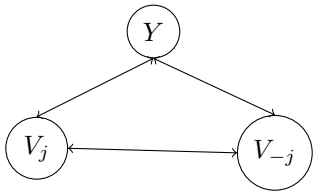
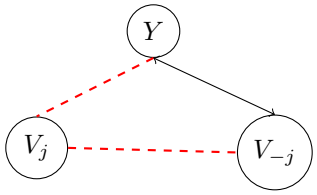
INFFOREST Variable Importance

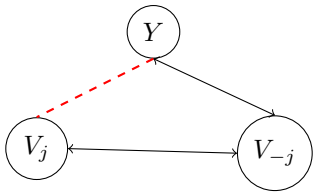
Variable importance measures must contend with the following relationships in the data, for each V_j in V: first, there is the relationship between Y and V_j, then the relationship between V_j and the other predictors, V_{-j}, and finally the relationship between Y and the other predictors V_{-j}.



In permuted variable importance, the null hypothesis is that Y is independent of V_j, regardless of the relationship between V_j and V_{-j}. By permuting V_j blindly and then calculating the RSS, the relationship between V_j and Y and the relationship between V_j and V_{-j} are broken. The other variables, V_{-j} are not permuted and since the RSS is calculated using the original model fitting $Y \sim V$, the relationship between Y and the V_{-j} is maintained.

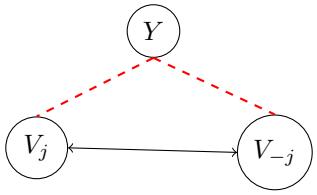


In conditional variable importance, however, the null hypothesis that is tested is that V_j is independent of Y given the relationship between V_j and V_{-j} and the relationship between V_{-j} and Y. Therefore, the permutations on V_j are done in such a way that the relationship between V_j and Y are broken, while approximately maintaining the relationships between V_j and V_{-j} and V_{-j} and Y.



This permutation structure has the following implications: 1. If V_j has a weak relationship with Y but a strong relationship with V_{-j}, the conditional variable importance value will be high. 2. If V_j is approximately independent of V_{-j}, but a good predictor of Y, then the conditional variable importance of V_j will be high.

Conditional variable importance provides a method of statistical inference on random forests, but it does not answer the same question as statistical inference on linear models. Namely, what is the relationship between V_j and Y given the other V_{-j} variables in the model? The INFFOREST method of variable importance permutes under the null hypothesis that V_j given V_{-j} is independent of Y. This leads us to break the relationships between Y and the V_j en mass, according to the respective structure of V_j and V_{-j}.



This has the following conclusions: 1. If V_j is a good predictor of Y but independent of the rest of the predictors, the INFFOREST variable importance will be high. 2. If V_j is a good predictor of Y but is heavily

correlated with at least one of the other predictors, the INFFOREST variable importance will be low. It is assumed that the information gained from adding V_j to the model could be gained from one of the other predictors.

Algorithm and Implementation

The INFFOREST variable importance is a method of permuted variable importance not unlike that of conditionally permuted variable importance. INFFOREST values are calculated at the tree level, using the partitions on V_j from a tree created to predict the model $V_j \sim V_1, ..., V_{j-1}, V_{j+1}, ..., V_p$. This auxiliary tree is fit by considering all $p-1$ predictors at each split and so may be quite large or quite small depending on the richness of the correlation structure around V_j. The auxiliary tree is also fit using the OOB sample, $\hat{B}_t$, for the tree at question. If the auxiliary tree results in a single leaf (i.e. there are no splits), then $\hat{B}_t$ is permuted blindly, without partitions. If the auxiliary tree results in two leaves, there will be two partitions on $\hat{B}_t$ to permute $\hat{B}_t$ within, and so on. After permuting $\hat{B}_t$ within these partitions, the RSS is calculated for that tree using the permuted dataset. The absolute difference of the RSS after permutation and the RSS before permuting the sample is INFFOREST variable importance for that tree.

