Bayesian Inference of the Number of Trees in the BART Model

Gavin Collins$^{1,2}$, Matthew Pratola$^{2}$, Radu Herbei$^{2}$, Robert McCulloch$^{3}$, and Edward George$^{4}$

Brigham Young University

45th Annual Summer Institute of Applied Statistics

June 22, 2023
The Bottom Line Up Front

• Prior distribution on the number of trees
• MH step to add/delete one tree at a time
• Takes longer
• Works well
• (Still a work in progress)
1. Recap of BART

2. Bayesian Inference of the Number of Trees
   i. Motivation
   ii. A Fully Bayesian Model
   iii. Sampling from the Posterior Distribution
   iv. Code
   v. Simulations
   vi. Application to Real Data

3. Conclusion
1. Recap of BART
2. Bayesian Inference of the Number of Trees
   i. Motivation
   ii. A Fully Bayesian Model
   iii. Sampling from the Posterior Distribution
   iv. Code
   v. Simulations
   vi. Application to Real Data
3. Conclusion
Bayesian Additive Regression Trees

Data

• \((x_1, y_1), \ldots, (x_n, y_n)\)
• \(\text{input } x_i \in \mathbb{R}^p \rightarrow \text{response } y_i \in \mathbb{R}\)

Regression Model

• \(y_i \mid x_i \sim N(f(x_i), \sigma^2), i = 1, \ldots, n \ (\text{ind})\)
• \(f: \mathbb{R}^p \rightarrow \mathbb{R} \) (mean function)
• \(\sigma^2 \geq 0 \) (residual variance)

“Branin” Example:

• \(p = 2\)
• \(f = \text{“The Branin Function”}\)
• \(n = 300\)
• \(x_1, \ldots, x_{300} \sim \text{Unif}(0,1)^2 (iid)\)
• \(\sigma^2 = 1\)
Bayesian Additive Regression Trees

Data
- \((x_1, y_1), ..., (x_n, y_n)\)
- input \(x_i \in \mathbb{R}^p \rightarrow \) response \(y_i \in \mathbb{R}\)

Regression Model
- \(y_i \mid x_i \sim N(f(x_i), \sigma^2), i = 1, ..., n\) (ind)
- \(f: \mathbb{R}^p \rightarrow \mathbb{R}\) (mean function)
- \(\sigma^2 \geq 0\) (residual variance)

Hurricane Example:
- \(p = 6\)
- \(f = \) Computer Model
- \(n = 4,000\)
- Goal: Infer \(f\) for sensitivity analysis, model calibration, etc.

Input \(x\)
- \(x_1 = \) Initial Sea Level
- \(x_2 = \) Hurricane Heading
- \(x_3 = \) Velocity of the Eye
- \(x_4 = \) Max Wind Speed
- \(x_5 = \) Min Pressure
- \(x_6 = \) Landfall Location

Response \(y\)
- \(y =\) Maximum Water Level During a Storm Surge
Bayesian Additive Regression Trees

\[ x = (x_1, x_2) \rightarrow T \]

- \( \eta_1 \): \( x_2 < 0.8 \)
- \( \eta_2 \): \( x_1 < 0.25 \)
- \( \eta_3 \): \( \mu_3 = -45 \)
- \( \eta_4 \): \( \mu_4 = 40 \)
- \( \eta_5 \): \( x_2 \geq 0.5 \)
- \( \eta_{10} \): \( \mu_{10} = -10 \)
- \( \eta_{11} \): \( \mu_{11} = 20 \)

\[ f(x) \]

- \( x_1 \) ranges from 0.0 to 1.0
- \( x_2 \) ranges from 0.0 to 1.0

Legend:
- Red: 40
- Blue: 0
- Green: -40
Bayesian Additive Regression Trees

- $T$ has $L$ “terminal nodes”
- Terminal node parameter $\mu \in \mathbb{R}^L$
- $f(x) = g(x; T, \mu)$
Bayesian Additive Regression Trees

- $T_{1:m} \equiv T_1, \ldots, T_m$ ($m \approx 200$)
- $\mu_{1:m} \equiv \mu_1, \ldots, \mu_m$
- $T_j$ has $L_j$ terminal nodes
- $\mu_j \in \mathbb{R}^{L_j}$
- $f(x) = \sum_{j=1}^{m} g(x; T_j, \mu_j)$

$L_j = 1 \Rightarrow “Stump”$
Bayesian Additive Regression Trees

Prior Distribution $\pi(T_{1:m}, \mu_{1:m}, \sigma^2) = \pi(\sigma^2) \prod_{j=1}^{m} \pi(T_j) \pi(\mu_j \mid T_j)$

• $T_j \sim$ Tree-Generating Stochastic Process

• $\mu_{j\ell} \mid T_j \sim N(0, \tau_m^2); \; \ell = 1, \ldots, L_j; \; j = 1, \ldots, m$ (iid)

• $\sigma^2 \sim$ Scaled-inv-$\chi^2(\nu, \lambda)$
Bayesian Additive Regression Trees

