Simulations and Comparisons

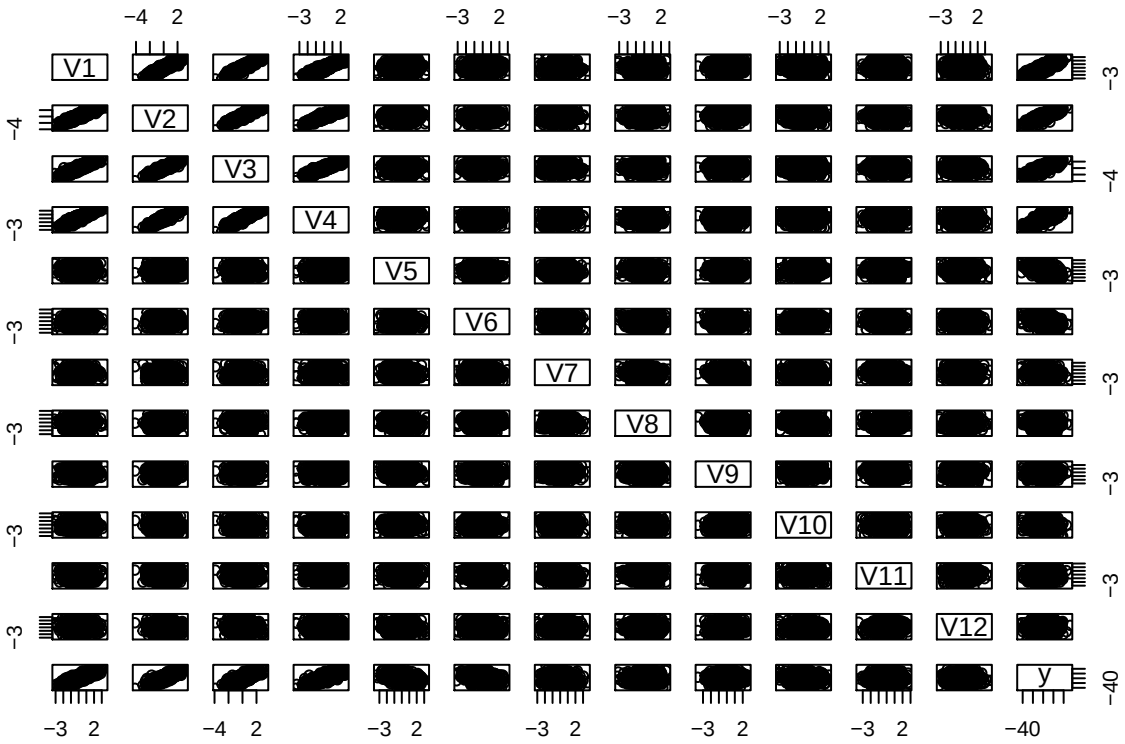
Simulated Data

- From Strobl et al:

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, 2008???). 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 that the coefficents for $V_7, ..., V_{12}$ are all zero.



- Non linearly correlated

Models and Comparisons

In 1984, Breiman et al introduces a revolutionary new algorithm for trees. **Need to acquire** *Classification and Regression Trees* **to make sure the method discused in MASS is the same that Breiman uses/is used in** `randomForest`

Tree Algorithm CART?

Begin by considering the entire feature space $X_1, ..., X_n$. Then:

1. Consider every possible pair of partitions of this feature space, P_1, P_2, so that if $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ where $x_1, ..., x_n \in P_1$ then our prediction is the mean value of y given $x_1, .., x_n \in P_1$.
2. Choose the partitions that minimize RSS

3. For each new partition, repeat steps 1 and 2 until some stopping condition is reached.

CI trees 1. For case weights w test the global null hypothesis of independence between any of the m covariates and the response. Stop if this hypothesis cannot be rejected. Otherwise select the j_{th} covariate Xj with strongest association to Y.

2. Choose a set $A \subset X_j$ in order to split X_j into two disjoint sets A and $X_j \ \ A$. The case weights w_{left} and w_{right} determine the two subgroups with $w_{left,i} = w_i I(X_{j \cdot i} \in A)$ and $w_{right,i} = w_i I(X_{ji} \in A)$ for all $i = 1, ..., n$ (I(û) denotes the indicator function).

3. Recursively repeat steps 1 and 2 with modified case weights w_left and w_right, respectively.

from https://eeecon.uibk.ac.at/~zeileis/papers/Hothorn+Hornik+Zeileis-2006.pdf

After step 1 is completed, any goodness of fit method can be used to generate the split and choose the set A. Note that in this method the splitting is done separately from the variable selection.