Pattern Recognition: validation and inference

Maria J. Rosa

PRoNTo course 2017

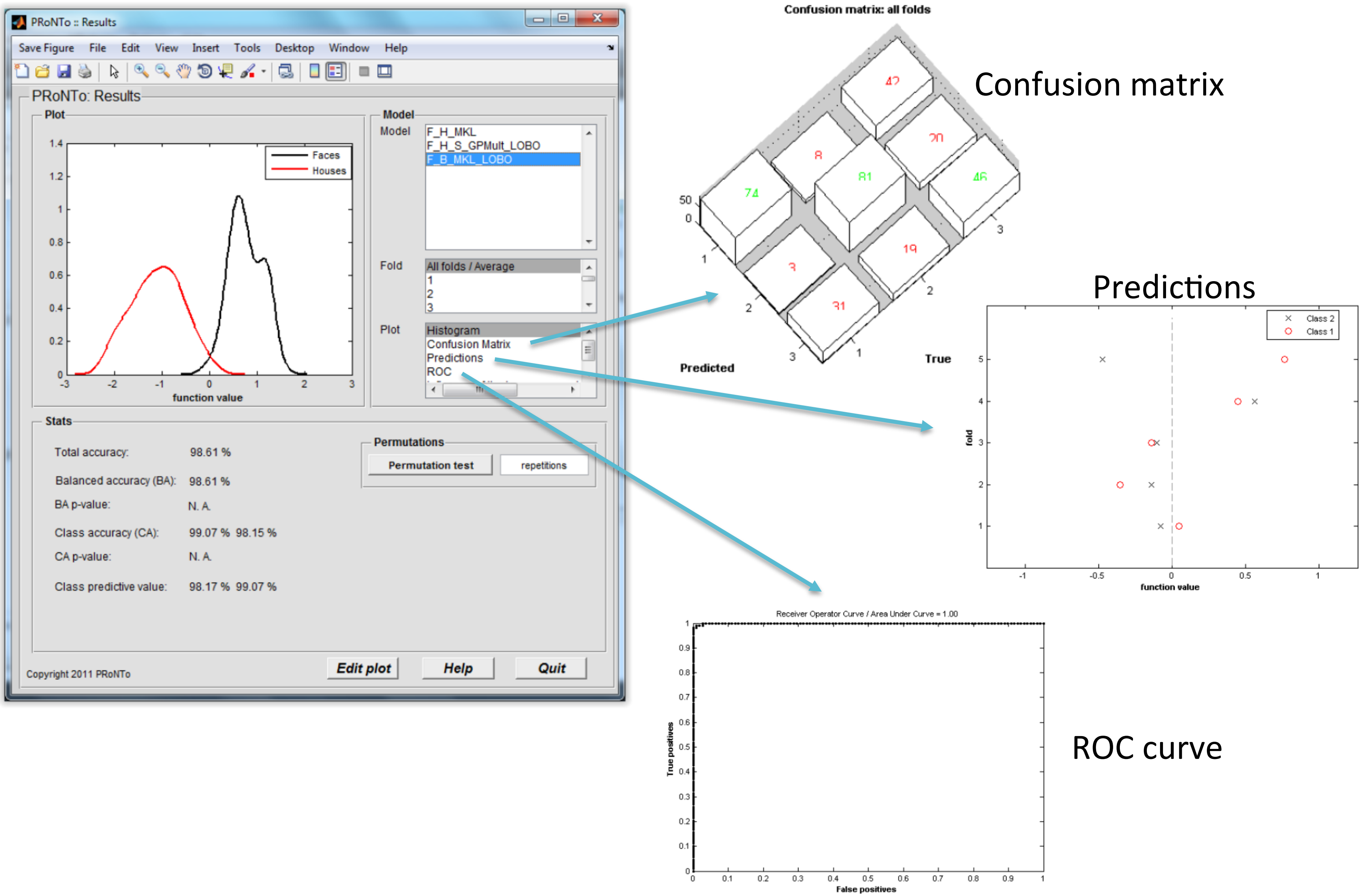
Outline

- Measures of performance for classification
- Measures of performance for regression
- Permutation tests
- Flexible cross-validation
- Nested cross-validation / parameter optimisation

Revision

- Classification: learn a function that predicts *C* discrete categories (**class labels**) from data
- Regression: learn a function that predicts a scalar value (**target**) from data
- Our model is an approximation to “real function”

Measures of performance for classification



Measures of performance for Classification

		Predicted label P	Predicted label N
True label	P	TP	FN
	N	FP	TN

- Accuracy statistics can be shown in a confusion matrix :

Class 1 accuracy = TP/(TP+FN)

Class 2 accuracy = TN/(TN+FP)

Total Accuracy = (TP+TN)/(TP+FP+FN+TN)

Balanced Accuracy (BA) = mean of class accuracy

Class 1 predictive value: TP/(TP+FP)

Class 2 predictive value: TN/(TN+FN)

Perfect: FN=FP=0. Be suspicious if this happens!

Random: TP=TN=FN=FP. Same as flipping a coin.

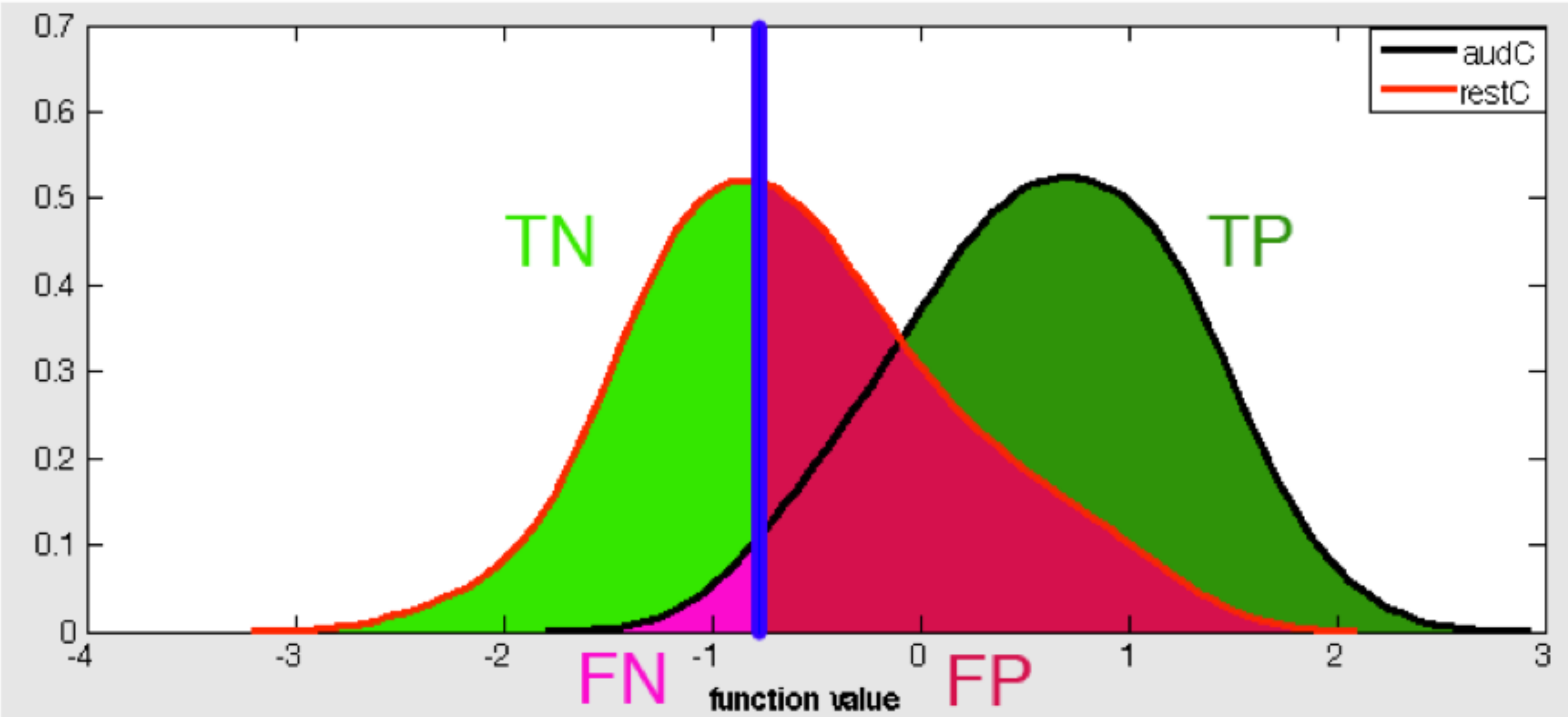
Measures of performance for Classification