Posterior Sampling MCMC Algorithm

\[ \text{Notation: } T_{-j} \equiv (T_1, ..., T_{j-1}, T_{j+1}, ..., T_m) \text{ and } \mu_{-j} \equiv (\mu_1, ..., \mu_{j-1}, \mu_{j+1}, ..., \mu_m) \]

For \( i = 1, ..., N_{m_{cmc}} \):

1. For \( j = 1, ..., m \)
   a. Sample \( T_j \mid (T_{-j}, \mu_{-j}, \sigma^2, \text{data}) \) (Metropolis–Hastings)
   b. Sample \( \mu_j \mid \cdot \) (Gibbs Step)

2. Sample \( \sigma^2 \mid \cdot \) (Gibbs Step)
Bayesian Additive Regression Trees

Branin Function $f(x)$

Posterior Mean

Posterior 0.025 Quantile

Posterior 0.975 Quantile
1. Recap of BART

2. Bayesian Inference of the Number of Trees
   i. Motivation
   ii. A Fully Bayesian Model
   iii. Sampling from the Posterior Distribution
   iv. Code
   v. Simulations
   vi. Application to Real Data

3. Conclusion
How Many Trees???

**Default** $m = 200$

**Large** $m$

- Flexible Estimation
- More computation
- Risk overfitting

**Small** $m$

- Improved variable selection
- Less computation
- Risk underfitting
Out-of-Sample Prediction

"Typical"

Small $m$ best

Large $m$ best

"RMSE for $f(x_{\text{test}})$" = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (f(x_{\text{test},i}) - \hat{f}(x_{\text{test},i}))^2}
Variable Selection

$$f_{\text{friedman}}(\mathbf{x}) = 10 \sin(\pi x_1 x_2) + 20 (x_3 - 0.5)^2 + 10 x_4 + 5 x_5 + \sum_{j=6}^{10} 0x_j$$

Friedman: $p = 10$

FVAR = Proportion of branches involving “false” input variables
Cross-Validation

• Pick a grid of $m$-values (e.g., $m = 1, 10, 20, 50, 100, 200, 300, 400$)
• For each value of $m$
  • Split data into train and test sets
  • Fit BART to the training set
  • Predict responses in the test set
• Compare out-of-sample RMSE across the grid
• Pick the value $m = m_{CV}$ that minimizes RMSE
• Fit a BART model to the full dataset, with $m = m_{CV}$
Cross-Validation

- Pick a grid of $m$-values (e.g., $m = 1, 10, 20, 50, 100, 200, 300, 400$)
- For each value of $m$
  - Split data into train and test sets
  - Fit BART to the training set
  - Predict responses in the test set
- Compare out-of-sample RMSE across the grid
- Pick the value $m = m_{CV}$ that minimizes RMSE
- Fit a BART model to the full dataset, with $m = m_{CV}$

How to pick the grid?

Expensive!

What about variable selection, computation time, etc.?
1. Recap of BART

2. Bayesian Inference of the Number of Trees
   i. Motivation
   ii. A Fully Bayesian Model
   iii. Sampling from the Posterior Distribution
   iv. Code
   v. Simulations
   vi. Application to Real Data

3. Conclusion
Fully Bayesian Inference of $m$

$$\pi(m, T_{1:m}, \mu_{1:m}, \sigma^2) = \pi(\sigma^2) \pi(m) \prod_{j=1}^{m} \pi(T_j) \pi(\mu_j | T_j, m)$$

- $m \sim \text{Poisson}(\theta)$ [Truncated]
  - Optionally, assign $\theta$ a hyperprior
- $T_j \sim \text{Tree-Generating Stochastic Process}$ (as before)
- $\mu_{j\ell} | (m, T_j) \sim N(0, \tau_m^2); \ell = 1, ..., L_j; \ j = 1, ..., m$ (iid) (as before)
- $\sigma^2 \sim \text{Scaled-inv-\chi^2}(\nu, \lambda)$ (as before)
Prior Distribution

\[
\pi(m) \propto \frac{\theta^m e^{-\theta}}{m!} I(1 \leq m \leq 1000) \text{ (Truncated Poisson)}
\]

- \( \Rightarrow E(m) \approx \theta \)
- Default \( \theta = 200 \)
- Optionally, assign \( \theta \) a hyperprior
  - \( \theta \sim \frac{\theta_0 \chi^2_{\kappa_0}}{\kappa_0} \)
  - \( E(\theta) = \theta_0 \)
  - Degree of Freedom \( \kappa_0 \)
  - Default \( \kappa_0 = 200 \)
Prior Distribution

\[ \mu_{j \ell} \mid (m, T_j) \sim N(0, \tau_m^2) \]

\[ \tau_m = \frac{\max y_i - \min y_i}{2k \sqrt{m}} \]

\[ \Rightarrow f(x) \sim N\left(0, \left(\frac{\max y_i - \min y_i}{2k}\right)^2\right) \text{ (for all } m) \]
1. Recap of BART

