Simulations and Comparisons

Our goal for this chapter is to compare trees, random forests, and linear models. We will leave our orange tree example behind for some simulated data. One reason for this is theoretical consistancy. One hopes that one's results will not be rendered null and void by a mistep in the data collection that comes to light. This also ensures that these simulations can be repeated by later researchers, but, granted, it does not make for the most exciting analysis. For now, consider Y to be our response variable. In the first simulation, V will be the set of our predictors, and V_j to be a predictor in V. The formula will be the same for each model: $Y \sim V$. In our second simulation, our set of predictors will be denoted as X and X_j will be a member of X. The formula in this case is $Y\ simX$.

Simulated Data

Random forests excel in predicting outcomes with correlated predictors, although these situations can pose problems for inference. In a situation with correlated predictors X_1 and X_2, and the tree model we are considering is $Y \sim X_1 + X_2$, it can be difficult to say if X_1 or X_2 is truely the better predictor. To illustrate this idea, compare a few existing methods, and explore methods of inference on tree based models two 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, $V_1, .., V_{12}$, 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:

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

Note in table 1, the coefficients for $V_7, ..., V_{12}$ are all zero.

Table 1: Empirical correlations and coefficients of the variables in the first simulated data set

	V1	V2	V3	V4	V5	V6	V7	y	beta
V1	1.000	0.915	0.908	0.907	-0.034	0.006	0.012	0.839	5
V2	0.915	1.000	0.914	0.914	-0.020	-0.001	-0.001	0.838	5
V3	0.908	0.914	1.000	0.903	-0.017	-0.007	0.007	0.818	2
V4	0.907	0.914	0.903	1.000	-0.002	-0.015	0.023	0.800	0
V5	-0.034	-0.020	-0.017	-0.002	1.000	0.044	0.005	-0.392	-5
V6	0.006	-0.001	-0.007	-0.015	0.044	1.000	-0.005	-0.368	-5
V7	0.012	-0.001	0.007	0.023	0.005	-0.005	1.000	0.004	0

In the last column of table 1, "beta", although $V4$ was not included in the model $Y \sim V1, ..V_{12}$, it has a strong correlation with more influential predictors $V_1, ..., V_3$ ensures 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]

[1] A 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.

[2] Another 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.

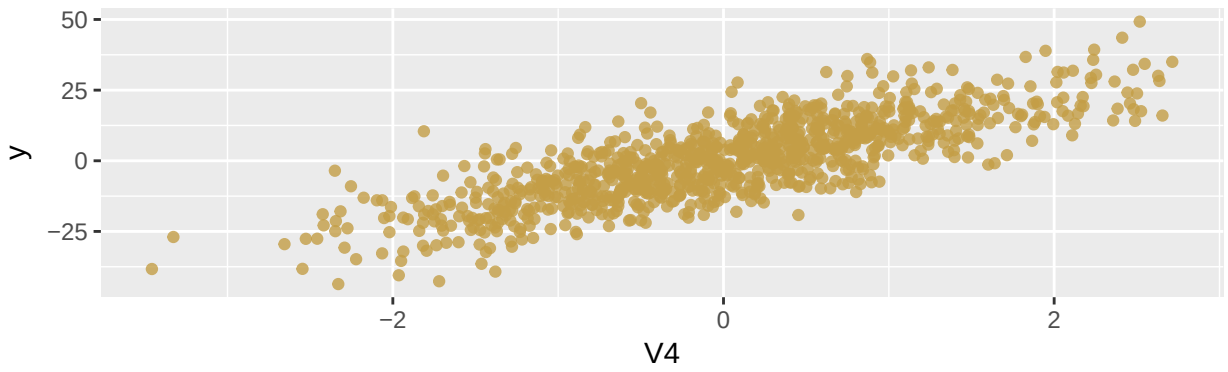

Figure 1: Relation between V4 and Y. This relation has empirical linear correlation = .789

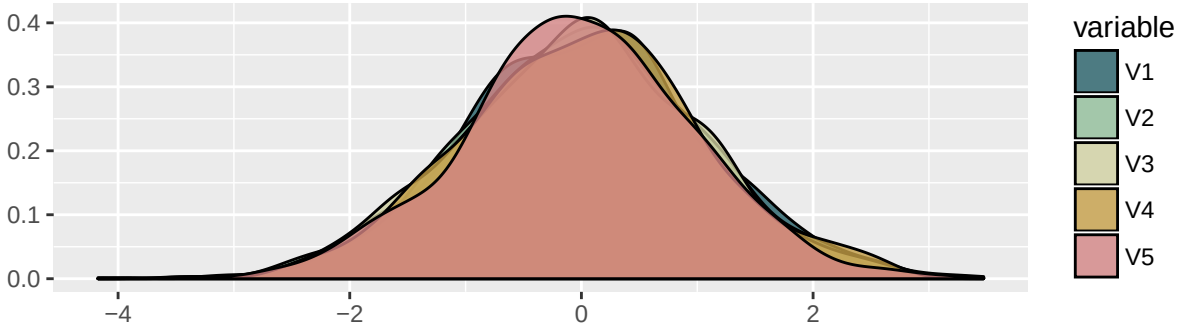

Figure 2: Empirical densities for V1 through V4

As in figure **??**, the densities of $V_1, ..., V_4$ are all very similar due to the way they were generated.

