Bayesian additive regression trees and the general BART model - Tan and Roy, 2019

Review for BART for continuous outcomes

이재용, 2023/01/22

Contents

BART for continuous outcomes

- Regression tree
- Formal definition
- Sum of regression trees
- Simple example
- The BART algorithm
- Performance of the BART

BART for continuous outcomes

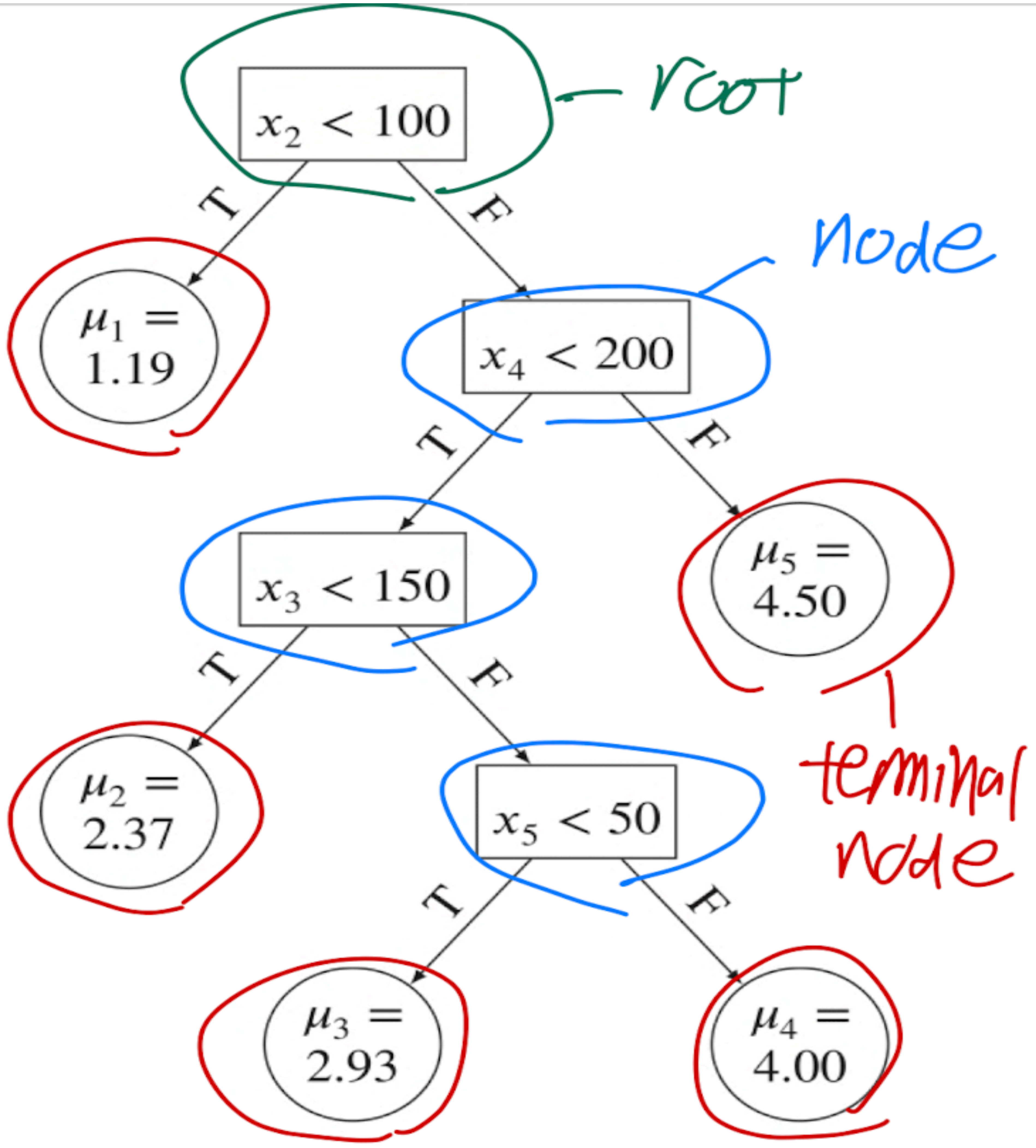
Regression tree

Example of $g(x; T, M)$, where the data looks like

x_1	$\cdots$	x_p	y
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Thus $g(x; T, M)$ is a function that assigns the value of μ_i to $E(Y|X)$ via binary decision rules denoted as T.

We can also view it as an ANOVA model =>



$$
\begin{aligned}
y = {} & \mu_1 I\{x_2 < 100\} + \mu_2 I\{x_2 \geq 100\} I\{x_4 < 200\} I\{x_3 < 150\} \\
& + \mu_3 I\{x_2 \geq 100\} I\{x_4 < 200\} I\{x_3 \geq 150\} I\{x_5 < 50\} \\
& + \mu_4 I\{x_2 \geq 100\} I\{x_4 < 200\} I\{x_3 \geq 150\} I\{x_5 \geq 50\} \\
& + \mu_5 I\{x_2 \geq 100\} I\{x_4 \geq 200\} + \varepsilon,
\end{aligned}
$$

BART for continuous outcomes

Formal definition

$y = f(x) + \epsilon = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon$, meaning $f(x)$ is estimated by $\sum_{j=1}^{m} g(x; T_j, M_j)$, sum of m regression trees,

m is usually set as 50, 100, 200

where $\epsilon \sim N(0,\sigma^2)$, $x = (x_1, \ldots, x_p)$

T_j : jth binary tree structure,

M_j : $\left\{\mu_{j1}, \ldots, \mu_{jb_j}\right\}$ (vector of terminal nodes of T_j),

b_j : number of terminal nodes in T_j.

