Conclusion

INFFOREST Comparisons With Other Methods

As discussed in the beginning of chapter 4, each type of permuted variable importance, permuted, conditionally permuted, and INFFOREST, operates on a slightly different null hypothesis. This explains the differences in the results when each method is run on the same random forest.

INFFOREST variable importance is not alone in methods that conduct statistical inference on random forests.To compare the results from INFFOREST, conditional variable importance, and permuted variable importance, a random forest was generated on the first 200 rows of the data set D_1, following the formula $Y \sim V$. This random forest considered 7 of the 12 predictors at each split and contained 200 trees. Then the INFFOREST, conditional permuted, and permuted variable importance distributions were calculated for each variable. These distributions are represented below in figure 1.

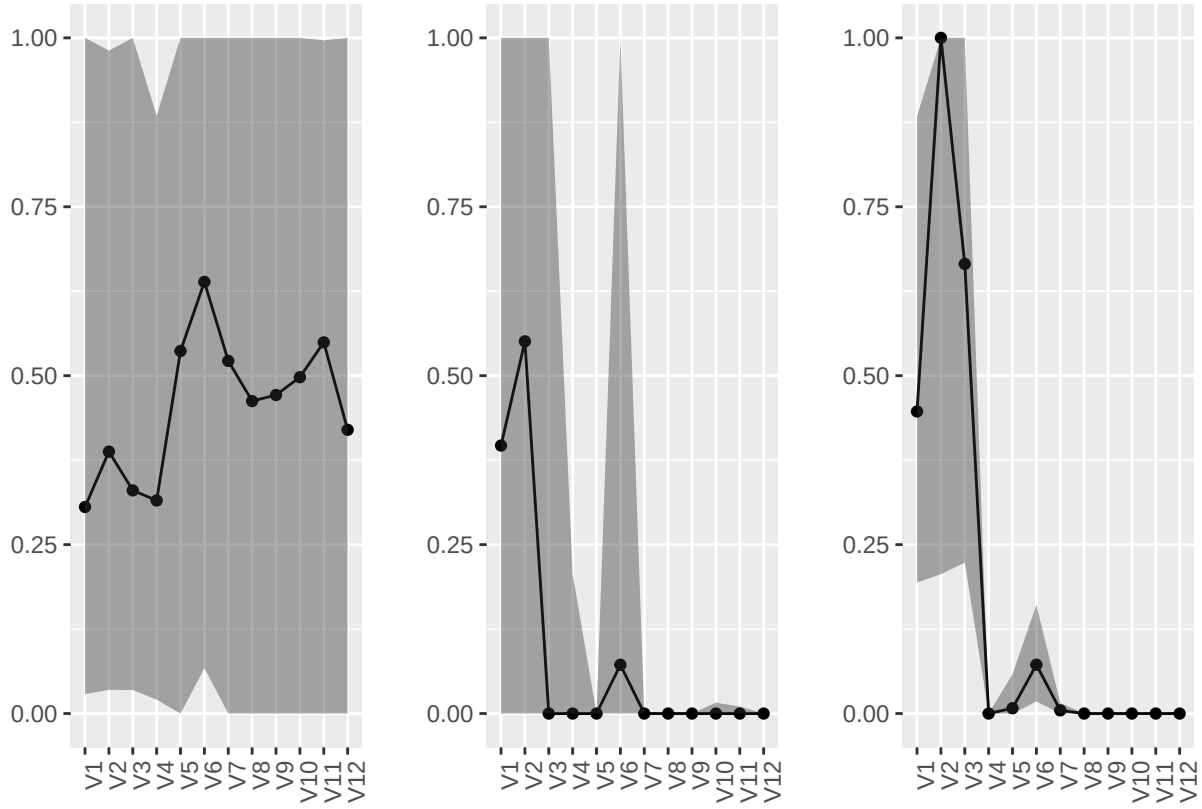

Figure 1: Median Values of INFFOREST, Conditionally Permuted, and Permuted Variable Importance

Mimicking the construction of the plot in chapter 4, the ribbon surrounding the average values is the 95% confidence interval constructed around the average importance values. The first main difference visible between INFFOREST and the other methods in figure 1 is that the median INFFOREST values are above zero, even for the variables that are not considered significant. The random forest is conducted in such a way that even predictors that may not be important in the overall model are important in a few trees. These are three different methods, following three different permutation schemes, and they are based on three different null hypotheses.

