

Simulations and Comparisons

Simulated Data

Tree-based methods shine in predictive situations with correlated predictors, although these situations can pose problems for inference. In a situation with correlated predictors X_1 and X_2 , and the treemodel we are considering is $Y \sim X_1 + X_2$, it is difficult to say how much of the modeled effect on Y is due to X_1 or X_2 . To illustrate this idea, compare a few existing methods, and explore methods of inference on tree based models three datasets will be simulated with different correlation structures. We will be focused more on the correlation structure between the predictors than on their relationships with the response and this will be reflected in the simulations.

To aid in comparisons between the methods, one of the simulated datasets considered in this paper will be generated from the same method as used in (Strobl et al, 2008b). Under this method, the 13 x 1000 data set, D_1 , has 12 predictors, V_1, \dots, V_{12} , where $V_j \sim N(0, 1)$. The first four are, however, block correlated to each other with $\rho = .9$. They are related to Y by the linear equation:

$$Y = 5 \cdot V_1 + 5 \cdot V_2 + 2 \cdot V_3 + 0 \cdot V_4 + -5 \cdot V_5 + -5 \cdot V_6 + 0 \cdot V_7 + 0 \cdot \dots + E, E \sim N(0, \frac{1}{2})$$

Note that the coefficients for V_7, \dots, V_{12} are all zero.

Table 1: Correlation of V_1, \dots, V_7 and Y

	V1	V2	V3	V4	V5	V6	V7	y	beta
V1	1.000	0.915	0.908	0.907	-0.034	0.006	0.012	0.829	5
V2	0.915	1.000	0.914	0.914	-0.020	-0.001	-0.001	0.830	5
V3	0.908	0.914	1.000	0.903	-0.017	-0.007	0.007	0.808	2
V4	0.907	0.914	0.903	1.000	-0.002	-0.015	0.023	0.789	0
V5	-0.034	-0.020	-0.017	-0.002	1.000	0.044	0.005	-0.388	-5
V6	0.006	-0.001	-0.007	-0.015	0.044	1.000	-0.005	-0.364	-5
V7	0.012	-0.001	0.007	0.023	0.005	-0.005	1.000	-0.141	-2

As can be seen from the last column in the table, “beta”, although V_4 was not included in the model $Y \sim V_1, \dots, V_{12}$, it has a strong correlation with more influential predictors V_1, \dots, V_3 insures that it still shows a strong, empirical linear correlation with Y . A linear model would likely *overstate* the effect of V_4 on Y .^{1 2}

As can be seen in Figure 1 the densities of V_1, \dots, V_5 are all very similar due to the way they were generated.

Models and Comparisons

CART: Regression Trees

A single CART tree representing the model $Y \sim X_1, \dots, X_{12}$ is easy enough to understand. Starting at the very top of the tree, predictions can be made based on the values of the leaves (or ending nodes) given the

¹A brief note on uncertainty is needed here. It’s true that in this setting we can say that V_4 is actually unimportant to understanding Y , but in situations with real data this is profoundly more difficult to parse. Often like in the social science situations that Morgan and Sonquist encountered, the real relationship between correlated predictors is complicated and often there is some theoretical backing or other insight that is gained to include variables that may not be important to the model.

²Another point that could be said is that, no V_4 is not unimportant, V_1, V_2 , and V_3 are just stand ins for the real star, V_4 , as they are nearly the same ($\rho \sim 1$). Then the real relationship represented here is $Y \sim (5 + 5 + 2) \cdot V_4 + -5 \cdot V_5 + -5 \cdot V_6 + -2 \cdot V_7$. This model is not unsuccessful in capturing the structure of the data, and this is typically the practice used to model data with highly correlated predictors. If this seems philosophically satisfying to you, the rest of this thesis may seem a bit inconsequential.