BART for continuous outcomes

Sum of regression trees

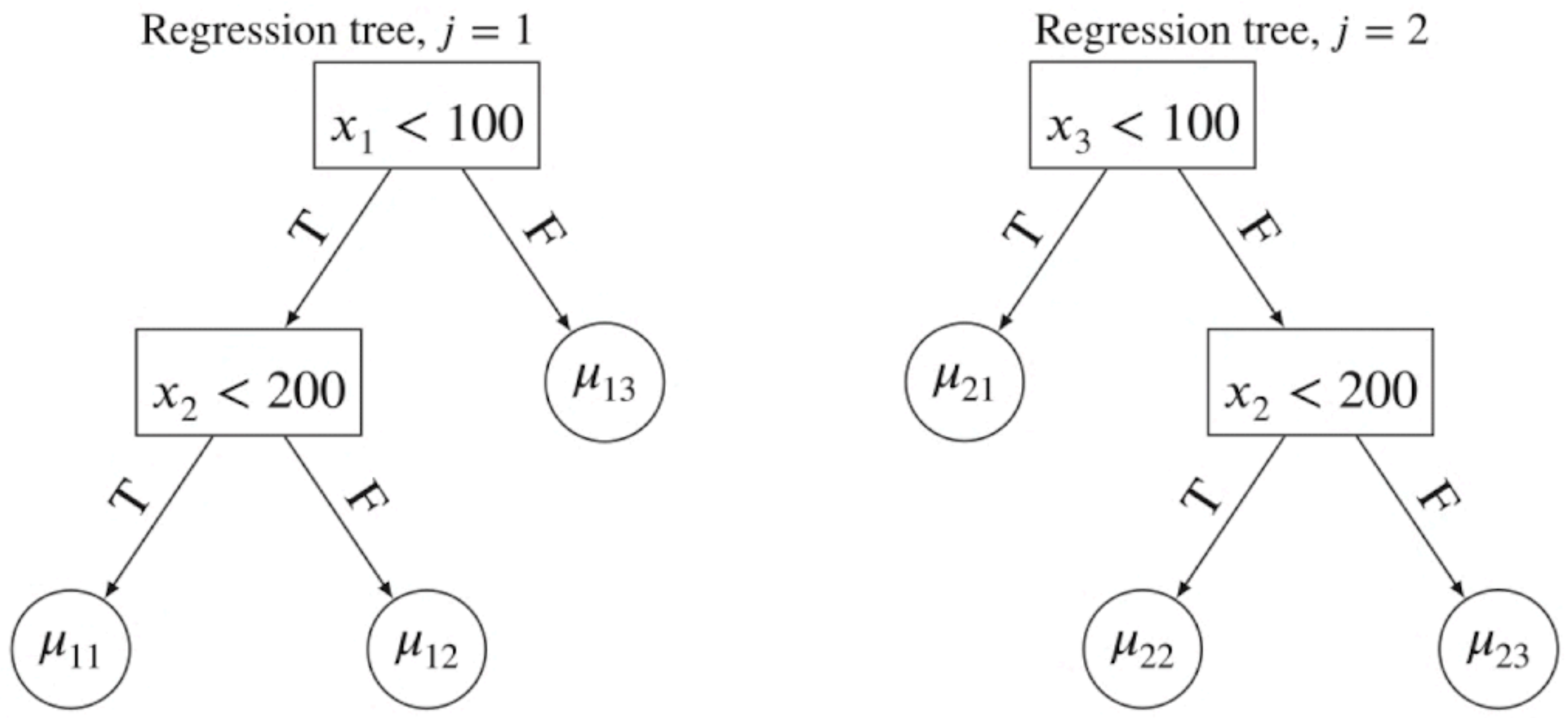
The following is an example of sum of regression trees for $m = 2$ and $p = 3$ (for x, the covariates).

In practice, each tree $g(x; T, M)$ is unknown so we need prior distributions for these functions

=> Bayesian additive regression trees (BART)

Advantage of BART:

The uncertainty about both the functional form $(g(\cdot\,; T))$ and the parameters (M) will be accounted for in the posterior predictive distribution of y.



$$\begin{aligned} y &= g(x; T_1, M_1) + g(x; T_2, M_2) + \varepsilon \\ &= \mu_{11} I\{x_1 < 100\} I\{x_2 < 200\} + \mu_{12} I\{x_1 < 100\} I\{x_2 \geq 200\} + \mu_{13} I\{x_1 \geq 100\} \\ &\quad + \mu_{21} I\{x_3 < 100\} + \mu_{22} I\{x_3 \geq 100\} I\{x_2 < 200\} + \mu_{23} I\{x_3 \geq 100\} I\{x_2 \geq 200\} + \varepsilon. \end{aligned}$$

Subject	y	x_1	x_2	x_3	$g(x; T_1, M_1)$	$g(x; T_2, M_2)$	$f(x)$
1	y_1	−182	235	−333	μ_{12}	μ_{21}	$\mu_{12} + \mu_{21}$
2	y_2	54	339	244	μ_{12}	μ_{23}	$\mu_{12} + \mu_{23}$
3	y_3	−106	−50	−682	μ_{11}	μ_{21}	$\mu_{11} + \mu_{21}$
4	y_4	−80	−62	−320	μ_{11}	μ_{21}	$\mu_{11} + \mu_{21}$
5	y_5	−123	198	−77	μ_{11}	μ_{21}	$\mu_{11} + \mu_{21}$
6	y_6	175	108	−46	μ_{13}	μ_{21}	$\mu_{13} + \mu_{21}$
7	y_7	−44	11	136	μ_{11}	μ_{22}	$\mu_{11} + \mu_{22}$
8	y_8	−131	−10	−70	μ_{11}	μ_{21}	$\mu_{11} + \mu_{21}$
9	y_9	−56	68	257	μ_{11}	μ_{22}	$\mu_{11} + \mu_{22}$
10	y_{10}	7	324	282	μ_{12}	μ_{23}	$\mu_{12} + \mu_{23}$

