Pattern Recognition: validation, inference and model interpretation

Jessica Schrouff

Course 2018

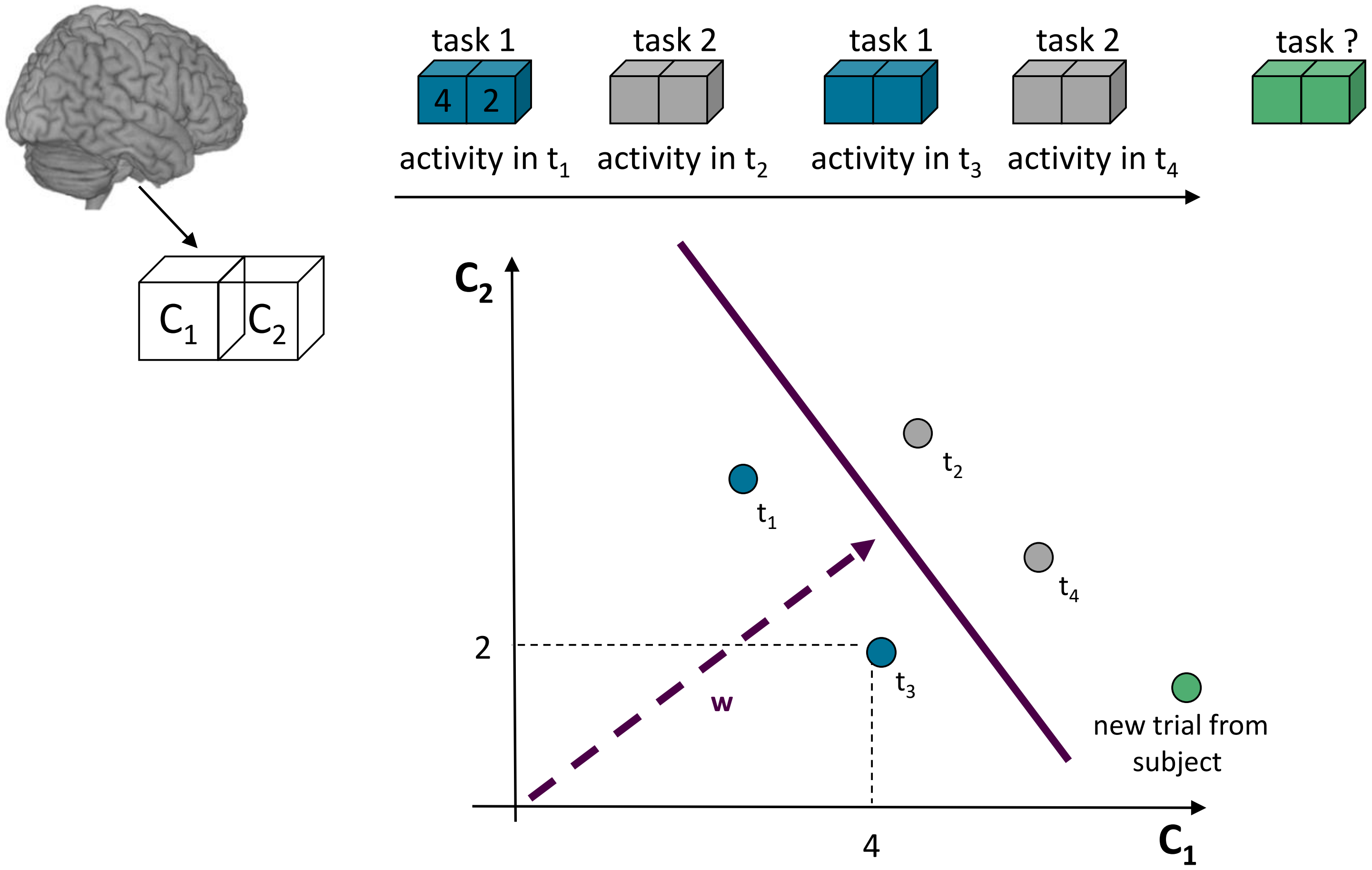
May 14^{th}- 15^{th}

UCL, London

Outline

- Is my model good?
  - ➢Measures of performance for classification
  - ➢Measures of performance for regression
  - ➢Validation set and cross-validation
  - ➢Nested cross-validation
  - ➢Assessing significance
- What does my model look like?
  - ➢Model interpretation

Classification: reminder



Train model on $t_1, \ldots, t_4$:

$X = (c_1,c_2)_{t1\text{-}4}$; y = task 1/2

Test on $t_1, \ldots, t_4$:

$X^* = (c_1,c_2)_{t1\text{-}4}$

Classification: confusion matrix

Accuracy statistics can be shown in a confusion matrix:

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

Class 1 (P) accuracy, sensitivity = TP/(TP+FN)

Class 2 (N) accuracy, specificity = TN/(FP+TN)

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

Balanced Accuracy (BA) = mean of classes accuracy

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

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

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

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

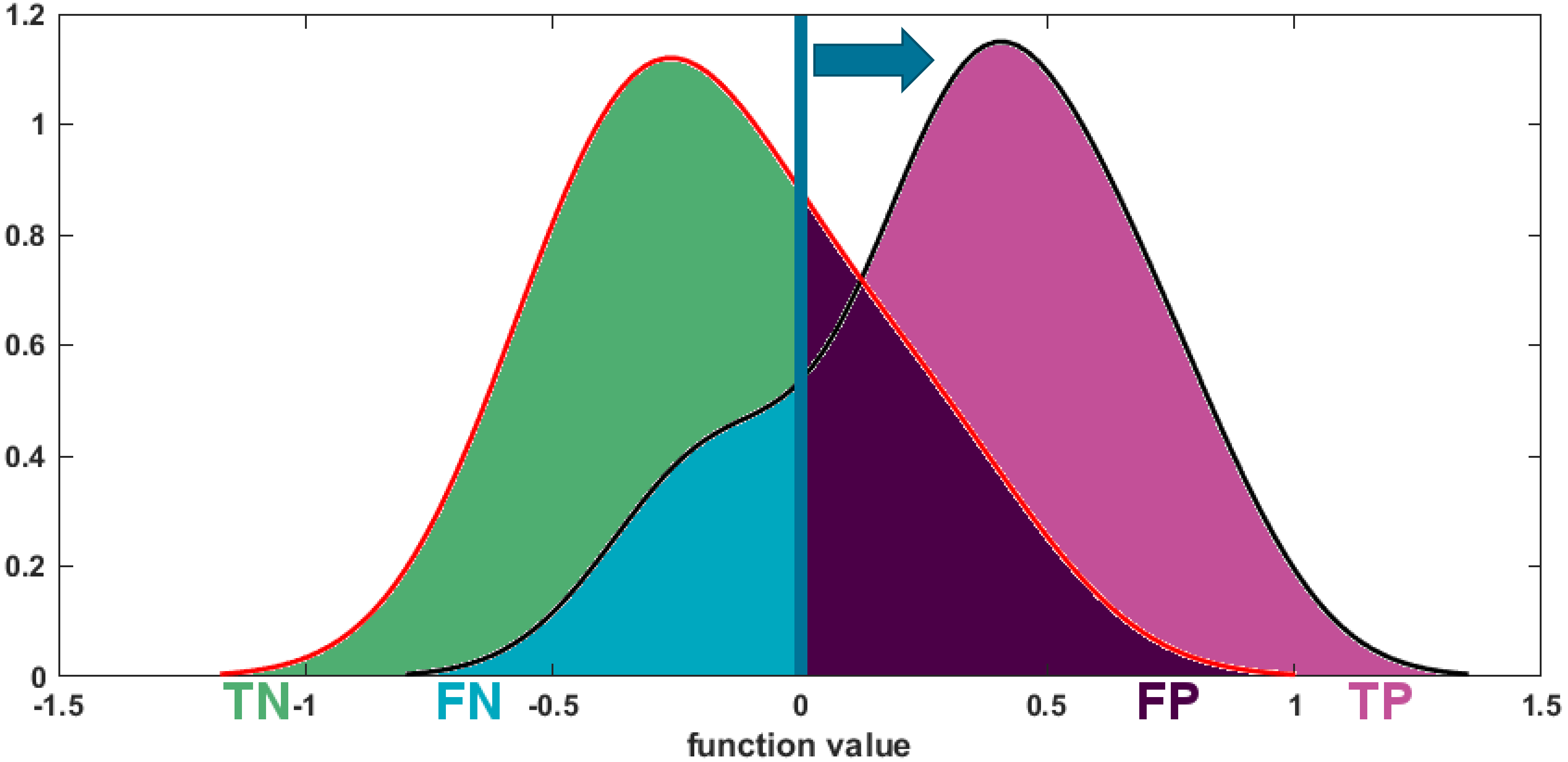
Classification: accuracy

Total accuracy vs. balanced accuracy

- If classes don’t have the same number of examples
- Total 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

