INFTrees and INFFOREST Variable Importance

Theory

While conditional variable importance (Strobl et al) conditionally permutes each variable given the structure signified by the model that predicts the response, $Y \sim X_1, ..., X_i, ..., X_p$, our method conditionally permutes each variable given the structure outlined in a new model with the variable of interest as the response, $X_i \sim X_1, ...X_{i-1}, X_{i+1}, ...X_p$. This is not the most straightforward process, as trees partition the sample space, however, in INFTrees these partitions on the variables $X_1, ...X_{i-1}, X_{i+1}, ...X_p$ are treated as pseudo partitions on the variable of interest, X_i. This is accomplished by first partitioning on the sample predictors $X_1, ...X_{i-1}, X_{i+1}, ...X_p$ and then inferring the partitions on X_i.

***ADD BETTER PLOT FOR EXAMPLE**

INFTrees

For a CART, T, representing the model $Y\ X_1, ..., X_p$ where $Y, X_1, ..., X_p$ are vectors of length n, the INFTrees algorithm proceeds as follows:

```
Algorithm 1 INFTree, $VI_{inf}(T)$
  for each $X_i \in X_1, ..., X_p$ do
      Calculate: $\Phi_o = RSS(T, (Y, X_1, ..X_p))$
      Fit the tree $T_{X_i}$, where $T_{X_i} : X_i \sim X_1, ..., X_{i-1}, X_{i+1}, ...X_p$
      Extract the set $P_{X_i}$ of partitions on $X_i$ from $T_{X_i}$
      Permute $X_i$ with respect to $P_{X_i}$
      Find $\Phi^* = RSS(T, (Y, X_1, ..., \bar{X}_i, ...X_p))$
      Repeat the above procedure to find the distribution of $\Phi^*$
      Test the null hypothesis that $\Phi_o$ is the likely value of $RSS(T, (Y, X_1, ..X_p))$
  end for
```

This procedure allows the null hypothesis that Y is independent of X_i given the values of $X_1, ...X_{i-1}, X_{i+1}, ...X_p$ to be tested. Therefor, values of VI_{inf} could be compared in a similar manner to the coefficients of linear regression.

INFForests

The algorithm for determining $VI_{inf}(R)$ follows similarly.

Implementation In `INFTREES` and Results

Notes on the Implemetation

Implementing the `INFFOREST` and therefor the `INFTREES` algorithms, required creating a suite of functions to create trees and random forests. The trees are fit following the standard two-part CART-like algorithm. [1] The function chooses a variable to split on with linear correlation with respect to Y, but instead of looking for correlations above a certain threshold which is common, it chooses the variable with the highest correlation when compared to its peers. This alleviates the situation where a variable with a non-linear relationship

[1] A great deal of effort was undertaken by the author to find the definitive, authentic CART algorithm. This implementation follows the rough strokes set out in the 1984 text *Classification and Regression Trees* to the best of the author's ability and may not be exactly the algorithm found in R packages like 'tree()'

Algorithm 2 INFForests, $VI_{inf}(R)$

```
Fit a random forest, $R$ on the dataset $D$ fitting the model $Y \sim X_1, ..., X_p$.
for each $X_i \in X_1, ..., X_p$ do
    for each $t \in R$ do
        Calculate: $\Xi_o = \frac{1}{\nu_t} RSS(t, \bar{B}^t)$
        Calculate a tree $T_i$ that predicts $X_i \sim X_1, ..., X_{i-1}, X_{i+1}, ...X_p$ using the subset of the observations used to fit $t$
        Permute the subset of $X_i$ contained in $\bar{B}_t$ with respect to the set of partitions $P_{xi}$ from $T_i$.
        Now find $\Xi^* = \frac{1}{\nu_t} RSS(t, \bar{B}_t^*)$
        The difference between these values, $\Xi^* - \Xi_o$, is the variable importance for $X_i$ on $t$
    end for
    Test the null hypothesis that $\Xi_o$ is the likely value of $\frac{1}{\nu_t} RSS(t, \bar{B}_t^*)$ using the distribution of values of $\Xi^*$ gathered from each tree in $R$
end for
```

would be passed over again and again. The splitting is done via minimization of the following function with respect to i:

$$RSS_{node}(i, X, Y) = RSS_{leaf}(Y|X < i) + RSS_{leaf}(Y|X \geq i)$$

$$RSS_{leaf} = \sum (y - \hat{y})^2$$

$$\hat{Y} : \hat{y} \in \hat{Y} : \hat{y} = E(Y),\ where\ |\hat{Y}| = |Y|$$

This function considers the regression case only, and only numeric predictors. Leafs are created when the resultant split would be unsatisfactory, i.e. at least one daughter node would have five members or less. This generates very large trees - a quality that is not an issue in random forests but may be problematic in a stand-alone setting. At this time, there is also no function to prune the trees.

