Improving k-fold cross-validation with anticlustering

by Tim Angelike and Martin Papenberg

Heinrich Heine Universität Düsseldorf
Institut für Experimentelle Psychologie

Problem

Low sample size (e.g. clinical contexts) may lead to noisy and biased performance estimates in cross-validation and humans are hard to predict.

Possible Solution

Partition data during *k*-fold cross-validation using anticlustering* for creating clusters of high between-group similiarity

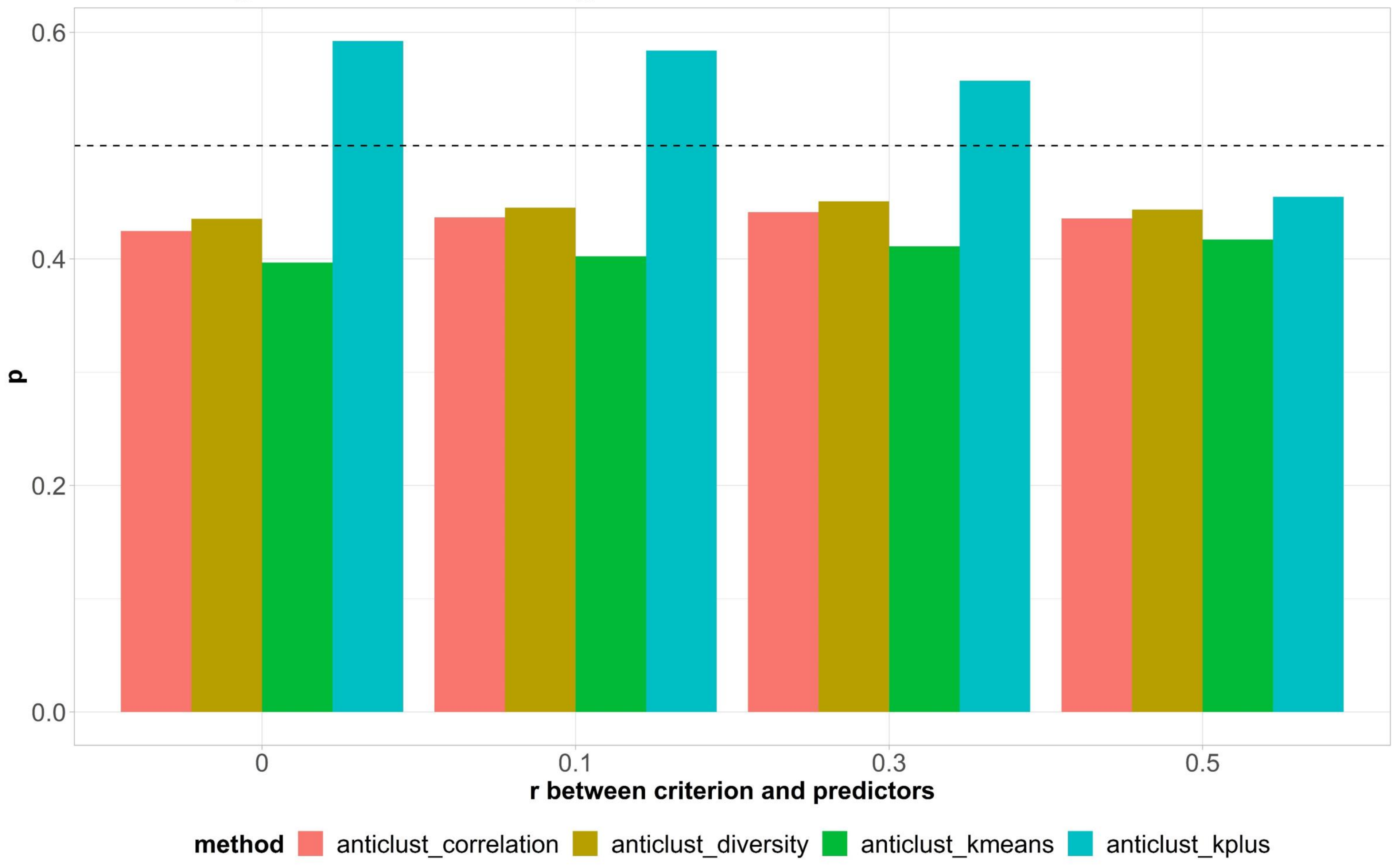
Goal

Compare prediction accuracy between classical cross-validation and anticlustering in 10 times repeated 10-fold cross-validation