Classification: ROC

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

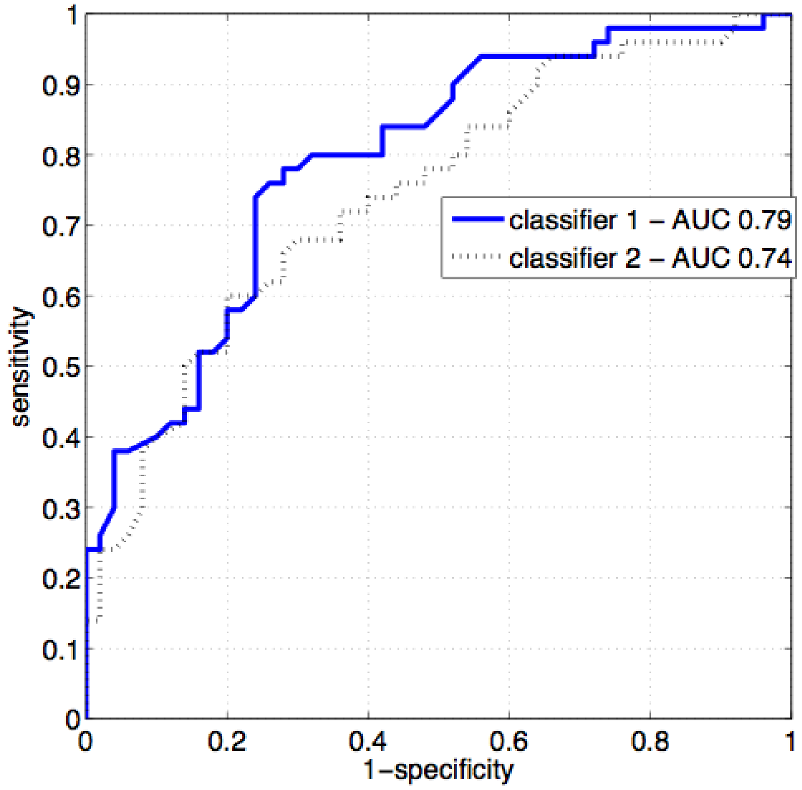


Classification: ROC

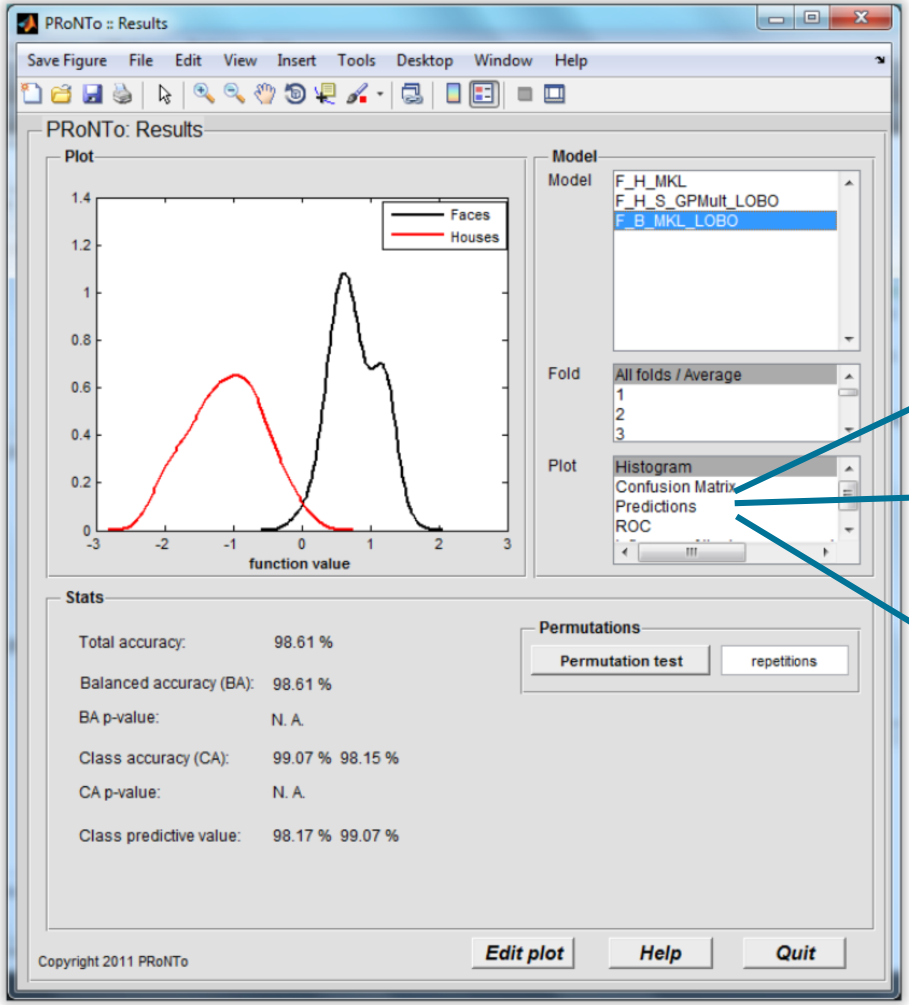
The **Receiver Operating Characteristic (ROC) curve** is a good way of seeing the sensitivity/specificity tradeoff over the operating range of a classifier.

Classifier comparison via **Area Under Curve (AUC)**

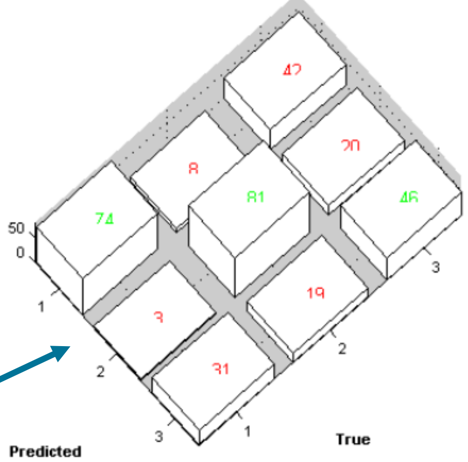
AUC = 1.0: perfect
AUC = 0.5: chance



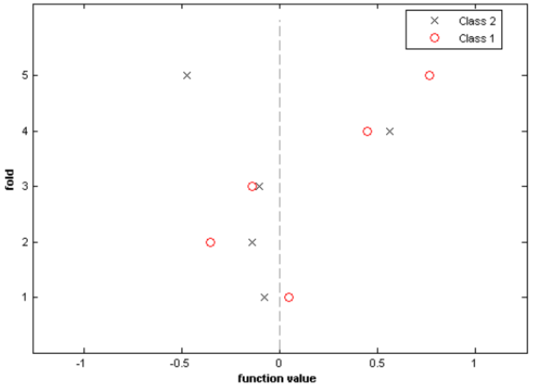
Classification: PRoNTo



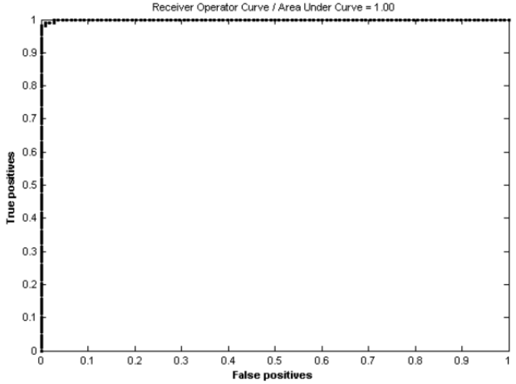
Confusion matrix



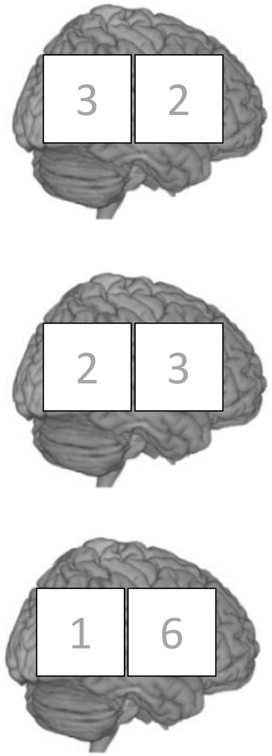
Predictions

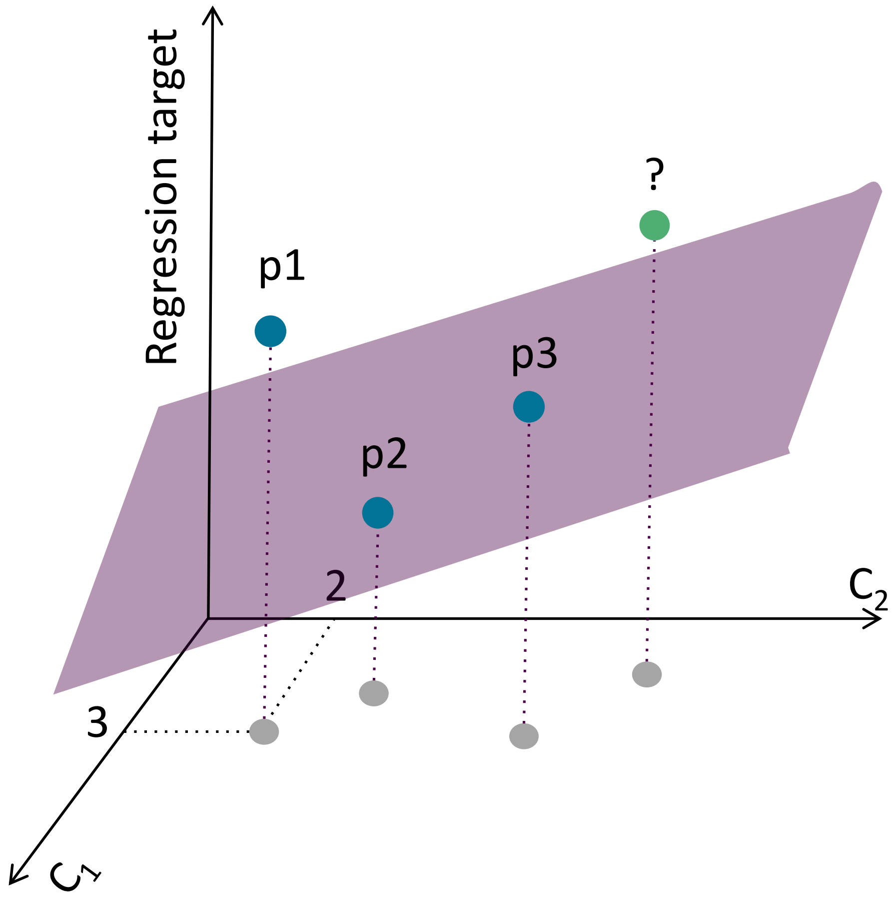


ROC curve



Regression: reminder

Pattern of brain activation	Target
	p1
	p2
	?



