Bayesian additive regression tree (BART)

- Nonparametric prediction tool
- Ensemble of 'weak learners' that leads to robust out-of-sample predictions
- Dynamically learns important predictors via a sparsity inducing prior

Mathematically, a BART model looks like:

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \sum_{t=1}^{T} \sum_{l \in L_t} \psi_{lt} I(x_i \rightsquigarrow (t, l))$$

$I(x_i \rightsquigarrow (t, l)) = 1$ if x_i falls into node (t, l) of tree t; 0 otherwise.

Shared Forest

- An extension of BART for multivariate responses
- Correlation between responses is utilized via shared tree structures
- Variables that predict Y_1 are likely the same variables that predict Y_2 (though the nature of the relationship may be different!)
- Information sharing $\Rightarrow$ better predictions for Y_1 *and* Y_2

1 binary, 1 continuous

$$Y_i \sim N(\mu_i, \sigma_i^2) \tag{1}$$
$$\delta_i \sim Bernoulli(\pi_i) \tag{2}$$

- $\begin{pmatrix} \mu_i \\ \pi_i \end{pmatrix}$ modeled jointly using a shared forest model

$$\begin{pmatrix} \mu_i \\ \Phi(\pi_i) \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{T} \sum_{l \in L_t} \psi_{lt} I(x_i \rightsquigarrow (t, l)) \\ \sum_{t=1}^{T} \sum_{l \in L_t} \theta_{lt} I(x_i \rightsquigarrow (t, l)) \end{pmatrix}$$

- Likelihood involves two dimensional response: $\{Y_i, \delta_i\}_{i=1}^N$

We run 100 simulations. In each:

- $n_{train} = 500$, $n_{test} = 500$
- We set the number of covariates to P = 150.
- $x_1, \ldots, x_{150} \sim Unif(0, 1)$
- True underlying means based on a modification of the "Friedman function"

$$y_i = 10\sin(\pi x_{1i}x_{2i}) + 20(x_{3i} - 0.5)^2 + 10x_{4i} + 5x_{5i} + \epsilon_i^Y$$

$$\delta_i = \begin{cases} 1 & \text{if} 5\sin(\pi x_{1i}x_{2i}) + 25(x_{3i} - 0.5)^2 + 5x_{4i} + 10x_{5i} + \epsilon_i^\delta > 0 \\ 0 & \text{otherwise} \end{cases}$$

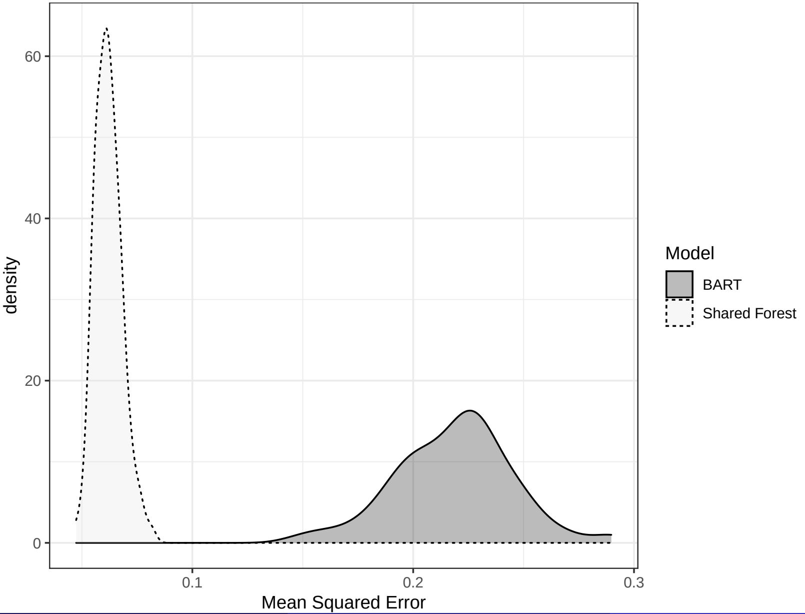
where $(\epsilon_i^Y, \epsilon_i^\delta) \overset{iid}{\sim} N(0, 1)$

