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

[1]Sandia National Laboratories, [2]The Ohio State University, [3]Arizona State University,
[4]Wharton School at the University of Pennsylvania

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

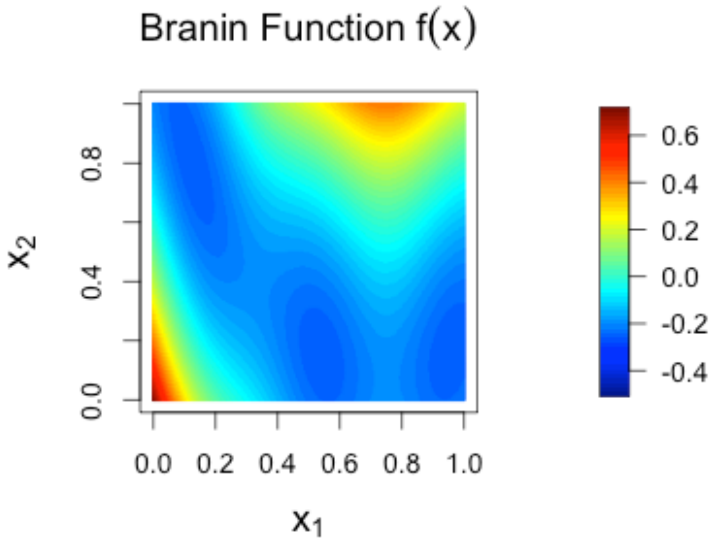
- $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$
- input $\boldsymbol{x}_i \in \mathbb{R}^p \rightarrow$ response $y_i \in \mathbb{R}$

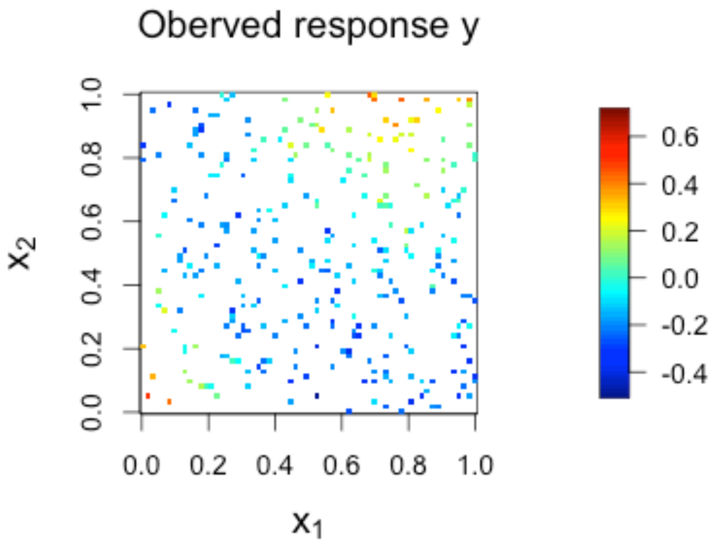
Regression Model

- $y_i \mid \boldsymbol{x}_i \sim N(f(\boldsymbol{x}_i), \sigma^2), i = 1, \ldots, n \ (ind)$
- $f: \mathbb{R}^p \rightarrow \mathbb{R}$ (mean function)
- $\sigma^2 \geq 0$ (residual variance)

"Branin" Example:

- $p = 2$
- $f =$ "The Branin Function"
- $n = 300$
- $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{300} \sim \text{Unif}(0,1)^2 (iid)$
- $\sigma^2 = 1$





Bayesian Additive Regression Trees

Data

- $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$
- input $\boldsymbol{x}_i \in \mathbb{R}^p \rightarrow$ response $y_i \in \mathbb{R}$

Regression Model

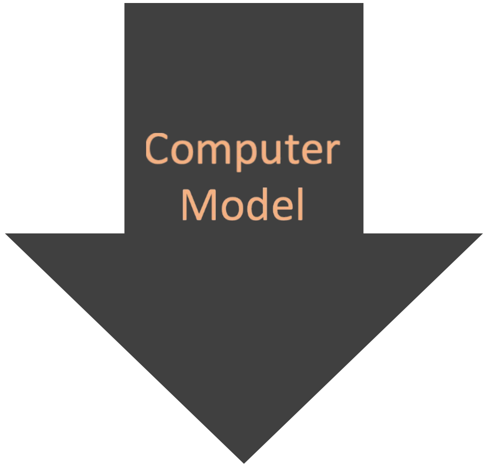
- $y_i \mid \boldsymbol{x}_i \sim N(f(\boldsymbol{x}_i), \sigma^2), \mathrm{i} = 1, \ldots, n$ (ind)
- $f: \mathbb{R}^p \rightarrow \mathbb{R}$ (mean function)
- $\sigma^2 \geq 0$ (residual variance)

Hurricane Example:

- $p = 6$
- $f =$ Computer Model
- $n = 4{,}000$
- Goal: Infer f for sensitivity analysis,
  model calibration, etc.

Input $\boldsymbol{x}$

- 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

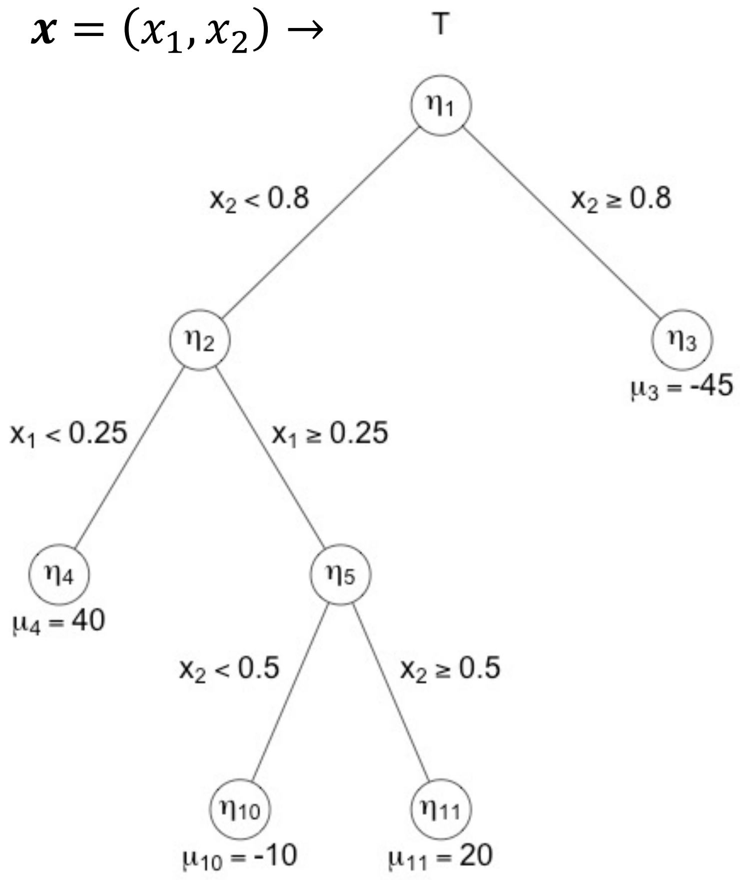


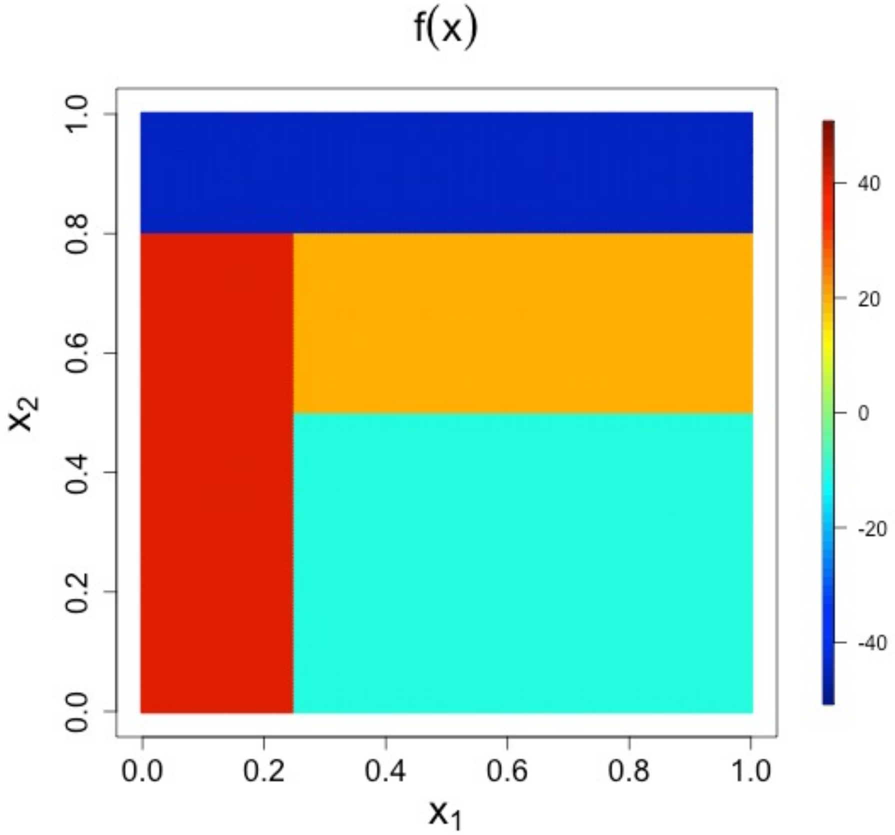
Response y

y = Maximum Water Level During a Storm Surge

Bayesian Additive Regression Trees

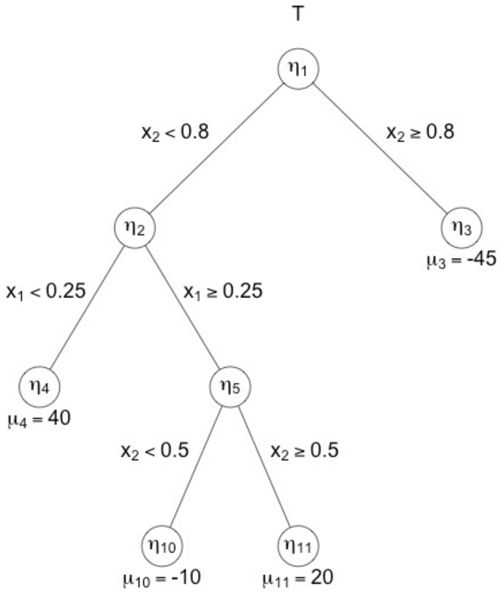
$\boldsymbol{x} = (x_1, x_2) \rightarrow$