Note that for this reason, the INFFOREST variable importance is always greater than or equal to zero, and is standardized by the max INFFOREST variable importance value given by that tree. As the variable importance values are calculated for each tree for each variable, once the method is completed there is a distribution of potential variable importance values for V_j, one for each tree. These distributions may or may not be normal, depending on the multicolinearity of the predictors. The INFFOREST variable importance algorithm works as follows:

Algorithm 1 INFForests, $VI_{inf}(R)$

1: Fit a random forest, R on the dataset D fitting the model $Y \sim V_1, ..., V_p$.
2: **for** each $V_i \in V_1, ..., V_p$ **do**
3: **for** each $t \in R$ **do**
4: Calculate: $\Xi_o = \frac{1}{\nu_t} RSS(t, \bar{B}_t)$
5: Calculate a tree T_i that predicts $V_i \sim V_1, ..., V_{i-1}, V_{i+1}, ...V_p$ using the subset of the observations used to fit t
6: Permute $\bar{B}_t$ with respect to the set of partitions P_{xi} from T_i.
7: **if then**T_i is a leaf
8: Permute the V_i values blindly with respect to no partitions. Set $\bar{B}_t^*$ to be equal to $\bar{B}_t$ except V_i is permuted.
9: **end if**
10: Now find $\Xi^* = \frac{1}{\nu_t} RSS(t, \bar{B}_t^*)$
11: The difference between these values, $\Xi^* - \Xi_o$, is the variable importance for V_i on t
12: **end for**
13: Test the null hypothesis that 0 is the likely value of $\frac{1}{\nu_t} RSS(t, \bar{B}_t^*)$ using the distribution of values of Ξ^* gathered from each tree in R
14: **end for**

INFFOREST variable importance operates under the null hypothesis that Y is independent of V_J given the correlation structure of V_j and the other V_{-j} predictors, or that the true INFFOREST variable importance for V_j is 0. The alternative hypothesis is that Y and V_j are not independent given the correlation structure of V_j and the other predictors or that the INFFOREST variable importance for V_j is greater than zero. After INFFOREST values have been computed for the entire forest, they are treated as samples from the population of possible INFFOREST values for V_j given the random forest R_f, a significance test can be run under the null hypothesis stated above.

Recall the data sets D_1 and D_2, introduced in chapter 2. Both datasets contain 12 predictors and one response, where there is some type of correlation structure between the first four variables. In D_1, this structure is linear, and in D_2 it is not. The median INFFOREST values for a random forest fit to a subset of

each data set are presented in the following tables. The p-values listed in table 1 are the observed proportion of INFFOREST values for that variable that are above zero.

Table 1: Median INFFOREST variable importance values from random forests of 100 trees fit on the first simulated data set

	median	Coefficient	pval
V1	0.3252743	5	0.0000000
V2	0.3325859	5	0.0000000
V3	0.3044358	2	0.0000000
V4	0.2967535	0	0.0000000
V5	0.4687643	-5	0.0505051
V6	0.5529826	-5	0.0202020
V7	0.4586385	0	0.1414141
V8	0.5320091	0	0.0505051
V9	0.5216065	0	0.0303030
V10	0.5001833	0	0.1717172
V11	0.4885839	0	0.1111111
V12	0.4824380	0	0.1414141

At significance level $\alpha = .05$, we would reject the null hypothesis that the true INFFOREST value for these variables is zero for variables $V_1, ..., V_4$ and V_6. We have found that in the context of the other predictors these predictors have a significant relationship with Y.

The parameters of a random forest are the data, the formula, the number of trees ($ntree$), and the number of variables to consider as possible candidates at each split ($mtry$). We'll investigate how INFFOREST variable importance on D_1 changes as the last two parameters are altered. First, random forests will be created for the following values of $mtry$: $mtry = 4, 7, 12$. The random forest behind table 1 was created using $mtry = 7$. These forests will all have the same number of trees, $ntree = 50$.