Regression: performance

- Correlation:

$$\mathrm{corr}\big(y, f(x)\big) = \frac{\sum_n (y_n - \mu_y)(f(x_n) - \mu_f)}{\sqrt{\sum_n (y_n - \mu_y)^2 \sum_n (f(x_n) - \mu_f)^2}}$$

- Coefficient of determination:

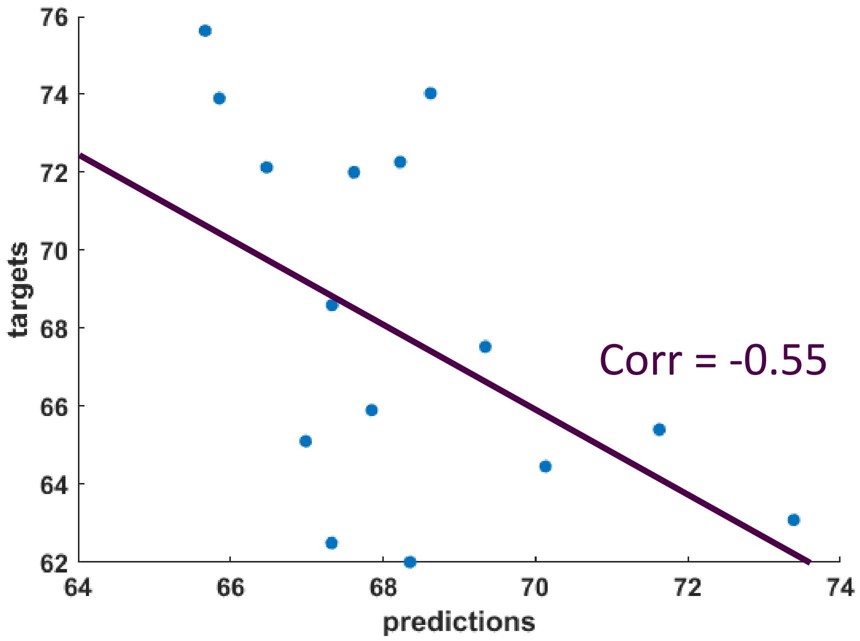
$$R^2 = \mathrm{corr}(y, f(x))^2$$

- Mean Squared Error:

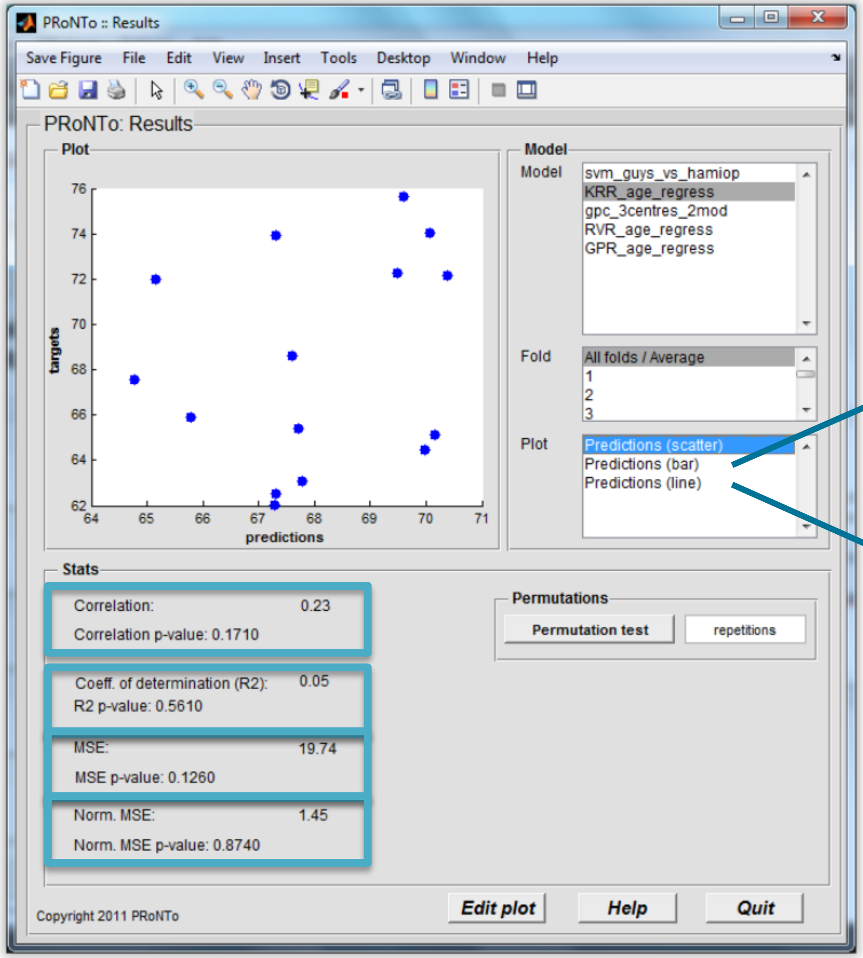
$$\mathrm{MSE} = \frac{1}{N} \sum_n (y_n - f(x_n))^2$$

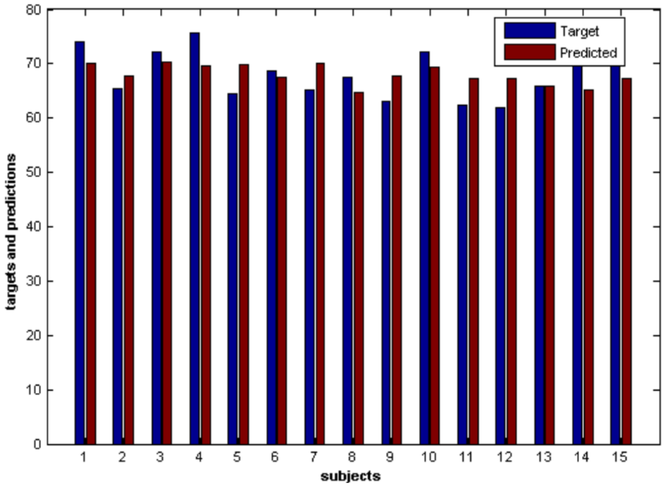
- Normalized MSE:

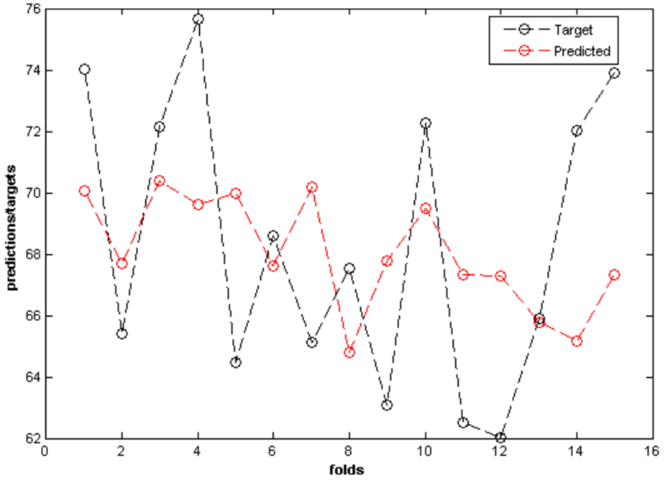
$$\mathrm{NMSE} = \mathrm{MSE}/(y_{max} - y_{min})$$



