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.

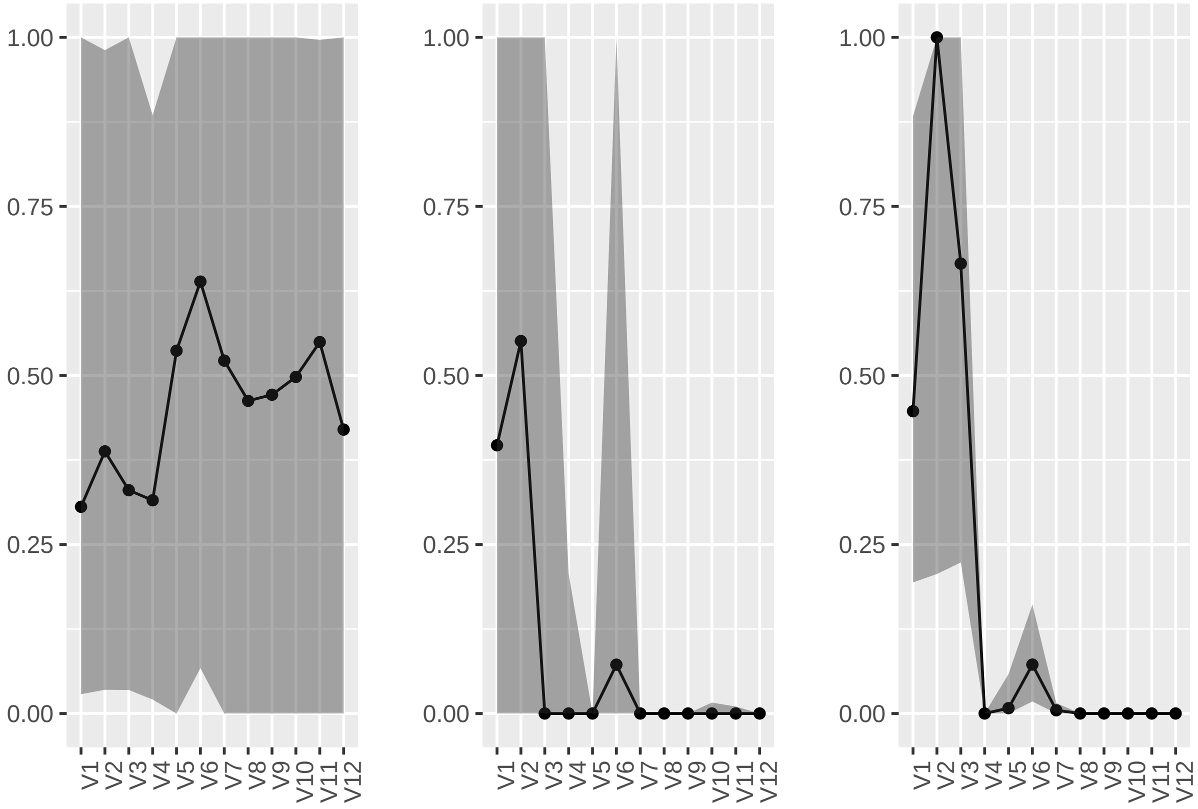

Figure 1: Median Values of INFFOREST, Conditionally Permuted, and Permuted Variable Importance

&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 1 demonstrates the implications of each permutation scheme.

Variable Importance Method	Ties between X_j and Y	Ties between X_j and X_{-j}	Ties between X_{-j} and Y
Permuted	Broken	Broken	Maintained
Conditional Permuted	Broken	Maintained	Maintained
INFFOREST	Broken	Maintained	Broken

Table 1: The permutation structure in each variable importance method functions to break one or more ties between the predictors and the response.

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 2: 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

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. As demonstrated by Hothorn et al., CART is biased toward variables with a great number of possible splits. This could explain our results in the following way: if a predictor has a strong relationship with another predictor, then it could make it a more attractive candidate for the splitting criteria. 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.

Breiman et al., in *Classification and Regression Trees*, begin an early chapter with: "the question raised in this section is: what is truth and how can it be estimated?" [@bibCART, p. 9]. This textbook was written in 1984, far before machine learning methods were used to analyze large amounts of data on private citizens across nearly every industry and even by the U.S. government. [@bibNYer]. The question of truth and the limits we face in estimating it, has ramifications far beyond what Breiman et al. could have envisioned.