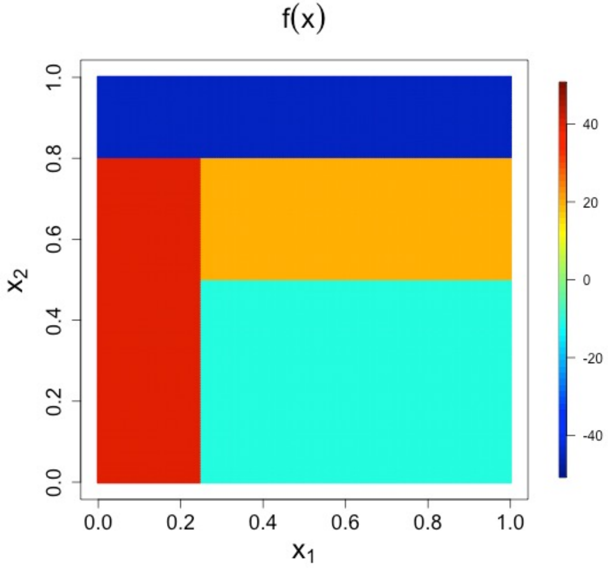




Bayesian Additive Regression Trees

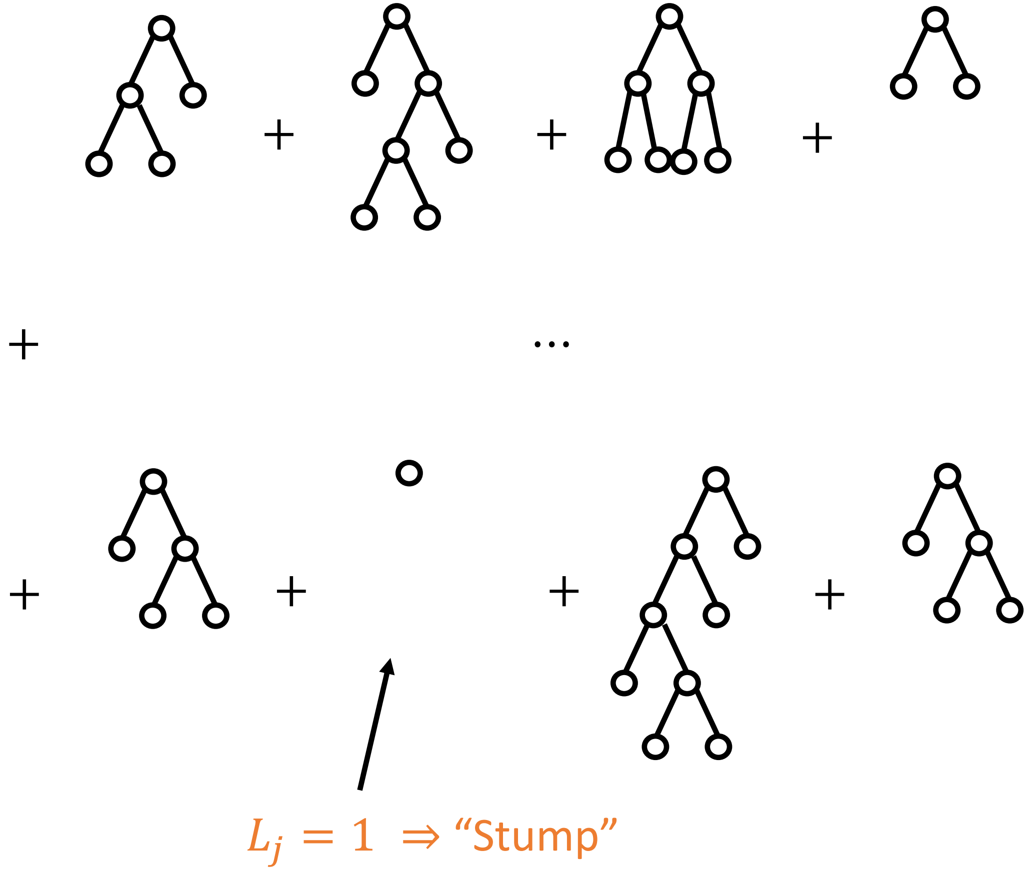
- T has L "terminal nodes"
- Terminal node parameter $\boldsymbol{\mu} \in \mathbb{R}^L$
- $f(\boldsymbol{x}) = g(\boldsymbol{x}; T, \boldsymbol{\mu})$





Bayesian Additive Regression Trees

- $T_{1:m} \equiv T_1, \ldots, T_m \;\; (m \approx 200)$
- $\boldsymbol{\mu}_{1:m} \equiv \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m$
- T_j has L_j terminal nodes
- $\boldsymbol{\mu}_j \in \mathbb{R}^{L_j}$
- $f(\boldsymbol{x}) = \sum_{j=1}^{m} g(\boldsymbol{x}; T_j, \boldsymbol{\mu}_j)$



Bayesian Additive Regression Trees

Prior Distribution $\pi(T_{1:m}, \boldsymbol{\mu}_{1:m}, \sigma^2) = \pi(\sigma^2) \prod_{j=1}^{m} \pi(T_j)\pi(\boldsymbol{\mu}_j \mid T_j)$

- $T_j \sim$ 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

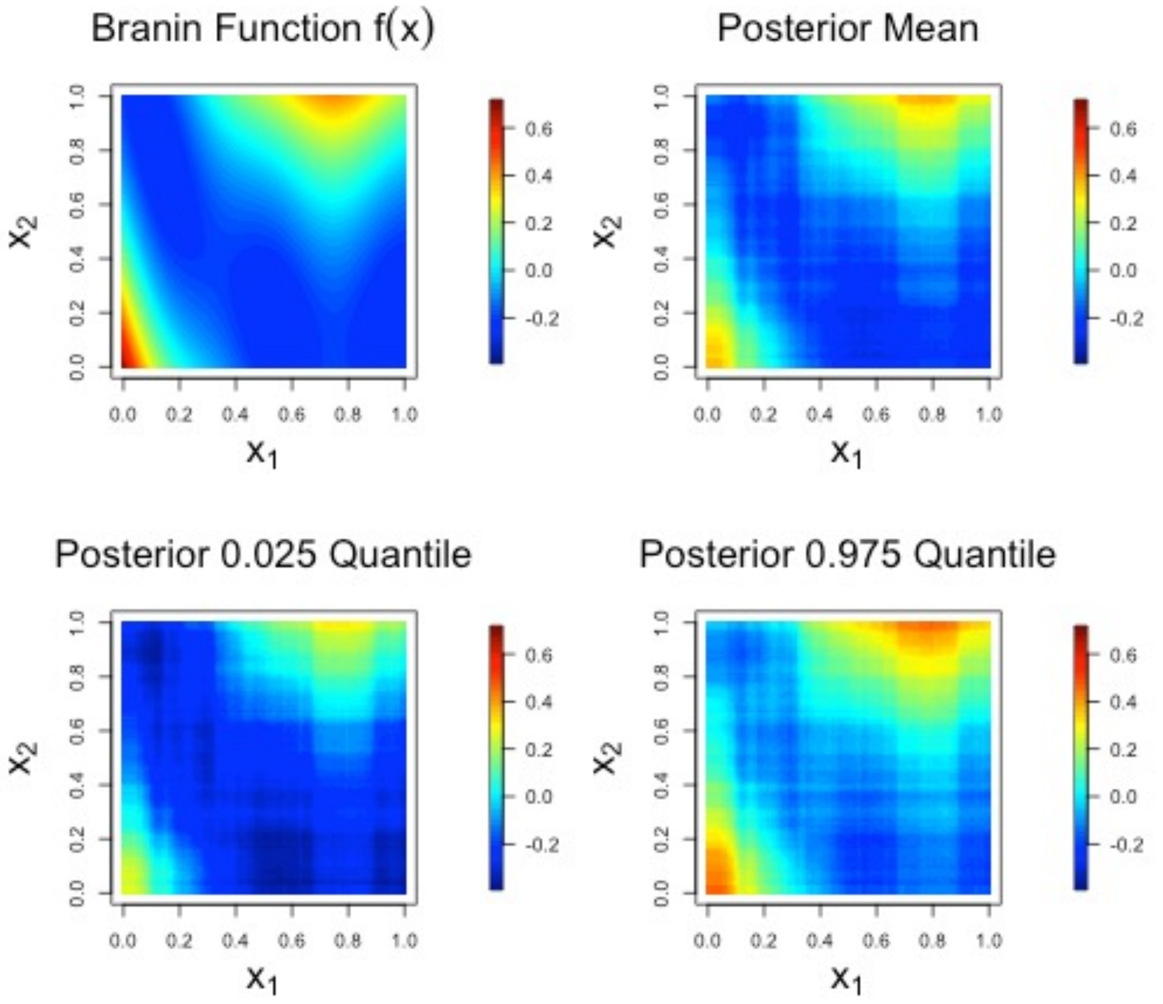
Posterior Sampling MCMC Algorithm

[Notation: $T_{-j} \equiv (T_1, \dots, T_{j-1}, T_{j+1}, \dots, T_m)$ *and* $\boldsymbol{\mu}_{-j} \equiv (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{j-1}, \boldsymbol{\mu}_{j+1}, \dots, \boldsymbol{\mu}_m)$ *]*

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

1. For $j = 1, \dots, m$
    a. Sample $T_j \mid (T_{-j}, \boldsymbol{\mu}_{-j}, \sigma^2, \text{data})$ (Metropolis–Hastings)
    b. Sample $\boldsymbol{\mu}_j \mid \cdot$ (Gibbs Step)
2. Sample $\sigma^2 \mid \cdot$ (Gibbs Step)