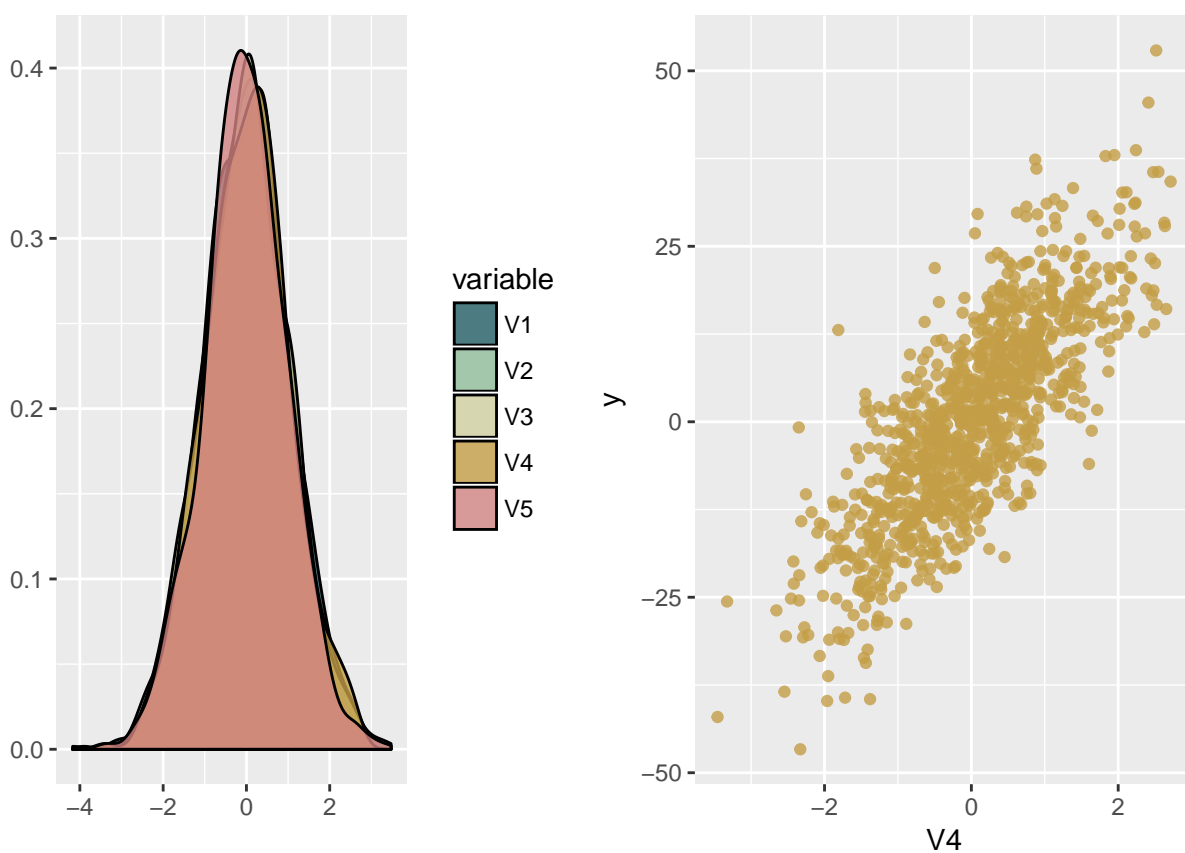


Figure 1: Density Graphs for V1 through V5 and a Plot of $Y \sim V4$, Correlation = .789

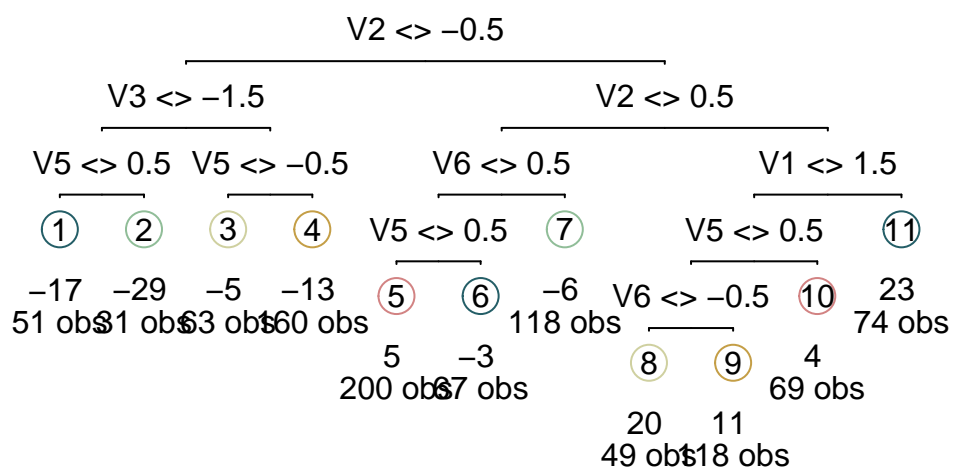


Figure 2: CART for the Model $Y \sim$, from D1

requirements of the path to get there. Trees can be quite variable, so to get a better idea of the differences between the methods let's run a simulation.