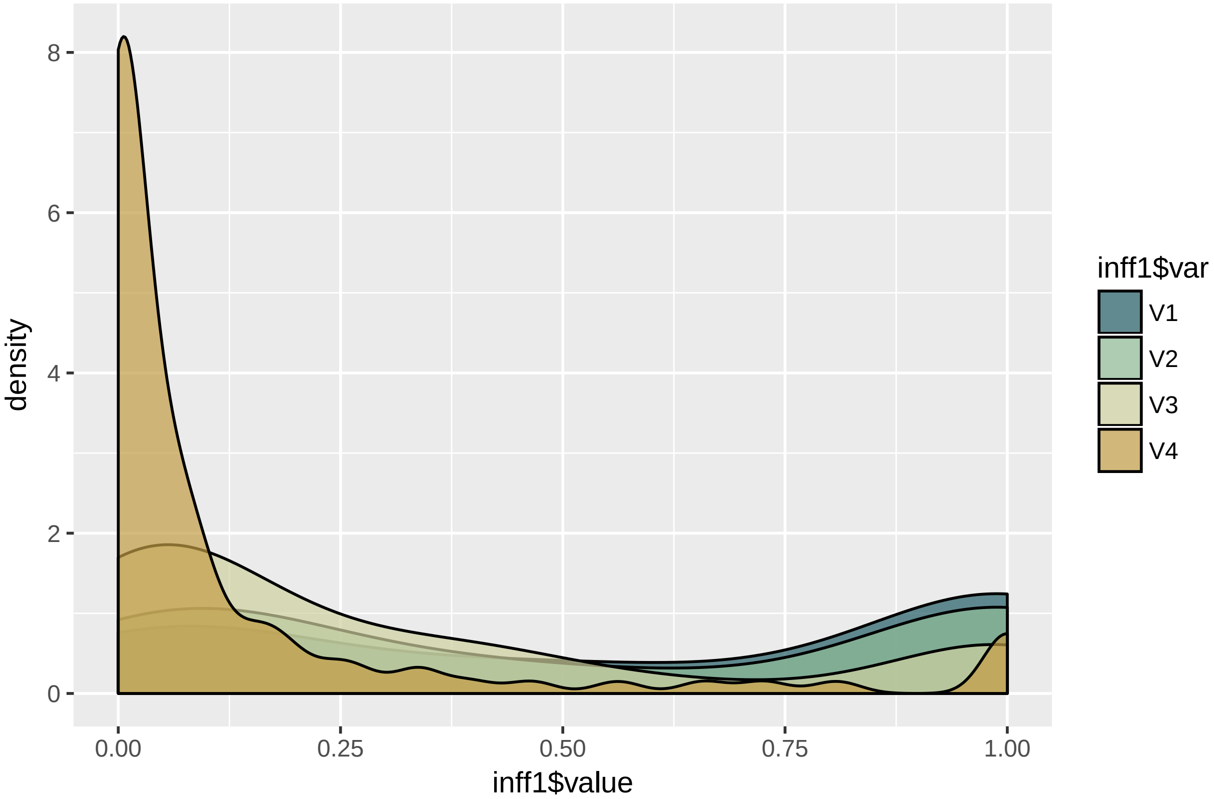
Table ___: A Home-Grown Tree on $Y\ X_1 + X_2 + X_3 + X_4$