Bayesian Additive Regression Trees



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 $m = 200$

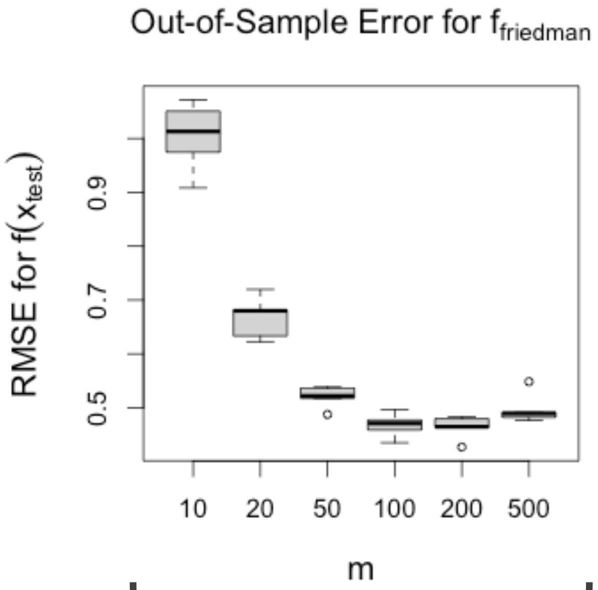
Large m

- Flexible Estimation
- More computation
- Risk overfitting

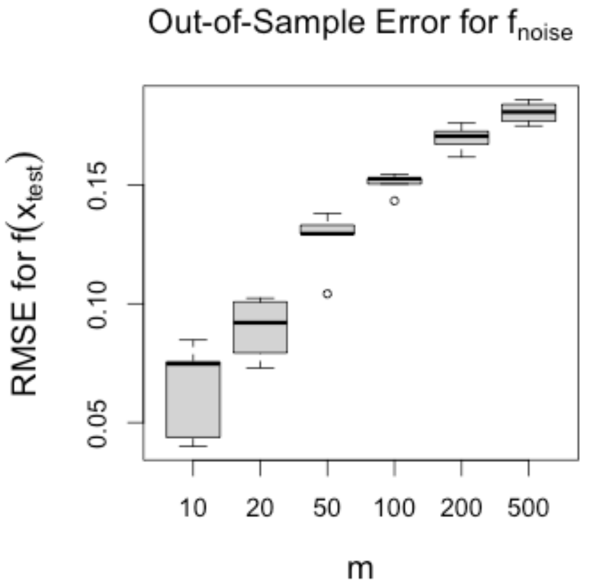
Small m

- Improved variable selection
- Less computation
- Risk underfitting

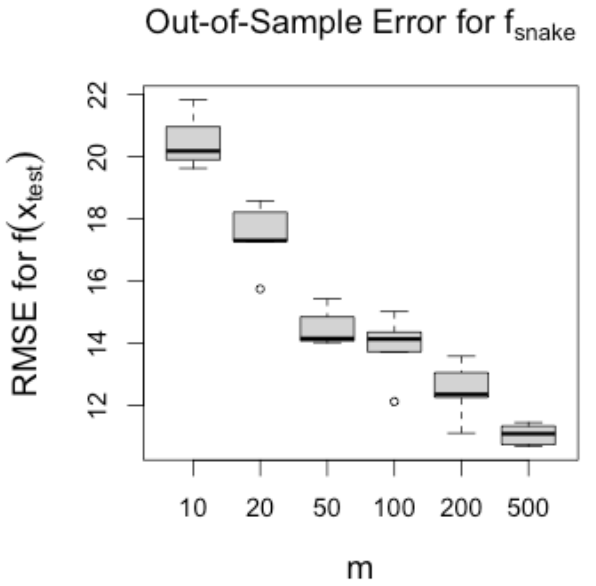
Out-of-Sample Prediction



"Typical"



Small m best



Large m best

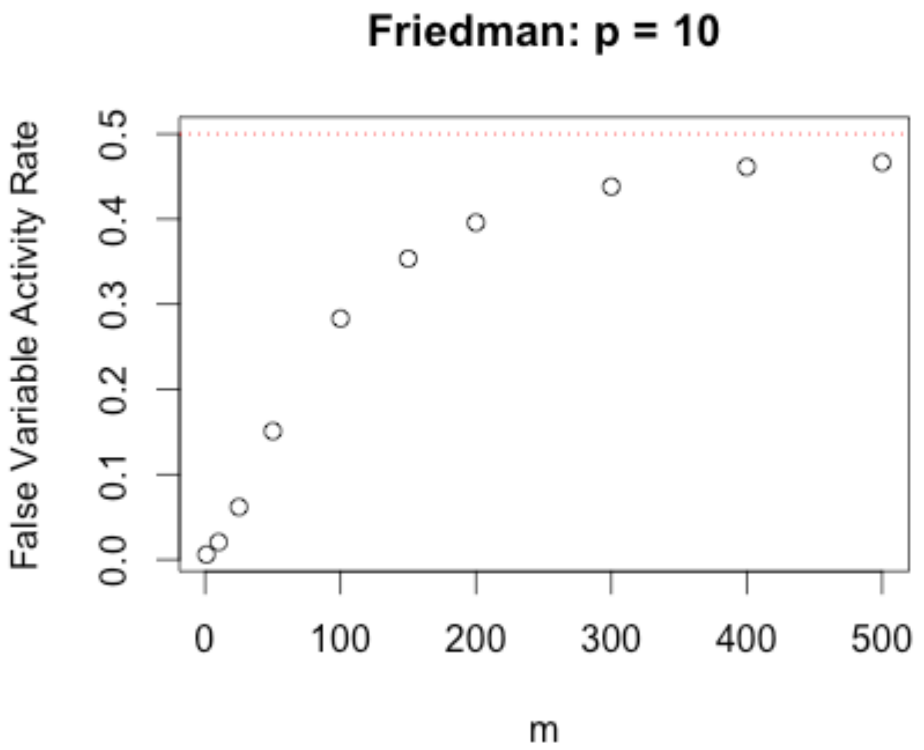
$$\text{“RMSE for } f(\boldsymbol{x}_{\text{test}})\text{”} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(f(\boldsymbol{x}_{\text{test},i}) - \hat{f}(\boldsymbol{x}_{\text{test},i}) \right)^{\mathbf{2}}}$$

Variable Selection

"Real" Inputs

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

"False" Inputs



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

How to pick the grid?

- 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

  } Expensive!
- 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 $m = m_{CV}$

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 $\boldsymbol{m}$

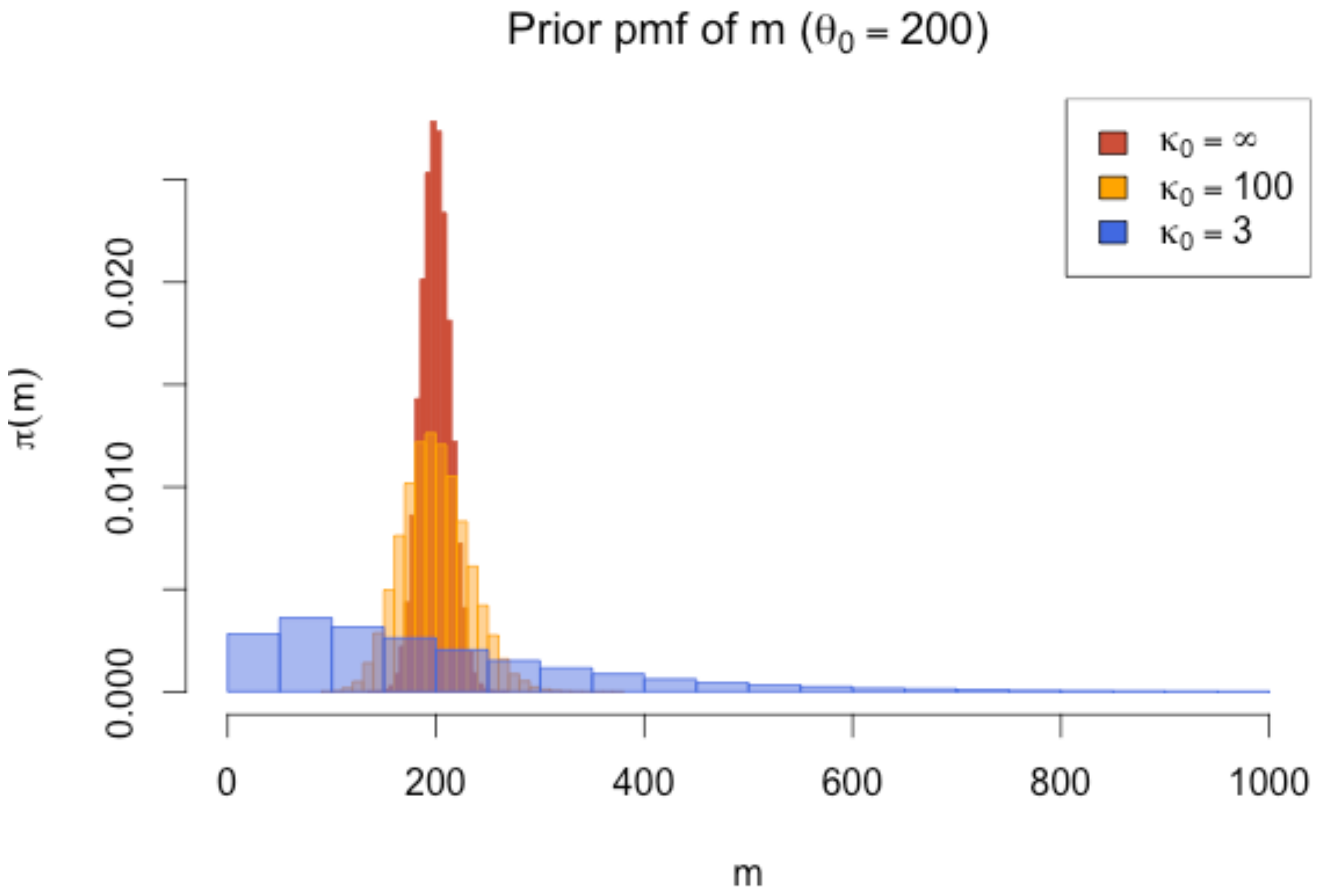
$$\pi(m, T_{1:m}, \boldsymbol{\mu}_{1:m}, \sigma^2) = \pi(\sigma^2)\pi(m) \prod_{j=1}^{m} \pi(T_j)\pi(\boldsymbol{\mu}_j \mid T_j, m)$$

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

Prior Distribution

$\pi(m) \propto \frac{\theta^m e^{-\theta}}{m!} \boldsymbol{I}(1 \leq m \leq 1000)$ *(Truncated Poisson)*

- $\Rightarrow E(m) \approx \theta$
- Default $\theta = 200$
- Optionally, assign θ a hyperprior
  - $\theta \sim \frac{\theta_0 \chi^2_{\kappa_0}}{\kappa_0}$
  - $E(\theta) = \theta_0$
  - Degree of Freedom κ_0
  - Default $\theta_0 = 200$



Prior Distribution

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

$$\tau_m = \frac{\max_i y_i - \min_i y_i}{2k\sqrt{m}}$$

$$\Rightarrow f(\boldsymbol{x}) \sim N\left(0, \left(\frac{\max_i y_i - \min_i y_i}{2k}\right)^2\right) \text{ (for all } m\text{)}$$

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 $i = 1, \dots, N_{mcmc}$:

1. Sample $\theta \mid m$ (Gibbs Step; if $\kappa_0 < \infty$)

Posterior Sampling MCMC Algorithm

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

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

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

Randomly select either

a) Birth or
b) Death

Posterior Sampling MCMC Algorithm

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

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

1. Sample $\theta \mid m$ (Gibbs Step; if $\kappa_0 < \infty$)
2. Sample $m \mid \cdot$ (Metropolis-Hastings)
    - If m was increased, sample new $\boldsymbol{\mu}_* \mid \cdot$ (Gibbs Step)