Total accuracy vs. balanced accuracy

- If classes don't have the same number of examples
- Accuracy may seem to be above chance whereas the minority classes are sacrificed and below chance
- A common strategy is to subsample the majority class, but data is lost
- Subsample many times (computationally intensive)
- Reporting class accuracies (p_0,..., p_C) is good practice
- Balanced accuracy is the average of class accuracies

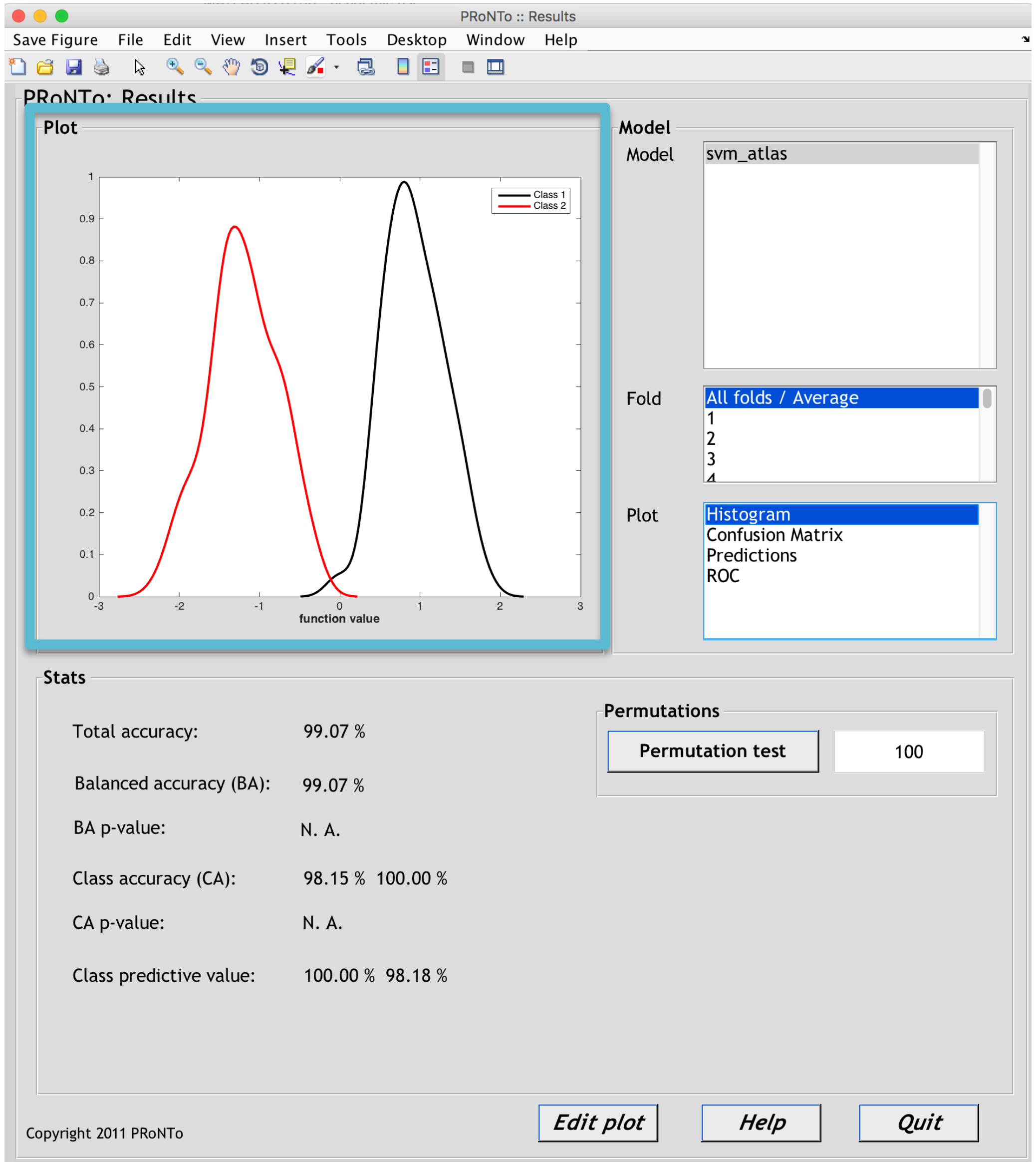
Measures of performance for Classification

For a fixed classifier, increasing sensitivity can only come at the cost of decreasing specificity, and vice-versa.



Measures of performance for Classification

Histograms



Measures of performance for Classification

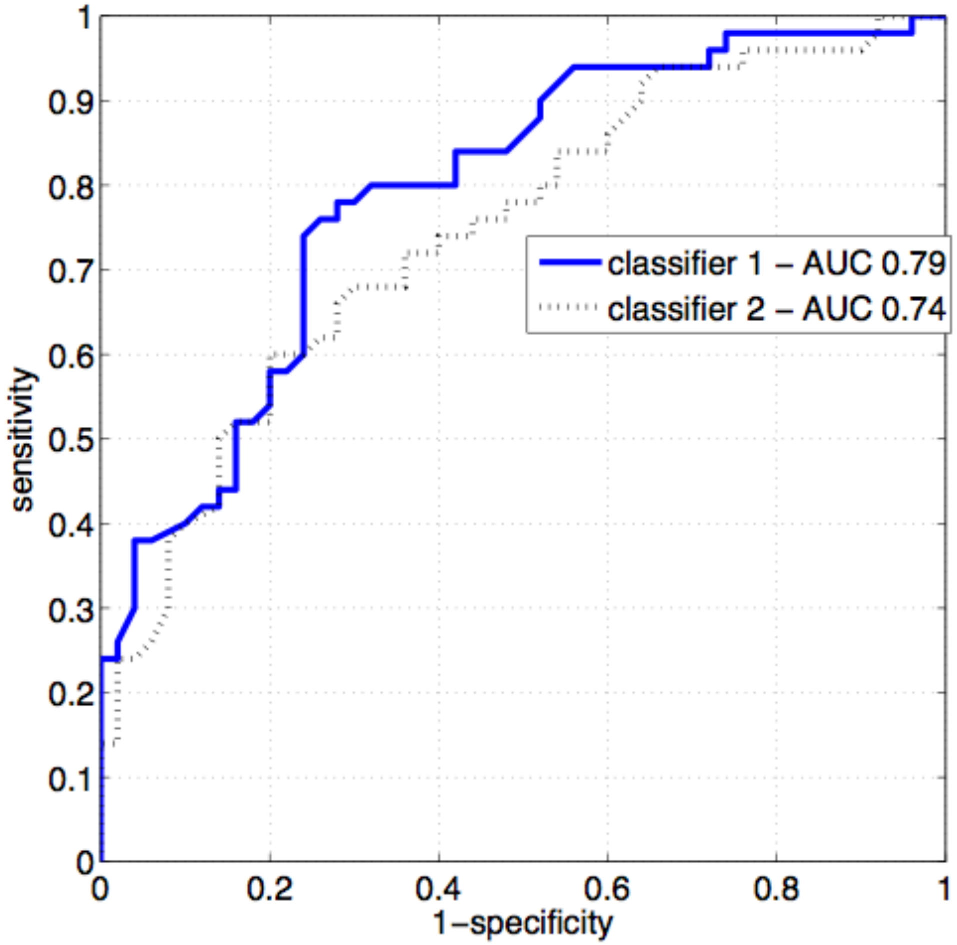
The **Receiver Operating Characteristic (ROC) curve** is a good way of seeing the sensitivity/specificity tradeoff over the operating range of a classifier.