Table 2: INFFOREST Variable Importance (left) for a random forest on the data set D2 with 300 trees is compared to the coefficeints of the linear model (right)

	median	pval	Estimate	Std. Error	t value	Pr(>|t|)
V1	0.54	0.13	-4.54	3.45	-1.32	0.19
V2	0.40	0.00	-13.98	1.94	-7.22	0.00
V3	0.41	0.02	2.59	1.60	1.62	0.11
V4	0.43	0.01	6.48	1.46	4.44	0.00
V5	0.54	0.07	-5.88	3.45	-1.71	0.09
V6	0.47	0.11	-8.94	3.28	-2.72	0.01
V7	0.52	0.09	-4.72	3.27	-1.44	0.15
V8	0.50	0.10	-6.39	3.47	-1.84	0.07
V9	0.49	0.14	2.06	3.38	0.61	0.54
V10	0.50	0.11	-0.56	3.14	-0.18	0.86
V11	0.55	0.10	-1.50	3.70	-0.41	0.69
V12	0.56	0.06	4.46	3.36	1.33	0.19

INFFOREST Conclusions

Using INFFOREST variable importance, we were able to demonstrate significance for all of the five predictors used to generate the response. More precisely, we were able to demonstrate that the INFFOREST values for $V_1, ..., V_6$ were significantly different from zero. Given the null hypothesis we considered, this allows us to claim that $V_1, ..., V_6$ are significant predictors of Y, given the other variables included in the model. We found one predictor to be significant, V_4, when it was not used to generate Y, and so did not quite achieve the original goal. For comparisons, these are the predictors deemed significant by the analogous linear model:

Table 1: The estimated coefficients and p-values for a linear model on the formula Y ~ V, D1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.01	0.04	-0.42	0.68
V1	4.83	0.11	44.85	0.00
V2	5.20	0.12	44.44	0.00
V3	2.07	0.10	19.91	0.00
V4	-0.09	0.09	-0.95	0.34
V5	-5.02	0.04	-131.79	0.00
V6	-5.01	0.04	-131.10	0.00
V7	0.03	0.03	1.05	0.30
V8	-0.04	0.04	-1.10	0.27
V9	0.07	0.04	1.84	0.07
V10	-0.01	0.03	-0.22	0.83
V11	0.04	0.03	1.28	0.20
V12	-0.05	0.04	-1.31	0.19

These results may not be as surprising as one would think. The response in the data set D_1 was generated linearly from the predictors. The relationship between the response and the predictors is better represented by the linear model, as it is a linear relationship. When the same procedure is repeated on D_2, seen in the table below, we do not find that the linear model is a better fit, as the structure in D_2 is not linear.

Areas for further study

There is clearly work to be done to recover the inferential and interpretable properties of the linear model in machine learning. As mentioned briefly in the introduction, random forests of CART trees may be biased. Hothorn et al. demonstrated that CART is biased toward variables with a great number of possible splits. An interesting follow up to this paper would be to apply INFFOREST variable importance to forests of the unbiased conditional inference trees, as discussed in Hothorn et al. As the `party` package in R created by Hothorn et al. that implements conditional inference trees, contains functions for accessing and creating custom random forests, this could be relatively simple.

On a theoretical level it would be interesting to see a proof that in the case where the linear model is appropriate, that CART does or does not approximate the linear model. From my reading and from my simulations, it does not seem like CART fits linear data very well, but a formalization of this phenomenon would be welcome.

Another area for further study is testing the INFFOREST code on situations with lots of data and lots of trees in the random forest. As *ntree* and n, the number of rows in the data set, increase, variance in the INFFOREST distribution seems to decrease. When n is increased, the average tree size increases. When *ntree* is increased, the size of our forests is increased. As it is written now, however, the random forest object requires quite a lot of memory and the code to perform operations on the trees is quite intensive. This would most likely require rewriting significant parts of the code, potentially in a language other than R. A full look at this property was unfortunately not possible in the given time frame due to these constraints.