BART for continuous outcomes

Simple example

For $x = (x_1, x_2, x_3)$ and $m = 4$

- Initiation

We start from $m = 4$ single-root nodes (as in the trees have only one terminal node), where

$\mu_{ji}^{(0)} = \dfrac{\bar{y}}{m}, j = 1,\ldots,m,\ i = 1,\ldots,b_j$ (b_j : number of terminal nodes in jth tree).

BART for continuous outcomes

Simple example

- MCMC iterations (explained more in detail on the next section)

We start with the first tree (note that the order of the tree doesn't matter).

For tree 1, we calculate the residual,

$$r_1 = y - \sum_{j \neq 1} g(x; T_j, M_j).$$

By MH algorithm, we compare the newly proposed tree 1, T_1^*, and the previous tree 1, T_1, and decide whether we accept T_1^* ($T_1 = T_1^*$) or not ($T_1 = T_1$).

We do this for $T_2, \ldots, T_m$ similarly.

BART for continuous outcomes

Simple example

- Posterior distribution of σ^2

After the MCMC iterations, and the posterior draws of the regression trees are complete, we draw the posterior distribution of σ^2.

- Prediction

With the posterior distribution of the trees and σ^2, we can obtain,

1. The predicted value of y for any x of interest (by summing the terminal nodes, μ_{ji}s, of interest).

2. 95% prediction interval for y

BART for continuous outcomes

Simple example

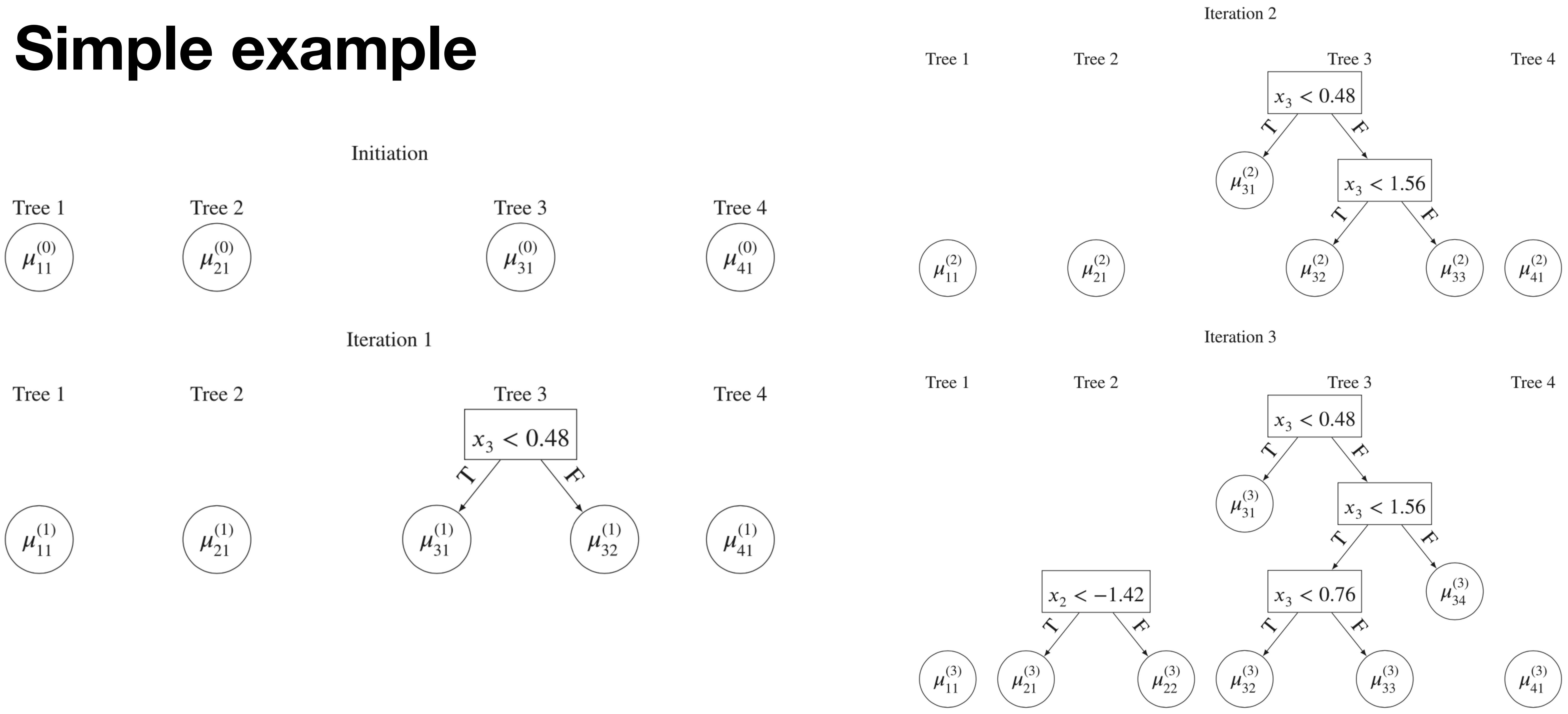


FIGURE 3 Initiation of BART to Iteration 3 of the MCMC steps within BART with $m = 4$. BART, Bayesian additive regression trees; MCMC, Monte Carlo Markov Chain

The regression trees are penalized by the prior to prevent a tree from growing too deep. This is a concept called boosting which we see a lot in the machine learning literature, where the performance of several weak models combined together is better than a single strong model.