Algorithm 1 Simulation Scheme 2.1

- 1: **for** $i \leq 1000$ **do**
 - 2: Randomly sample $\frac{2}{3}$ of the observations in D_1 to a training set, $D_{1,train}^i$. The other observations, $x \in D_1, x \notin D_{1,train}^i$ form the testing set $D_{1,test}^i$
 - 3: Fit a tree, T^i , to the data under the model $Y \sim X_1, \dots, X_2$ using the observations in $D_{1,train}^i$
 - 4: Calculate the MSE_{test} of the model using the equation: $MSE_{test} = \frac{1}{n} \sum (y_j - \hat{y}_j)^2$
 - 5: **end for**
-

Note that n is the number of observations in $D_{1,test}^i$, $y_j \in D_{1,test}^i$, $\hat{y}_j \in T^i(D_{1,test}^i)$ for $1 \leq j \leq n$. This produces one distribution of MSE_{test} for CART.

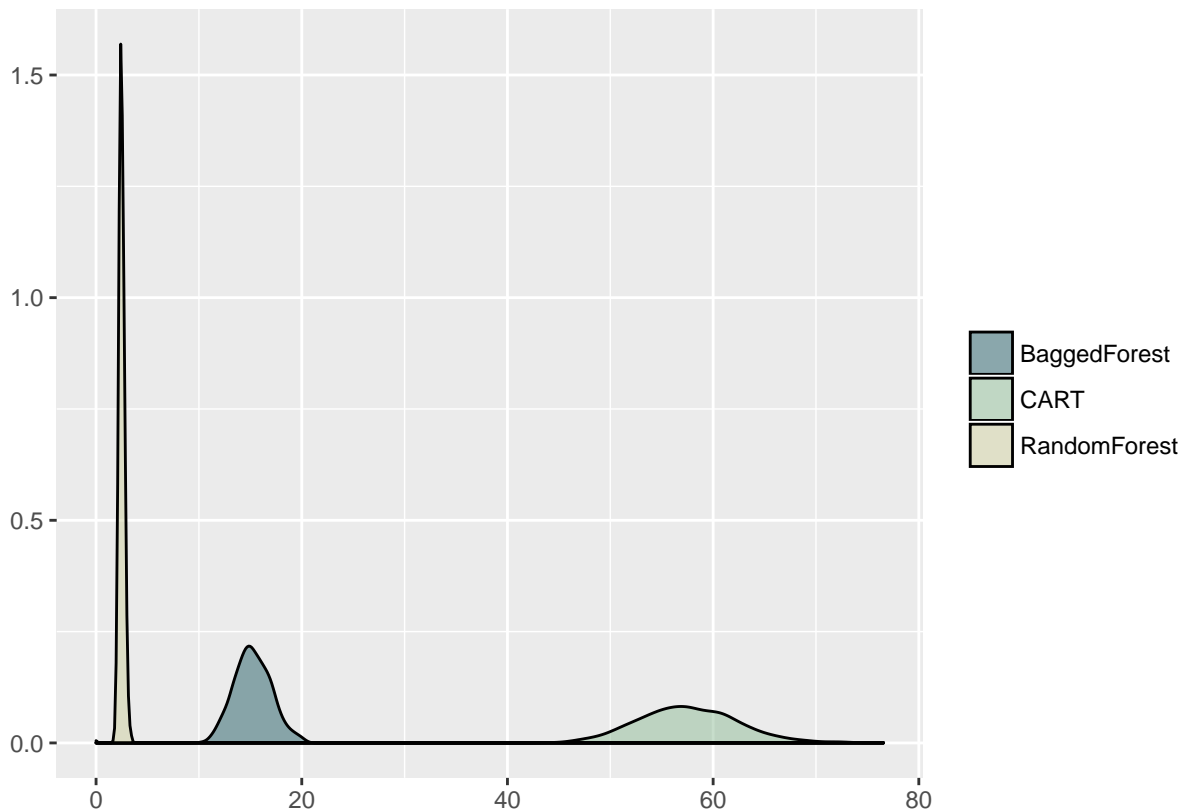


Figure 3: Simulated MSE_{test} Distributions of CART, Random, and Bagged Forests

The distribution of 100 CART trees' MSE_{test} in the above figure is roughly normal with a variance of `var(testmseC)`. There is a fair amount of variability in a single tree, they are heavily dependent on fluctuations in the starting data set. As mention briefly in the introduction, bagged forests present one solution to this problem. To create a bagged forest, as outlined in *An Introduction to Statistical Learning* by James, Witten, Hastie and Tibshirani, 2013, many bootstrapped samples are taken from the initial dataset and trees are fitted to them. The final predictions are, then, averaged over all of the trees. This ensures that while each tree has high variance, when they are aggregated the variance will decrease. As one can see, the values of MSE_{test} for the bagged forest were entirely below the MSE_{test} for the trees and the variance was much smaller. As random forests are unbiased, they can be much smaller than their bagged forest cousins without sacrificing accuracy.