Regression performance in PRoNTo

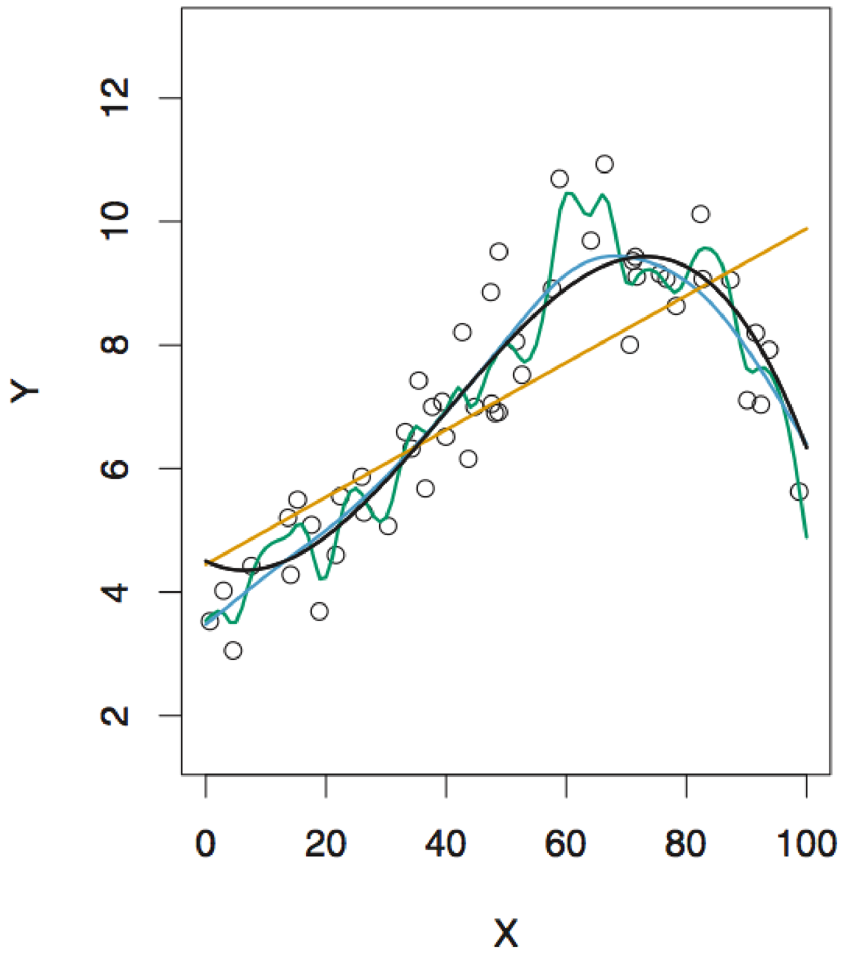




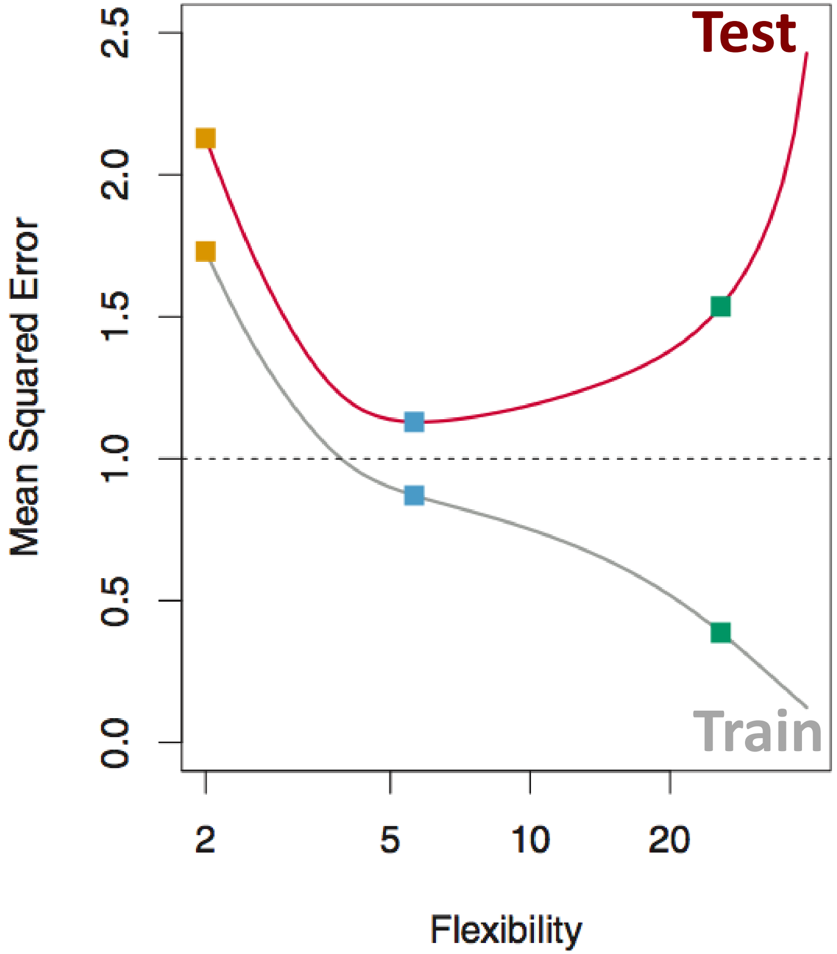


Train and test error

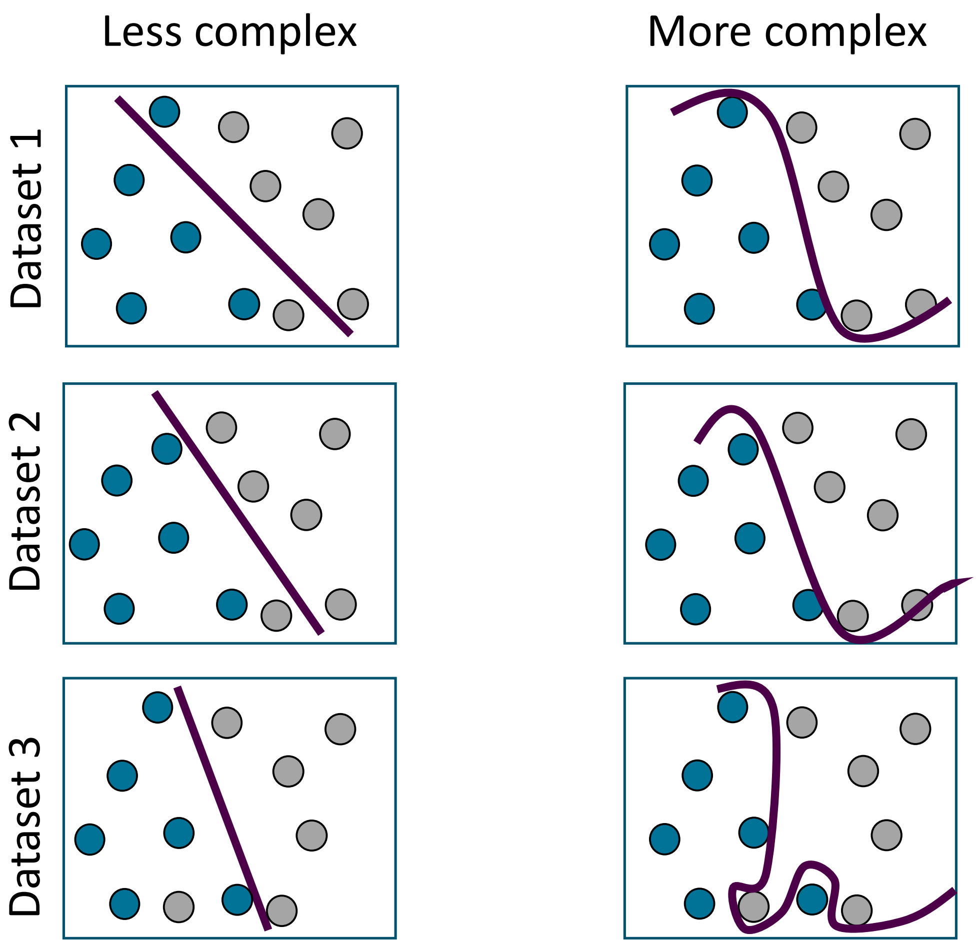
Different models

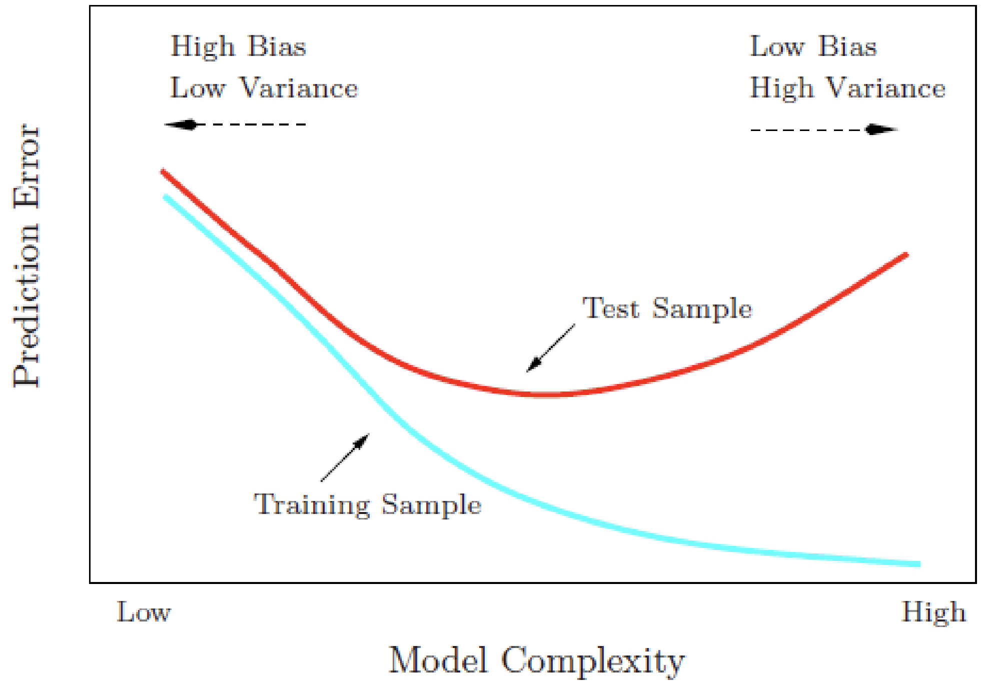


Prediction Error



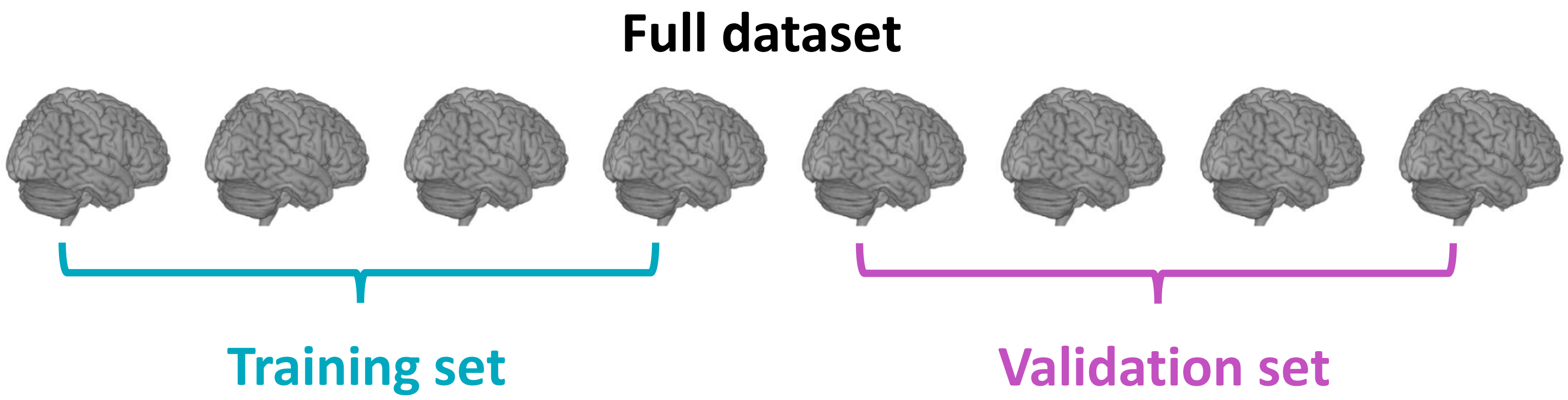
Bias-variance trade-off





Variance: variations in decision functions when the data set is modified (over-fitting)
Bias: error caused by model assumption (under-fitting)