It is also used for classifier comparison

We can compute the **Area Under Curve (AUC)** as a summary measure of performance

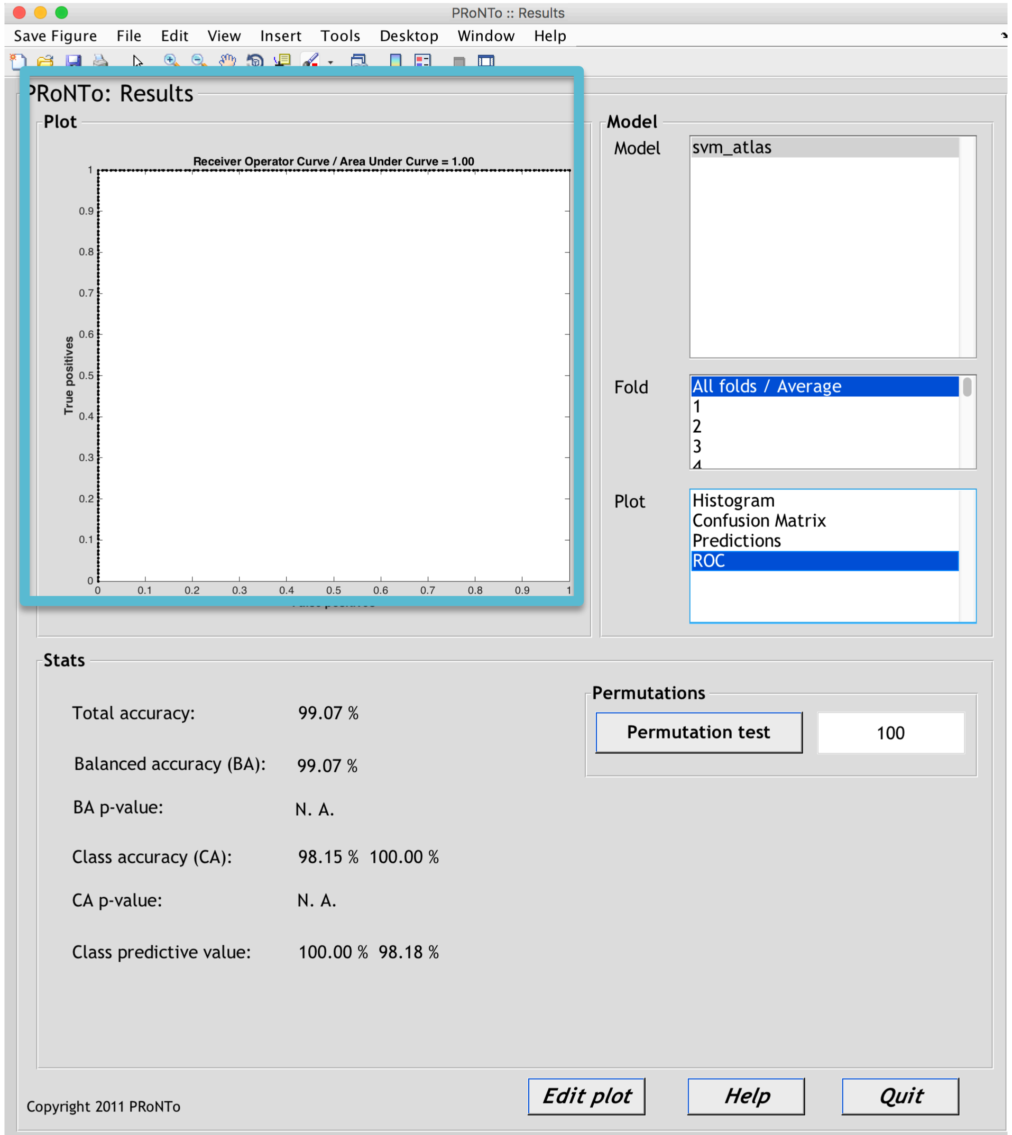
AUC = 1.0: perfect

AUC = 0.5: chance

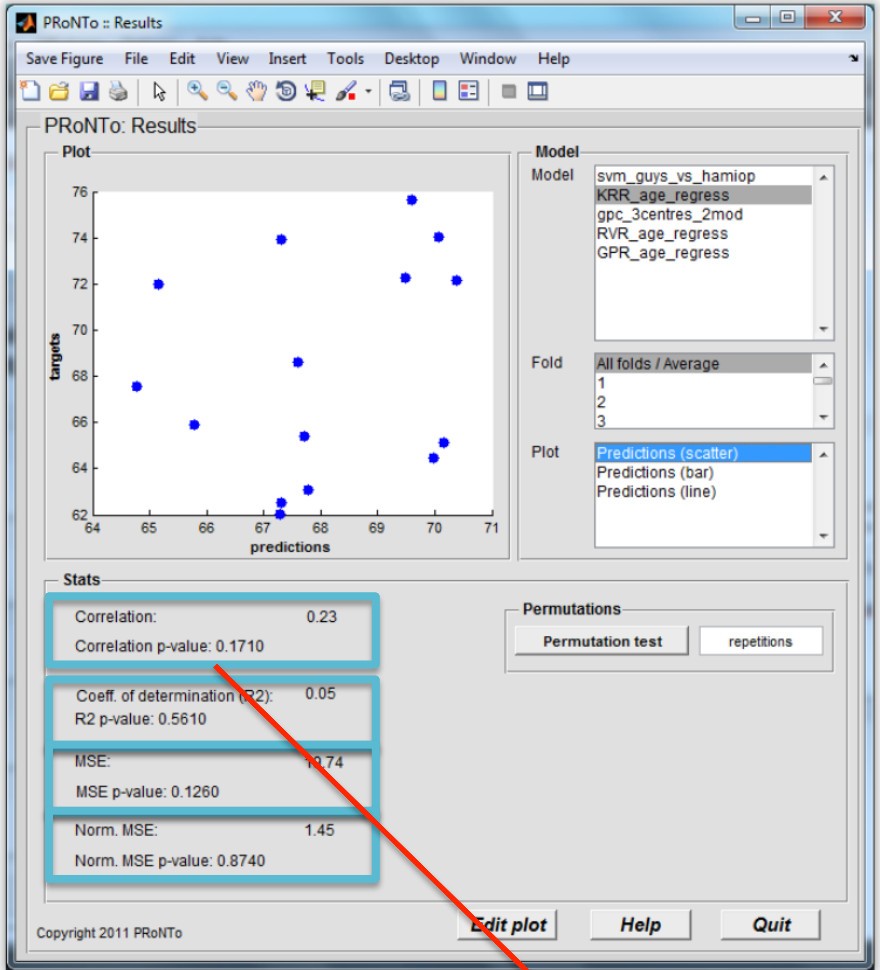


Measures of performance for Classification

ROC curve



Measures of performance for regression



Correlation

$$CORR = \frac{\sum_n (y_n - \mu_y)(f(\boldsymbol{x}_n) - \mu}{\{\sum_n (y_n - \mu_y)^2 \sum_n (f(\boldsymbol{x}_n) -}$$

Coefficient of determination

$$R^2 = CORR^2$$

Mean Squared Error (MSE)

$$MSE = \frac{1}{N} \sum_n (y_n - f(\mathbf{x}_n))^2$$

Normalised MSE

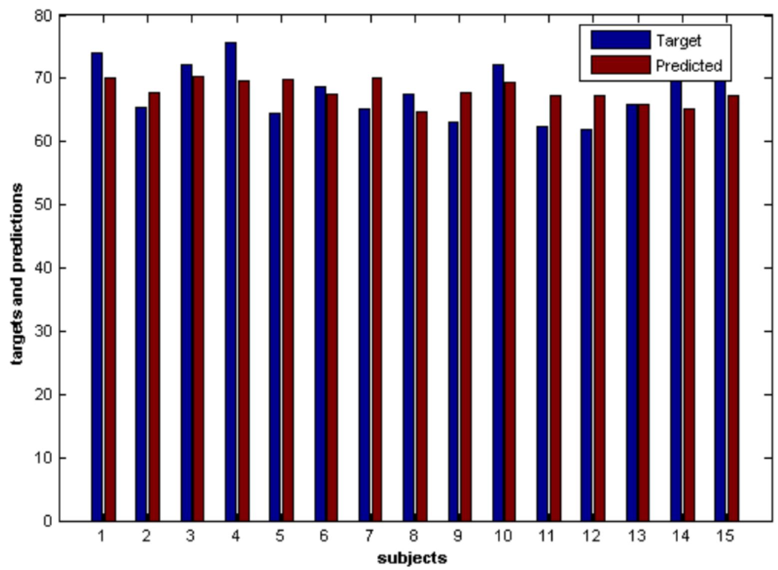
$$Norm.MSE = \frac{MSE}{(y_{max} - y_{min})}$$

