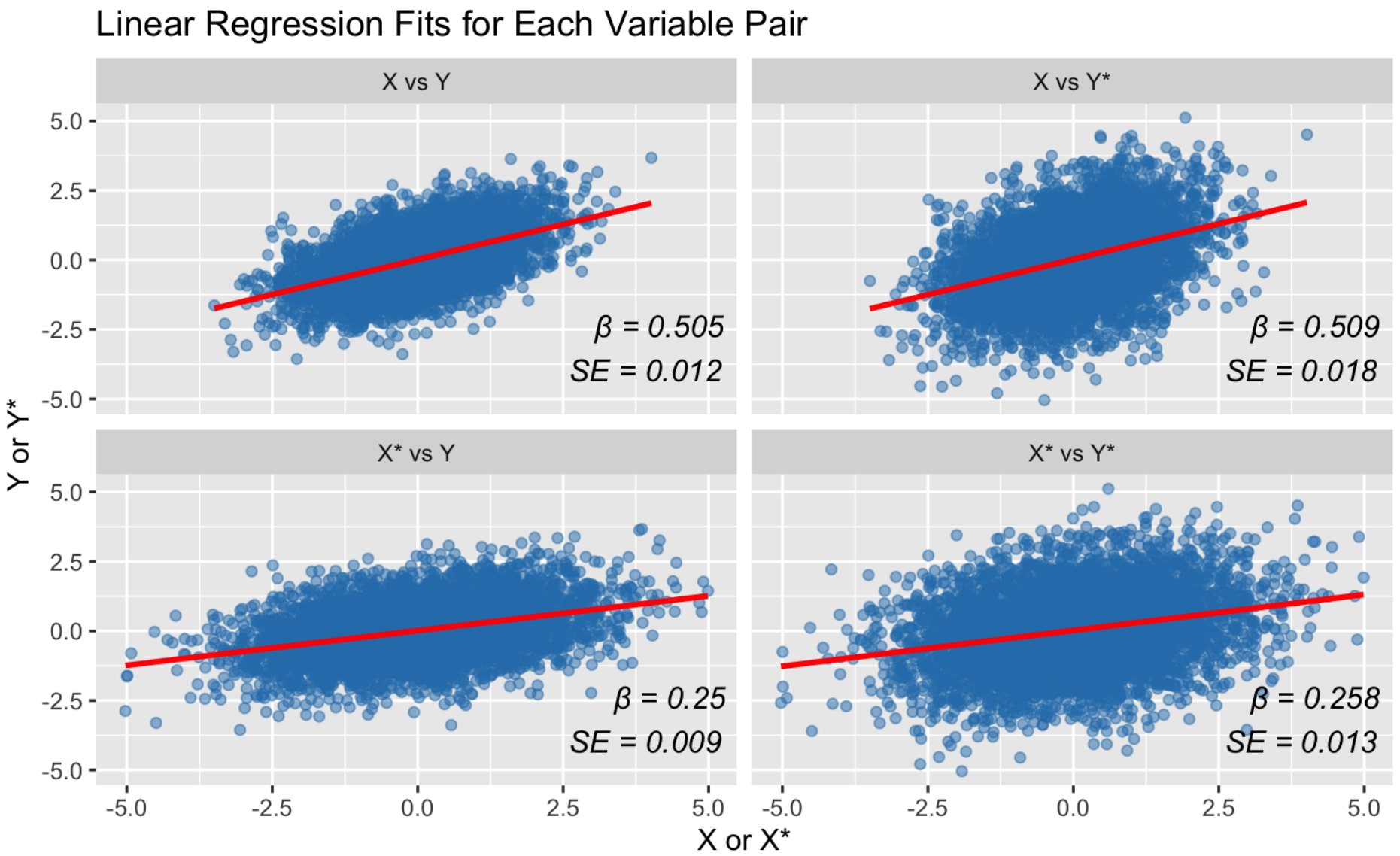
Example

Generate:

$$(X_i, y_i) \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

Calculate (X_i^*, y_i^*) such that $\sigma = \sigma_e = 1$.

Ignoring Noisy Input Data Can Have Profound Effects



Developing measurement error BART (meBART)

Goal

In this work, we investigate modeling measurement error in Bayesian additive regression trees (BART), with the goal of improving predictive performance, parameter estimation, and uncertainty quantification on complex datasets.

Developing measurement error BART (meBART)

Goal

In this work, we investigate modeling measurement error in Bayesian additive regression trees (BART), with the goal of improving predictive performance, parameter estimation, and uncertainty quantification on complex datasets.