Randomly select either

a) Birth or

b) Death

Posterior Sampling MCMC Algorithm

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

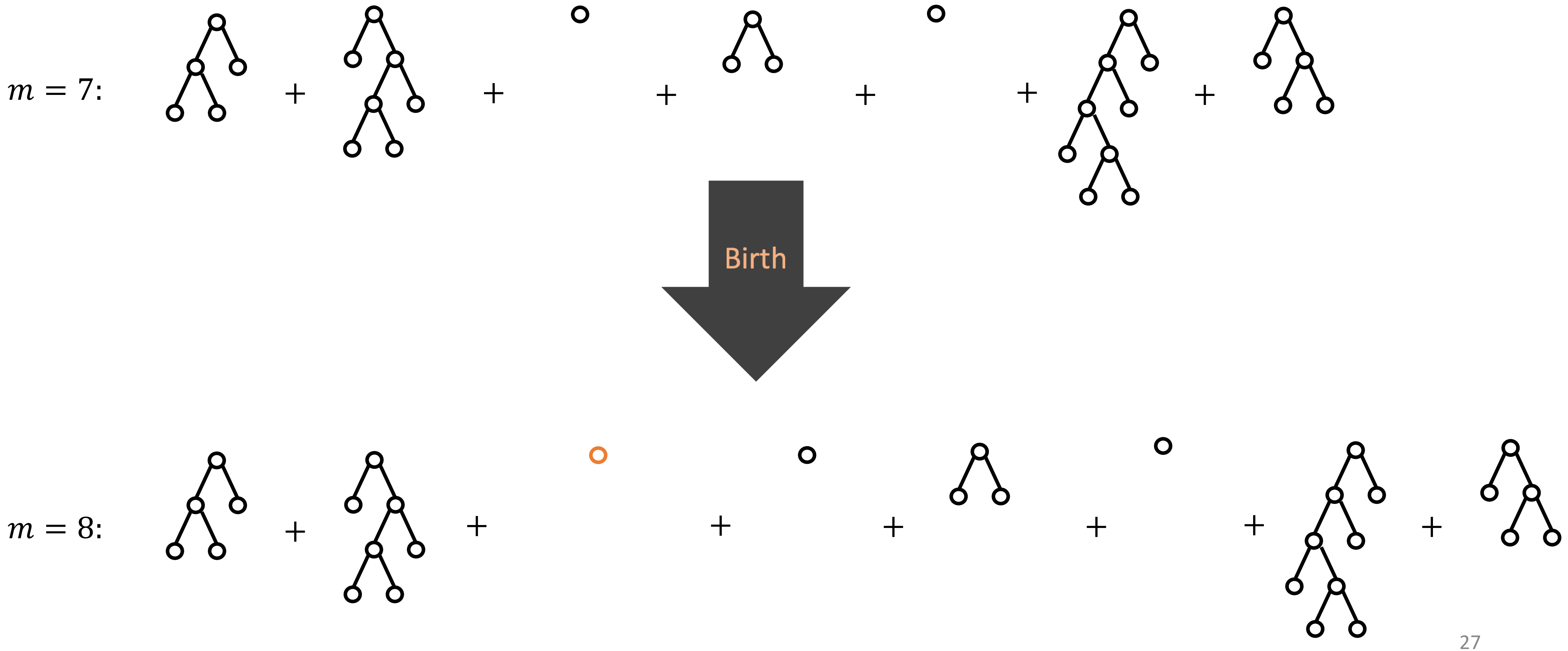
For $i = 1, \dots, N_{mcmc}$:

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

Randomly select either
a) Birth or
b) Death

Same as Standard BART

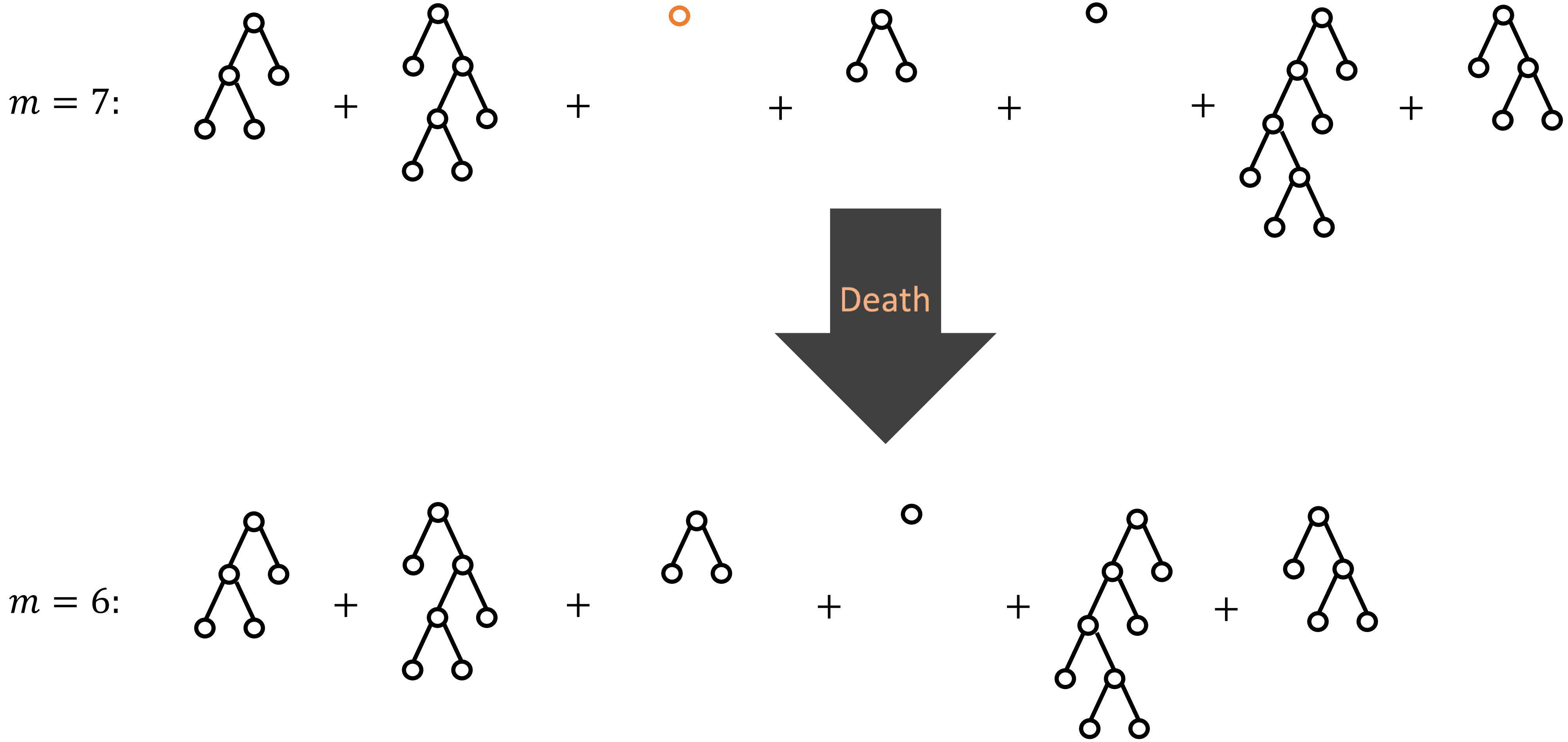
Birth Transition



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



MH Transition

- Current parameters $\psi = (m, T_{1:m}, \boldsymbol{\mu}_{1:m}, \sigma^2)$
- Randomly select either birth or death transition ($\Pr(\text{birth}) = \Pr(\text{death}) = 0.5$)
  - **Birth Transition**
    - Select a location to insert stump $T^\star$
    - Update to $\psi \to \psi^{\text{birth}} = (m + 1, T^\star_{1:(m+1)}, \boldsymbol{\mu}_{1:m}, \sigma^2)$

$$q(\psi \to \psi^{\text{birth}}) = 0.5 \times \frac{1}{m+1}$$

  - **Death Transition (if there are any "stumps")**
    - Select a stump $T^\star$ to delete
    - Update to $\psi \to \psi^{\text{death}} = (m - 1, T^\star_{1:(m-1)}, \boldsymbol{\mu}_{1:(m-1)}, \sigma^2)$

$$q(\psi \to \psi^{\text{death}}) = 0.5 \times \frac{1}{m_{\text{stumps}}}$$

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

MH Acceptance Probability: $\min\{1, \text{MH Ratio}\}$

MH Ratio for Birth

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

Ratio of Transition Probabilities

Ratio of Posterior Distributions

MH Acceptance Probability: $\min\{1, \text{MH Ratio}\}$

MH Ratio for Birth

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

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(\sigma^2)}{\pi(\sigma^2)} \times \frac{\pi(m+1)}{\pi(m)} \times \frac{\pi(T^\star)\prod_j^m \pi(T_j)}{\prod_j^m \pi(T_j)} \times \frac{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m+1, T_j)}{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m, T_j)} \times \frac{0.5/(m_{\text{stumps}}+1)}{0.5/(m+1)}$$

Likelihood

Prior Distribution

Ratio of Transition Probabilities

MH Acceptance Probability: $\min\{1, \text{MH Ratio}\}$

MH Ratio for Birth

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

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(\sigma^2)}{\pi(\sigma^2)} \times \frac{\pi(m+1)}{\pi(m)} \times \frac{\pi(T^\star)\prod_j^m \pi(T_j)}{\prod_j^m \pi(T_j)} \times \frac{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m+1, T_j)}{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m, T_j)} \times \frac{0.5/(m_{\text{stumps}}+1)}{0.5/(m+1)}$$

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(m+1)}{\pi(m)} \times \pi(T^\star) \times \frac{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m+1, T_j)}{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m, T_j)} \times \frac{1/(m_{\text{stumps}}+1)}{1/(m+1)}$$

MH Acceptance Probability: $\min\{1, \text{MH Ratio}\}$

MH Ratio for Birth

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

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(\sigma^2)}{\pi(\sigma^2)} \times \frac{\pi(m+1)}{\pi(m)} \times \frac{\pi(T^\star)\prod_j^m \pi(T_j)}{\prod_j^m \pi(T_j)} \times \frac{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m+1, T_j)}{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m, T_j)} \times \frac{0.5/(m_{\text{stumps}}+1)}{0.5/(m+1)}$$

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(m+1)}{\pi(m)} \times \pi(T^\star) \times \frac{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m+1, T_j)}{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m, T_j)} \times \frac{1/(m_{\text{stumps}}+1)}{1/(m+1)}$$

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\theta}{m+1} \times$$

MH Acceptance Probability: $\min\{1, \text{MH Ratio}\}$

MH Ratio for Birth

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

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(\sigma^2)}{\pi(\sigma^2)} \times \frac{\pi(m+1)}{\pi(m)} \times \frac{\pi(T^\star)\prod_j^m \pi(T_j)}{\prod_j^m \pi(T_j)} \times \frac{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m+1, T_j)}{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m, T_j)} \times \frac{0.5/(m_{\text{stumps}}+1)}{0.5/(m+1)}$$

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(m+1)}{\pi(m)} \times \pi(T^\star) \times \frac{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m+1, T_j)}{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m, T_j)} \times \frac{1/(m_{\text{stumps}}+1)}{1/(m+1)}$$

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\theta}{m+1} \times \pi(T^\star) \times$$