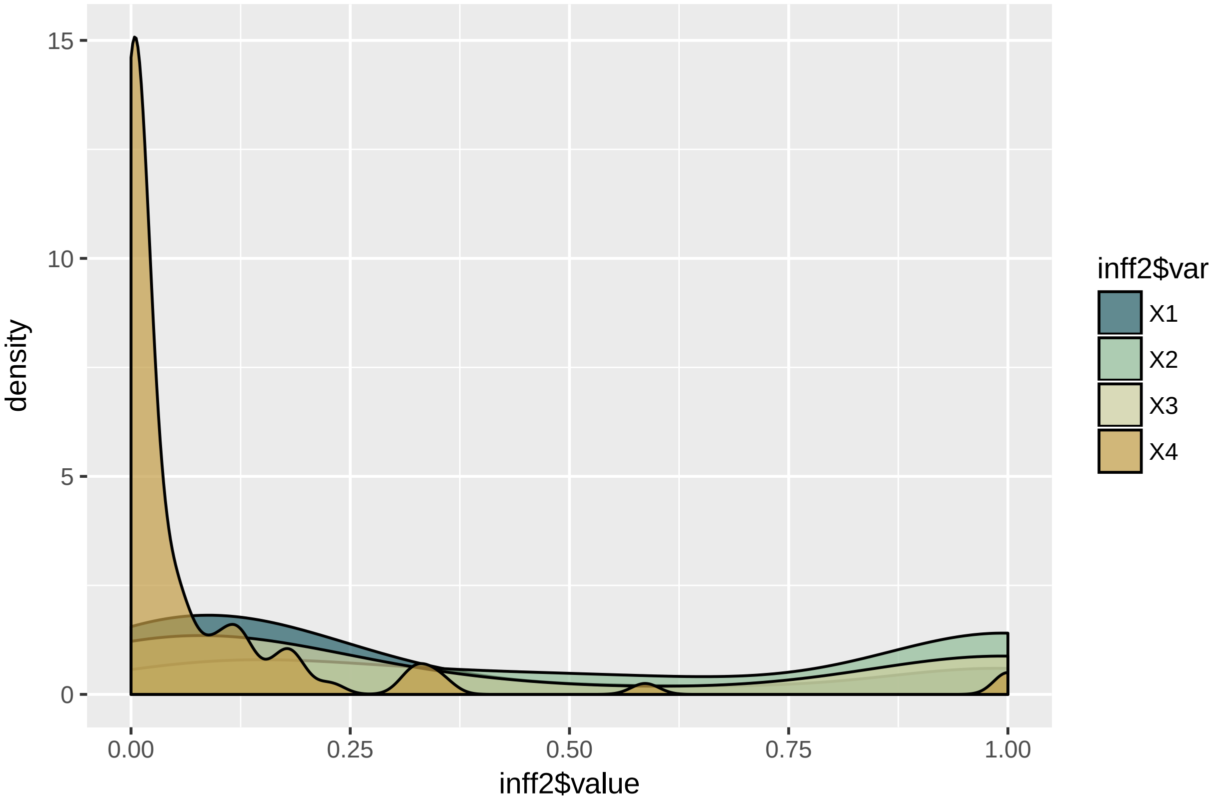
var	n	dev	ypred	split.cutleft
X2	100	13533.7097509203	-1.86611156597531	2.08291569465727
	20	3454.06120779419	14.2994680819369	0
X1	80	7113.12663449009	-5.90750647795336	-0.215201407955814
X1	49	3302.40662074389	-1.06398094549242	0.727851085907896
	10	641.178801737327	6.66470237216789	0
X3	39	1846.6927300008	-3.04569461668737	2.71188521864226
	11	609.989312132866	4.245610434698	0
X4	28	1146.50420509996	-5.91013588687448	3.85780171652793
	5	3.73292210024685	-9.75960885048842	0
X3	23	1072.0255738871	-5.07329393826275	1.72276855437662
	7	94.5037759591882	-7.72482735819737	0
	16	977.52179792791	-3.91324806704135	0
	31	3060.23418158996	-13.5634016744239	0

