Equations

Eric Scott

5/10/2019

## R2X

: “Cumulative SS of all the Xs explained by all extracted components”. From Eriksson et al. **This is the same for PCA and PLS.**

## R2Y

: “Cumulative SS of all the Ys explained by all extracted components”

**For PLS:**

**For PCA-DA:** It’s actually the coefficient of discrimination from Tjur 2009

Where and are mean fitted values for successes and failures.

## Q2

: “The fraction of the total variation of the Ys that can be predited by a component according to cross-validation”

Predicted Error Sum of Squares (PRESS) is squared differences between observed Y and predicted values when observations were kept out (because of cross-validation).

So…

where = the product of PRESS/SS for each individual component a.

## Permutation p-values

and are derived from permutation testing and are the fraction of values of and from permuted data that are greater than or equal to the values from the real data

## RMSEP

I’m now wondering if I did this correctly… <https://stats.stackexchange.com/questions/85507/what-is-the-rmse-of-k-fold-cross-validation>

What I did was calculate RMSEP for each fold and then just take the mean of them all. But the link above seems to indicate I can’t just take the mean across all folds for some reason.

So, I first created the splits. The data is split into 7 pieces and for each CV instance, one of those pieces is randomly left out to build the model. That missing piece is then used to calculate RMSE

where is the estimation of and is the number of observations of CV instance .

Then, I just took the mean of these for all of the RMSEjs 🤷

The stack exchange link I found just now says I need to take the square root only *after* averaging the mean squared errors across all folds

Or just grouping all the squared errors from all the CV instances together before taking the mean and square root.