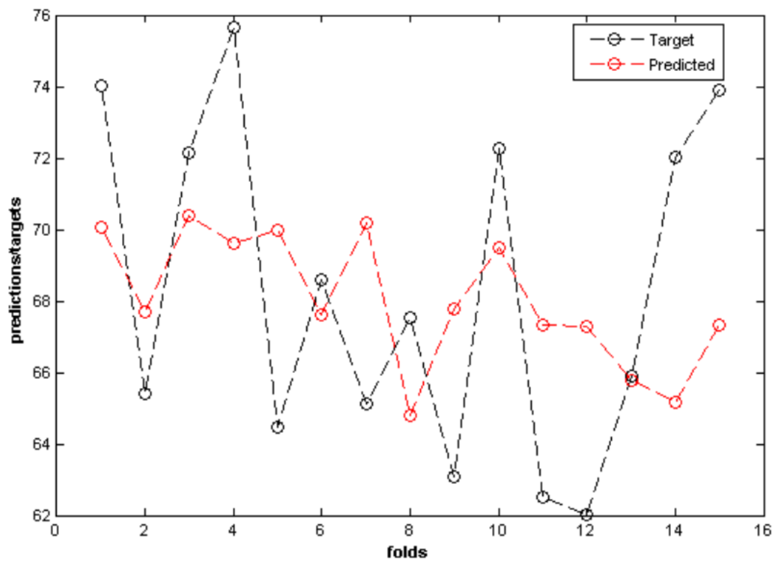
Negative correlation between the real and predicted labels are meaningless!

Measures of performance for regression

Regression plots:

- Scatter plot
- Predictions (bar)
- Predictions (line)





Confidence intervals

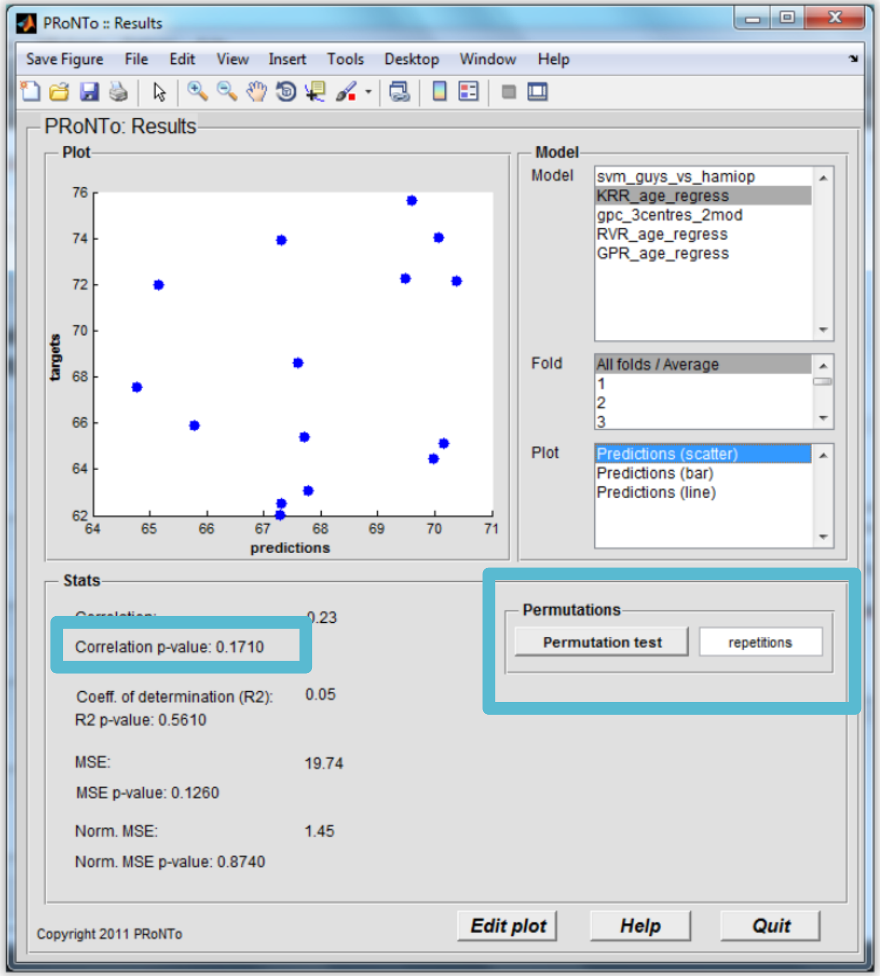
Parametric tests

- Binomial test
  - Model decision in two-class problem modeled as Bernoulli trials
  - Probability of *k* successes out of *n* trials follows binomial distribution

Not a good idea:

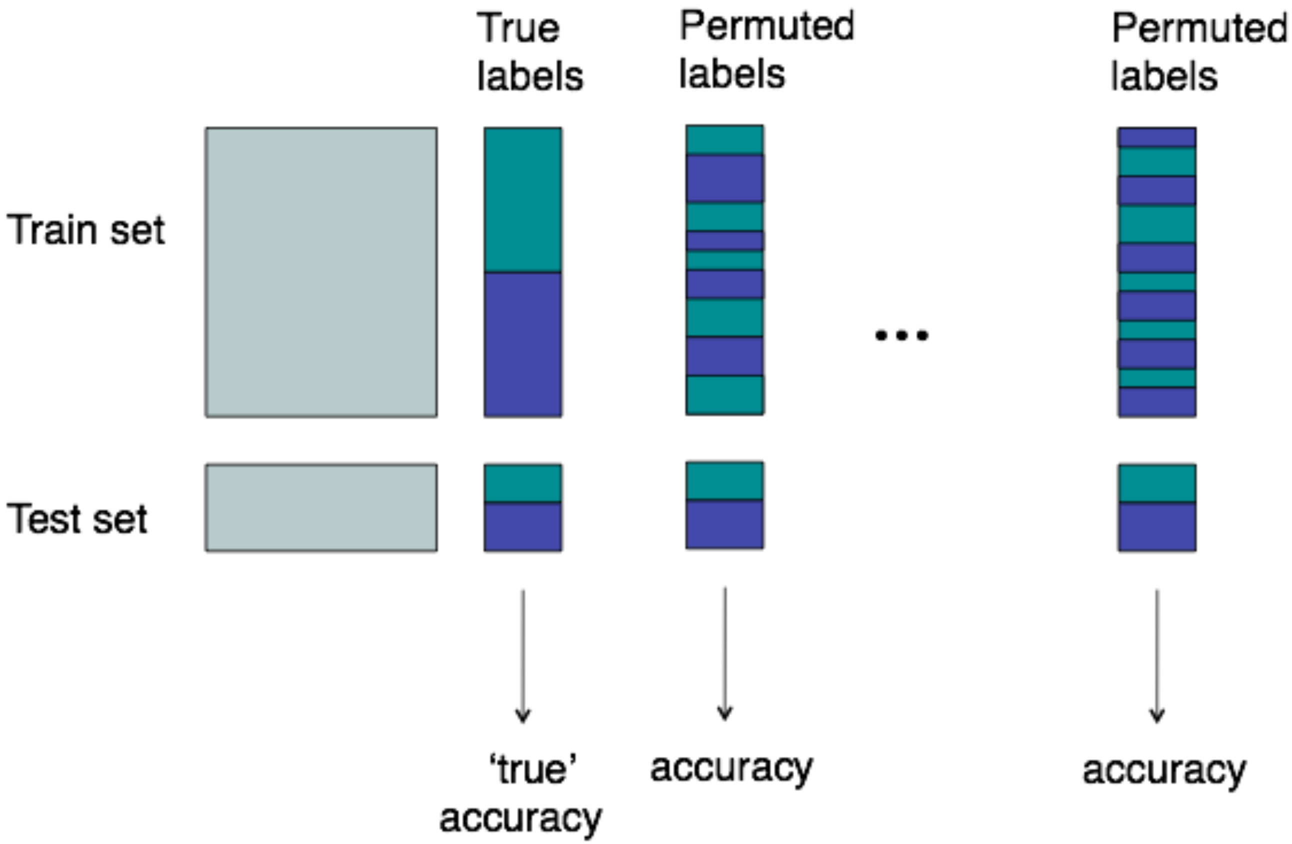
- Assumes IID samples
- Accuracy from cross-validated random data does not follow the binomial distribution (Noirhomme et al. 2014)