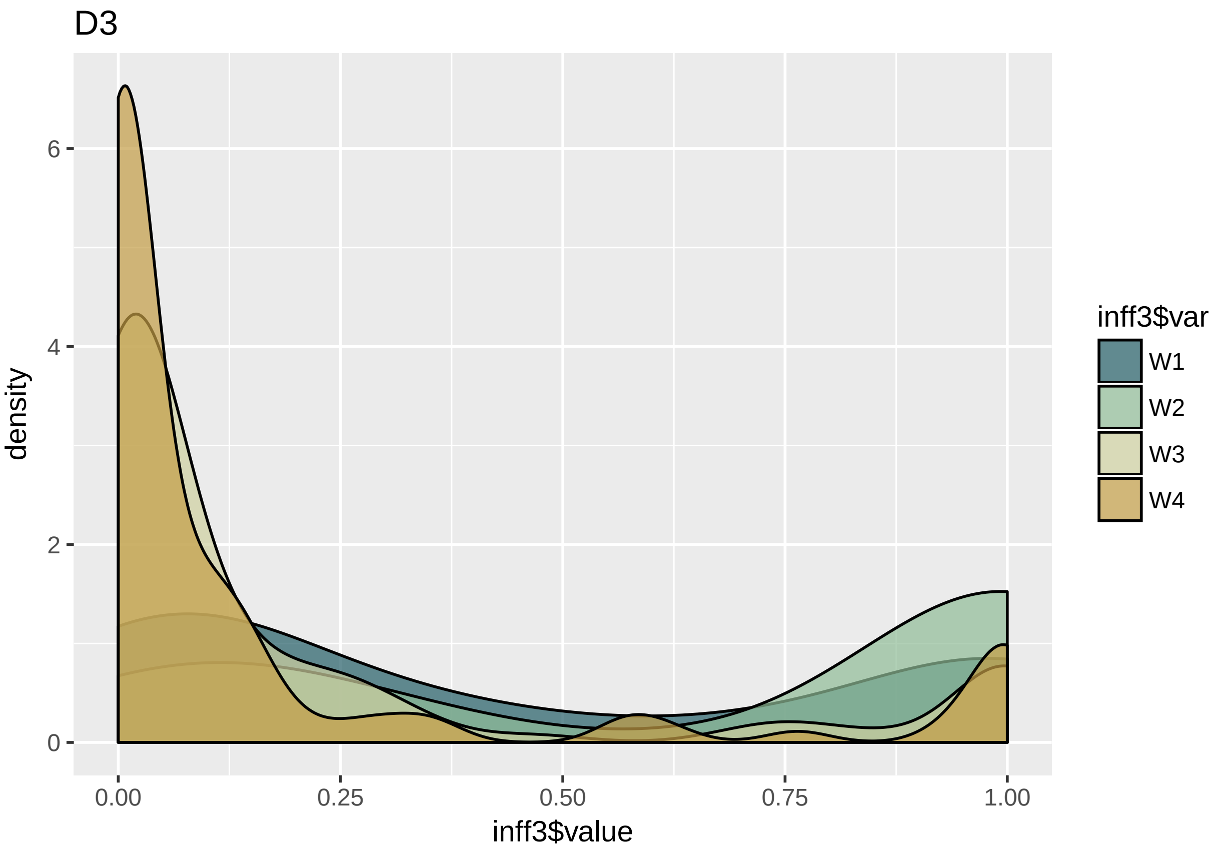
The INFTREE function follows the algorithm above *reference*. The partitions on X_j are generated by fitting a tree, T, to the model $X_j \sim X_1, ..., X_{j-1}, X_{j+1}, ..X_p$ and calculating the predictions $T(X_1, ..., X_{j-1}, X_{j+1}, ..X_p)$. Then permuting X_j with respect to the partitions on X_j given by those predictions. For example, if $x_j \in X_j$ and the value of $T(x_1, ..., x_{j-1}, x_{j+1}, ..x_p)$ corresponding to x_j is α, x_j is permuted along with the other values of X_j that also have $T(x_1, ..., x_{j-1}, x_{j+1}, ..x_p)$ corresponding to α.

The values of $INFFOREST(X_j)$ are scaled in the following way: since the INFFOREST function computes the INFFTREES, (or the difference in post and pre permutation RSS), values in a tree-wise manner, each tree's values are divided by the maximum value. This ensures that the values are between zero and one, and that in each tree one variable is clearly deemed the *most important*.