BART for continuous outcomes

The BART algorithm - prior distribution

Prior for $Y = f(x) + \varepsilon = \sum_{j=1}^{m} g(x; T_j, M_j) + \varepsilon$ is

$$P[(T_1, M_1), \dots, (T_m, M_m), \sigma]$$

$$= P[(T_1, M_1), \dots, (T_m, M_m)]\, P(\sigma) \quad (\because \text{independence is assumed})$$

$$= \left[\prod_{j=1}^{m} P(T_j, M_j)\right] P(\sigma)$$

$$= \left[\prod_{j=1}^{m} P(M_j \mid T_j)\, P(T_j)\right] P(\sigma)$$

$$= \left[\prod_{j=1}^{m} \left\{\prod_{i=1}^{b_j} P(\mu_{ji} \mid T_j)\right\} P(T_j)\right] P(\sigma) \quad (\because \text{assuming each terminal nodes are independent})$$

b_j : total number of terminal nodes on j^{th} tree

Thus we have 3 prior distributions.

BART for continuous outcomes

The BART algorithm - prior distribution

$$\mu_{ji} \mid T_j \sim N(\mu_\mu, \sigma_\mu^2)$$

$$\sigma^2 \sim IG\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right)$$

BART for continuous outcomes

The BART algorithm - prior distribution

(2) means to give equal probability to select one of x_i for an internal node.

(3) means to give equal probability to c for the binary decision rule, for the selected x_i from (2), $\{x_i < c\}$ and $\{x_i \geq c\}$.

priors for T_j : (1) × (2) × (3) ?

(1) $\frac{\alpha}{(1+d)^{\beta}}$: The Prob. that a node at depth d would split

, $\alpha \in \{0,1\}$: how likely a node would split

, $\beta > 0$: larger values of β reduces the number of terminal nodes

(2) Uniform dstn to select the covariates to split upon in an internal node

(3) Uniform dstn to select the cutoff point in an internal node once the covariate is selected

BART for continuous outcomes

The BART algorithm - prior distribution

The hyperparameters for the prior distributions are as follows: $\alpha, \beta, \mu_\mu, \sigma_\mu, \nu, \lambda$.

- $\alpha = 0.95$ and $\beta = 2$ provide a balanced penalizing effect for the probability of a node splitting.
- μ_μ, σ_μ are set such that $E(Y|X) \sim N(m\mu_\mu, m\sigma_\mu^2)$ assigns a high probability to the interval $(\min(y), \max(y))$.
- For ease of posterior calculations, y is transformed as, $\tilde{y} = \dfrac{y - \frac{\max(y) + \min(y)}{2}}{\max(y) - \min(y)}$, which results in $\tilde{y} \in (-0.5, 0.5)$. This allows us to set $\mu_\mu = 0,\ \sigma_\mu = \dfrac{0.5}{\nu\sqrt{m}}$, where ν is to be chosen.
- For $\nu = 2$, $N(m\mu_\mu, m\sigma_\mu^2)$ assigns prior probability of 0.95 to the interval $(\min(y), \max(y))$.
- λ is set so that $P(\sigma^2 < s^2; \nu, \lambda) = 0.95$, where s^2 is the estimated variance of the residuals from the multiple linear regression (MLR).

BART for continuous outcomes

The BART algorithm - posterior distribution

Such prior distributions induce the following posterior distribution.

$$p[(T_1, M_1), \cdots, (T_m, M_m), \sigma \mid Y]$$

$$\propto p[Y \mid (T_1, M_1), \cdots, (T_m, M_m), \sigma] \times p[(T_1, M_1), \cdots, (T_m, M_m), \sigma]$$

The posterior draws can be obtained by Gibbs sampling from

$p[(T_j, M_j) \mid T_{-j}, M_{-j}, Y, \sigma] = ⊛$ ($-j$ means all except the jth)

and then

$p[\sigma \mid (T_1, M_1), \cdots, (T_m, M_m), Y]$

from $IG\left(\frac{v+n}{2}, \{v\lambda + \Sigma[Y - f(x)]^2\}/2\right)$

BART for continuous outcomes

The BART algorithm - posterior distribution

Derivation of the posterior distribution of σ is as follows.

Let $y = (y_1, \ldots, y_n)^T$ with $\sigma^2 \sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2})$. We obtain the posterior draw of σ as follows:

$$
\begin{aligned}
p(\sigma^2|(T_1, M_1), \ldots, (T_m, M_m), y) &\propto p(y|(T_1, M_1), \ldots, (T_m, M_m), \sigma)p(\sigma^2) \\
&= \left\{ \prod (\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{(y - f(x))^2}{2\sigma^2} \right] \right\} (\sigma^2)^{-(\frac{\nu}{2}+1)} \exp\left(-\frac{\nu\lambda}{2\sigma^2} \right) \\
&= (\sigma^2)^{-(\frac{\nu+n}{2}+1)} \exp\left[-\frac{\nu\lambda + \sum (y - f(x))^2}{2\sigma^2} \right].
\end{aligned}
$$

BART for continuous outcomes

The BART algorithm - posterior distribution

Since $\circledast$ depends on $(T_{-j}, M_{-j}, Y, \sigma)$ through

$r_j = Y - \sum_{h \neq j} g(x; T_h, M_h) \iff r_j = g(x; T_j, M_j) + \varepsilon$

$\left(\because Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \varepsilon \right)$