MH Acceptance Probability: $\min\{1, \text{MH Ratio}\}$

MH Ratio for Birth

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

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(\sigma^2)}{\pi(\sigma^2)} \times \frac{\pi(m+1)}{\pi(m)} \times \frac{\pi(T^\star)\prod_j^m \pi(T_j)}{\prod_j^m \pi(T_j)} \times \frac{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m+1, T_j)}{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m, T_j)} \times \frac{0.5/(m_{\text{stumps}}+1)}{0.5/(m+1)}$$

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(m+1)}{\pi(m)} \times \pi(T^\star) \times \frac{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m+1, T_j)}{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m, T_j)} \times \frac{1/(m_{\text{stumps}}+1)}{1/(m+1)}$$

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\theta}{m+1} \times \pi(T^\star) \times \left(\frac{m}{m+1}\right)^{-\sum_{j=1}^m \frac{L_j}{2}} \exp\left(-\frac{1}{2m\tau_m^2}\sum_j^m \sum_{\ell=1}^{L_j} \mu_{j\ell}^2\right)$$

MH Acceptance Probability: $\min\{1, \text{MH Ratio}\}$

MH Ratio for Birth

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

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(\sigma^2)}{\pi(\sigma^2)} \times \frac{\pi(m+1)}{\pi(m)} \times \frac{\pi(T^\star)\prod_j^m \pi(T_j)}{\prod_j^m \pi(T_j)} \times \frac{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m+1, T_j)}{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m, T_j)} \times \frac{0.5/(m_{\text{stumps}}+1)}{0.5/(m+1)}$$

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\pi(m+1)}{\pi(m)} \times \pi(T^\star) \times \frac{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m+1, T_j)}{\prod_j^m \pi(\boldsymbol{\mu}_j \mid m, T_j)} \times \frac{1/(m_{\text{stumps}}+1)}{1/(m+1)}$$

$$= \frac{\pi(\text{data} \mid \psi^{\text{birth}})}{\pi(\text{data} \mid \psi)} \times \frac{\theta}{m+1} \times \pi(T^\star) \times \left(\frac{m}{m+1}\right)^{-\sum_{j=1}^m \frac{L_j}{2}} \exp\left(-\frac{1}{2m\tau_m^2}\sum_j^m \sum_{\ell=1}^{L_j} \mu_{j\ell}^2\right) \times \frac{m+1}{m_{\text{stumps}}+1}$$

Likelihood Ratio

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

Likelihood Ratio

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

$$= \frac{\pi(\text{data} \mid m+1, T^{\star}_{1:(m+1)}, \boldsymbol{\mu}_{1:m}, \sigma^2)}{\pi(\text{data} \mid m, T_{1:m}, \boldsymbol{\mu}_{1:m}, \sigma^2)}$$

Marginal Likelihood (m+1 trees)

Full likelihood (m trees)

Likelihood Ratio

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

$$= \frac{\pi(\text{data} \mid m+1, T^{\star}_{1:(m+1)}, \boldsymbol{\mu}_{1:m}, \sigma^2)}{\pi(\text{data} \mid m, T_{1:m}, \boldsymbol{\mu}_{1:m}, \sigma^2)}$$

$$= \frac{\int_{\mathbb{R}} \pi(\text{data} \mid m+1, T^{\star}_{1:(m+1)}, \boldsymbol{\mu}_{1:m}, \mu^{\star}, \sigma^2)\pi(\mu^{\star} \mid m+1, T^{\star})\, d\mu^{\star}}{\pi(\text{data} \mid m, T_{1:m}, \boldsymbol{\mu}_{1:m}, \sigma^2)}$$

Full Likelihood (m+1)

Prior Distribution of $\mu^{\star}$

Likelihood Ratio

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

$$= \frac{\pi(\text{data} \mid m+1, T^{\star}_{1:(m+1)}, \boldsymbol{\mu}_{1:m}, \sigma^2)}{\pi(\text{data} \mid m, T_{1:m}, \boldsymbol{\mu}_{1:m}, \sigma^2)}$$

$$= \frac{\int_{\mathbb{R}} \pi(\text{data} \mid m+1, T^{\star}_{1:(m+1)}, \boldsymbol{\mu}_{1:m}, \mu^{\star}, \sigma^2)\pi(\mu^{\star} \mid m, T^{\star})\, d\mu^{\star}}{\pi(\text{data} \mid m, T_{1:m}, \boldsymbol{\mu}_{1:m}, \sigma^2)}$$

$$= \frac{\int_{\mathbb{R}} \prod_{i=1}^{n} N\left(y_i; \mu^{\star} + \sum_{j=1}^{m} g(\boldsymbol{x}_i; T_j, \boldsymbol{\mu}_j), \sigma^2\right) N(\mu^{\star}; 0, \tau_m^2)\, d\mu^{\star}}{\prod_{i=1}^{n} N(y_i; \sum_{j=1}^{m} g(\boldsymbol{x}_i; T_j, \boldsymbol{\mu}_j), \sigma^2)}$$

Likelihood Ratio

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

$$= \frac{\pi(\text{data} \mid m+1, T^{\star}_{1:(m+1)}, \boldsymbol{\mu}_{1:m}, \sigma^2)}{\pi(\text{data} \mid m, T_{1:m}, \boldsymbol{\mu}_{1:m}, \sigma^2)}$$

$$= \frac{\int_{\mathbb{R}} \pi(\text{data} \mid m+1, T^{\star}_{1:(m+1)}, \boldsymbol{\mu}_{1:m}, \mu^{\star}, \sigma^2)\pi(\mu^{\star} \mid m, T^{\star})\, d\mu^{\star}}{\pi(\text{data} \mid m, T_{1:m}, \boldsymbol{\mu}_{1:m}, \sigma^2)}$$

$$= \frac{\int_{\mathbb{R}} \prod_{i=1}^{n} N\left(y_i; \mu^{\star} + \sum_{j=1}^{m} g(\boldsymbol{x}_i; T_j, \boldsymbol{\mu}_j), \sigma^2\right) N(\mu^{\star}; 0, \tau_m^2)\, d\mu^{\star}}{\prod_{i=1}^{n} N(y_i; \sum_{j=1}^{m} g(\boldsymbol{x}_i; T_j, \boldsymbol{\mu}_j), \sigma^2)}$$

$$= \left(\frac{\sigma^2}{n\tau_{m+1}^2 + \sigma^2} + 1\right)^{1/2} \exp\left(\frac{n^2\tau_{m+1}^2\left(\sum_{i=1}^{n}[y_i - \sum_{j=1}^{m} g(\boldsymbol{x}_i; T_j, \boldsymbol{\mu}_j)]\right)^2}{2\sigma^2(n\tau_{m+1}^2 + \sigma^2)}\right)$$

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:
  - bart(X, y, learnntree = TRUE, ntreemean = 200, ntreedf = Inf)

$$m \sim Pois(\theta) I(1 \leq m \leq 1000)$$

$$\theta \sim \frac{\theta_0 \chi^2_{\kappa_0}}{\kappa_0}$$

Simulation Setup

Fully Bayesian Inference for m

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

Cross-Validation

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

Compare accuracy using "RMSE for $f(\boldsymbol{x}_{\text{test}})$" $= \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(f(\boldsymbol{x}_{\text{test},i}) - \hat{f}(\boldsymbol{x}_{\text{test},i}) \right)^{\mathbf{2}}}$