2. Bayesian Inference of the Number of Trees
   i. Motivation
   ii. A Fully Bayesian Model
   iii. Sampling from the Posterior Distribution
   iv. Code
   v. Simulations
   vi. Application to Real Data

3. Conclusion
Posterior Sampling MCMC Algorithm

Initialize $m = m_0$ (default $m_0 = \theta_0$)

For $i = 1, \ldots, N_{mcmc}$:

1. Sample $\theta | m$ (Gibbs Step; if $\kappa_0 < \infty$)
Posterior Sampling MCMC Algorithm

Initialize $m = m_0$ (default $m_0 = \theta_0$)

For $i = 1, \ldots, N_{m_{\text{mcmc}}}$:

1. Sample $\theta | m$ (Gibbs Step; if $\kappa_0 < \infty$)
2. Sample $m | \cdot$ (Metropolis-Hastings)

Randomly select either
   a) Birth or
   b) Death
Posterior Sampling MCMC Algorithm

Initialize $m = m_0$ (default $m_0 = \theta_0$)

For $i = 1, \ldots, N_{\text{mcmc}}$:

1. Sample $\theta | m$ (Gibbs Step; if $\kappa_0 < \infty$)
2. Sample $m | \cdot$ (Metropolis-Hastings)
   - If $m$ was increased, sample new $\mu_* | \cdot$ (Gibbs Step)

Randomly select either
a) Birth or
b) Death
Posterior Sampling MCMC Algorithm

Initialize $m = m_0$ (default $m_0 = \theta_0$)

For $i = 1, \ldots, N_{\text{mcmc}}$:

1. Sample $\theta | m$ (Gibbs Step; if $\kappa_0 < \infty$)
2. Sample $m | \cdot$ (Metropolis-Hastings)
   - If $m$ was increased, sample new $\mu_* | \cdot$ (Gibbs Step)
3. For $j = 1, \ldots, m$
   a. Sample $T_j | (T_{-j}, \mu_{-j}, \sigma^2, \text{data})$ (Metropolis–Hastings)
   b. Sample $\mu_j | \cdot$ (Gibbs Step)
4. Sample $\sigma^2 | \cdot$ (Gibbs Step)

Randomly select either
a) Birth or
b) Death

Same as Standard BART
Birth Transition

$m = 7$:  +  +  +  +  +  +  +

$m = 8$:  +  +  +  +  +  +  +
Birth Transition

• Why stumps?
• Why randomize the location of the new tree?
  • Trees are exchangeable, but ordered in the prior distribution
  • Need reversibility
Death Transition

$m = 7$: 

$m = 6$: 

Death
MH Transition

• Current parameters $\psi = (m, T_{1:m}, \mu_{1:m}, \sigma^2)$

• Randomly select either birth or death transition ($Pr(\text{birth}) = Pr(\text{death}) = 0.5$)

  • Birth Transition
    • Select a location to insert stump $T^*$
    • Update to $\psi \rightarrow \psi^{\text{birth}} = (m + 1, T^*_{1:(m+1)}, \mu_{1:m}, \sigma^2)$

  • Death Transition (if there are any “stumps”)
    • Select a stump $T^*$ to delete
    • Update to $\psi \rightarrow \psi^{\text{death}} = (m - 1, T^*_{1:(m-1)}, \mu_{1:(m-1)}, \sigma^2)$

• Accept with the MH acceptance probability: $\min\{1, \text{MH Ratio}\}$
MH Acceptance Probability: $\min\{1, \text{MH Ratio}\}$

**MH Ratio for Birth**

$$
\frac{\pi(\psi^{\text{birth}} \mid \text{data})}{\pi(\psi \mid \text{data})} \times \frac{q(\psi^{\text{birth}} \rightarrow \psi)}{q(\psi \rightarrow \psi^{\text{birth}})} 
$$

Ratio of Posterior Distributions

Ratio of Transition Probabilities
MH Acceptance Probability: \( \min\{1, \text{MH Ratio}\} \)

**MH Ratio for Birth**

\[
\begin{align*}
\text{MH Ratio} &= \pi(\psi_{\text{birth}} \mid \text{data}) \times \frac{q(\psi_{\text{birth}} \rightarrow \psi)}{q(\psi \rightarrow \psi_{\text{birth}})} \\
&= \frac{\pi(\text{data} \mid \psi_{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(\sigma^2)}{\pi(\sigma^2)} \times \frac{\pi(m + 1)}{\pi(m)} \times \frac{\pi(T^*) \prod_j^m \pi(T_j)}{\prod_j^m \pi(T_j)} \times \frac{\prod_j^m \pi(\mu_j \mid m + 1, T_j)}{\prod_j^m \pi(\mu_j \mid m, T_j)} \times \frac{0.5/(m_{\text{stumps}} + 1)}{0.5/(m + 1)}
\end{align*}
\]

- **Likelihood**
- **Prior Distribution**
- **Ratio of Transition Probabilities**
MH Acceptance Probability: \( \text{min}\{1, \text{MH Ratio}\} \)