D_1 represents the case where some of the predictors are linearly correlated with each other, but that is not the only possible correlation structure. The data set D_2 is simulated similarly to D_2 in that D_2 contains twelve predictors and one response variable. The first four variables are generated in the following way:

$$X_1 \sim N(1, 0)$$

$$X_2 = log(X_1) + E, E \sim N(1, 0)$$

$$X_3 = log(X_2) + E, E \sim N(1, 0)$$

$$X_4 = log(X_4) + E, E \sim N(1, 0)$$

This simulation scheme leads to the first four variables having an obvious relationship between each other, but relatively low linear correlations. (See figure **??**). Predictors are sampled by $X_5, ..., X_12 \sim N(0, 1)$. The Y values are generated according to the following formula:

$$Y = 5 \cdot (X_1)^2 + 5 \cdot (X_2)^2 + 2 \cdot (X_3)^2 + 0 \cdot X_4 + -5 \cdot X_5 + -5 \cdot X_6 + 0 \cdot X_7 + 0 \cdot ..... + E, E \sim N(0, \frac{1}{2})$$

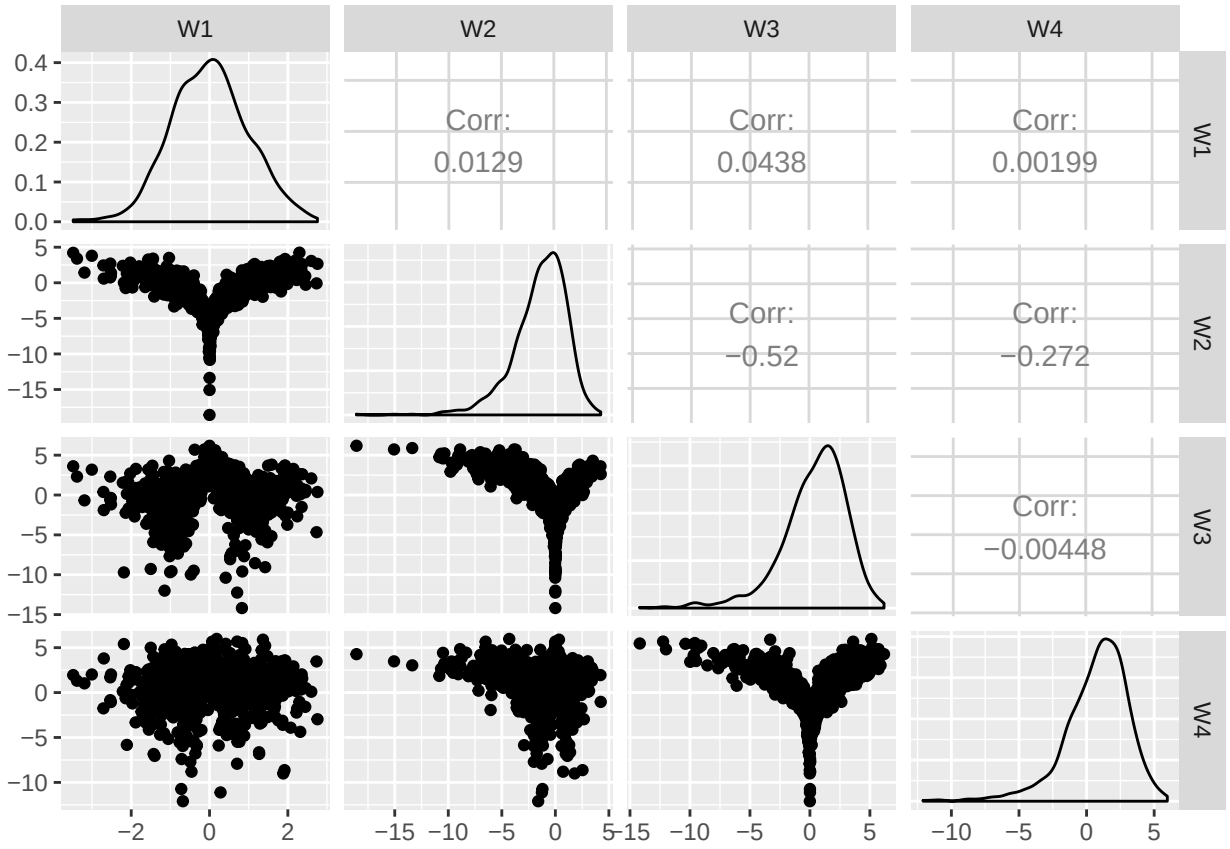

Figure 3: Correlation structure of the first four variables in d2

The linear correlation structure in D_2 is not as striking as in D_1. The two strongest linear relationships are between X_2 and X_3 with $\rho = -.534$ and between Y and X_2 with $\rho = .700$.