Simulation Setup

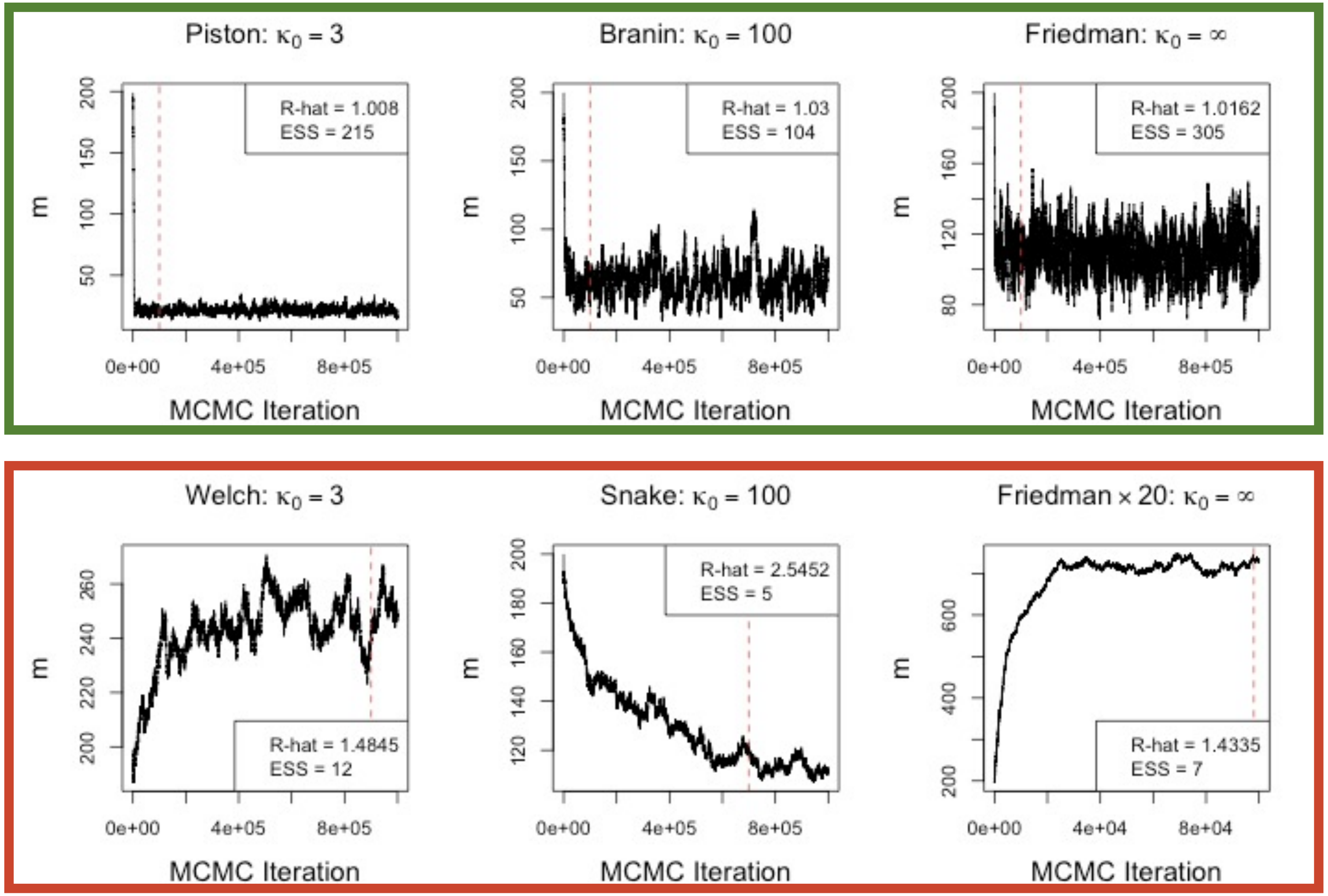
Bayesian Inference

Cross-Validation

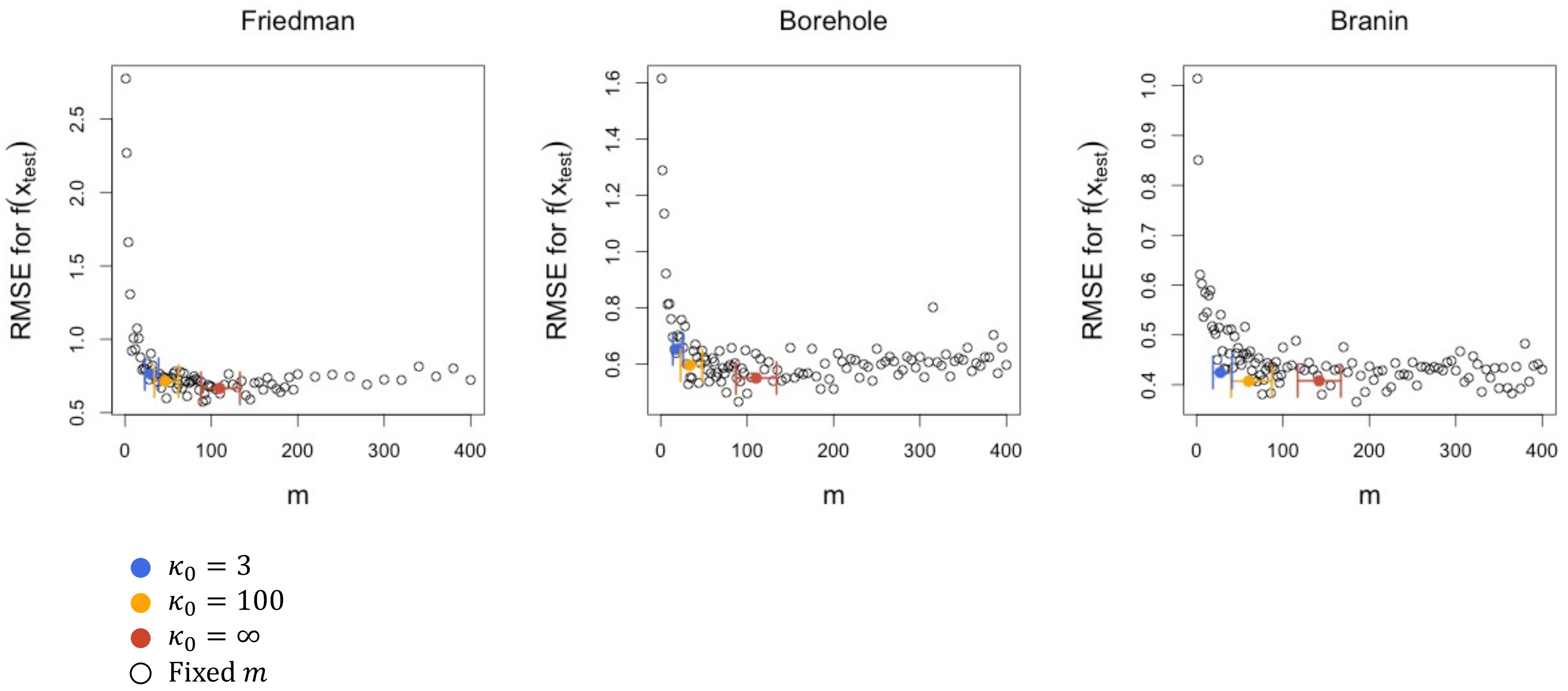
Simulations

f	n_{train}	n_{test}	p	SNR	N_{mc} (infer m)	N_{mc} (fix m)
Friedman	500	1,000	10	23.8	1,000,000 (✓)	3,000
Borehole	500	1,000	8	20.9	1,000,000 (✓)	3,000
Branin	1,000	2,000	2	18.2	1,000,000 (✓)	3,000
Piston	1,500	3,000	7	20.6	1,000,000 (✓)	3,000
Snake	10,000	10,000	2	2930	1,000,000 (✕)	100,000
Welch	10,000	10,000	20	2809	1,000,000 (✕)	100,000
Friedman×20	10,000	10,000	100	23.8	100,000 (✕)	50,000
300-Step	12,000	12,000	300	100	200,000 (✕)	10,000
100-Step	4,000	8,000	100	100	1,000,000 (✕)	10,000
1-Step	100	200	1	100	1,000,000 (✓)	3,000
T_4	800	1,000	15	340	1,000,000 (✕)	30,000

Convergence of m

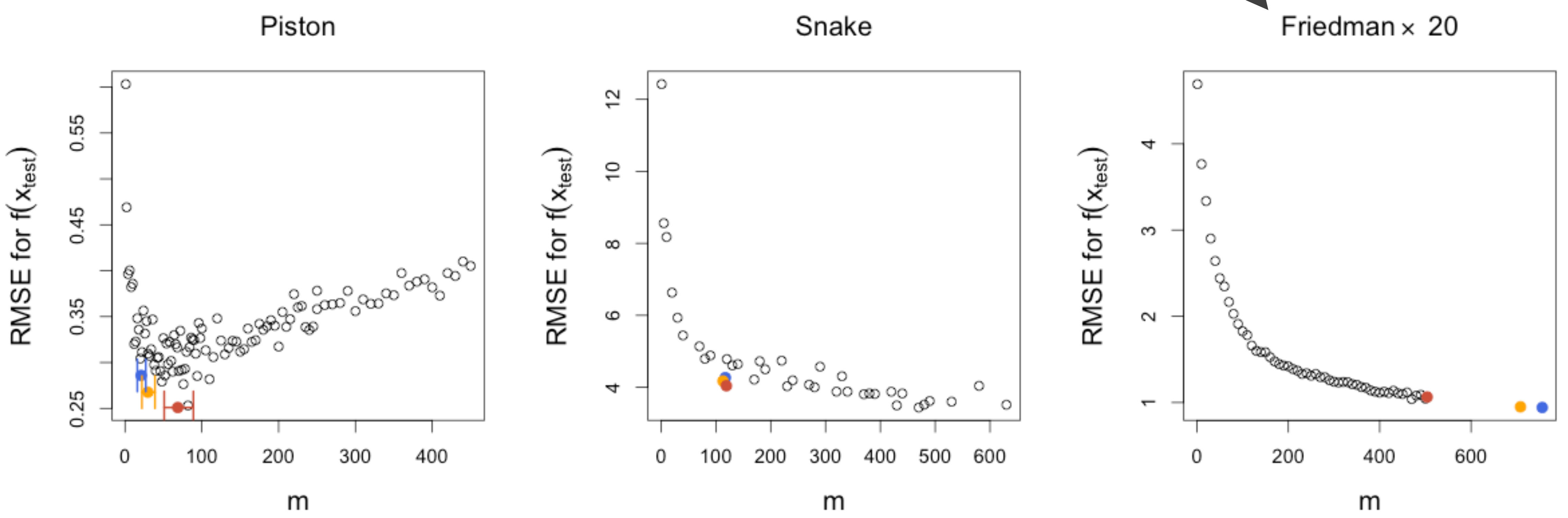


Results


Friedman
RMSE for f(x_test)
0.5
1.0
1.5
2.0
2.5
0
100
200
300
400
m
Borehole
RMSE for f(x_test)
0.6
0.8
1.0
1.2
1.4
1.6
0
100
200
300
400
m
Branin
RMSE for f(x_test)
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
100
200
300
400
m
$\kappa_0 = 3$
$\kappa_0 = 100$
$\kappa_0 = \infty$
Fixed m

Results

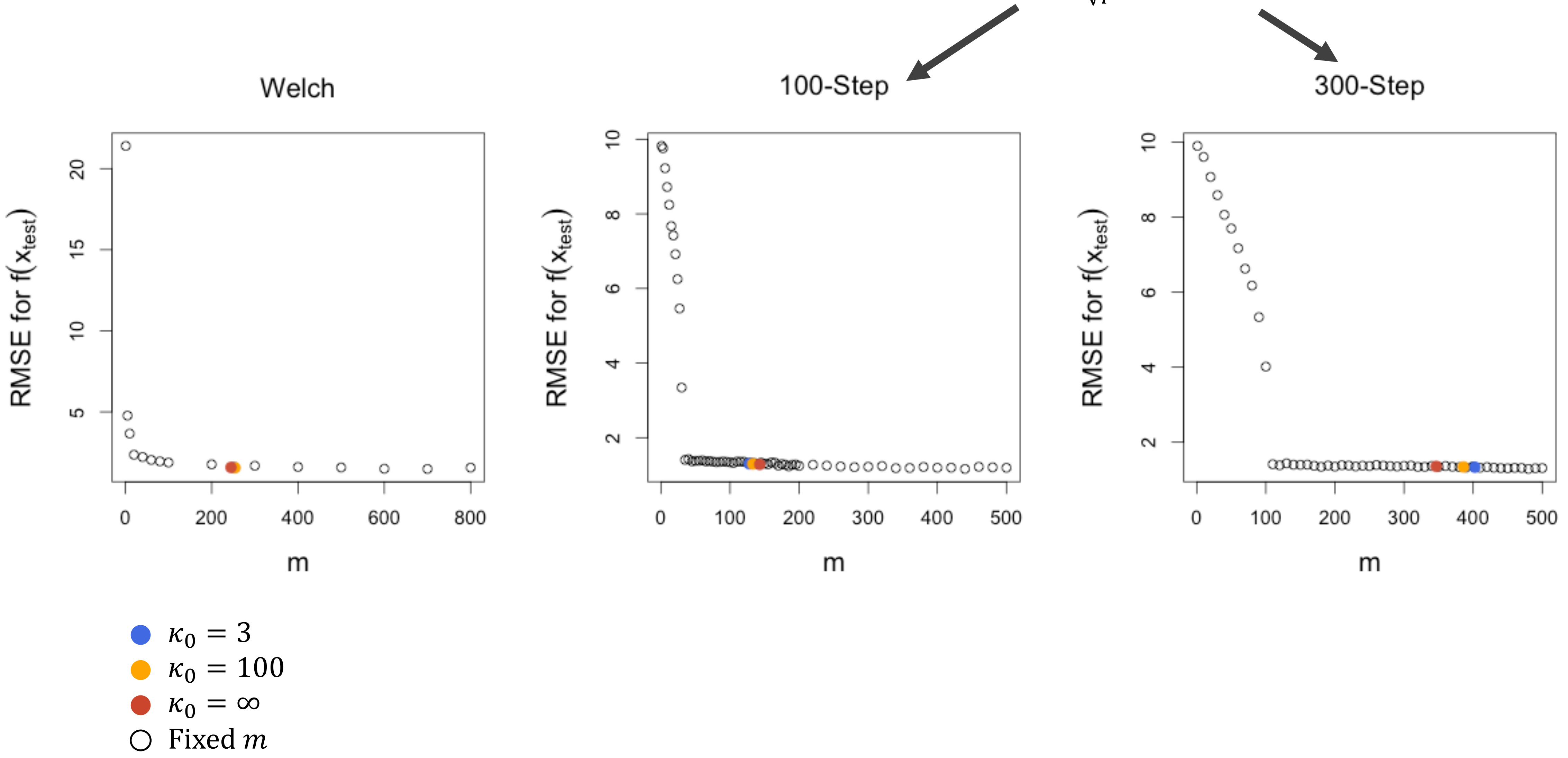
p = 100 inputs: all active!
f has an additive structure