- IID sequence of random variables $\{(X_i, y_i) : i = 1, \ldots n\}$
- However, instead of measuring X_i directly, we can only definitively measure X_i^*, a noisy realization of the true X_i.
- Thus, (y_i, X_i^*) are observed variables, and X_i will be considered an unknown latent variable that must be estimated.

Developing meBART

Full Likelihood

$$p(y_i, X_i^* | \theta) = \mathcal{N}\left(\sum_{h=1}^{m} g(X_i; T_h, M_h), \sigma^2 \right) \times \mathcal{N}(X_i, \boldsymbol{\Sigma}_e)$$

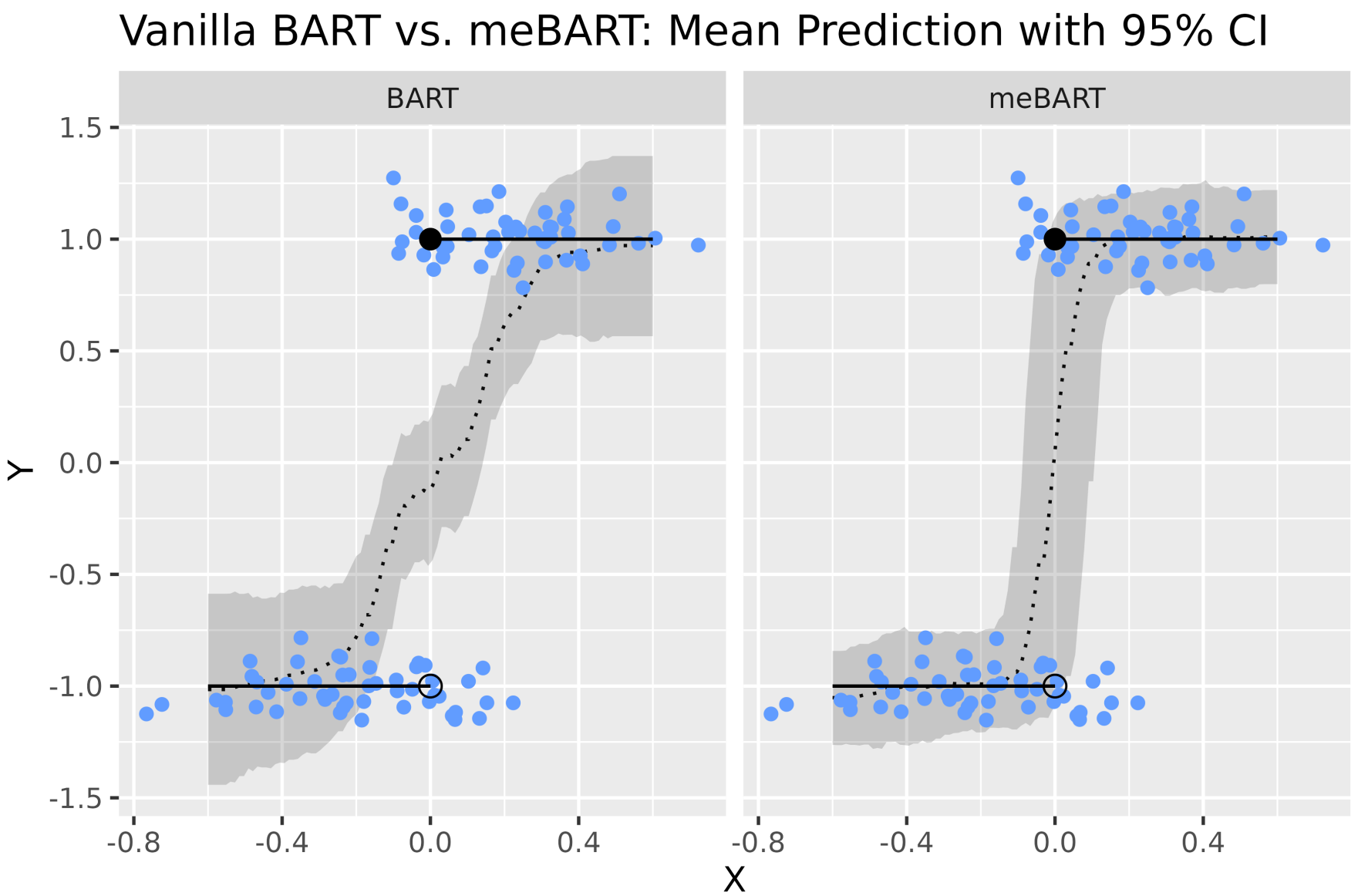
We then set the following prior on X_i:

Prior on X_i

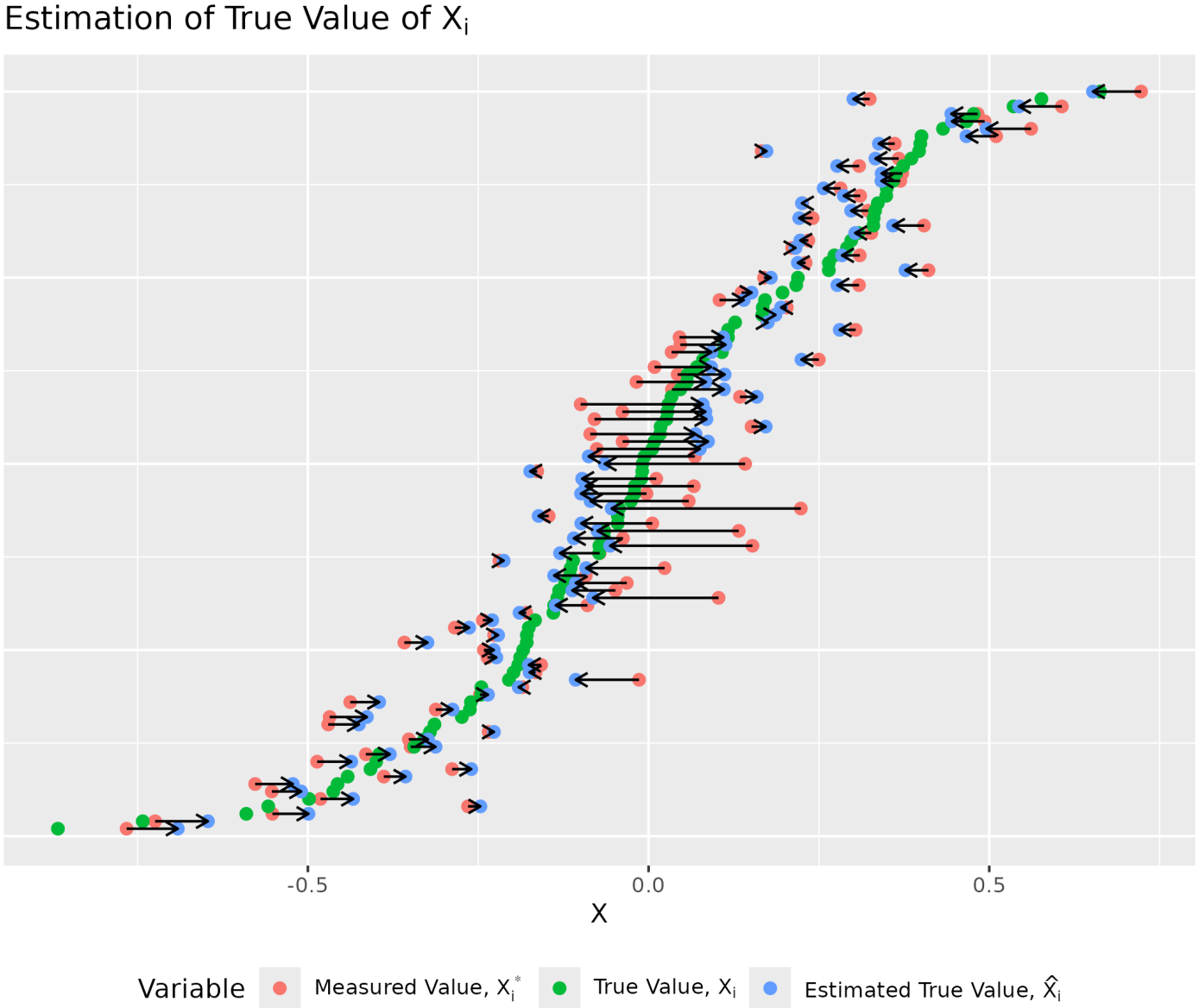
$$X_i \sim \mathcal{N}_p(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$$

We can now sample from the full posterior as a series of samples from each parameter's full conditional posterior distribution. The new X_i term is sampled using a Metropolis-Hastings step.