MH Ratio for Birth

\[
\frac{\pi(\psi^{\text{birth}} \mid \text{data})}{\pi(\psi \mid \text{data})} \times \frac{\frac{q(\psi^{\text{birth}} \rightarrow \psi)}{q(\psi \rightarrow \psi^{\text{birth}})}}{\pi(\sigma^2) \times \pi(m + 1) \times \frac{\pi(T^*) \prod_j^m \pi(T_j)}{\prod_j^m \pi(T_j)} \times \frac{\prod_j^m \pi(\mu_j | m + 1, T_j)}{\prod_j^m \pi(\mu_j | m, T_j)} \times \frac{0.5/(m_{\text{stumps}} + 1)}{0.5/(m + 1)}}
\]

\[
\frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(m + 1)}{\pi(m)} \times \pi(T^*) \times \frac{\prod_j^m \pi(\mu_j | m + 1, T_j)}{\prod_j^m \pi(\mu_j | m, T_j)} \times \frac{1/(m_{\text{stumps}} + 1)}{1/(m + 1)}
\]
MH Acceptance Probability: \( \min\{1, \text{MH Ratio}\} \)

**MH Ratio for Birth**

\[
\frac{\pi(\psi^{\text{birth}} | \text{data})}{\pi(\psi | \text{data})} \times \frac{q(\psi^{\text{birth}} \rightarrow \psi)}{q(\psi \rightarrow \psi^{\text{birth}})} = \frac{\pi(\text{data} | \psi^{\text{birth}})}{\pi(\text{data} | \psi)} \times \frac{\pi(\sigma^2)}{\pi(\sigma^2)} \times \frac{\pi(m + 1)}{\pi(m)} \times \frac{\pi(T^*) \prod_j^m \pi(T_j)}{\prod_j^m \pi(T_j)} \times \frac{\prod_j^m \pi(\mu_j | m + 1, T_j)}{\prod_j^m \pi(\mu_j | m, T_j)} \times \frac{0.5/(m_{\text{stumps}} + 1)}{0.5/(m + 1)}
\]

\[
= \frac{\pi(\text{data} | \psi^{\text{birth}})}{\pi(\text{data} | \psi)} \times \frac{\pi(m + 1)}{\pi(m)} \times \frac{\pi(T^*)}{\prod_j^m \pi(T_j)} \times \frac{\prod_j^m \pi(\mu_j | m + 1, T_j)}{\prod_j^m \pi(\mu_j | m, T_j)} \times \frac{1/(m_{\text{stumps}} + 1)}{1/(m + 1)}
\]

\[
= \frac{\pi(\text{data} | \psi^{\text{birth}})}{\pi(\text{data} | \psi)} \times \frac{\theta}{m + 1} \times \frac{1}{m + 1}
\]
MH Acceptance Probability: \( \min\{1, \text{MH Ratio} \} \)

**MH Ratio for Birth**

\[
\frac{\pi (\psi^{\text{birth}} | \text{data})}{\pi (\psi | \text{data})} \times \frac{q(\psi^{\text{birth}} \rightarrow \psi)}{q(\psi \rightarrow \psi^{\text{birth}})}
\]

\[
= \frac{\pi (\text{data} | \psi^{\text{birth}})}{\pi (\text{data} | \psi)} \times \frac{\pi (\sigma^2)}{\pi (\sigma^2)} \times \frac{\pi (m + 1)}{\pi (m)} \times \frac{\pi (T^*) \prod_j^m \pi (T_j)}{\prod_j^m \pi (T_j)} \times \frac{\prod_j^m \pi (\mu_j | m + 1, T_j)}{\prod_j^m \pi (\mu_j | m, T_j)} \times \frac{0.5/(m_{\text{stumps}} + 1)}{0.5/(m + 1)}
\]

\[
= \frac{\pi (\text{data} | \psi^{\text{birth}})}{\pi (\text{data} | \psi)} \times \frac{\pi (m + 1)}{\pi (m)} \times \frac{\pi (T^*)}{\prod_j^m \pi (\mu_j | m + 1, T_j)} \times \frac{\prod_j^m \pi (\mu_j | m, T_j)}{\prod_j^m \pi (\mu_j | m, T_j)} \times \frac{1/(m_{\text{stumps}} + 1)}{1/(m + 1)}
\]

\[
= \frac{\pi (\text{data} | \psi^{\text{birth}})}{\pi (\text{data} | \psi)} \times \frac{\theta}{m + 1} \times \pi (T^*) \times
\]
MH Acceptance Probability: min{1, MH Ratio}