Permutation tests



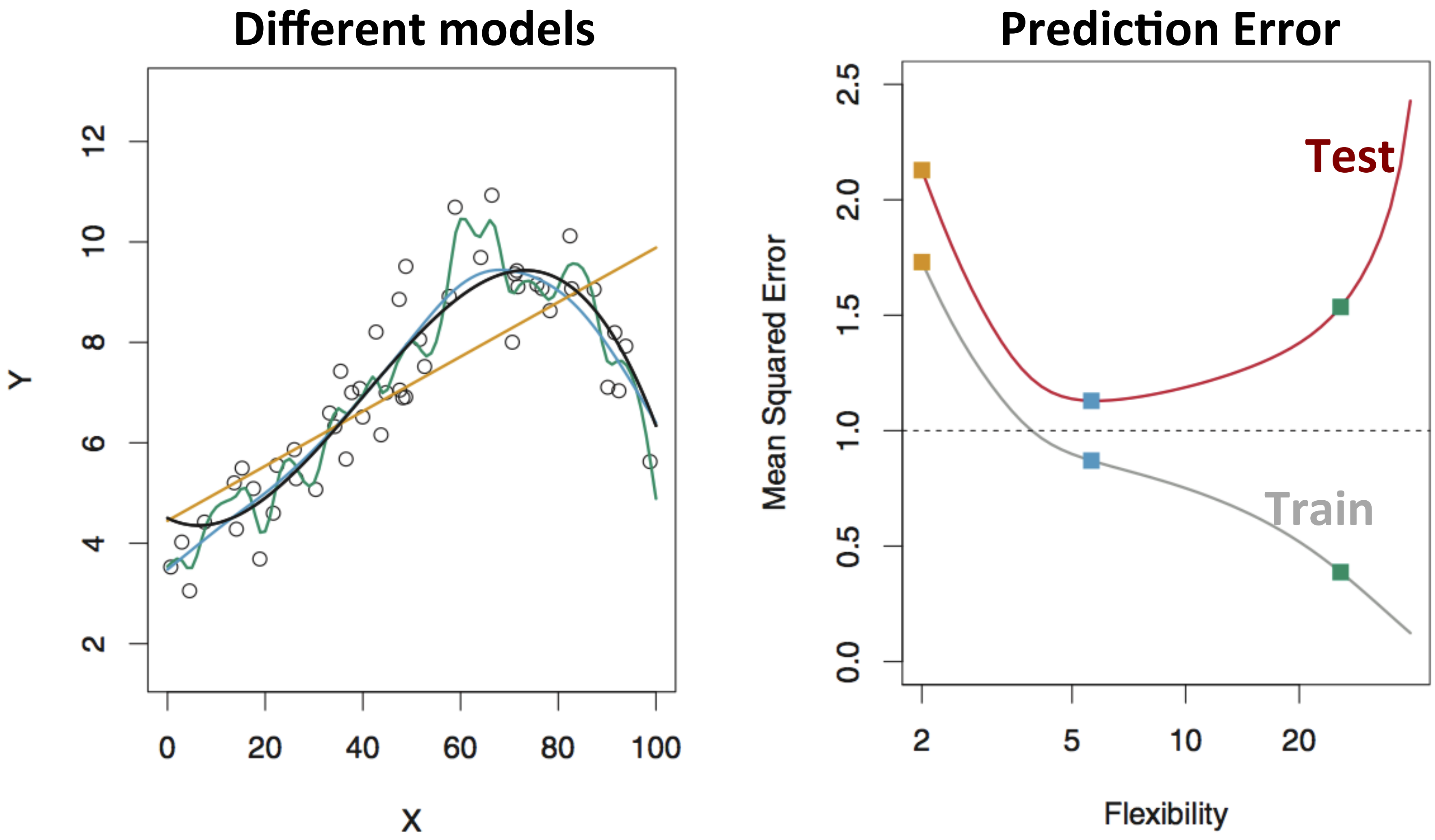
- No hypotheses on data distribution
- H_0: “targets are non-informative”
- Test statistic: CV accuracy / MSE / R^2
- Estimate the distribution of the test statistic under H_0 by randomly permuting targets M times, and running the full CV experiment