$\circledast$ is equivalent to

$P[(T_j, M_j) \mid r_j, \sigma]$

We integrate out M_j to obtain

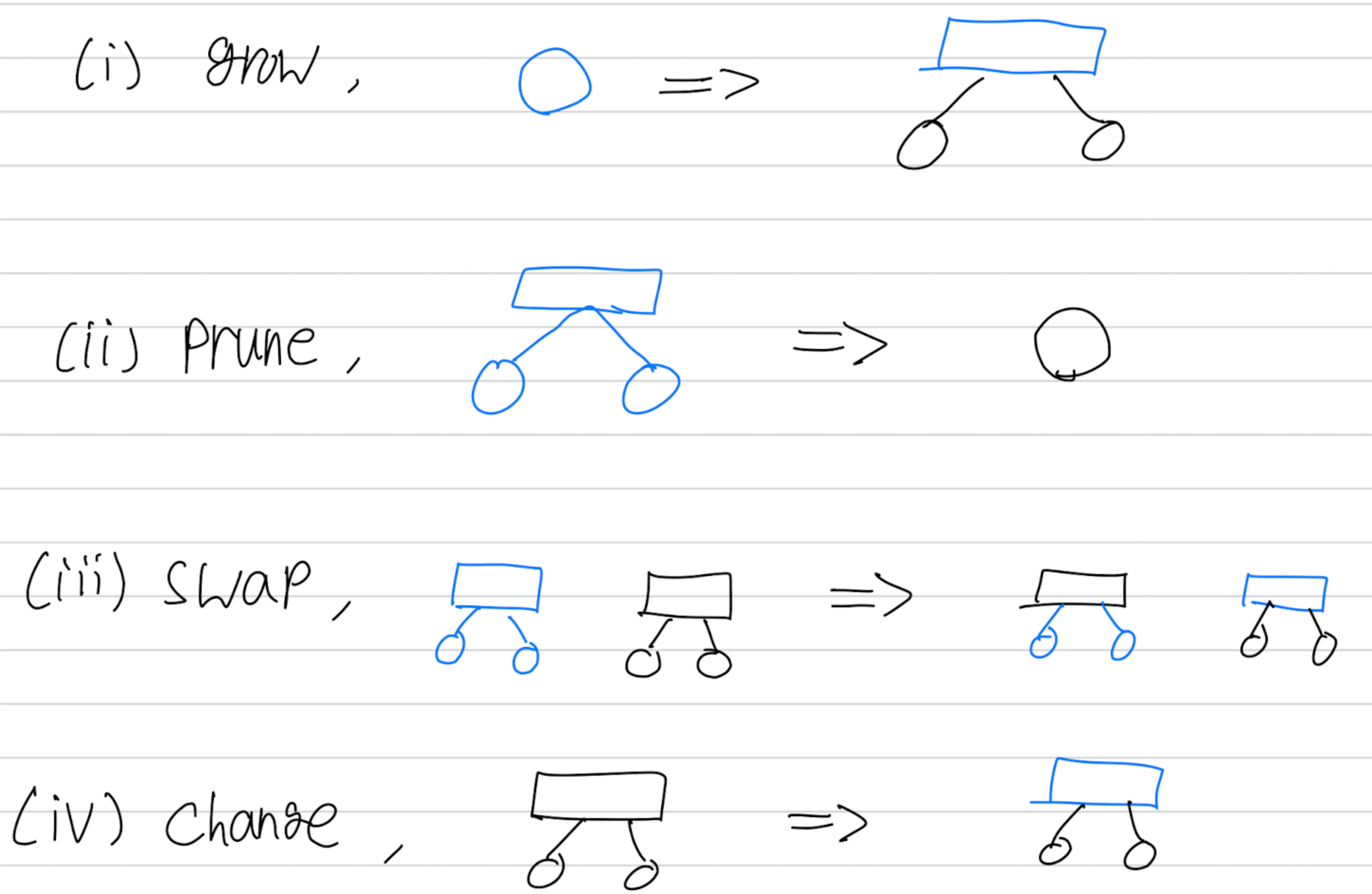
$P(T_j \mid r_j, \sigma)$

Since we used a conjugate normal prior on M_{ji}, (?)

BART for continuous outcomes

The BART algorithm - the MH algorithm

The new tree T_j^* can be proposed given the previous tree T_j by the following four local steps:



BART for continuous outcomes

The BART algorithm - the MH algorithm

We draw from $P(T_j \,|\, r_j, \sigma)$ by the MH algorithm with the acceptance ratio,

$$\alpha(T_j, T_j^*) = \min\left(1, \underbrace{\frac{q(T_j^*, T_j)}{q(T_j, T_j^*)}}_{\text{transition ratio}} \underbrace{\frac{P(R_j \,|\, x, T_j^*, M_j)}{P(R_j \,|\, x, T_j, M_j)}}_{\text{likelihood ratio}} \underbrace{\frac{P(T_j^*)}{P(T_j)}}_{\text{tree structure ratio}}\right)$$

BART for continuous outcomes

The BART algorithm - the MH algorithm

- Transition ratio for the “grow” proposal

$q(T_j^*, T_j) = P(T_j^* \mid T_j)$: the probability of moving from T_j to T_j^*, i.e., selecting a terminal node and growing two children from T_j.