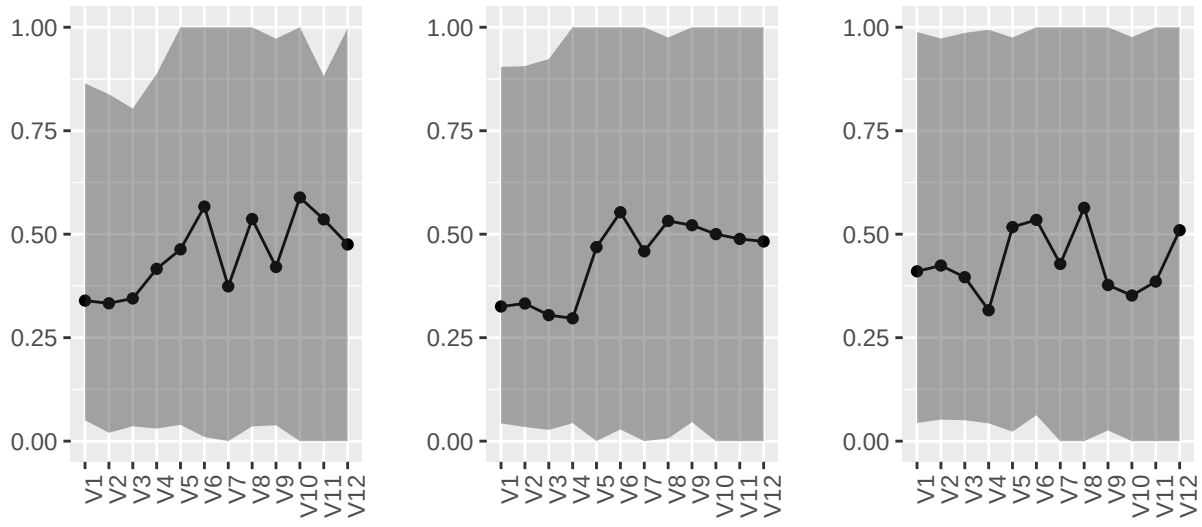


Figure 1: Distribution of INFFOREST variable importance values for data set D1 in random forests with mtry = 4,7,12.

In figure 1 the shading represents the most common values of INFFOREST for that variable. The shaded area represents a 95% confidence interval around the average value. As $mtry$ approaches the full number of

Table 3: INFFOREST Variable Importance for random forests with 50, 100, and 200 trees

	median	pval	median	pval	median	pval
V1	0.39	0.00	0.33	0.00	0.32	0.01
V2	0.36	0.00	0.33	0.00	0.35	0.00
V3	0.30	0.00	0.30	0.00	0.29	0.00
V4	0.29	0.00	0.30	0.00	0.36	0.00
V5	0.52	0.06	0.47	0.05	0.51	0.03
V6	0.52	0.00	0.55	0.02	0.55	0.03
V7	0.35	0.18	0.46	0.14	0.49	0.14
V8	0.55	0.02	0.53	0.05	0.51	0.06
V9	0.51	0.06	0.52	0.03	0.50	0.07
V10	0.48	0.16	0.50	0.17	0.48	0.16
V11	0.52	0.10	0.49	0.11	0.50	0.08
V12	0.53	0.14	0.48	0.14	0.50	0.16

predictors for the data set, the distributions become less variable. The parameter *mtry* is generally taken to be between one third and just over one half of the predictors in the data set, but should be optimized.

Table 2: Out of bag RSS values for random forests on data set D1 with mtry equal to 4, 7, or 12

mtry	RSS
4	21770.58
7	21485.52
12	24640.66

While the significance of the variables changed slightly for different values of *mtry*, in applications this may not be anything more than a practical inconvenience. The value of *mtry* that optimizes the tree is $mtry = 7$ as seen in table 2, and this is the model that would be used both for prediction and for inference.

Unlike *mtry* which must be optimized for each data set manually, traditionally the number of trees to fit in a forest follows a simpler rule: more trees are better. There is some risk that, given a high enough value of *mtry*, after a certain number of trees are fit, they will begin to be more correlated with each other than they would have been otherwise. To demonstrate the consistency in the INFFOREST variable importance significance testing, two new random forests will be constructed. Each will follow the same formula as the random forest from table 1, $Y \sim V$, and will have $mtry = 7$, and $ntree = 50, 200$ (the random forest where $ntree = 100$ was fit in the previous example).

```
## Warning in noder(y = yhatl, xs = xsl, mtry, min): NAs introduced by
## coercion
```

In the first we found V_8 to be significant and in the second forest, with 100 trees we found that V_9 was significant. Neither of these variables were used to generate Y and are roughly uncorrelated with the other predictors. As $ntree \to \infty$, the INFFOREST values become more consistant. Consider table **??**, that represents the INFFOREST values from a random forest fit on $Y \sim V$ with $mtry = 7$ and $ntree = 500$.

```
## Warning in noder(y = yhatl, xs = xsl, mtry, min): NAs introduced by
## coercion
```

```
## Warning in matrix(value, n, p): data length [11] is not a sub-multiple or
## multiple of the number of columns [5]
```

```
## Warning in noder(y = yhatl, xs = xsl, mtry, min): NAs introduced by
```

```
## coercion
```

Table 4: INFFOREST Variable Importance for a random forest on the data set D1 with 500 trees

	median	pval
V1	0.32	0.00
V2	0.35	0.00
V3	0.28	0.00
V4	0.36	0.00
V5	0.52	0.03
V6	0.56	0.02
V7	0.46	0.15
V8	0.53	0.05
V9	0.52	0.06
V10	0.44	0.17
V11	0.49	0.08
V12	0.46	0.16