1 binary, 1 continuous

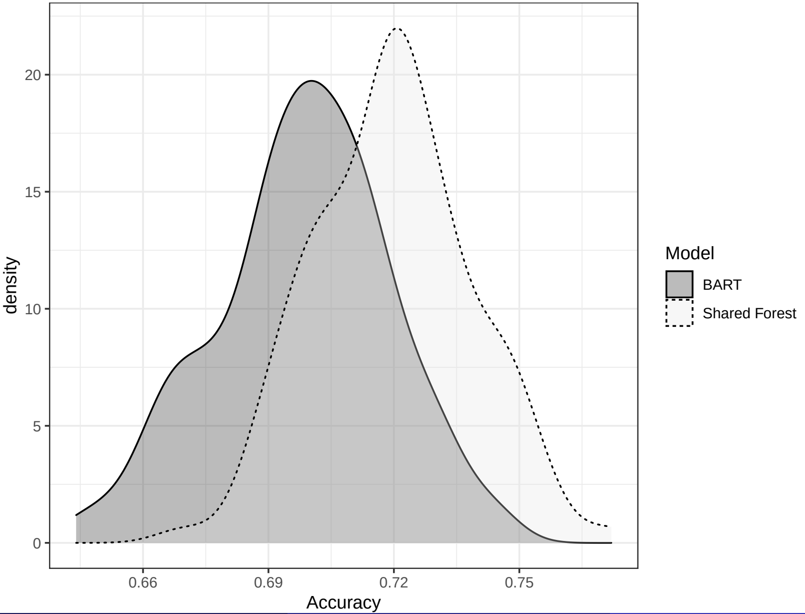
Two quantities may be of interest for prediction

1. $E(\delta^*, y^* \mid x)$, where * indicates those observations come from a test / hold-out sample
   - ex) Predict salary of individual, conditional on cellular traits x_i
   - ex) Predict gender of individual, conditional on cellular traits x_i
2. $E(y^* \mid \delta)$
   - ex) Predict mean salary of females (avg. across traits observed for females)

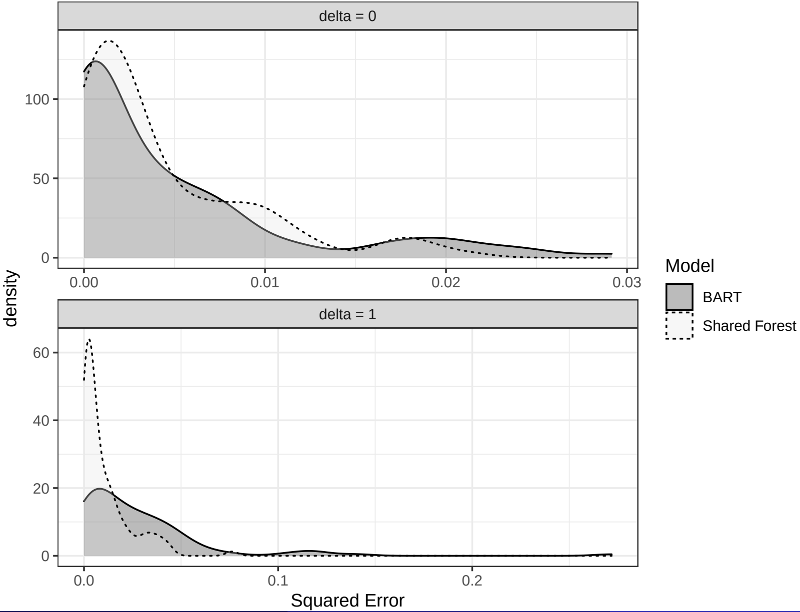
Predicting Individual Y^*



Predicting Individual $\delta*$



Predicting $E(Y^* \mid \delta^*)$



Predicting $E(Y^* \mid \delta^*)$

δ	Model	MSE	L_bound	U_bound
0	BART	0.004875	0.000004	0.023207
0	Shared Forest	0.004330	0.000007	0.017697
1	BART	0.028464	0.000042	0.119264
1	Shared Forest	0.010399	0.000011	0.041591

Table: Mean of the squared errors (across simulations) comparing the true $E(Y^* \mid \delta^*)$ to the estimated $\hat{E}(Y^* \mid \hat{\delta})$.

2 binary responses

The model structure is the same as described before; however, the likelihood reflects that:

$$\delta_{1i} \sim \mathit{Bernoulli}(\pi_{1i}) \tag{3}$$
$$\delta_{2i} \sim \mathit{Bernoulli}(\pi_{2i}) \tag{4}$$

As before, the tree structures will be built using information shared between δ_1 and δ_2, while the location parameters are estimated separately.

2 binary responses

In this context, we are interested in predicting $P(\delta_2^* = 1 \mid \delta_1^*)$.

ex) Probability a randomly selected female is employed.
(δ_1 = gender, δ_2 = employment)

We measure model performance by looking at the squared errors comparing the true population (conditional) means to the estimated.

Predicting $E(\delta_2^* \mid \delta_1^*)$