$P(T_j^* \mid T_j) = P(\text{grow})$
$\times P(\text{selecting terminal node to grow from})$
$\times P(\text{selecting covariate to split from})$
$\times P(\text{selecting value to split on})$
$= P(\text{grow}) \frac{1}{b_j} \frac{1}{p} \frac{1}{\eta}$

$P(\text{grow}) = 0.25$ (default)

b_j : number of terminal nodes in T_j

p : number of x variables left in the partition of the chose terminal node

η : number of unique values left in the chosen variable after adjusting for the parents splits

BART for continuous outcomes

The BART algorithm - the MH algorithm

- Transition ratio for the "grow" proposal

$q(T_j, T_j^*) = P(T_j \mid T_j^*)$: the probability of selecting the correct internal node to prune on such that T_j^* becomes T_j.

$$P(T_j \mid T_j^*) = P(\text{prune})\, P(\text{selecting the correct internal node to prune})$$
$$= P(\text{prune}) \frac{1}{w_2^*}$$

where w_2^* denotes the number of internal nodes that have only two children terminal nodes.

BART for continuous outcomes

The BART algorithm - the MH algorithm

- Transition ratio for the “grow” proposal

Therefore,

This gives a transition ratio of

$$\frac{q\left(T_j^*, T_j\right)}{q\left(T_j, T_j^*\right)} = \frac{P\left(T_j^* | T_j\right)}{P\left(T_j | T_j^*\right)} = \frac{P(prune)}{P(grow)} \frac{b_j p\eta}{w_2^*}.$$

반대 아닌가?

If there are no x variables with two or more unique values, this transition ratio will be set to 0.

BART for continuous outcomes

The BART algorithm - the MH algorithm

- Likelihood ratio for the “grow” proposal

Since the rest of the tree structure will be the same between T_j and T_j^* except for the terminal node where the two children are grown, we only need to concentrate on this terminal node.

Let l be the terminal node and l_L and l_R be the two children of the grow step. Then,

$$\frac{P\left(r_j|x, T_j^*, M_j\right)}{P\left(r_j|x, T_j, M_j\right)} = \frac{P\left(r_{l_{(L,1)},j}, \ldots, r_{l_{(L,n_L)},j}|\sigma^2\right) P\left(r_{l_{(R,1)},j}, \ldots, r_{l_{(R,n_R)},j}|\sigma^2\right)}{P\left(r_{1,j}, \ldots, r_{n_l,j}|\sigma^2\right)}.$$

$$= \sqrt{\frac{\sigma^2\left(\sigma^2 + n_l\sigma_\mu^2\right)}{\left(\sigma^2 + n_L\sigma_\mu^2\right)\left(\sigma^2 + n_R\sigma_\mu^2\right)}} \exp\left[\frac{\sigma_\mu^2}{2\sigma^2}\left(\frac{\left(\sum_{k=1}^{n_L} r_{l_{(L,k)},j}\right)^2}{\sigma^2 + n_L\sigma_\mu^2} + \frac{\left(\sum_{k=1}^{n_R} r_{l_{(R,k)},j}\right)^2}{\sigma^2 + n_R\sigma_\mu^2} - \frac{\left(\sum_{k=1}^{n_l} r_{l_{(l,k)},j}\right)^2}{\sigma^2 + n_l\sigma_\mu^2}\right)\right].$$

BART for continuous outcomes

The BART algorithm - the MH algorithm

- Tree structure ratio for the “grow” proposal

T_j can be specified by,

$P_{\text{SPLIT}}(\theta) \propto \dfrac{\alpha}{(1+d_\theta)^\beta}$: probability of the selected node θ will split, and

$P_{\text{RULE}}(\theta) \propto \dfrac{1}{p}\dfrac{1}{\eta}$: probability of a certain variable and value is selected.

Since T_j and T_j^* only differ at the children nodes,

$$\begin{aligned}
\frac{P(T_j^*)}{P(T_j)} &= \frac{\prod_{\theta \in H^*_{terminals}}(1-P_{SPLIT}(\theta))\prod_{\theta \in H^*_{internals}} P_{SPLIT}(\theta)\prod_{\theta \in H^*_{internals}} P_{RULE}(\theta)}{\prod_{\theta \in H_{terminals}}(1-P_{SPLIT}(\theta))\prod_{\theta \in H_{internals}} P_{SPLIT}(\theta)\prod_{\theta \in H_{internals}} P_{RULE}(\theta)} \\
&= \frac{[1-P_{SPLIT}(\theta_L)][1-P_{SPLIT}(\theta_R)]P_{SPLIT}(\theta)P_{RULE}(\theta)}{1-P_{SPLIT}(\theta)} \\
&= \frac{\left(1-\frac{\alpha}{(1+d_{\theta_L})^\beta}\right)\left(1-\frac{\alpha}{(1+d_{\theta_R})^\beta}\right)\frac{\alpha}{(1+d_\theta)^\beta}\frac{1}{p}\frac{1}{\eta}}{1-\frac{\alpha}{(1+d_\theta)^\beta}} \\
&= \alpha\frac{\left(1-\frac{\alpha}{(2+d_\theta)^\beta}\right)^2}{[(1+d_\theta)^\beta-\alpha]p\eta}
\end{aligned}$$

because $d_{\theta_L} = d_{\theta_R} = d_\theta + 1.$

BART for continuous outcomes

The BART algorithm - the MH algorithm

Once we have the draw of $P(T_j \mid r_j, \sigma)$, we then draw

$$P(\mu_{ji} \mid T_j, r_j, \sigma) \sim N\Big([\sigma_\mu^2 \sum r_{ji}]/[n_i\sigma_\mu^2 + \sigma^2], [\sigma^2\sigma_\mu^2]/[n_i\sigma_\mu^2 + \sigma^2]\Big)$$

where r_{ji} is the subset of elements in r_j allocated to the terminal node parameter μ_{ji} and n_i is the number of r_{ji}'s allocated to μ_{ji}.