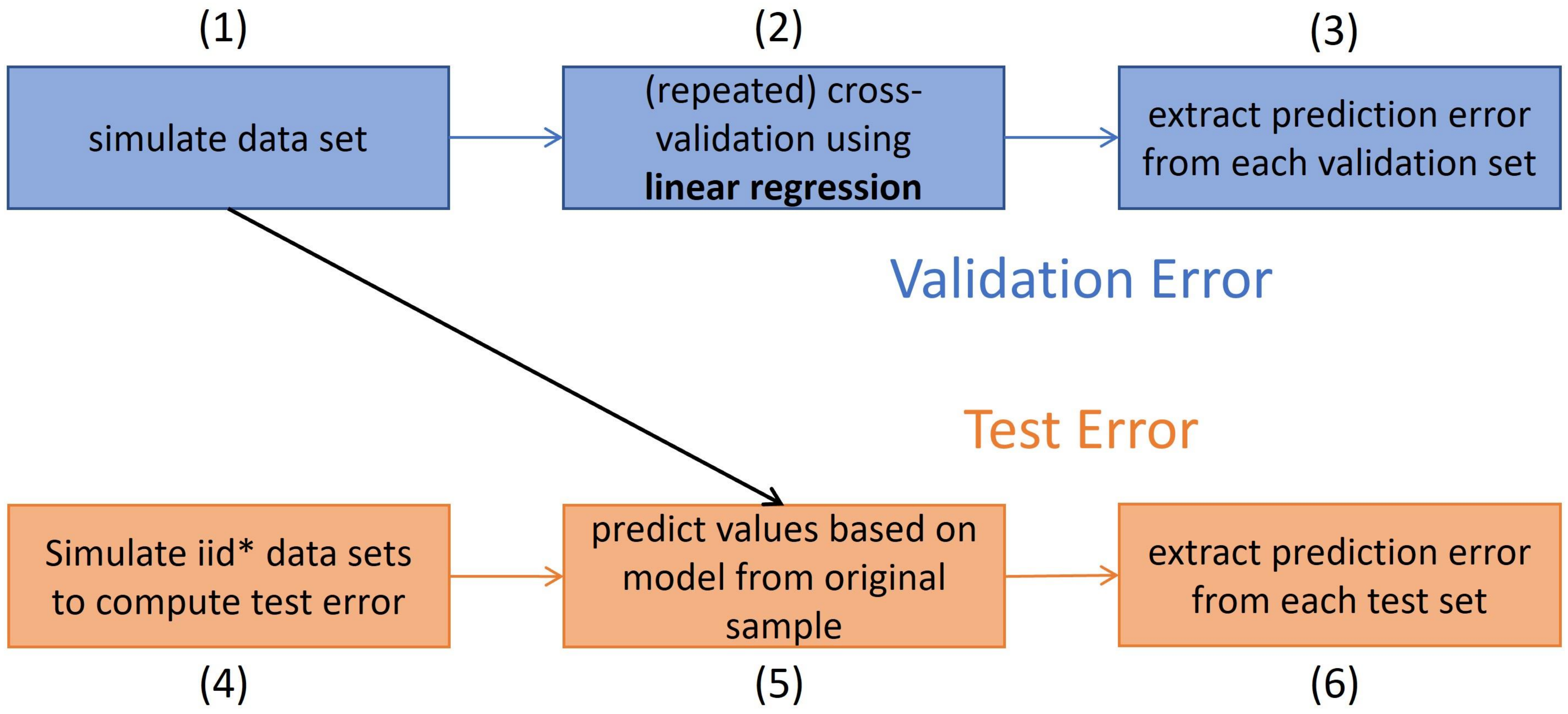
Anticlustering Methods

- (reversed) *kmeans*: creates clusters of similar means
- *kplus*: creates clusters of similar means and variances
- *correlation*: creates clusters of similar means, variances, and covariance structure
- *diversity*: maximizes sum of pairwise dissimilarities within clusters