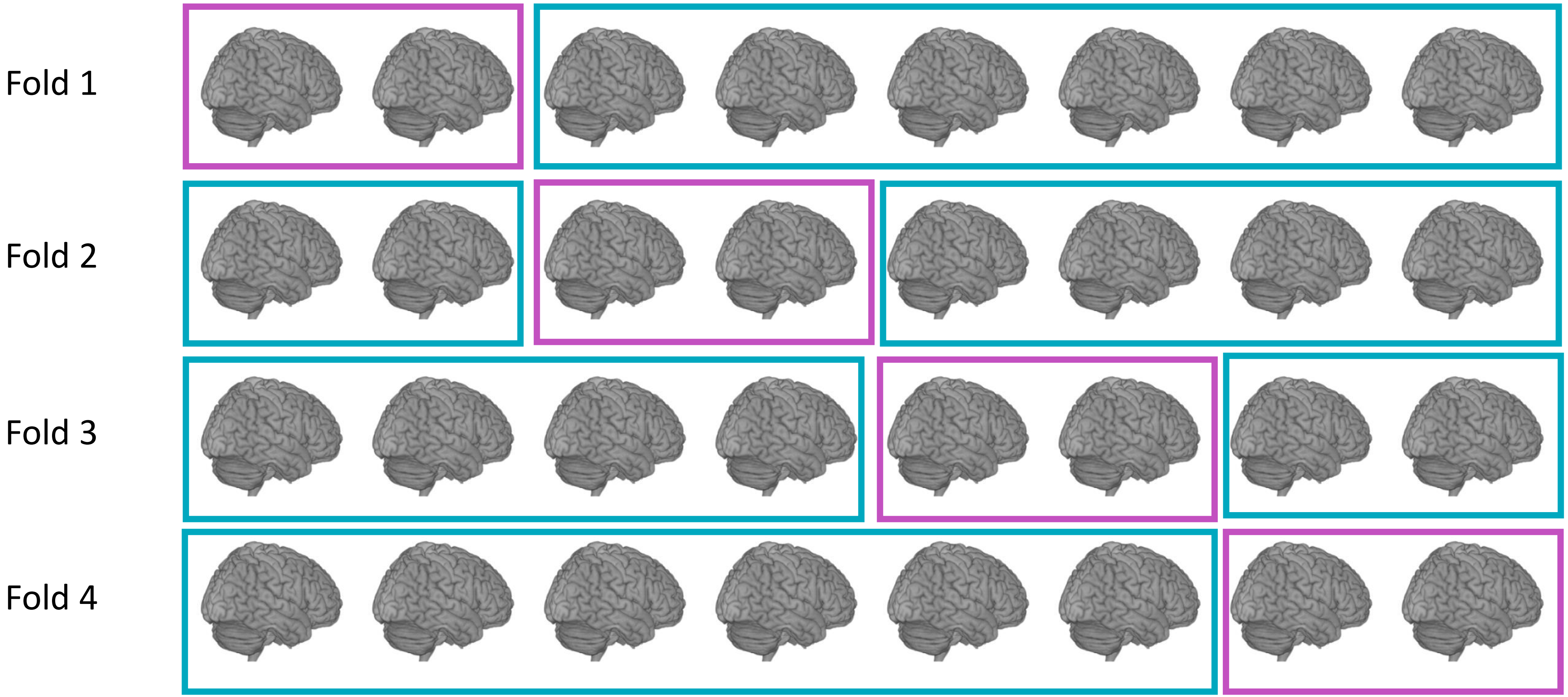
Validation: validation set



Drawbacks:

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

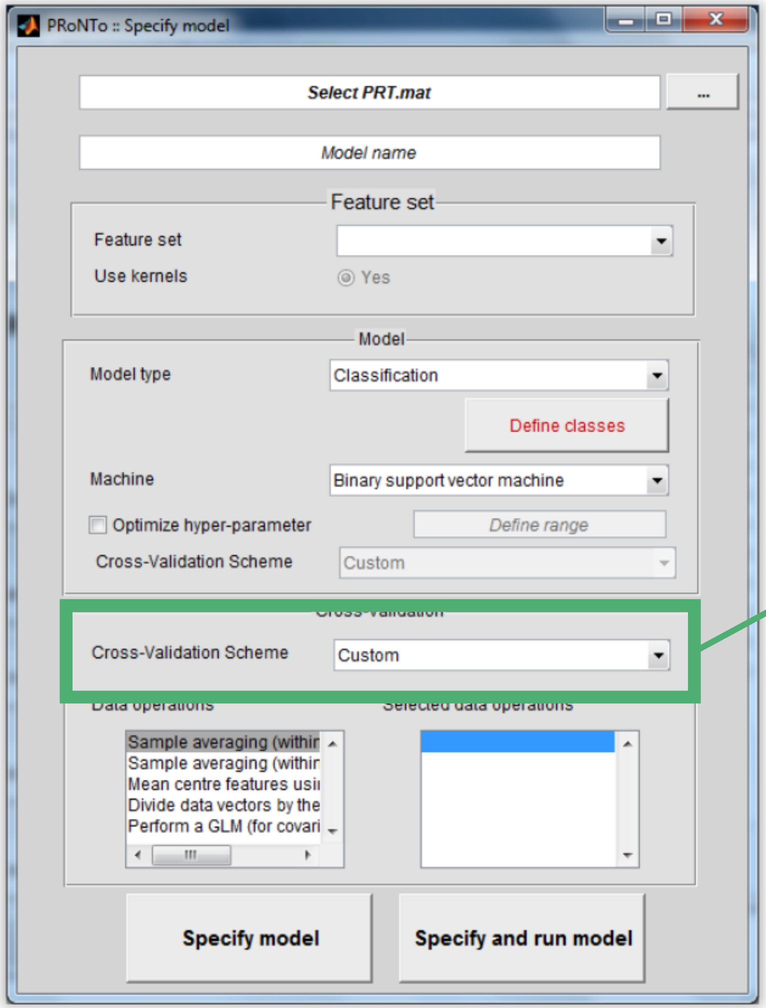
Validation: cross-validation



Validation: cross-validation

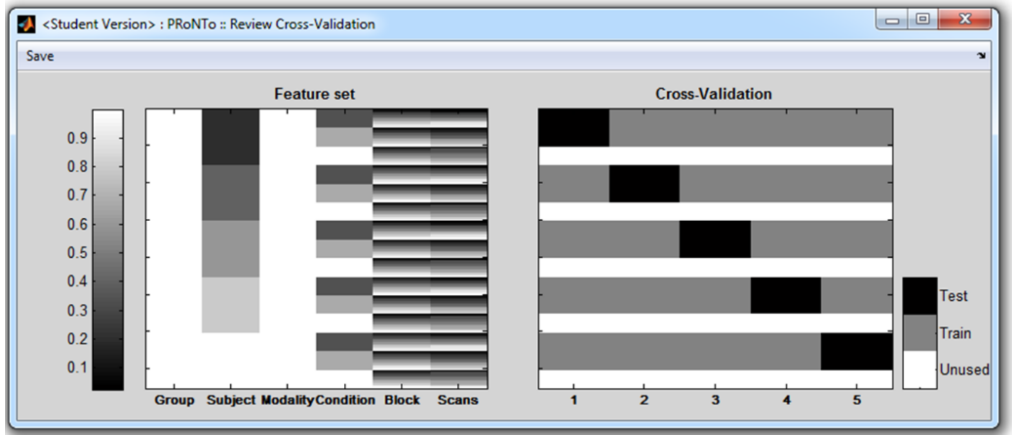
- Number of folds:
  - = number of samples: Leave-One-Out (but see (Varoquaux, 2017))
  - = user based: typically, leave 10 to 20% of data out
- Data in each fold:
  - Regression: are samples sorted?
  - Classification: Leave-per-Class-Out, keeping frequency distributions in each fold
  - Structured data: correlated blocks in test set
- Results will depend on chosen cross-validation, no cherry picking!
- Good practice to report model performance in average and std

Validation: PRoNTo

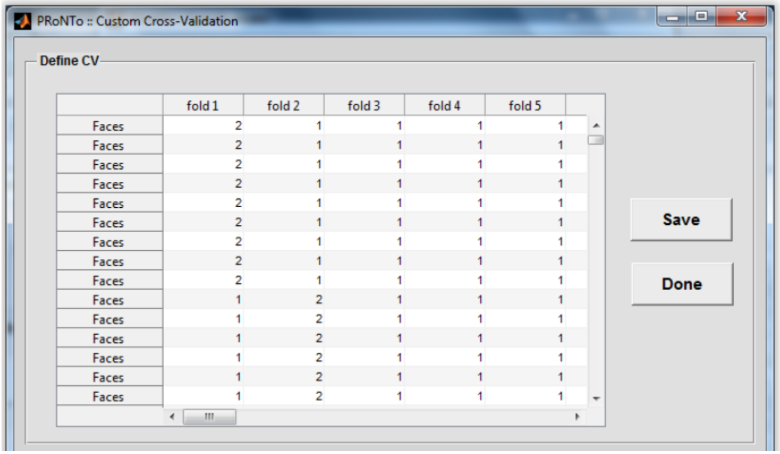


Standard approaches:

- LOSO
- LOBO
- LORO
- LOSCO
- k-fold CV

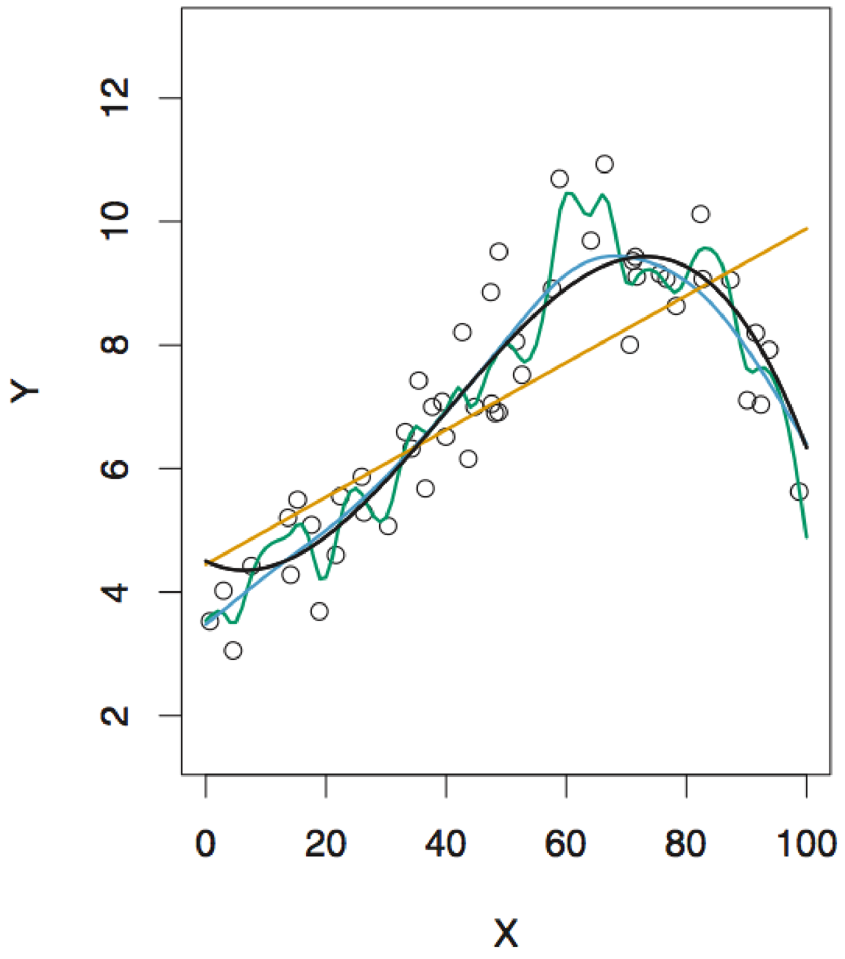


Flexible CV schemes allowed

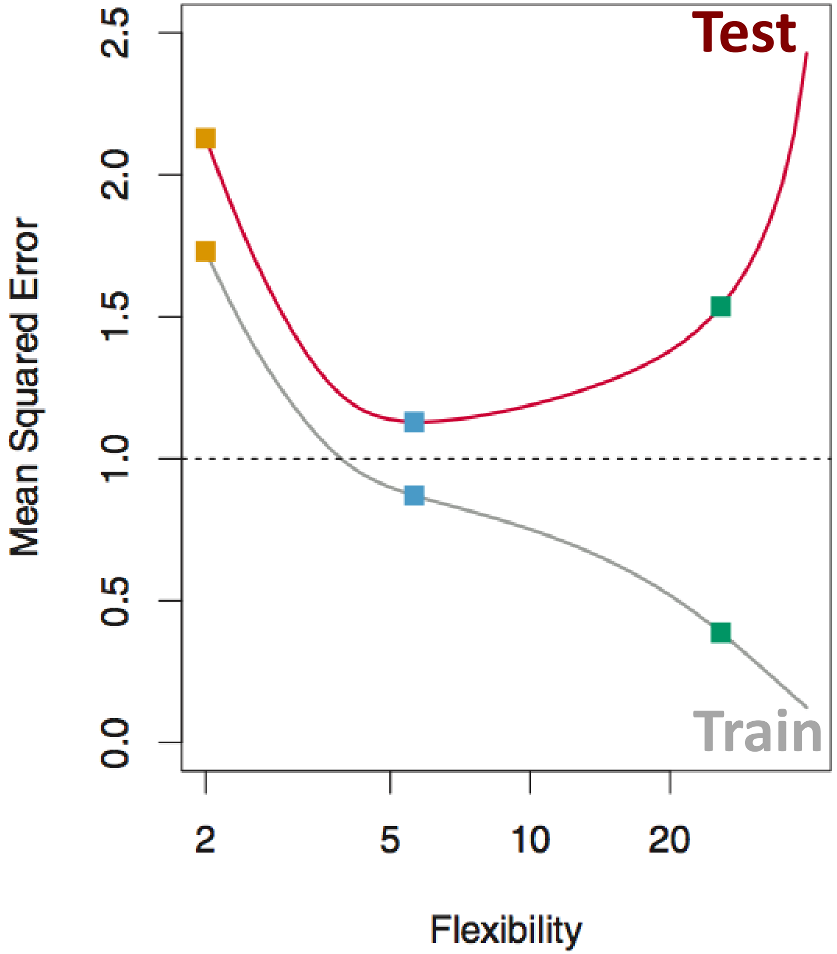


Hyper-parameters

Different models

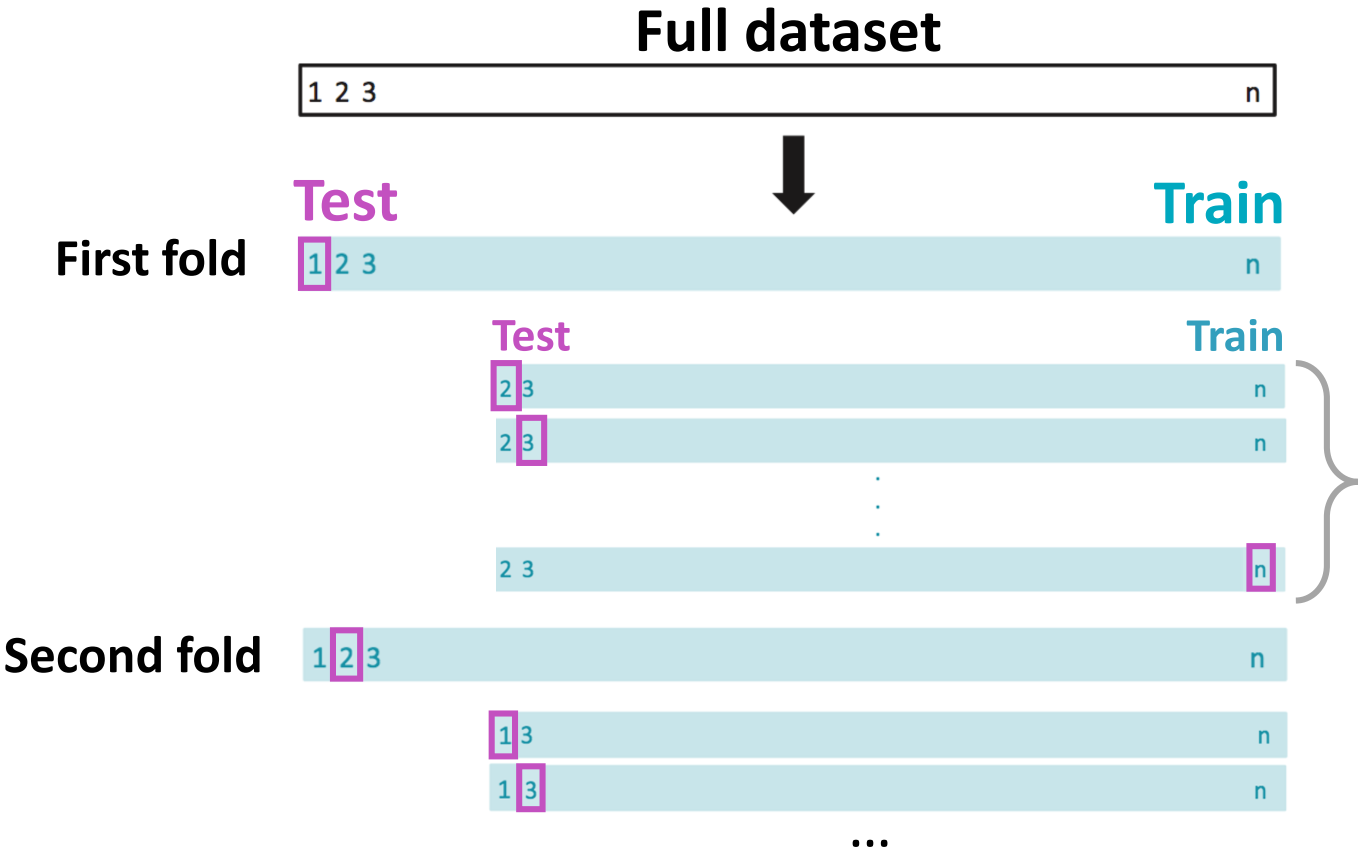


Prediction Error



Nested cross-validation

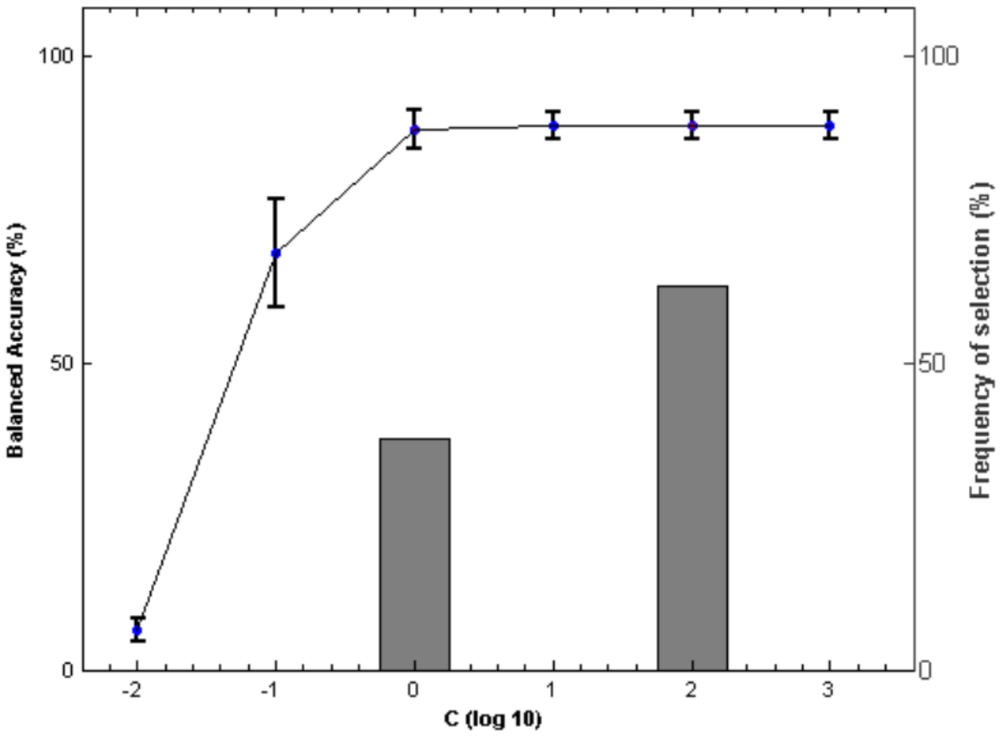
- Problem: use CV to select best model and assess model performance (test error)
- Solution: Run CV inside CV for model or feature selection / Bayesian Models



Nested CV:
Select model **hyper-parameters / feature selection**

Model selection in PRoNTo

If hyper-parameter optimisation was performed using nested CV:

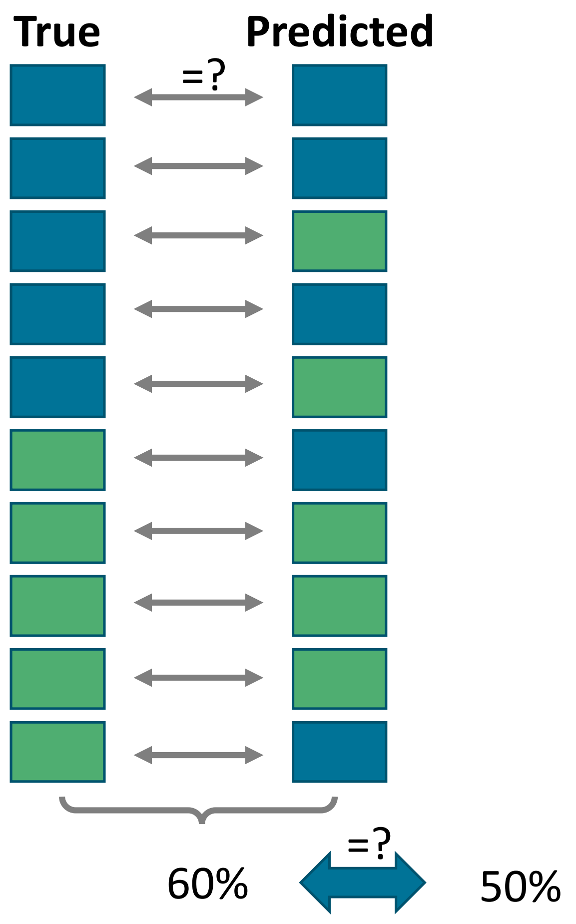


Assessing significance

Parametric tests

e.g. Binomial test

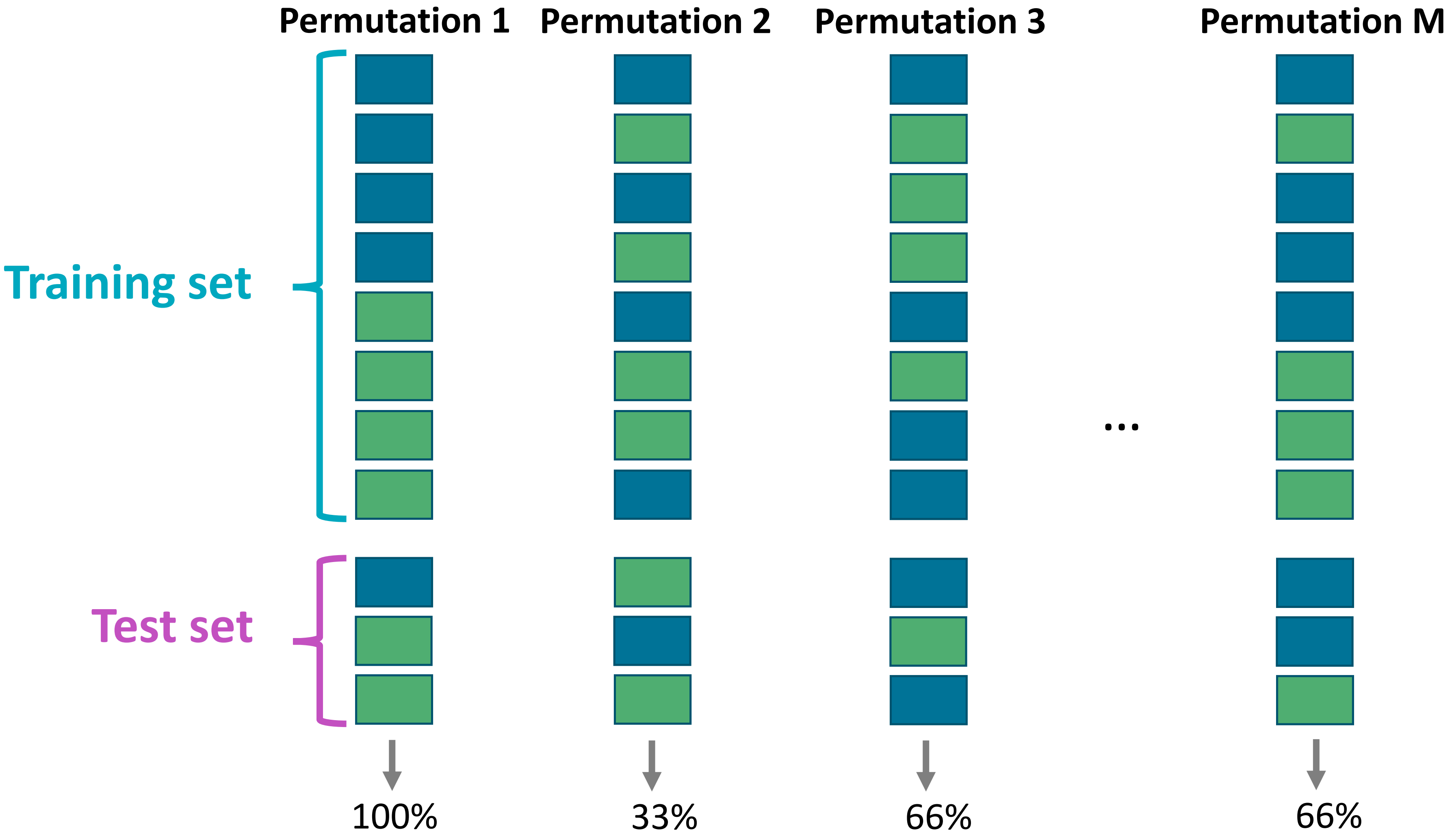
- Model decision in two-class problem modeled as Bernoulli trials
- Probability of *k* successes out of *n* trials follows binomial distribution



Not a good idea:

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

Assessing significance



p-value:

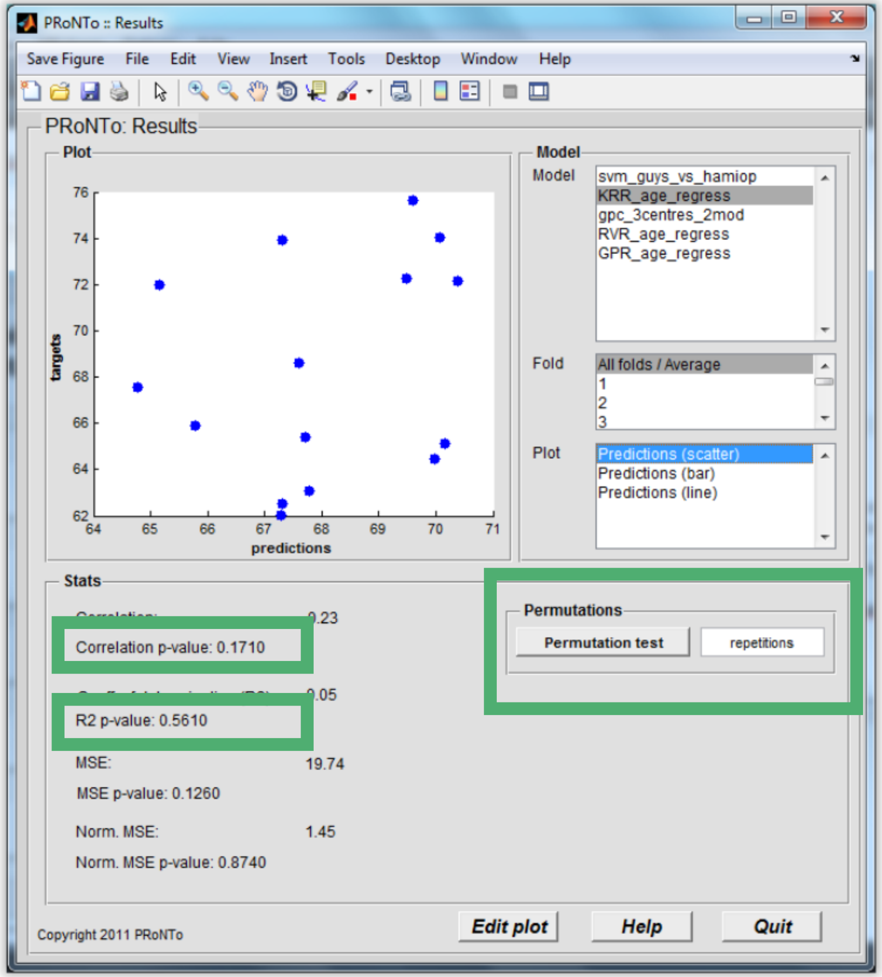
$$p = \frac{1}{M}\sum_{m=1}^{M}\left(acc_m^{perm} \geq acc_1\right)$$

Minimum: $p = \frac{1}{M}$

Assessing significance

- No hypotheses on data distribution
- H_0: “targets are non-informative”
- Test statistic: balanced and class accuracy / MSE / R^2
- Estimate the distribution of the test statistic under H_0 by randomly permuting targets M-1 times, and running the full CV experiment

Assessing significance



In PRoNTo:

User-input = M-1

Take-home on performance

- Always separate data into 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

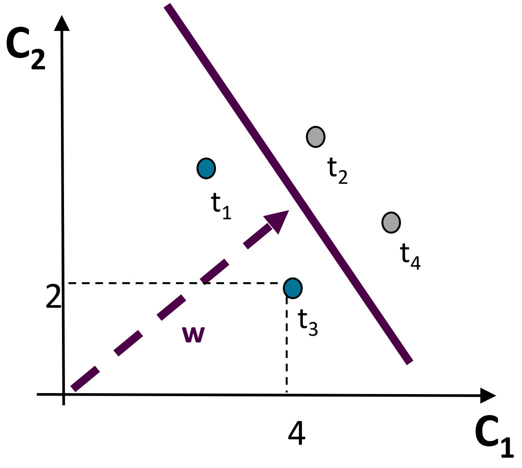
Outline

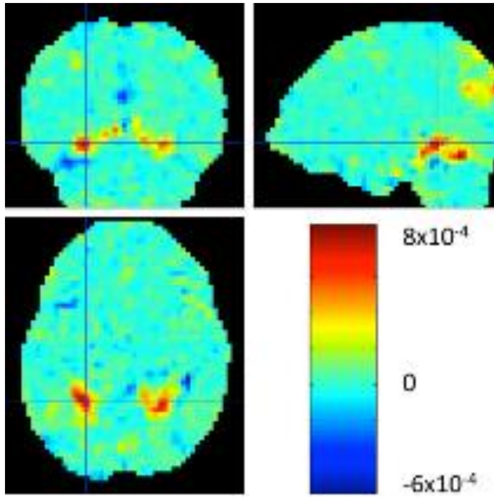
- Is my model good?
  - ➢Measures of performance for classification
  - ➢Measures of performance for regression
  - ➢Validation set and cross-validation
  - ➢Nested cross-validation
  - ➢Assessing significance
- What does my model look like?
  - ➢Model interpretation

Interpretation: weights

- Linear predictive models (classifier or regression) are parameterized by a weight vector **w** and a bias term *b*.