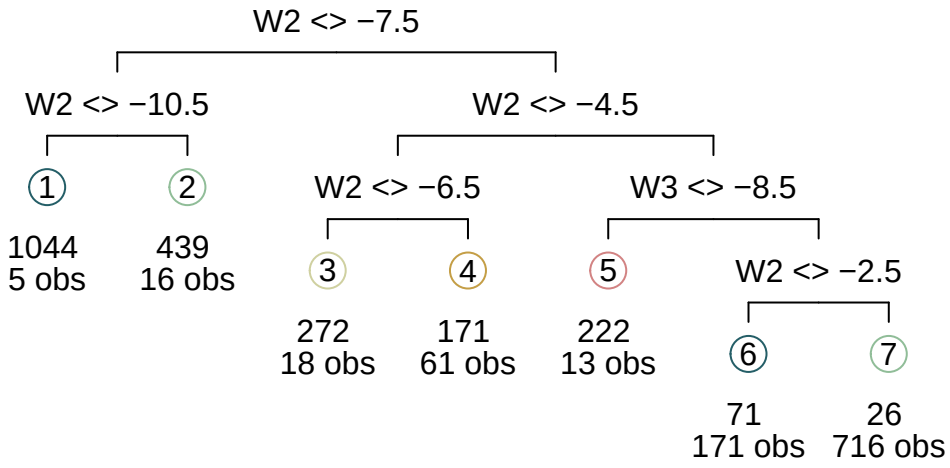

Figure 4: CART representing Y~ X, from D2

Models and Comparisons

CART: Regression Trees

The CART tree representing the model $Y \sim X$ in figure **??** 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 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. This simulation scheme will take advantage of the non linearity present in D_2.

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, $D^i_{1,train}$. The other observations, $x \in D_2, x \notin D^i_{2,train}$ form the testing set $D^i_{2,test}$
3: Fit a tree, T^i, to the data under the model $Y \sim X_1, ..., X_2$ using the observations in D^i_2
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 $D^i_{1,test}$, $y_j \in D^i_{2,test}, \hat{y_j} \in T^i(D^i_{2,test})$ for $1 \leq j \leq n$ This produces one distribution of MSE_{test} for CART. This simulation scheme will be repeated for the linear model and the random forest and the MSE_{test} distributions are compared in figure **??**.

The linear model is characteristically less flexible and less prone to overfitting than either of the tree-based methods, CART and random forests, and has a MSE_{test} distribution that is quite peaked. CART is flexible and suffers from high variance. The random forest models perform much better on average than either the CART or the linear model, due to both the non-linear relationships between Y and the predictors and the random forest's ability to decorrelate each of the trees by restricting the variables available on each split. See chapter 3 for more discussion on the inforced hetergenaity of trees in the random forest.

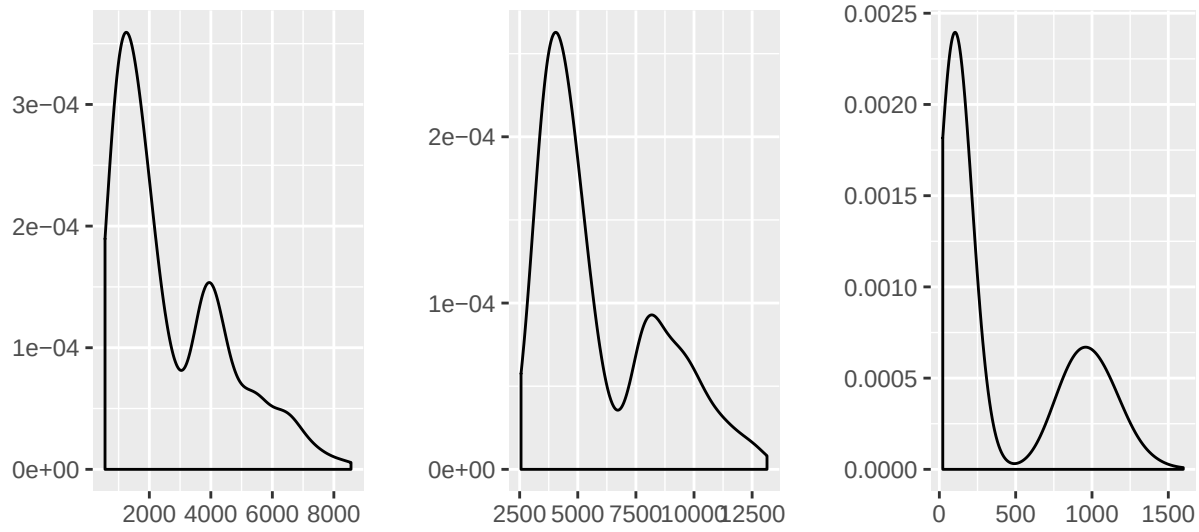

Figure 5: The simulated MSE distributions of CART, linear model, and the random forest on D2