Permutation tests

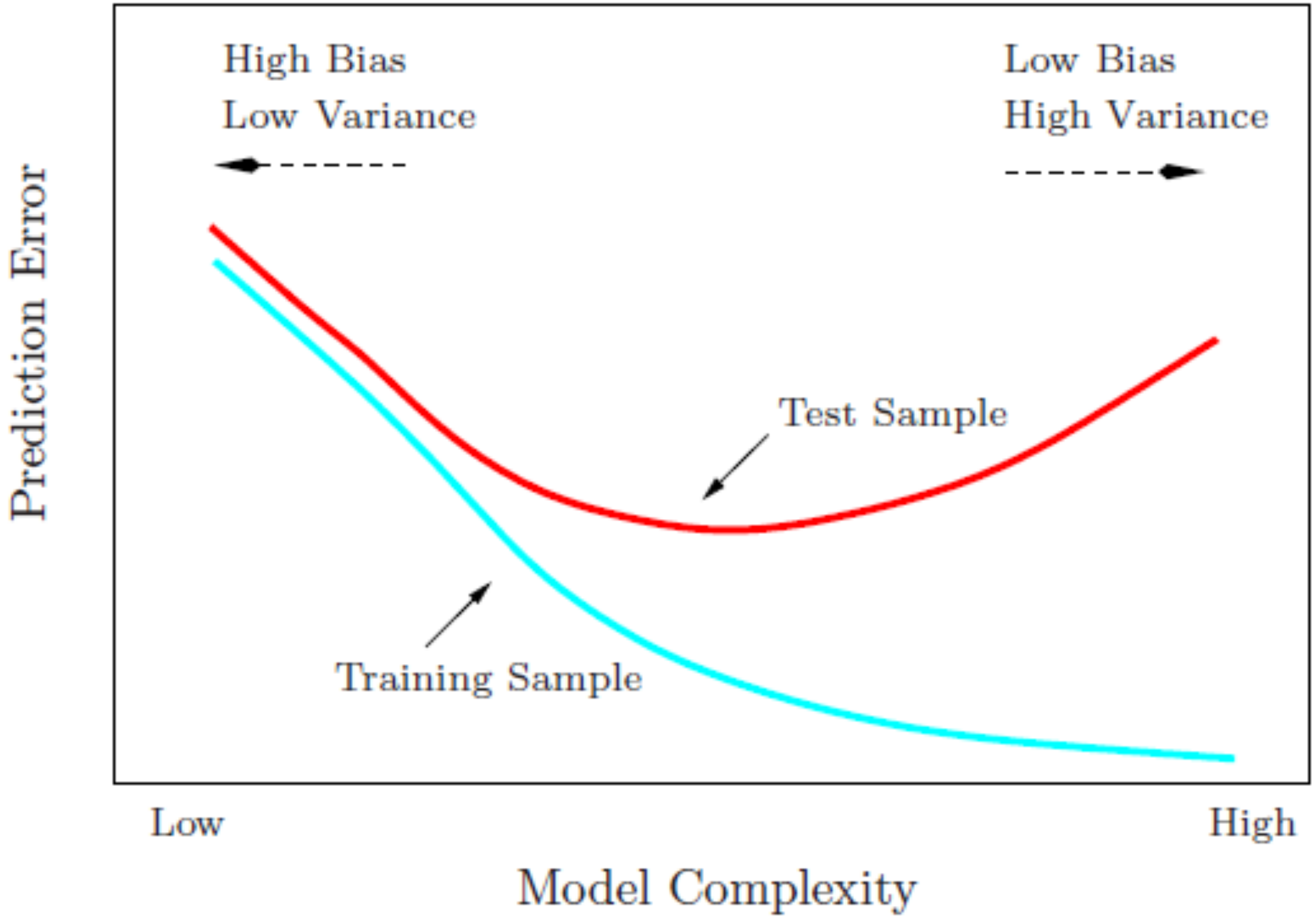


P-value: $$\frac{1}{M}\sum_{m}^{M}(p_m^{perm} \geq p^{real})$$

Train and test error



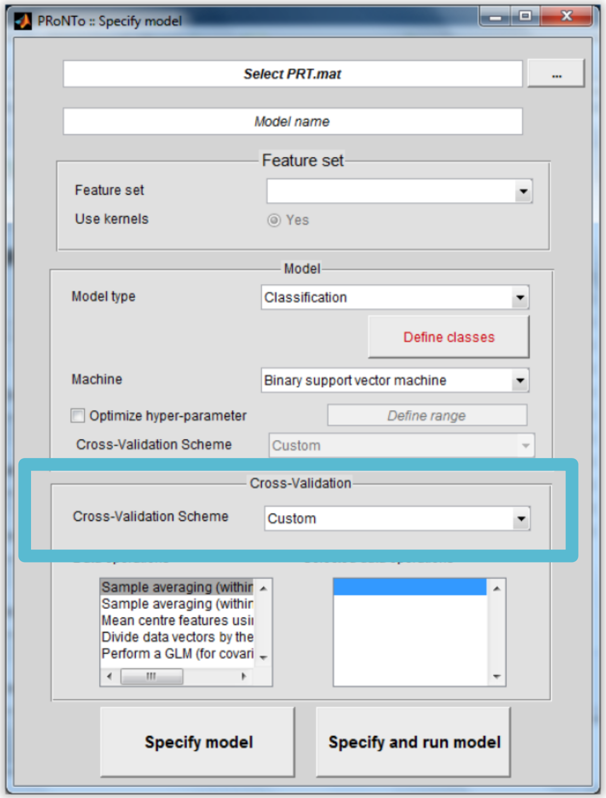
Bias-Variance tradeoff



Cross-validation (CV)

- Allows us to estimate test error of the model using available data
- Partition data into training and testing sets

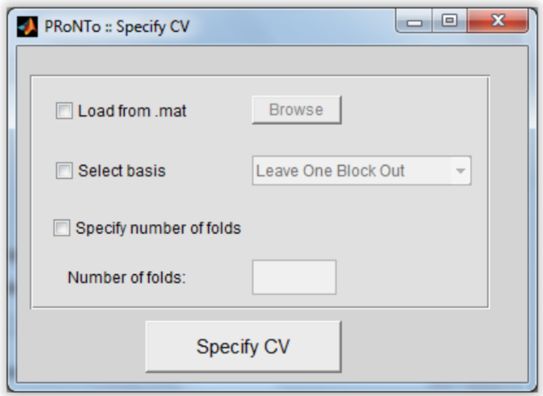
In PRoNTo:



Flexible CV schemes allowed

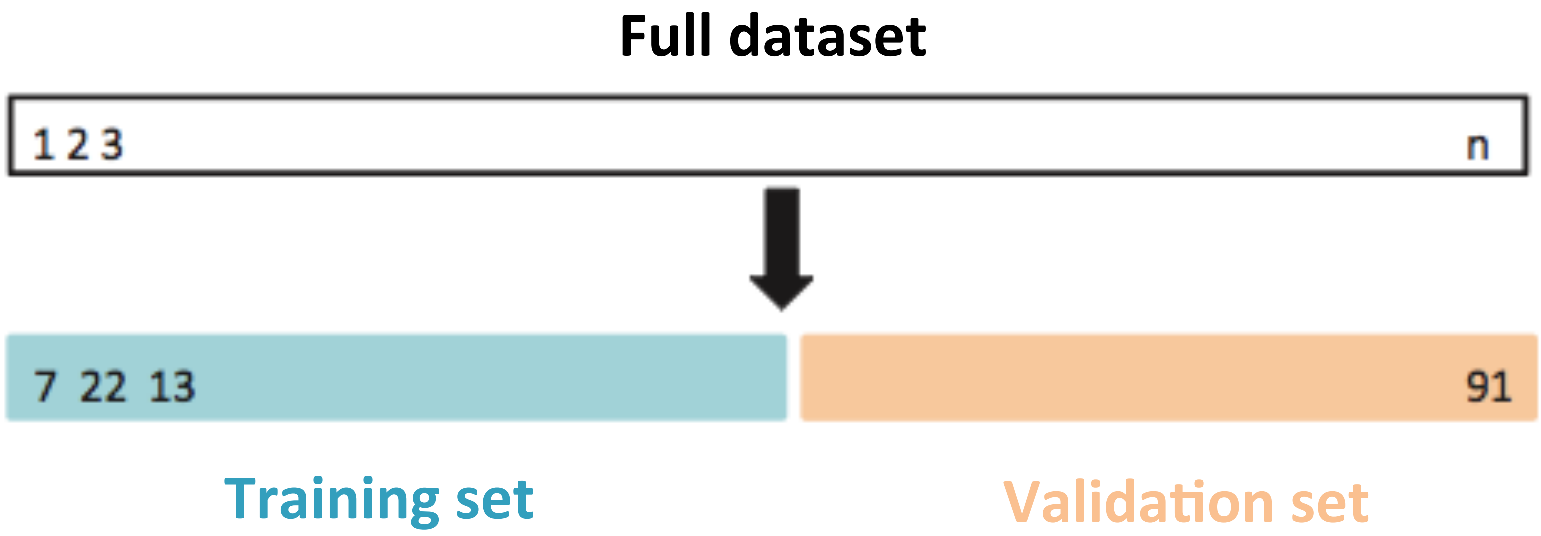
Standard approaches:

- LOSO
- LOBO
- LORO
- LOSGO
- k-fold CV



Cross-validation (CV)

- Validation set approach

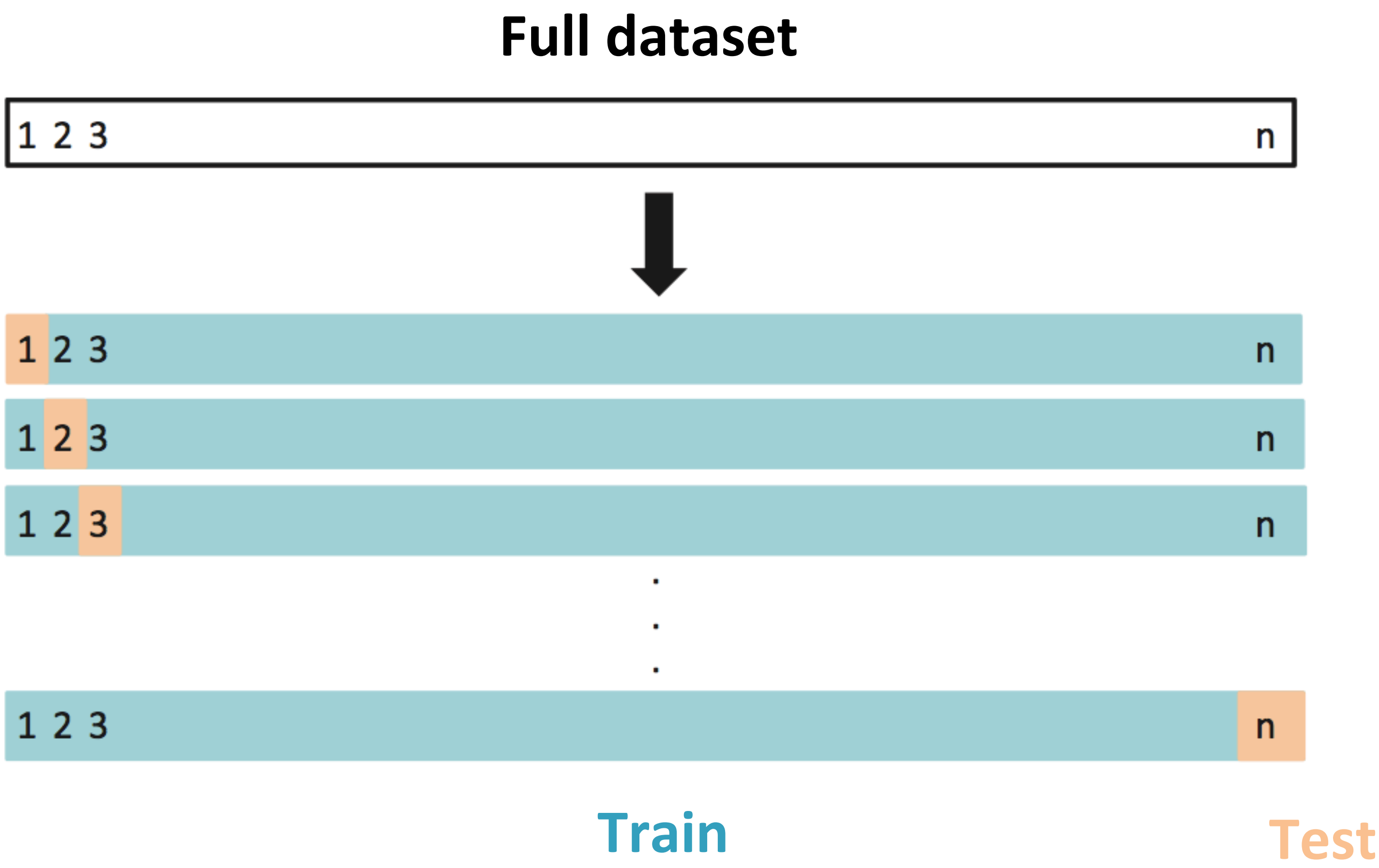


Drawbacks:

- Uses few observations and tends to overestimate the test error
- Test error estimates are highly variable

Cross-validation

Leave-one-out (LOO)



Leave each sample out for testing and use the rest for training. Repeat n times.