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

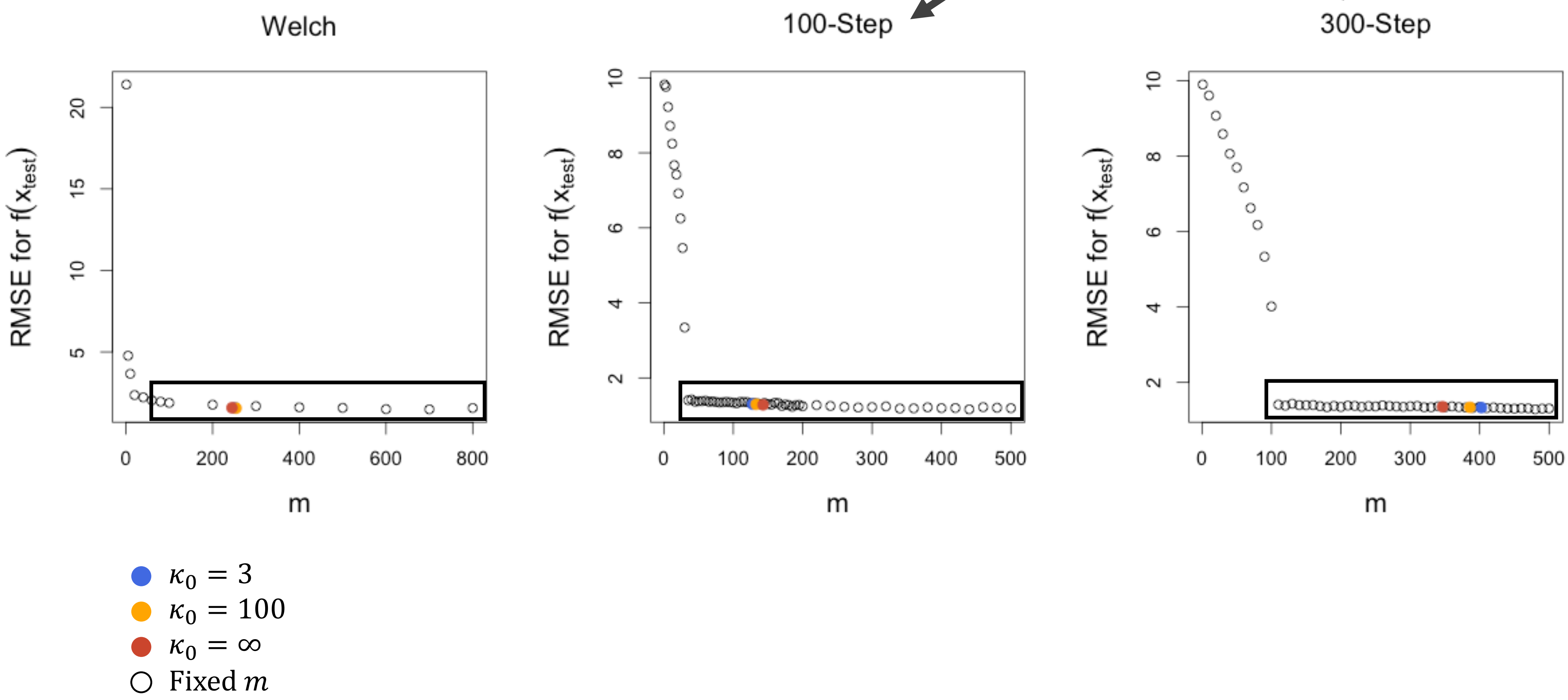
Results

p-step: $f(\boldsymbol{x}) = \frac{20}{\sqrt{p}} \sum_{j=1}^{p} \boldsymbol{I}(\boldsymbol{x}_j \geq 0.5)$



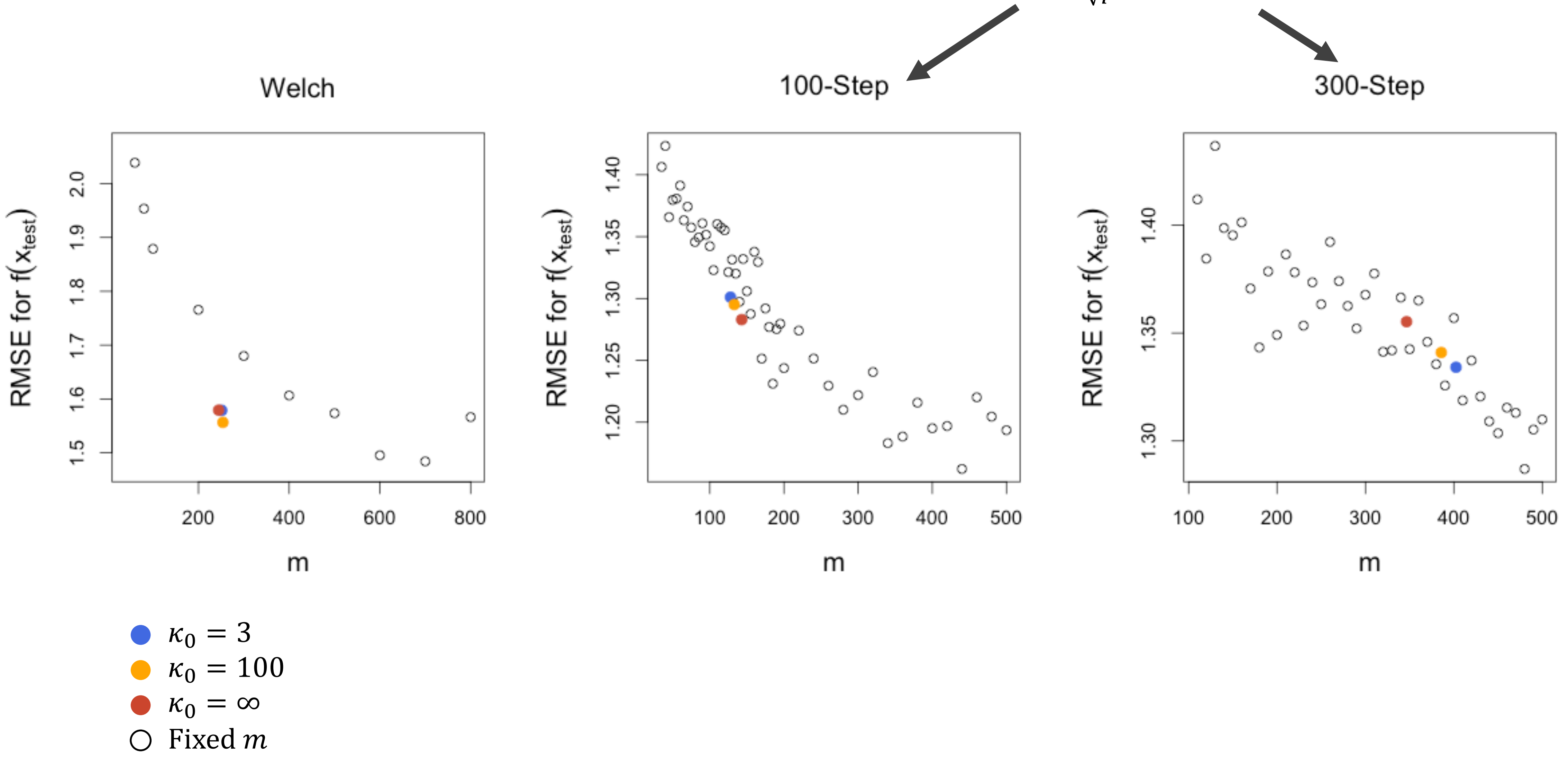
Results

p-step: $f(\boldsymbol{x}) = \frac{20}{\sqrt{p}} \sum_{j=1}^{p} \boldsymbol{I}(\boldsymbol{x}_j \geq 0.5)$