- **w** has the same dimensionality of the input data and can be plotted as an image.





(Haxby, S1, Faces vs Houses)

Interpretation: definition

- In machine learning:
  - Identifying a subset of relevant features
  - Feature selection or regularization

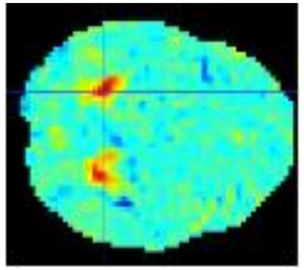
- In neuroscience:
  - Why is a feature relevant?
  - Comparing highest weights with literature or GLM results

Interpretation: decision function

Predictive function

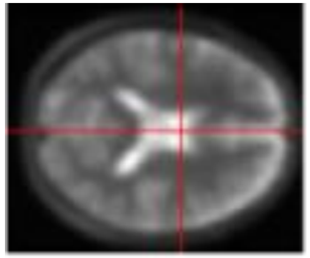
$$f(\mathbf{x}_*) = \mathbf{w} \times \mathbf{x}_* + b$$

Weight map (**w**)



5	2	-6	-1
2	1	2	-1

New example (**x***)



$$f(\mathbf{x}_*) = (5 \times 2) + (2 \times 1) + (-6 \times 2) + (-1 \times -1) + 0$$

$$f(\mathbf{x}_*) = 10 + 2 - 12 + 1 = 1$$

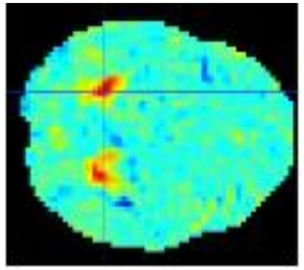
$f(\mathbf{x}_*)$ is the predicted score for regression or the distance to the decision boundary for classification models.

Interpretation: decision function

Predictive function

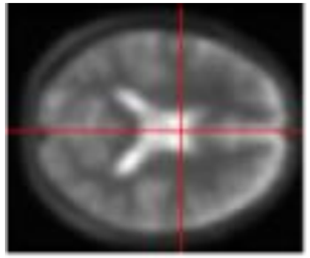
$$f(\mathbf{x}_*) = \mathbf{w} \times \mathbf{x}_* + b$$

Weight map (**w**)



5	0	-6	0

New example (**x***)



2	1	2	-1

$$f(\mathbf{x}_*) = (5\times 2) + (0\times 1) + (-6\times 2) + (0\times -1) + 0$$

$$f(\mathbf{x}_*) = 10 + 0 - 12 + 0 = -2$$

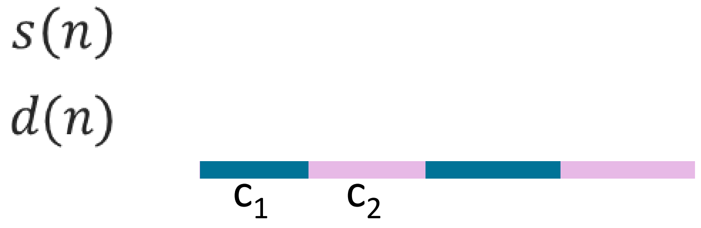
$f(\mathbf{x}_*)$ truncated does not correspond to $f(\mathbf{x}_*)$!

Interpretation: weight amplitude

- What do weights represent?

  Assume:
  - Signal in voxel 1: $s(n) + d(n)$
  - Signal in voxel 2: $d(n)$

  Weights:
  - Voxel 1: $w = 1$
  - Voxel 2: $w = -1$



(Haufe et al., 2014)

- Not only (neural) signal can lead to high weight amplitude in a voxel!
- Also, weight=0 does not necessarily mean no signal (depends on regularization)!

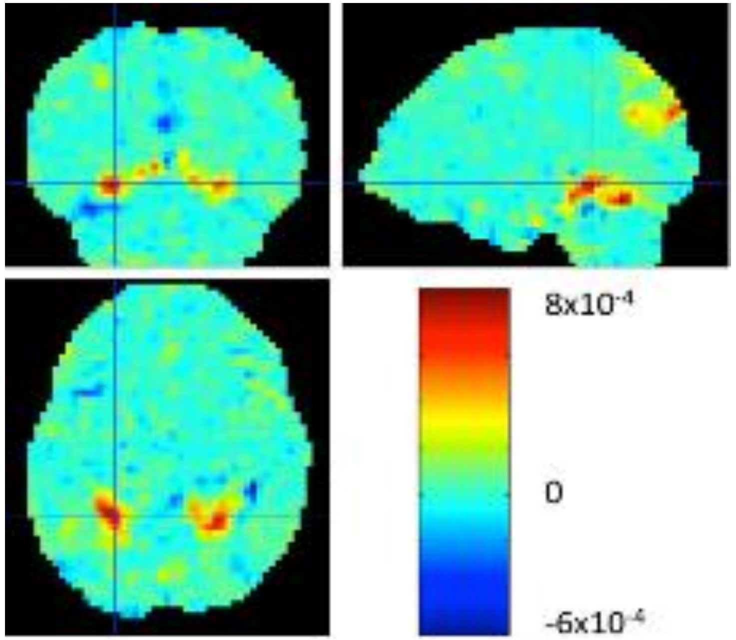
Interpretation: strategies

- **A priori**
    1. Masking
    2. Searchlight mapping
- **During model estimation**
    3. Feature selection
    4. Sparse algorithms
    5. Atlas based Multiple Kernel Learning (MKL)
    6. Using weight stability in model selection
- **A posteriori**
    7. Atlas based weight summarization
    8. Permutation test
    9. Transforming weights into activation patterns

Interpretation in PRoNTo

- **A priori**
  1. Masking
  2. Searchlight mapping (with extra code)
- **During model estimation**
  3. Feature selection
  4. Sparse algorithms (v3)
  5. Atlas based Multiple Kernel Learning (MKL)
  6. Using weight stability in model selection
- **A posteriori**
  7. Atlas based weight summarization
  8. Permutation test (building weight maps for permutation, no second-level in PRoNTo)
  9. Transforming weights into activation patterns

Take home on interpretation



- ✓ Spatial representation of the predictive function.
- ✓ Shows the contribution of each feature/voxel to the prediction.
- ✓ Multivariate pattern -> All voxels with weights different from zero contribute to the final prediction (no arbitrary threshold should be applied).
- ✓ Mixture of signal of interest and noise, but also depends on input neural signal SNR and sparsity.
- ✓ Strategies available to help, each with their pros and cons.

Recommended reading: performance

- 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.
- Varoquaux, *Cross-validation failure: Small sample sizes lead to large error bars*, NeuroImage, 2017.

Recommended reading: weights

- Baldassarre, L., Pontil, M. & Mourão-Miranda, J.. «Sparsity is better with stability: combining accuracy and stability for model selection in brain decoding. » Frontiers in Neuroscience: Brain Imaging Methods. (2017)
- De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. «Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. » NeuroImage. (2008) 43(1), 44–58.
- Gaonkar, B. & Davatzikos, C. «Analytic estimation of statistical significance maps for support vector machine based multivariate image analysis and classification.» NeuroImage. (2013) 78: 270-283
- Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B. & Taylor, J.E. «Interpretable whole-brain prediction analysis with GraphNet.» NeuroImage. (2013) 72, 304–321
- Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B. & Bießmann, F. «On the interpretation of weight vectors of linear models in multivariate neuroimaging.» NeuroImage. (2014) 87, 96–110.
- Kia, S.M., Vega-Pons, S., Weisz, N. & Passerini, A. «Interpretability of Multivariate Brain Maps in Linear Brain Decoding: Definition, and Heuristic Quantification in Multivariate Analysis of MEG Time-Locked Effects.» Frontiers in Neuroscience. (2017) 10.3389/fnins.2016.00619
- Kriegeskorte, N., Rainer, G. & Bandettini, P. «Information-based functional brain mapping.» *PNAS* 103 (2006): 3863-3868.
- Michel, V., Gramfort, A., Varoquaux, G., Eger, E. & Thirion, B. «Total variation regularization for fmri-based prediction of behavior.» IEEE Trans Med Imaging. (2011) 30, 1328 –1340.
- Rakotomamonjy, A., Bach, F., Canu, S. & Grandvalet, Y. «SimpleMKL» Journal of Machine Learning 9 (2008): 2491-2521.

Recommended reading: weights

- Rondina J., Hahn T., de Oliveira L., Marquand A., Dresler T., Leitner T., Fallgatter A., Shawe-Taylor J. & Mourao-Miranda J. «SCoRS - a method based on stability for feature selection and mapping in neuroimaging.» IEEE Trans Med Imaging. (2014) Jan:33(1).
- Sato, J.R., Mourao-Miranda, J., Morais Martin Mda G., Amaro E. Jr., Morettin P.A. & Brammer M.J. «The impact of functional connectivity changes on support vector machines mapping of fMRI data.» J Neurosci Methods. (2008) 172(1):94-104
- Schrouff, J., Cremers, J., Garraux, G., Baldassarre, L., Mourão-Miranda, J. & Phillips, C. «Localizing and comparing weight maps generated from linear kernel machine learning models.» Proceedings of the 3rd workshop on Pattern Recognition in NeuroImaging. 2013.
- Schrouff, J., Rosa, M. J., Rondina, J., Marquand, A., Chu, C., Ashburner, J., Phillips, C., Richiardi, J., & Mourão-Miranda, J. «PRoNTo: Pattern Recognition for Neuroimaging Toolbox. » Neuroinformatics. (2013) 11(3): 319-337.
- Schrouff, J., Mourao-Miranda, J., Phillips, C., & Parvizi, J. «Decoding intracranial EEG data with multiple kernel learning method.» Journal of Neuroscience Methods. (2016) 261: 19-28.
- Tibshirani, R. «Regression shrinkage and selection via the lasso.» Journal of the Royal Statistical Society. 58 (1996): 267-288.
- Tzourio-Mazoyer, N., et al. «Automated Anatomical Labeling of activations in SPM using a Macroscopic Anatomical Parcellation of the MNI MRI single-subject brain.» NeuroImage 15 (2002): 273-289.
- Zou, H., & Hastie, T. «Regularization and variable selection via the elastic net.» J. R. Statist. Soc. B 67 (2005): 301-320.

Thank you!

Questions?