Results


D1
8
6
4
2
0
density
0.00
0.25
0.50
0.75
1.00
inff1$value
inff1$var
V1
V2
V3
V4

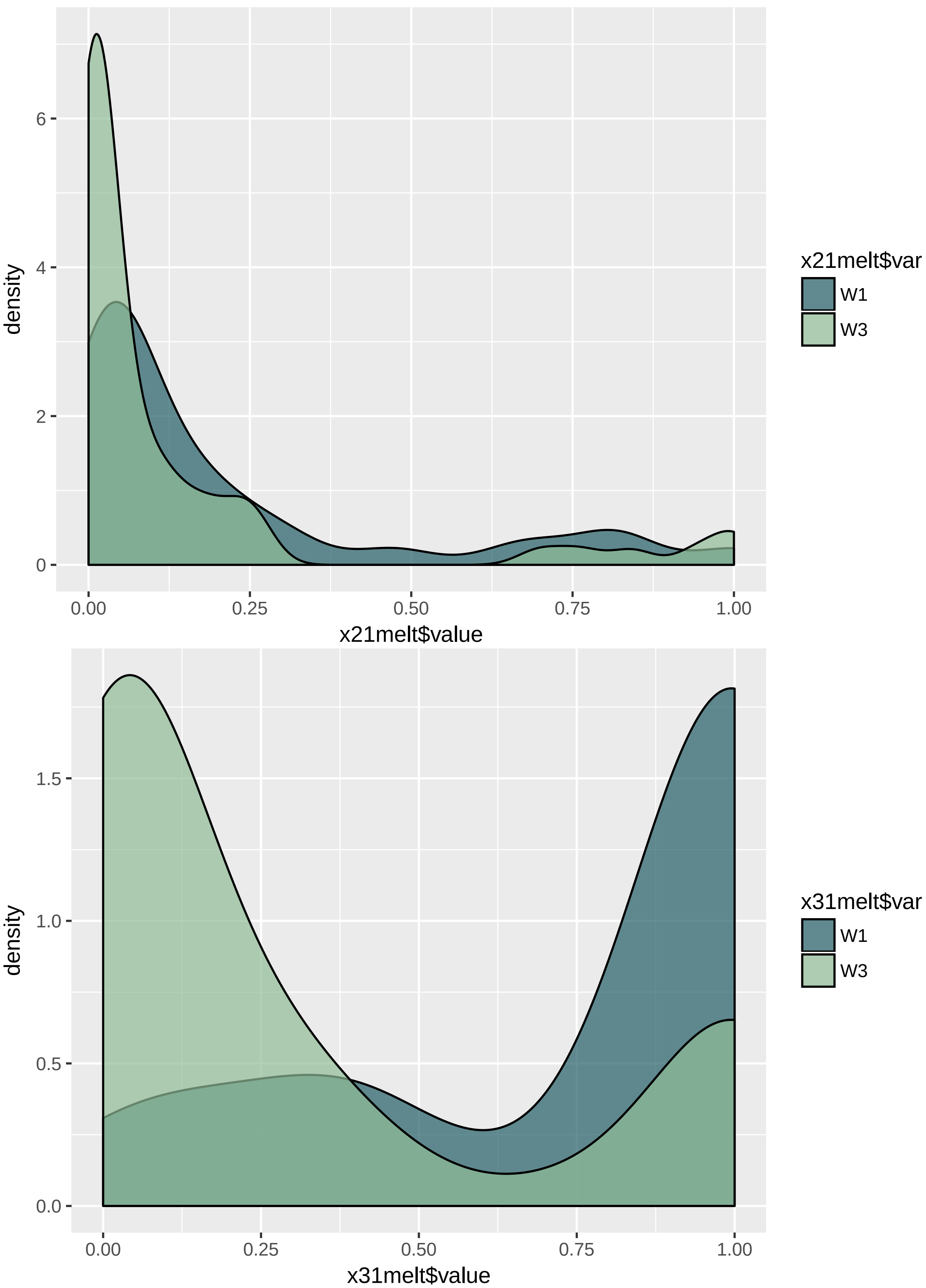

D2
15
10
5
0
density
0.00
0.25
0.50
0.75
1.00
inff2$value
inff2$var
X1
X2
X3
X4



In the situation where there is little correlation between the predictors, the distribution of the INFFOREST output is a sharp peak ending at one of the end points, zero or one. When there are, however, strong correlations between the predictor variables, and `mtry` is suitably large but smaller than `p`, the trees in the forest must decide between them. In these situations, the INFFOREST distribution is multimodal, with one peak at one end of the interval, $INFFOREST(X_i) = 1$ and another when $INFFOREST(X_i) = 0$.

To demonstrate this situation, take the dataset $D2$, as described above. In the random forest corresponding to this model, the variables $X2$ and $X3$ are considered substitutes for each other. In the trees where $X2$ has $INFFOREST = 1$, $X3$ has $INFFOREST <<$ and visa versa.

```
## Using  as id variables
## Using  as id variables
```



(i.e. the INFFOREST distributions of X2 and X3 in the trees where X3< .5)

Of course, one may be inclined to infer a p-value for the null hypothesis that $INFFOREST = 0$ for each of these variables. This could be done straight-forwardly enough in situations where there is not strong multicolinearity within the predictors as the distributions are reliably half of the familiar bell shaped curve centered around either zero or one. It would be quite difficult, however, for INFFOREST alone to test the significance of the INFFOREST distribution corresponding to correlated, paired predictors and it may not makes sense to do so at all. *talk with Andrew about fixing this?*