MH Ratio for Birth

\[
\text{MH Ratio} = \frac{\pi(\psi^{\text{birth}} \mid \text{data})}{\pi(\psi \mid \text{data})} \times \frac{q(\psi^{\text{birth}} \to \psi)}{q(\psi \to \psi^{\text{birth}})}
\]

\[
= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(\sigma^2)}{\pi(\sigma^2)} \times \frac{\pi(m + 1)}{\pi(m)} \times \frac{\pi(T^*) \prod_{j=1}^m \pi(T_j)}{\prod_{j=1}^m \pi(T_j)} \times \frac{\prod_{j=1}^m \pi(\mu_j \mid m + 1, T_j)}{\prod_{j=1}^m \pi(\mu_j \mid m, T_j)} \times \frac{0.5/(m_{\text{stumps}} + 1)}{0.5/(m + 1)}
\]

\[
= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(m + 1)}{\pi(m)} \times \frac{\pi(T^*)}{\prod_{j=1}^m \pi(\mu_j \mid m + 1, T_j)} \times \frac{\prod_{j=1}^m \pi(\mu_j \mid m, T_j)}{1/(m_{\text{stumps}} + 1)} \times \frac{1}{1/(m + 1)}
\]

\[
= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\theta}{m + 1} \times \pi(T^*) \times \left(\frac{m}{m + 1}\right)^{-\sum_{j=1}^m \frac{L_j}{2}} \exp\left(-\frac{1}{2m\tau_m^2} \sum_j \sum_{\ell=1}^{L_j} \mu_{j,\ell}^2\right)
\]
MH Acceptance Probability: \( \min\{1, \text{MH Ratio}\} \)

**MH Ratio for Birth**

\[
\frac{\pi(\psi^{\text{birth}} | \text{data})}{\pi(\psi | \text{data})} \times \frac{q(\psi^{\text{birth}} \rightarrow \psi)}{q(\psi \rightarrow \psi^{\text{birth}})} \]

\[
= \frac{\pi(\text{data} | \psi^{\text{birth}})}{\pi(\text{data} | \psi)} \times \frac{\pi(\sigma^2)}{\pi(\sigma^2)} \times \frac{\pi(m + 1)}{\pi(m)} \times \frac{\pi(T^*) \Pi_j^m \pi(T_j)}{\Pi_j^m \pi(T_j)} \times \frac{\Pi_j^m \pi(\mu_j | m + 1, T_j)}{\Pi_j^m \pi(\mu_j | m, T_j)} \times \frac{0.5/(m_{\text{stumps}} + 1)}{0.5/(m + 1)}
\]

\[
= \frac{\pi(\text{data} | \psi^{\text{birth}})}{\pi(\text{data} | \psi)} \times \frac{\pi(m + 1)}{\pi(m)} \times \frac{\pi(T^*)}{\Pi_j^m \pi(T_j)} \times \frac{\Pi_j^m \pi(\mu_j | m + 1, T_j)}{\Pi_j^m \pi(\mu_j | m, T_j)} \times \frac{1/(m_{\text{stumps}} + 1)}{1/(m + 1)}
\]

\[
= \frac{\pi(\text{data} | \psi^{\text{birth}})}{\pi(\text{data} | \psi)} \times \frac{\theta}{m + 1} \times \frac{\pi(T^*)}{\Pi_j^m \pi(T_j)} \times \left(\frac{m}{m + 1}\right)^{-\sum_j L_j} \times \exp\left(-\frac{1}{2m\tau_m} \sum_{j=1}^m \sum_{\ell=1}^{L_j} \mu_{j\ell}^2 \right) \times \frac{m + 1}{m_{\text{stumps}} + 1}
\]
Likelihood Ratio

\[
\text{Likelihood Ratio} = \frac{\pi \left( \text{data} \mid \psi^{\text{birth}} \right)}{\pi \left( \text{data} \mid \psi \right)}
\]
Likelihood Ratio

\[
\text{Likelihood Ratio} = \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)}
\]

\[
= \frac{\pi(\text{data} \mid m + 1, T^*_1, m, \mu_{1:m}, \sigma^2)}{\pi(\text{data} \mid m, T^*_1, m, \mu_{1:m}, \sigma^2)}
\]

Marginal Likelihood (m+1 trees)

Full likelihood (m trees)
Likelihood Ratio

\[
\text{Likelihood Ratio} = \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)}
\]

\[
= \frac{\pi(\text{data} \mid m + 1, T_{1:(m+1)}^*, \mu_{1:m}, \sigma^2)}{\pi(\text{data} \mid m, T_{1:m}, \mu_{1:m}, \sigma^2)}
\]

\[
= \int_{\mathbb{R}} \frac{\pi(\text{data} \mid m + 1, T_{1:(m+1)}^*, \mu_{1:m}, \mu^*, \sigma^2) \pi(\mu^* \mid m + 1, T^*) d\mu^*}{\pi(\text{data} \mid m, T_{1:m}, \mu_{1:m}, \sigma^2)}
\]