Cross-validation (CV)

LOO-CV

Main advantages:

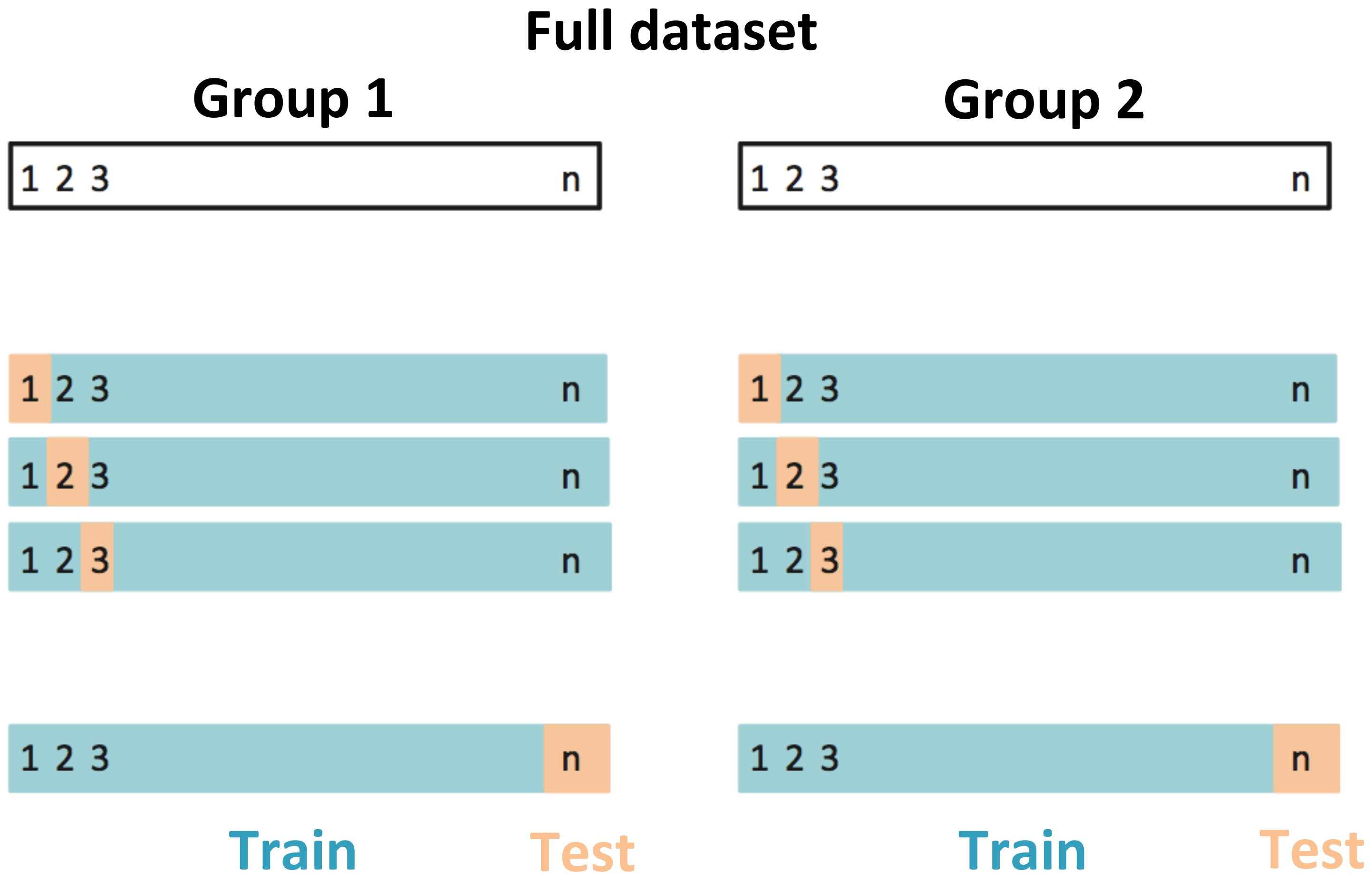
- Better use of data than half-split approach for small sample-sized data
- Almost unbiased test error estimate

Main disadvantages:

- Computationally intensive
- Test error estimate has high variance

Cross-validation

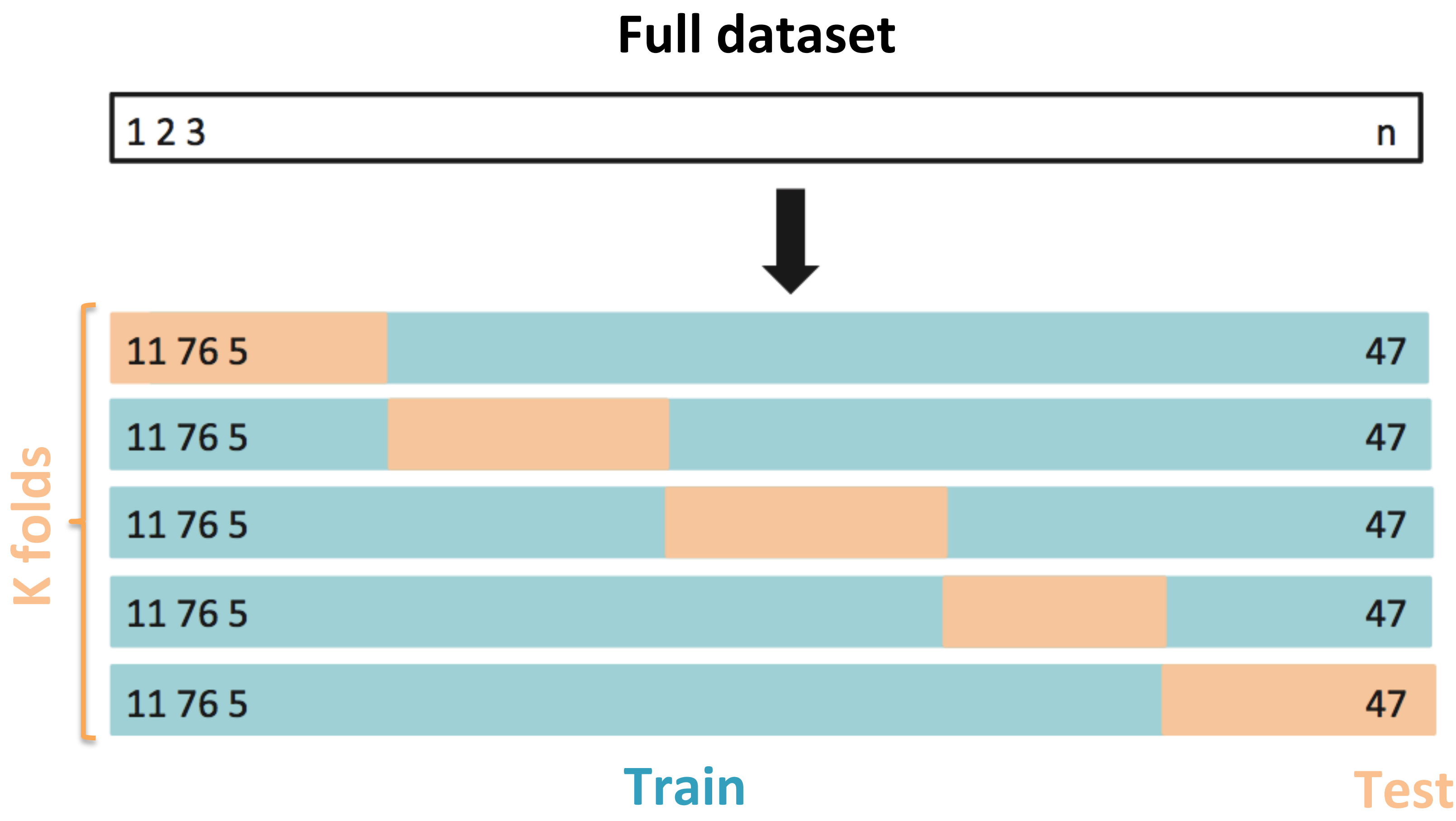
Leave-one-sample-per-group-out (LOSGO)