BART for continuous outcomes

The BART algorithm - the MH algorithm

The derivation of the posterior distribution of μ_{ji} is as follows.

Let $r_{ji} = (r_{ji1}, \ldots, r_{jin_i})^T$ be a subset from r_j where n_i is the number of r_{jih}'s allocated to the terminal node with parameter μ_{ji} and h indexes the subjects allocated to the terminal node with parameter μ_{ji}. We note that $r_{ji} \mid T_j, \mu_{ji}, \sigma \sim N(\mu_{ji}, \sigma^2)$ and $\mu_{ji} \mid T_j \sim N(\mu_\mu, \sigma_\mu^2)$. Then, the posterior distribution of μ_{ji} is given by

$$
\begin{aligned}
p(\mu_{ji}|T_j,\sigma,r_j) &\propto p(r_{ji}|T_j,\mu_{ji},\sigma)p(\mu_{ji}|T_j) \\
&\propto \exp\left[-\frac{\sum_h (r_{jih}-\mu_{ji})^2}{2\sigma^2}\right]\exp\left[-\frac{(\mu_{ji}-\mu_\mu)^2}{2\sigma_\mu^2}\right] \\
&\propto \exp\left[-\frac{\left(n_i\sigma_\mu^2+\sigma^2\right)\mu_{ji}^2 - 2\left(\sigma_\mu^2 \sum_h r_{jih} + \sigma^2\mu_\mu\right)\mu_{ji}}{2\sigma^2\sigma_\mu^2}\right] \\
&\propto \exp\left[-\frac{\left(\mu_{ji} - \frac{\sigma_\mu^2\sum_h r_{jih}+\sigma^2\mu_\mu}{n_i\sigma_\mu^2+\sigma^2}\right)^2}{2\frac{\sigma^2\sigma_\mu^2}{n_i\sigma_\mu^2+\sigma^2}}\right],
\end{aligned}
$$

where $\sum_h (r_{jih} - \mu_{ji})^2$ is the summation of the squared difference between the parameter μ_{ji} and r_{jih}'s allocated to the terminal node with parameter μ_{ji}

BART for continuous outcomes

Performance of BART - synthetic data

The point estimates of BLR were far way from the true values and many of the true values were not covered by the 95% credible intervals.

For BART, as m (number of trees) increased, there was a significant improvement in point estimates and the credible intervals were also narrowed. Note that there was no signficant improvement in result by increasing m after 50.