Full Likelihood (m+1)  
Prior Distribution of \( \mu^* \)
Likelihood Ratio

\[
\text{Likelihood Ratio} = \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)}
\]

\[
= \frac{\pi(\text{data} \mid m + 1, T_{1:(m+1)}^*, \mu_{1:m}, \sigma^2)}{\pi(\text{data} \mid m, T_{1:m}, \mu_{1:m}, \sigma^2)}
\]

\[
= \int \pi(\text{data} \mid m + 1, T_{1:(m+1)}^*, \mu_{1:m}, \mu^*, \sigma^2) \pi(\mu^* \mid m, T^*) \, d\mu^* \frac{\pi(\text{data} \mid m, T_{1:m}, \mu_{1:m}, \sigma^2)}
\]

\[
= \int \prod_{i=1}^{n} N(y_i; \mu^* + \sum_{j=1}^{m} g(x_i; T_j, \mu_j), \sigma^2) \frac{N(\mu^*; 0, \tau_m^2)}{\prod_{i=1}^{n} N(y_i; \sum_{j=1}^{m} g(x_i; T_j, \mu_j), \sigma^2)}
\]
Likelihood Ratio

\[ \text{Likelihood Ratio} = \frac{\pi(\text{data} | \psi^{\text{birth}})}{\pi(\text{data} | \psi)} \]

\[ = \frac{\pi(\text{data} \mid m + 1, T_{1:(m+1)}^*, \mu_{1:m}, \sigma^2)}{\pi(\text{data} \mid m, T_{1:m}, \mu_{1:m}, \sigma^2)} \]

\[ = \int_{\mathbb{R}} \frac{\pi(\text{data} \mid m + 1, T_{1:(m+1)}^*, \mu_{1:m}, \mu^*, \sigma^2) \pi(\mu^* \mid m, T^*) d\mu^*}{\pi(\text{data} \mid m, T_{1:m}, \mu_{1:m}, \sigma^2)} \]

\[ = \int_{\mathbb{R}} \frac{\prod_{i=1}^n N(y_i; \mu^* + \sum_{j=1}^m g(x_i; T_j, \mu_j), \sigma^2) N(\mu^*; 0, \tau_{m}^2) d\mu^*}{\prod_{i=1}^n N(y_i; \sum_{j=1}^m g(x_i; T_j, \mu_j), \sigma^2)} \]

\[ = \left( \frac{\sigma^2}{n \tau_{m+1}^2 + \sigma^2} + 1 \right)^{1/2} \exp \left( \frac{n^2 \tau_{m+1}^2 \left( \sum_{i=1}^n y_i - \sum_{j=1}^m g(x_i; T_j, \mu_j) \right)^2}{2 \sigma^2 (n \tau_{m+1}^2 + \sigma^2)} \right) \]
1. Recap of BART

2. Bayesian Inference of the Number of Trees
   i. Motivation
   ii. A Fully Bayesian Model
   iii. Sampling from the Posterior Distribution
   iv. Code
   v. Simulations
   vi. Application to Real Data

3. Conclusion
Code

• R implementation forthcoming:

  • \texttt{bart(X, y, learnntree = TRUE, ntreemean = 200, ntree\_df = Inf)}

\[
m \sim Pois(\theta)I(1 \leq m \leq 1000)
\]

\[
\theta \sim \frac{\theta_0 \chi^2_{\kappa_0}}{\kappa_0}
\]
1. Recap of BART

2. Bayesian Inference of the Number of Trees
   i. Motivation
   ii. A Fully Bayesian Model
   iii. Sampling from the Posterior Distribution
   iv. Code
   v. Simulations
   vi. Application to Real Data

3. Conclusion
Simulation Setup

**Fully Bayesian Inference for m**
- Generate training and test data
  - \( x_i \sim \text{Unif}(0,1)^p \) and
  - \( y_i \mid x_i \sim N(f(x_i), 1) \)
- For \( \kappa_0 \in \{3, 100, \infty\} \) (with \( \theta_0 = 200 \))
  - Fit BART to training set with Bayesian inference for \( m \)

**Cross-Validation**
- For \( m \in \text{Grid} \):
  - Generate training and test sets
    - \( x_i \sim \text{Unif}(0,1)^p \) and
    - \( y_i \mid x_i \sim N(f(x_i), 1) \)
  - Fit BART to training set with \( m \) trees

Compare accuracy using “RMSE for \( f(x_{\text{test}}) \)” = \[
\sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (f(x_{\text{test},i}) - \hat{f}(x_{\text{test},i}))^2}
\]
Simulation Setup