Percentage of Methods being closer to Test Error Than Cross-Validation



Simulation Design



(1) simulate data set

(2) (repeated) cross-validation using **linear regression**

(3) extract prediction error from each validation set

Validation Error

Test Error

(4) Simulate iid* data sets to compute test error

(5) predict values based on model from original sample

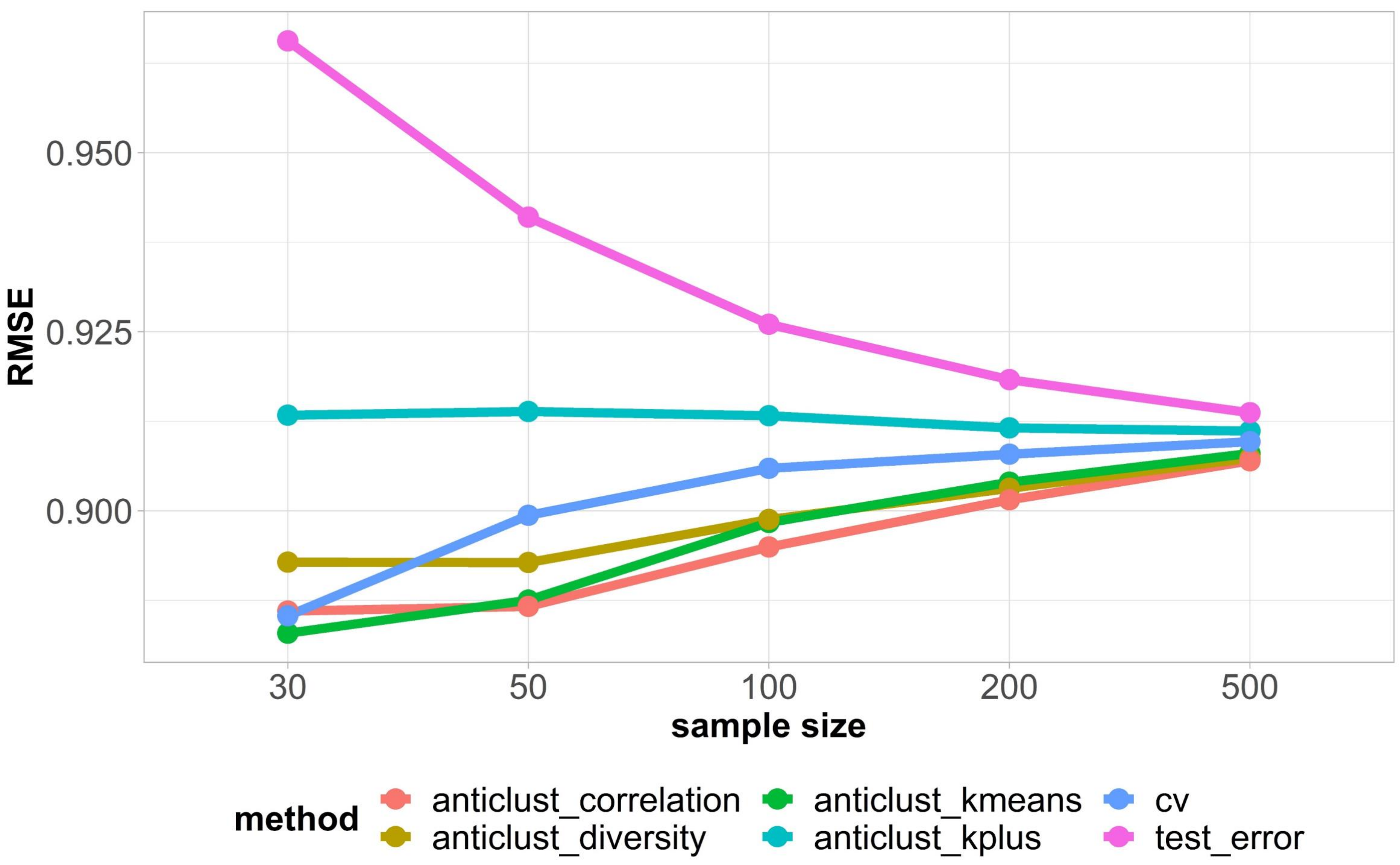
(6) extract prediction error from each test set

*iid = independent and identically distributed

Simulation Variables

sample size (*n*), *r* between criterion and predictors, *r* between predictors, #predictors

Mean Validation and Test Error



Conclusion

Cross-validation using splits based on anticlustering instead of random splits seems to provide more realistic validation error estimates

Some qualifications:
a) Depends on anticlustering method
b) Not for high predictive accuracy
c) Advantage diminished with large *N*

*Papenberg & Klau (2021)