If subjects/samples in each group are paired (e.g. repeated measures)

Cross-validation

K-fold cross validation



Cross-validation (CV)

- K-fold CV

Main advantages:

- Test error estimate has less variance than LOO-CV
- Computationally less intensive

Main disadvantages:

- Higher bias of test error estimate than LOO-CV

Common *k* choices: 5 and 10

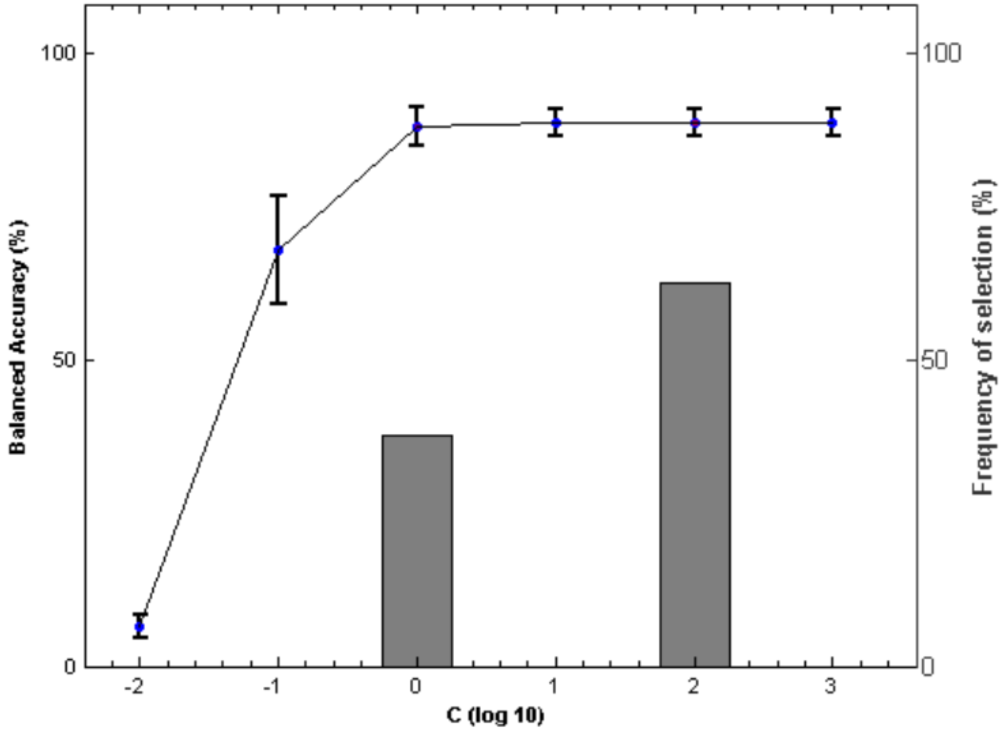
Nested cross-validation

- Problem: use CV to select best model and assess model performance (test error)
- Solution:

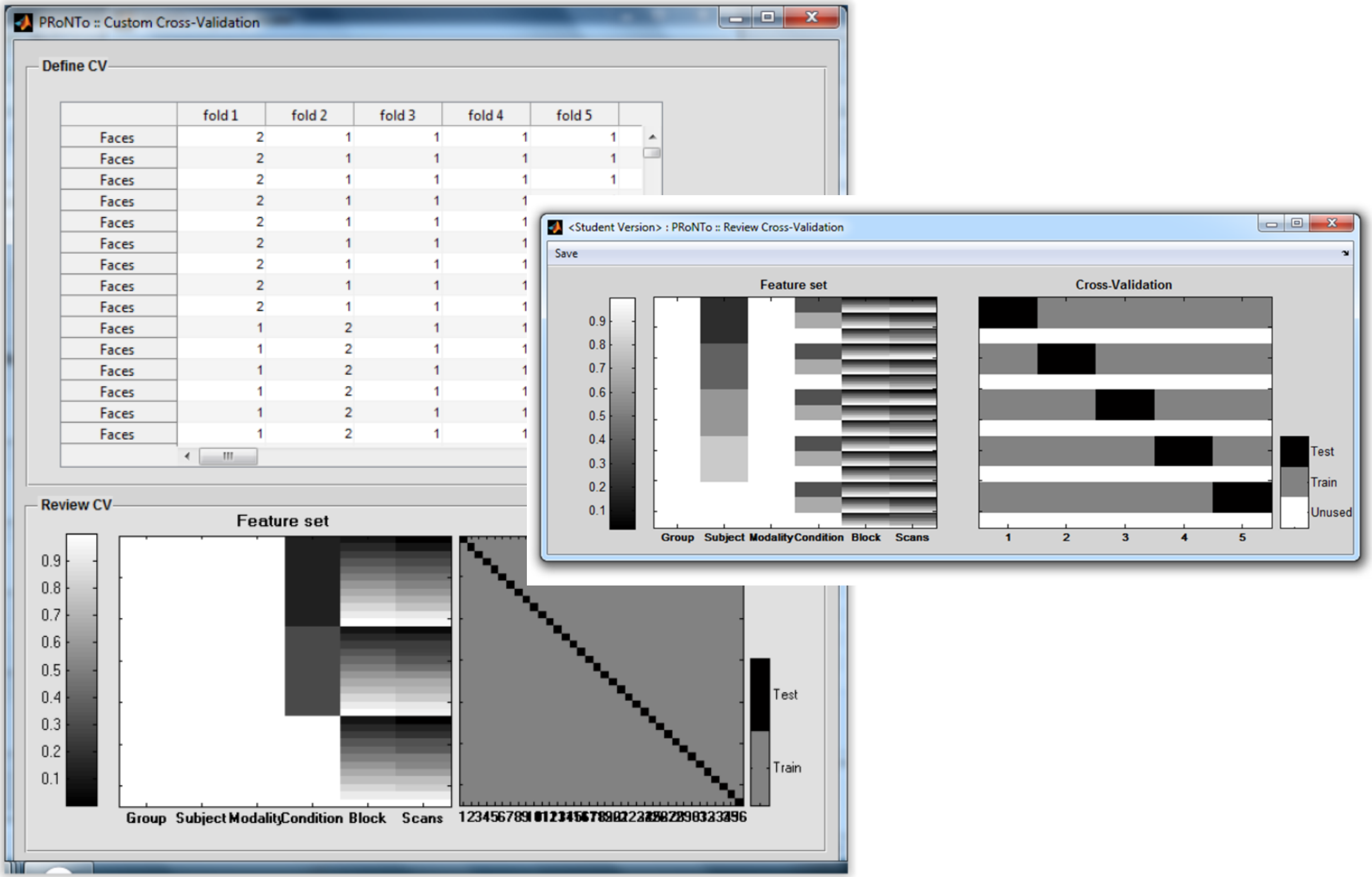
  Run CV inside CV for model or feature selection / Bayesian Models

Model selection

If hyper-parameter optimisation was performed using nested CV:



Cross-validation matrix



Take-home messages

- Always separate data intro training and testing sets
- Use cross-validation
- Be careful with correlated data (e.g. fMRI)
- Use nested cross-validation for model or feature selection
- Use permutation tests to assess significance of performance measure

Recommended reading

- James et al., *Introduction to Statistical Learning,* Springer, 2014.
- Duda et al., *Pattern Recognition*, Wiley, 2001.
- Hastie et al., *The elements of statistical learning*, Springer, 2009.
- Pereira et al., *Machine learning classifiers and fMRI: A tutorial overview*, NeuroImage 45, 2009.
- Kriegeskorte et al., *Circular analysis in systems neuroscience: the dangers of double dipping*, Nature Neuroscience 12, 2009.
- Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, IJCAI, 1995.