Simulations

<table>
<thead>
<tr>
<th>$f$</th>
<th>$n_{train}$</th>
<th>$n_{test}$</th>
<th>$p$</th>
<th>SNR</th>
<th>$N_{mc}$ (infer $m$)</th>
<th>$N_{mc}$ (fix $m$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friedman</td>
<td>500</td>
<td>1,000</td>
<td>10</td>
<td>23.8</td>
<td>1,000,000 (√)</td>
<td>3,000</td>
</tr>
<tr>
<td>Borehole</td>
<td>500</td>
<td>1,000</td>
<td>8</td>
<td>20.9</td>
<td>1,000,000 (√)</td>
<td>3,000</td>
</tr>
<tr>
<td>Branin</td>
<td>1,000</td>
<td>2,000</td>
<td>2</td>
<td>18.2</td>
<td>1,000,000 (√)</td>
<td>3,000</td>
</tr>
<tr>
<td>Piston</td>
<td>1,500</td>
<td>3,000</td>
<td>7</td>
<td>20.6</td>
<td>1,000,000 (√)</td>
<td>3,000</td>
</tr>
<tr>
<td>Snake</td>
<td>10,000</td>
<td>10,000</td>
<td>2</td>
<td>2930</td>
<td>1,000,000 (×)</td>
<td>100,000</td>
</tr>
<tr>
<td>Welch</td>
<td>10,000</td>
<td>10,000</td>
<td>20</td>
<td>2809</td>
<td>1,000,000 (×)</td>
<td>100,000</td>
</tr>
<tr>
<td>Friedman×20</td>
<td>10,000</td>
<td>10,000</td>
<td>100</td>
<td>23.8</td>
<td>100,000 (×)</td>
<td>50,000</td>
</tr>
<tr>
<td>300-Step</td>
<td>12,000</td>
<td>12,000</td>
<td>300</td>
<td>100</td>
<td>200,000 (×)</td>
<td>10,000</td>
</tr>
<tr>
<td>100-Step</td>
<td>4,000</td>
<td>8,000</td>
<td>100</td>
<td>100</td>
<td>1,000,000 (×)</td>
<td>10,000</td>
</tr>
<tr>
<td>1-Step</td>
<td>100</td>
<td>200</td>
<td>1</td>
<td>100</td>
<td>1,000,000 (√)</td>
<td>3,000</td>
</tr>
<tr>
<td>$T_4$</td>
<td>800</td>
<td>1,000</td>
<td>15</td>
<td>340</td>
<td>1,000,000 (×)</td>
<td>30,000</td>
</tr>
</tbody>
</table>

Bayesian Inference

Cross-Validation
Convergence of $m$

Piston: $\kappa_0 = 3$
R-hat = 1.008
ESS = 215

Branin: $\kappa_0 = 100$
R-hat = 1.03
ESS = 104

Friedman: $\kappa_0 = \infty$
R-hat = 1.6162
ESS = 305

Welch: $\kappa_0 = 3$
R-hat = 1.4845
ESS = 12

Snake: $\kappa_0 = 100$
R-hat = 2.5452
ESS = 5

Friedman x 20: $\kappa_0 = \infty$
R-hat = 1.4335
ESS = 7
Results

\[ \frac{\kappa}{\mu} = 3 \]
\[ \frac{\kappa}{\mu} = 100 \]
\[ \frac{\kappa}{\mu} = \infty \]

Fixed \( m \)

\[ 49 \]
Results

$\frac{\kappa_0}{\kappa_0} = 3$

$\frac{\kappa_0}{\kappa_0} = 100$

$\frac{\kappa_0}{\kappa_0} = \infty$

$p = 100$ inputs: all active!

$f$ has an additive structure

Piston

Snake

Friedman $\times$ 20

RMSE for $f(x_{\text{test}})$

$m$

RMSE for $f(x_{\text{test}})$

$m$

RMSE for $f(x_{\text{test}})$

$m$

- $\kappa_0 = 3$
- $\kappa_0 = 100$
- $\kappa_0 = \infty$
- Fixed $m$
Results

\[ \frac{\kappa}{\sqrt{m}} = 3 \]
\[ \frac{\kappa}{\sqrt{m}} = 100 \]
\[ \frac{\kappa}{\sqrt{m}} = \infty \]

Fixed \( m \)

\[ p\text{-step: } f(x) = \frac{20}{\sqrt{p}} \sum_{j=1}^{p} I(x_j \geq 0.5) \]
Results

\[ \frac{\gamma_0}{\gamma} = 3 \]
\[ \frac{\gamma_0}{\gamma} = 100 \]
\[ \frac{\gamma_0}{\gamma} = \infty \]

Fixed \( m \)-step:

\[ f(x) = \frac{20}{\sqrt{p}} \sum_{j=1}^{p} I(x_j \geq 0.5) \]

\[ \sum_{j=1}^{p} I(x_j \geq 0.5) \]

\( \sum_{j=1}^{p} \)

\[ \sum_{j=1}^{p} \]

\( I(x_j \geq 0.5) \]

\( f(x) \]

\( \infty \]

\( \gamma_0 \]

\( \gamma_0 \]

\( \infty \]

\( \sum_{j=1}^{p} \]

\( \sum_{j=1}^{p} \]

\( I(x_j \geq 0.5) \]

\( f(x) \]

\( \infty \]

\( \gamma_0 \]

\( \gamma_0 \]

\( \infty \]

\( \sum_{j=1}^{p} \]

\( \sum_{j=1}^{p} \]

\( I(x_j \geq 0.5) \]

\( f(x) \]

\( \infty \]

\( \gamma_0 \]

\( \gamma_0 \]

\( \infty \]

\( \sum_{j=1}^{p} \]

\( \sum_{j=1}^{p} \]

\( I(x_j \geq 0.5) \]

\( f(x) \]

\( \infty \]