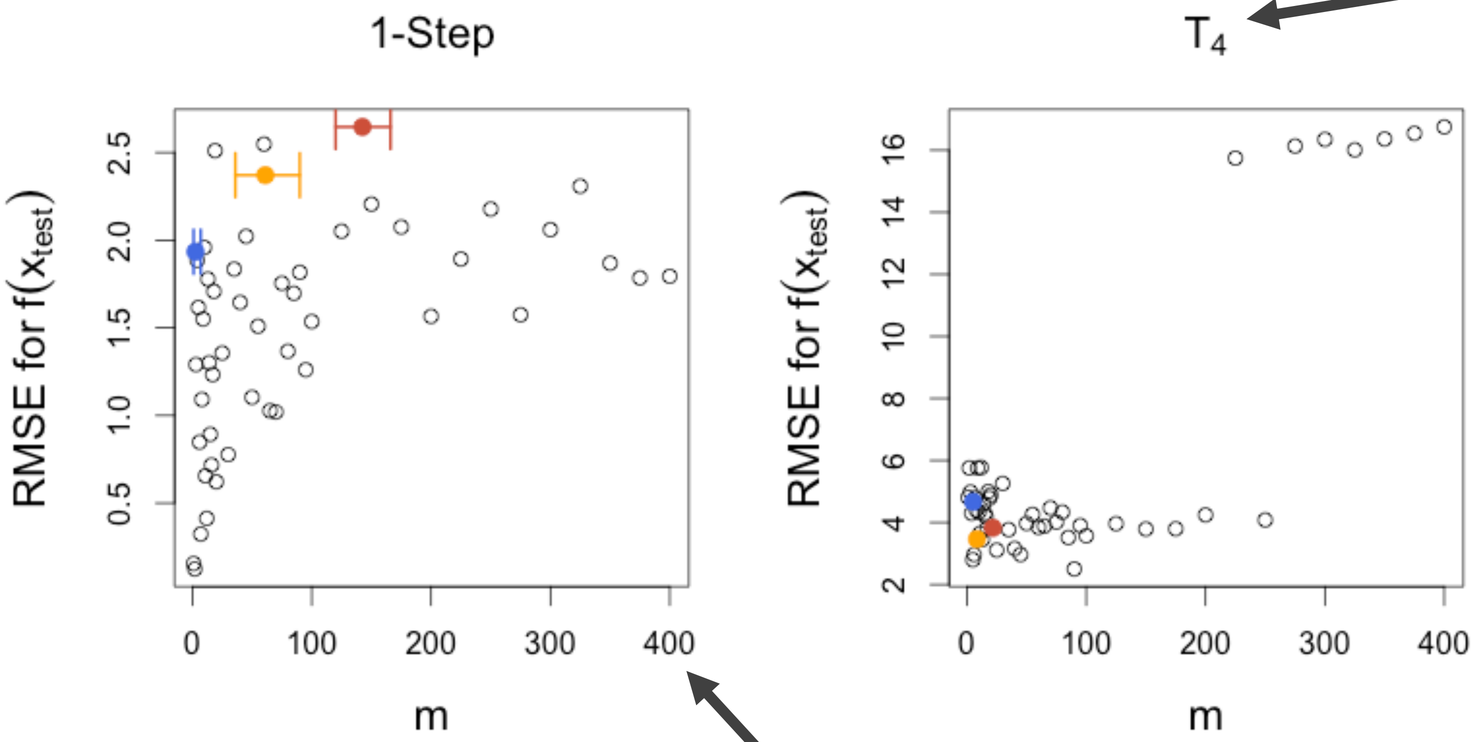


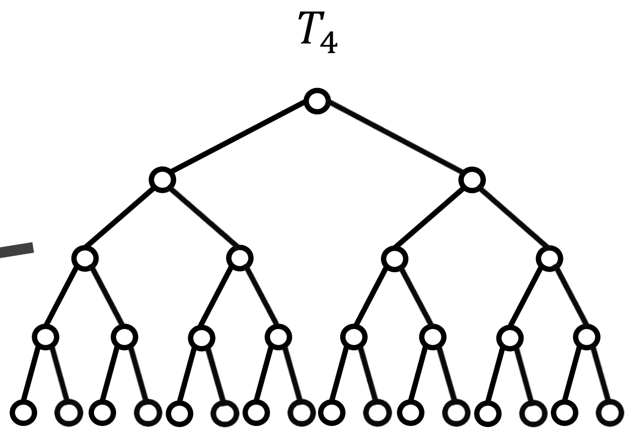
Results

p-step: $f(\boldsymbol{x}) = \frac{20}{\sqrt{p}} \sum_{j=1}^{p} \boldsymbol{I}(\boldsymbol{x}_j \geq 0.5)$



Results



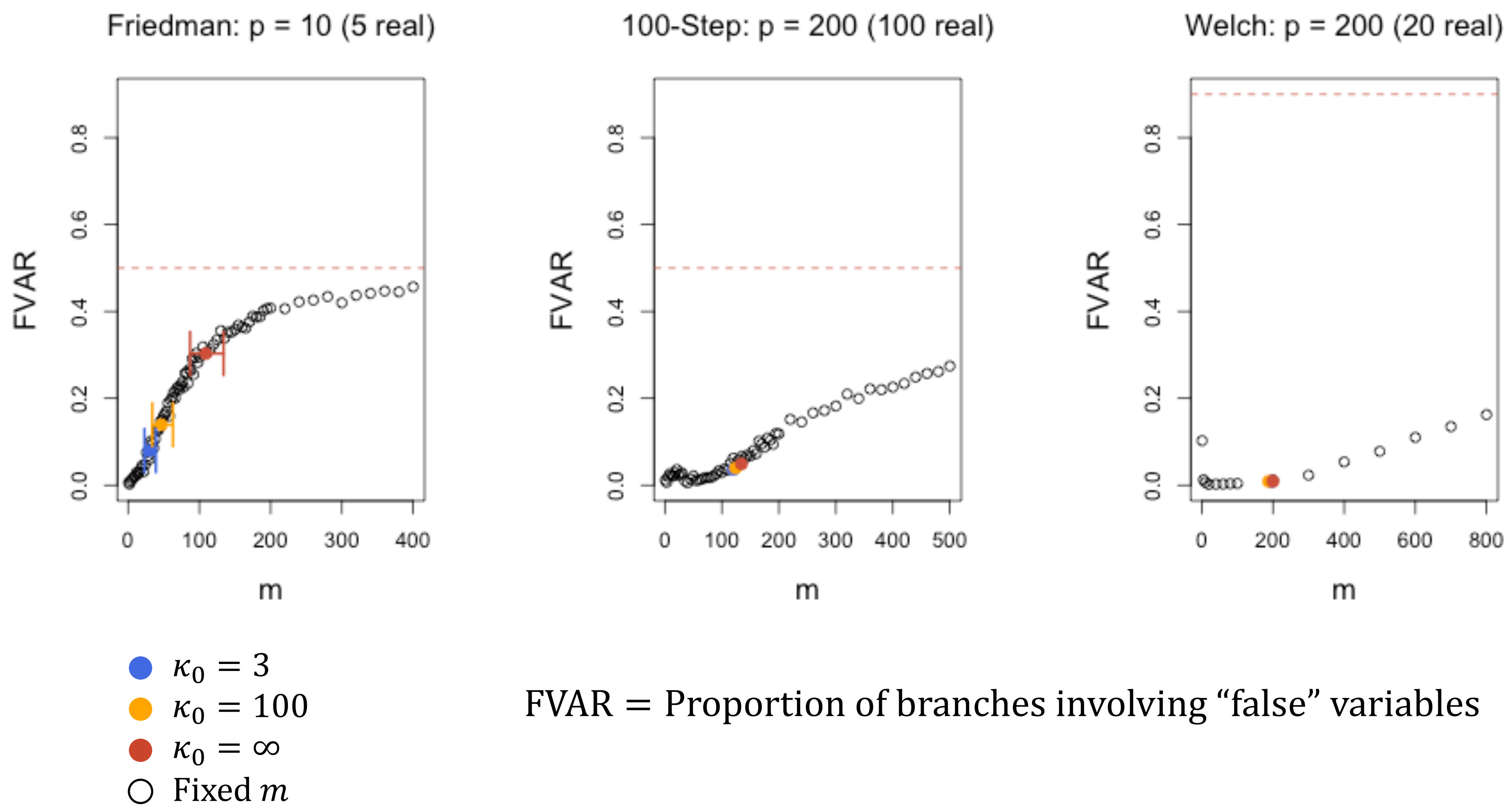


$p = 15$ inputs (each a different branch)
f not additive at all!

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

1-step: $f(x) = 20\boldsymbol{I}(x \geq 0.5)$ $(p = 1)$

Variable Selection



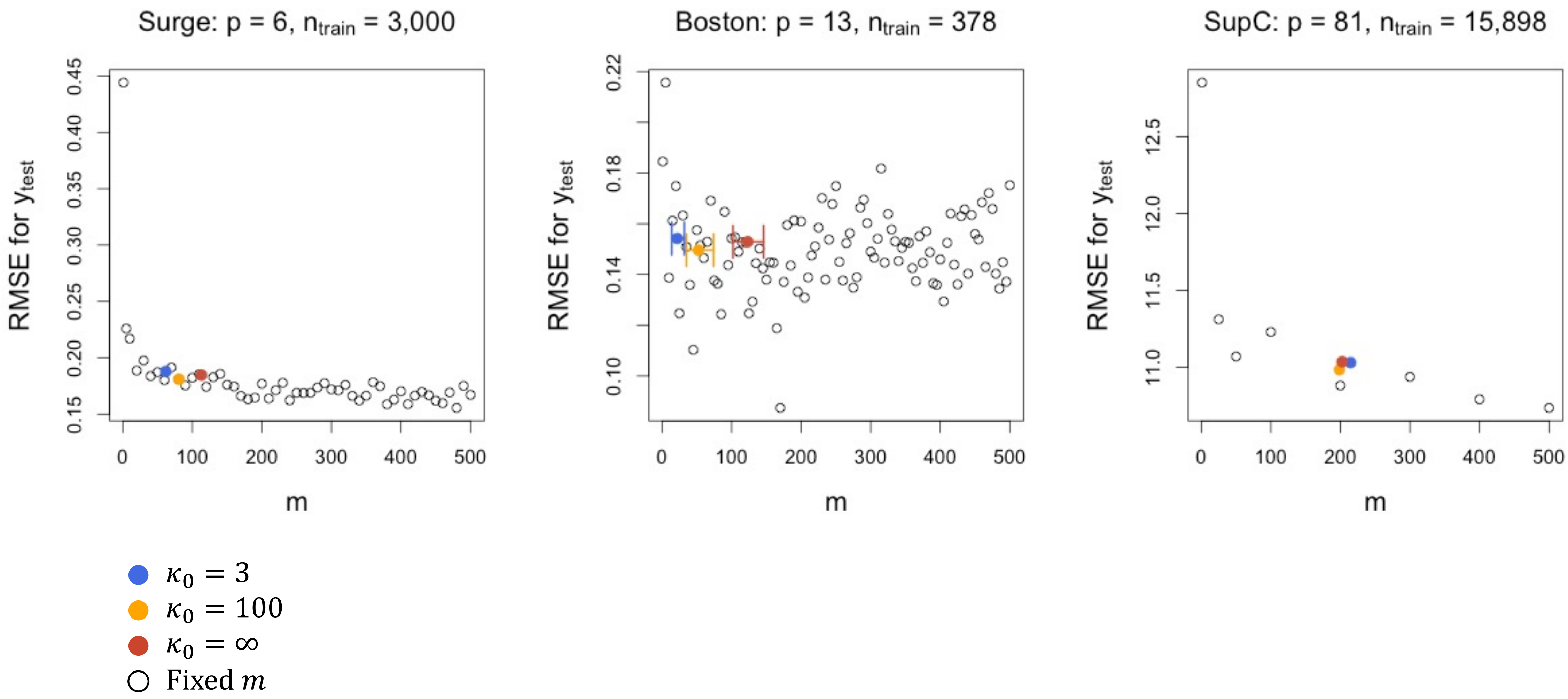
FVAR = Proportion of branches involving "false" variables

Real Datasets

Real Data

Dataset	$\boldsymbol{n}_{\text{train}}$	$\boldsymbol{n}_{\text{test}}$	$\boldsymbol{p}$	$\boldsymbol{N}_{\text{mc}}$ **(infer** $\boldsymbol{m}$**)**	$\boldsymbol{N}_{\text{mc}}$ **(fix** $\boldsymbol{m}$**)**
Surge	3,000	1,000	6	62,000 (✕)	22,000
Boston	378	128	13	1,000,000 (✓)	3,000
Superconductor	15,898	5,299	81	100,000 (✕)	100,000

Results


Surge: p = 6, n_{train} = 3,000
RMSE for y_{test}
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0
100
200
300
400
500
m
Boston: p = 13, n_{train} = 378
RMSE for y_{test}
0.10
0.14
0.18
0.22
0
100
200
300
400
500
m
SupC: p = 81, n_{train} = 15,898
RMSE for y_{test}
11.0
11.5
12.0
12.5
0
100
200
300
400
500
m
$\kappa_0 = 3$
$\kappa_0 = 100$
$\kappa_0 = \infty$
Fixed m

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 κ_0
- Boosting

Thank You!