

## Is there a method similar to commonality analysis but can be applied beyond the scope of multiple regression?

I got one dependent variable and six independent variables. All of them are continuous.

First, I built a linear regression model, but the  $R^2$  was only 0.22.

Then I tried to build a random forest model with the same data. The  $R^2$  increased to 0.78.

After comparing the results of both models on an independent test set, I think random forest model was not over-fitted. And I'm satisfied with the result.


When I built a linear regression model, I could use commonality analysis to decompose  $R^2$  into unique and common variance of independent variables. Now the question is which method should I use to decompose the  $R^2$  explained by the random forest model in a similar way?

My first thought was to build a full model and a model without some independent variable (Let's say x). The difference of  $R^2$  between these two models should be the part explained by x. However, it turned out some model with less independent variables had higher  $R^2$  than full model. So this is an invalid method.

Could anyone give some suggestions on this? Any help would be appreciated.

random-forest variance-decomposition

asked Nov 25 '15 at 14:19

 **gh2017554**  
30 4

Did you look at importance for random forests in R? See for example: [r-bloggers.com/variable-importance-plot-and-variable-selection](http://r-bloggers.com/variable-importance-plot-and-variable-selection) – [spdml](#) Nov 25 '15 at 16:22

@spdml. Yes, I did. But the some independent variables' importance were negative. If I understand it correctly, I should remove those variables from the random forest model. However, I can use these variables which have negative importance to build a linear regression model. So they can explain some part of the dependent variable's variance. And this part of variance cannot be reflected by the variable's importance. – [gh2017554](#) Nov 26 '15 at 1:05

"The  $R^2$  increased to 0.78" is that OOB cross-validated? – [Soren Havelund Welling](#) Nov 26 '15 at 9:18

What is the purpose of your analysis: Decomposition, exploratory, or optimizing predictiveness? I just read a random paper on CA "[link.springer.com/article/10.3758%2FBRM.40.2.457](http://link.springer.com/article/10.3758%2FBRM.40.2.457)". As RF is non-linear and allow interactions between variables I don't think commonality analysis will provide a meaningful answer. Two variables could simply be complimentary, such that either variable alone would be useless, but brought together an interaction term could be learned by the model fit. I guess that would challenge the additive book-keeping principle of explained variance of commonality analysis. – [Soren Havelund Welling](#) Nov 26 '15 at 9:36

@Soren Havelund Welling. Thank you for your comments. The increase of  $R^2$  is not OOB cross-validated. I agree with you about the commonality analysis and interactions between terms. What I want is a method that can evaluate variance explained by each predictor, commonality analysis mentioned in my question is just a metaphor to illustrate the effects I want to achieve. – [gh2017554](#) Nov 27 '15 at 1:47

### 1 Answer

Your tool of choice is most likely to be the variable importance (VI) measure. That is the loss of cross-validated model predictive performance if a given variable is permuted after training.

I did develop something quite similar commonality analysis for random forest. The remaining answer, was my thoughts on that.

Regular VI do not reveal any information whether variables are redundant or complimentary. A year ago, I defined, a two-way VI term. But I'm probably not the first to get that idea. Two-way VI can under favorable conditions (n.obs, n.var, etc.) describe variable redundancy and complementarity.

Two-way VI is defined as the change of regular VI of one variable (A) conditioned by the permutation of second variable (B). - If the two-way VI is positive, this points to the model relied more on A when missing B. This points to A is redundant to B. - If the two-way VI is negative, this points to the model could not make as much use of A when missing B. This points to A being complimentary to B.

Two-way VI can be computed with and without retraining of model.

- If retraining is performed after permuting B, redundant A will improve its oneway VI. If retraining is not performed after permuting B, redundant A will only improve oneway VI slightly, as the model have not adapted to rely more on A.
- If retraining is not performed after permuting B, complimentary A will decrease its oneway VI. The model have not adapted to replace B to form a meaningful interaction effect with A. If retraining is performed after permuting B, complimentary A may only slightly decrease oneway VI, as the model can replace B to make use of A.

Retraining is best to spot variables being redundant. No retraining is best to show variables being complimentary, due to a learned interaction effect. Two variables can be both complimentary and redundant to each other.

The drawbacks of twoway-VI:

- It is not three-, four- or five-way. Thus redundancy and complementarity can be masked by variable C and D also being either redundant or complimentary. It would take too much time to computing higher orders VI and too uncertain.
- Only strong redundant/complimentary variables were identified, even with many simulation reruns to reduce the background noise.
- Variable importance is a global term generalizing the influence of a variable to the entire feature space. This may not be useful. A model may learn a interaction effect between variable in one part of feature space (making the variables locally complimentary), while in another place learn a weighted average of the variables (making the variables partly redundant). Local VI has been invented, but it would likely be too uncertain to combine with twoway-VI.

To better understand when a RF model fit make use of weighted averages of variables and when it make use of interactions effects I shamelessly recommend my own R package, [forestFloor](#), described in this [post](#).

A sketch on how to compute variable importance.

**oneway-VI** is computed by this sequence: Train model -> compute OOB.CV.1 -> Permute variable A -> compute OOB.CV.2

**oneway-VI** = OOB.CV.1 - OOB.CV.2

**twoway-VI.v1** is computed by this sequence: Train model -> compute OOB.CV.0 -> Permute variable A -> compute OOB.CV.1 -> Permute variable B -> compute OOB.CV.2

**twoway-VI** = (OOB.CV.0 - OOB.CV.1) - (OOB.CV.0 - OOB.CV.2) = OOB.CV.2 - OOB.CV.1

**twoway-VI.v2** is computed by this sequence: Train model -> compute OOB.CV.0 -> Permute variable A -> compute OOB.CV.1 -> Reconstruct variable A -> Permute variable B -> retrain model -> compute OOB.CV.2 -> Permute variable A -> compute OOB.CV.3

**twoway-VI.v2** = OOB.CV.0 - (OOB.CV.1) - (OOB.CV.2 - OOB.CV.3)

answered Nov 27 '15 at 16:09



[Soren Havelund Welling](#)

2,871 5 17

Thank you for sharing your thought. But one more thing confused me. In my experience, when a linear regression model was built, people claimed that the R2 was the proportion of variance 'explained' by the model. And this R2 was calculated on the same data which was used to build the model. The process of cross validation seemed not involved. So why the R2 generated on the training set by RF cannot be used as the explained variance? This is a little digressive, so I wonder if you can kindly explain that a little more. Thanks again. – [gh2017554](#) Nov 28 '15 at 11:34

fully grown trees are alone very overfitted. Training error to do not show overfitting. Try read this tutorial on overfitting, bias and variance: [theclevermachine.wordpress.com/2013/04/21/...](#) – [Soren Havelund Welling](#) Nov 28 '15 at 14:03

[Add Another Answer](#)