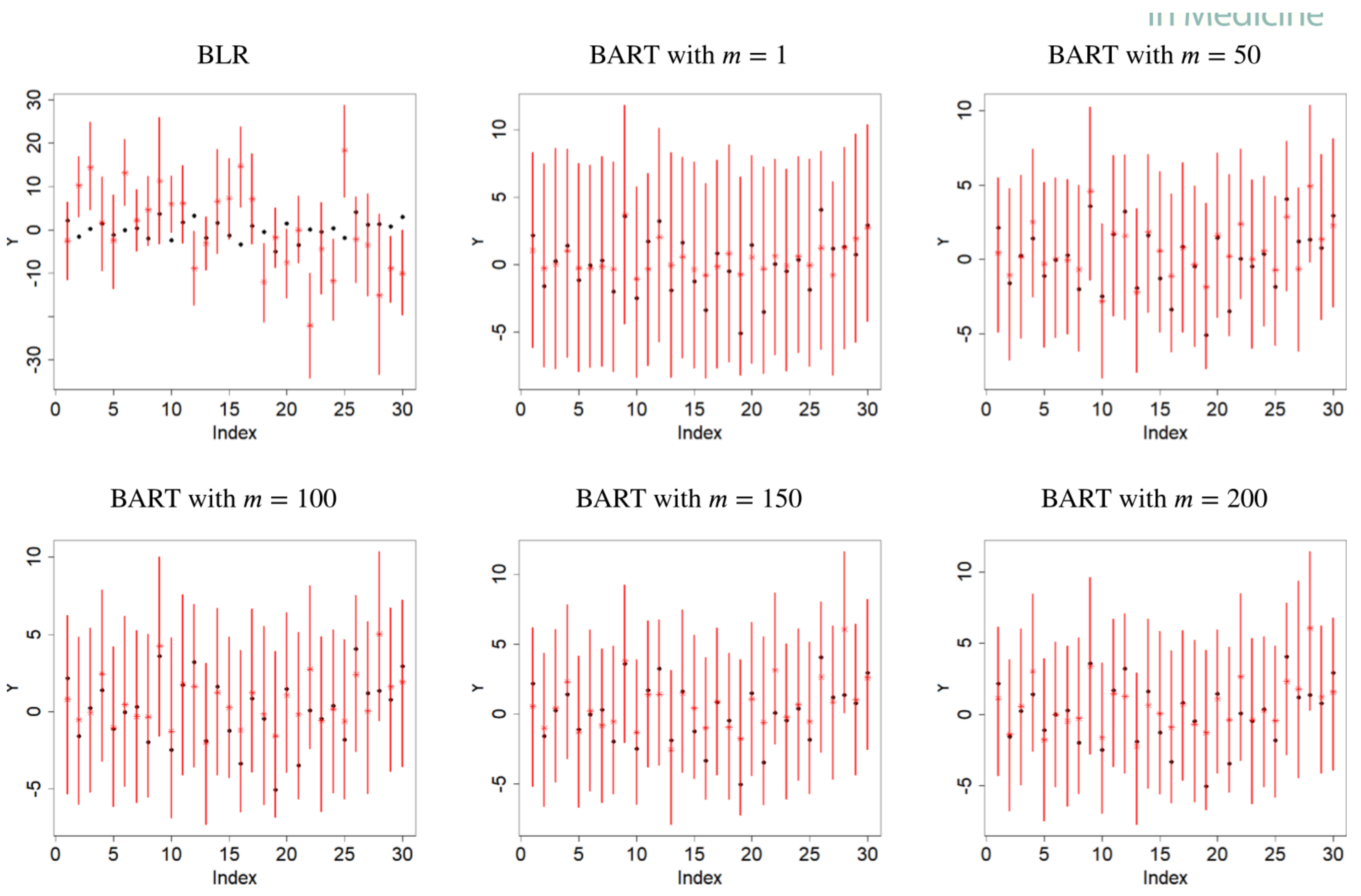


FIGURE 5 Posterior mean and 95% credible interval of Bayesian linear regression (BLR) and Bayesian additive regression trees (BART) with $m = 1, 50, 100, 150, 200$ for 30 randomly selected testing set outcomes. $n = 1000$, black = true value, red = model estimates [Colour figure can be viewed at wileyonlinelibrary.com]

BART for continuous outcomes

Performance of BART - real data

The figure shows the 10 RMSEs produced by each method from the 10-fold cross-validation. Both BART and RF produced very similar prediction performances and are better compared to MLR. MLR produced a mean of the RMSE of 0.24 while BART and RF produced a mean of 0.23.

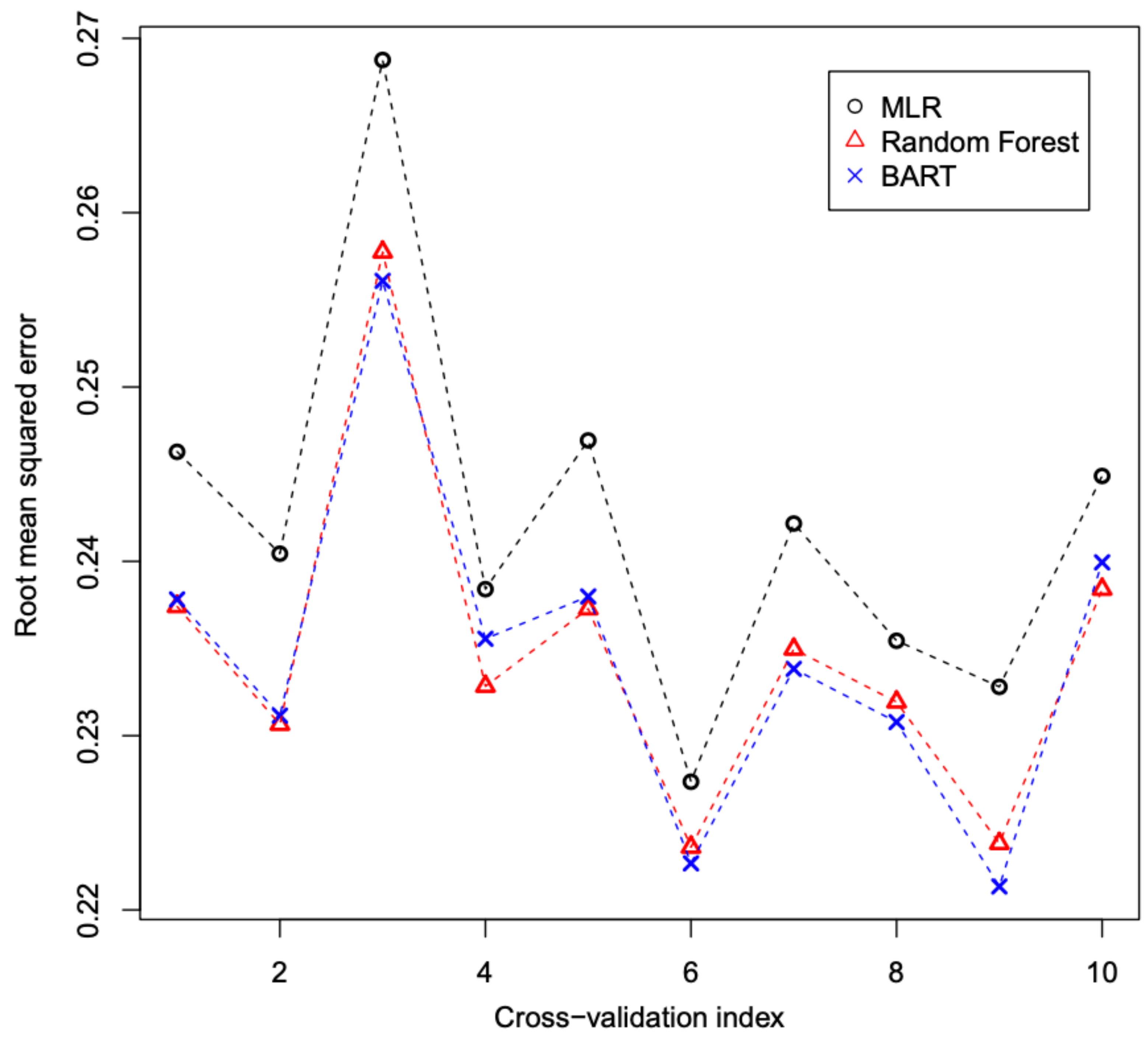


FIGURE 6 Root mean squared error (RMSE; *y*-axis) for the 10-fold cross-validation of multiple linear regression (MLR), random forest, and Bayesian additive regression trees (BART) of log transformed standardized hospitalization ratio (SHR). *x*-axis indicates the RMSE for the *x*th fold [Colour figure can be viewed at wileyonlinelibrary.com]