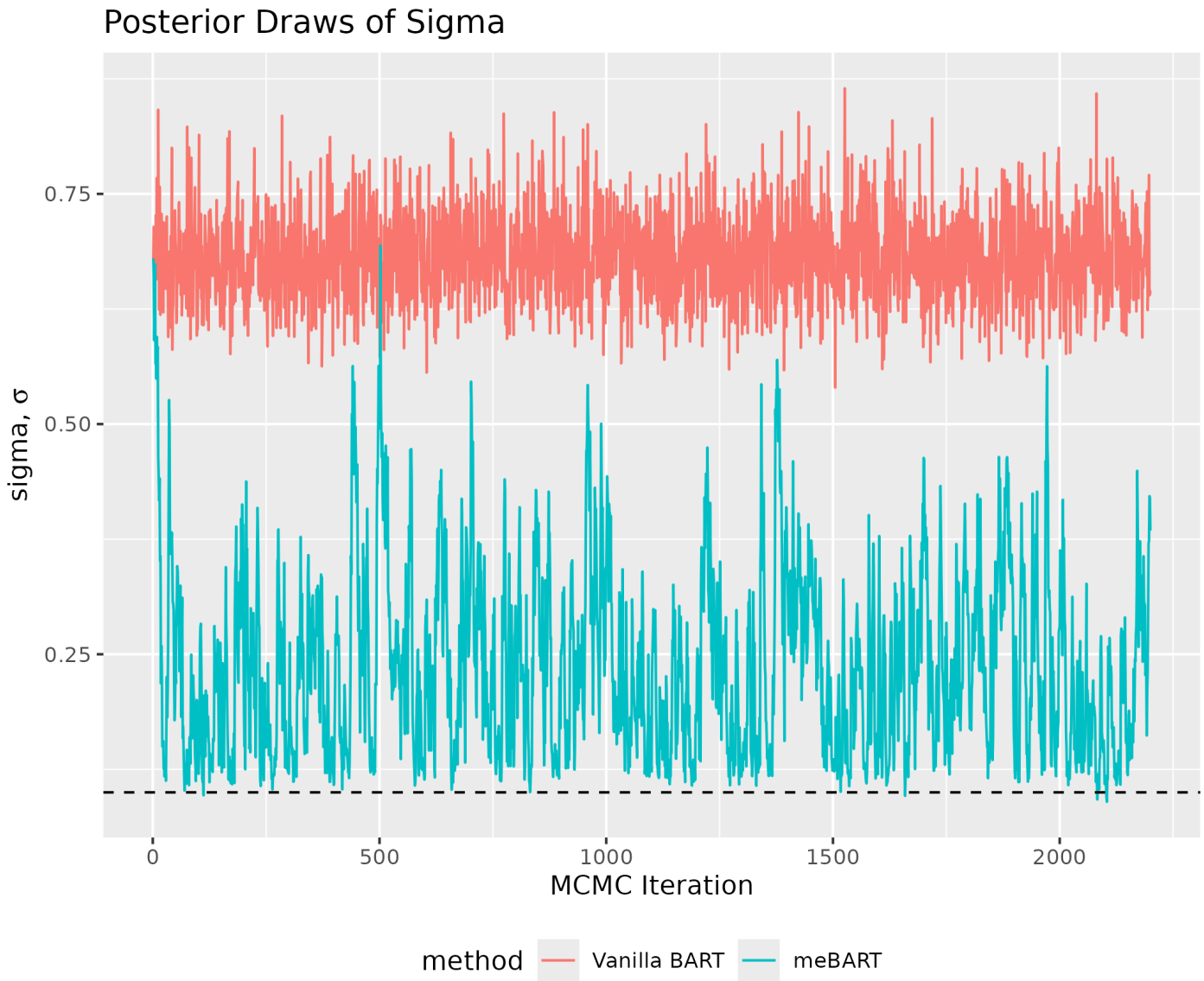
meBART Performance on Synthetic Data



meBART Performance on Synthetic Data



meBART Performance on Synthetic Data



meBART Summary

- Major benefits include:
  - Better recovery of the underlying function $f(x)$ and noise parameters.
  - More accurate empirical coverage probabilities of the mean function $f(x)$ and test predictions y_{test}.
  - Updated estimates of the underlying true values of X.
- In the future we can continue investigating other priors for X_i and likelihoods for X_i^* (such as asymmetric error or data with limits of detection).

Q & A

Funding Acknowledgments

- National Science Foundation Graduate Research Fellowship Program (Grant No. 1842494)
- The Ken Kennedy Institute Computational Science and Engineering Recruiting Fellowship

Collaborators

Christine Peterson
Rice University

Zachary Wooten
St. Jude Children's Research Hospital

Check out our preprint! →

Connect with me on LinkedIn! →