\( \gamma_0 \]

\( \gamma_0 \]

\( \infty \]

\( \sum_{j=1}^{p} \]

\( \sum_{j=1}^{p} \]

\( I(x_j \geq 0.5) \]

\( f(x) \]

\( \infty \]

\( \gamma_0 \]

\( \gamma_0 \]

\( \infty \]

\( \sum_{j=1}^{p} \]

\( \sum_{j=1}^{p} \]

\( I(x_j \geq 0.5) \]

\( f(x) \]
Results

\[ \frac{\kappa}{\nu} = 3 \]
\[ \frac{\kappa}{\nu} = 100 \]
\[ \frac{\kappa}{\nu} = \infty \]

Fixed \( m \)-step:
\[ f(x) = \frac{20}{\sqrt{p}} \sum_{j=1}^{p} I(x_j \geq 0.5) \]

\[ p \text{-step: } f(x) = \frac{20}{\sqrt{p}} \sum_{j=1}^{p} I(x_j \geq 0.5) \]
Results

\[ \frac{\kappa}{\delta} = 3 \]
\[ \frac{\kappa}{\delta} = 100 \]
\[ \frac{\kappa}{\delta} = \infty \]

Fixed \( m \)

1-step: \( f(x) = 20I(x \geq 0.5) \) (\( p = 1 \))

\( p = 15 \) inputs (each a different branch)
\( f \) not additive at all!
Variable Selection

Friedman: $p = 10$ (5 real)

100-Step: $p = 200$ (100 real)

Welch: $p = 200$ (20 real)

FVAR = Proportion of branches involving “false” variables
1. Recap of BART

2. Bayesian Inference of the Number of Trees
   i. Motivation
   ii. A Fully Bayesian Model
   iii. Sampling from the Posterior Distribution
   iv. Code
   v. Simulations
   vi. Application to Real Data

3. Conclusion
# Real Datasets

## Real Data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$n_{\text{train}}$</th>
<th>$n_{\text{test}}$</th>
<th>$p$</th>
<th>$N_{\text{mc (infer m)}}$</th>
<th>$N_{\text{mc (fix m)}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surge</td>
<td>3,000</td>
<td>1,000</td>
<td>6</td>
<td>62,000 (✗)</td>
<td>22,000</td>
</tr>
<tr>
<td>Boston</td>
<td>378</td>
<td>128</td>
<td>13</td>
<td>1,000,000 (✓)</td>
<td>3,000</td>
</tr>
<tr>
<td>Superconductor</td>
<td>15,898</td>
<td>5,299</td>
<td>81</td>
<td>100,000 (✗)</td>
<td>100,000</td>
</tr>
</tbody>
</table>
Results

Surge: $p = 6$, $n_{\text{train}} = 3,000$

Boston: $p = 13$, $n_{\text{train}} = 378$

SupC: $p = 81$, $n_{\text{train}} = 15,898$

$\frac{\kappa}{\alpha} = 3$

$\frac{\kappa}{\alpha} = 100$

$\frac{\kappa}{\alpha} = \infty$

Fixed $m$
1. Recap of BART

2. Bayesian Inference of the Number of Trees
   i. Motivation
   ii. A Fully Bayesian Model
   iii. Sampling from the Posterior Distribution
   iv. Code
   v. Simulations
   vi. Application to Real Data

3. Conclusion
Conclusions

• Bayesian Inference of $m$ generally works well
  • Accurate predictions
  • Variable selection
  • Convenience
Conclusions

• Bayesian Inference of $m$ generally works well
  • Accurate predictions
  • Variable selection
  • Convenience
  • Sometimes underfit ($m$ too small)
  • Computation time?
Conclusions

• Bayesian Inference of $m$ generally works well
  • Accurate predictions
  • Variable selection
  • Convenience
  • Sometimes underfit ($m$ too small)
  • Computation time?

• Just fix $m = 200$?
Conclusions

• Bayesian Inference of $m$ generally works well
  • Accurate predictions
  • Variable selection
  • Convenience
  • Sometimes underfit ($m$ too small)
  • Computation time?

• Just fix $m = 200$?

• Recommendation: Truncated Poisson prior with $\theta = 200$ ($\kappa_0 = \infty$)
Conclusions

• Bayesian Inference of $m$ generally works well
  • Accurate predictions
  • Variable selection
  • Convenience
  • Sometimes underfit ($m$ too small)
  • Computation time?
• Just fix $m = 200$?
  • Recommendation: Truncated Poisson prior with $\theta = 200$ ($\kappa_0 = \infty$)
  • Maybe also try $\kappa_0 = 3$
Conclusions

• Bayesian Inference of $m$ generally works well
  • Accurate predictions
  • Variable selection
  • Convenience
  • Sometimes underfit ($m$ too small)
  • Computation time?

• Just fix $m = 200$?

• Recommendation: Truncated Poisson prior with $\theta = 200$ ($\kappa_0 = \infty$)

• Or try two values of $\kappa_0$

• Boosting
Thank You!