

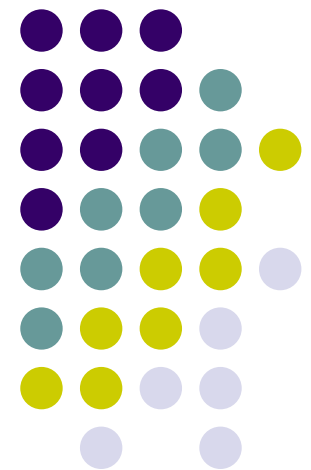
Statistical Validation of Metabolic Differences

Johan Westerhuis

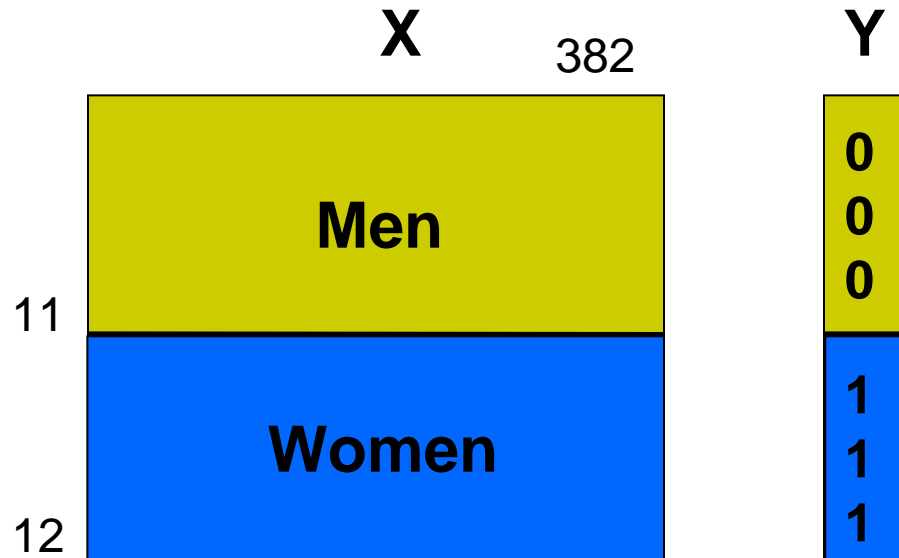
Biosystems Data Analysis

Swammerdam Institute for Life Sciences

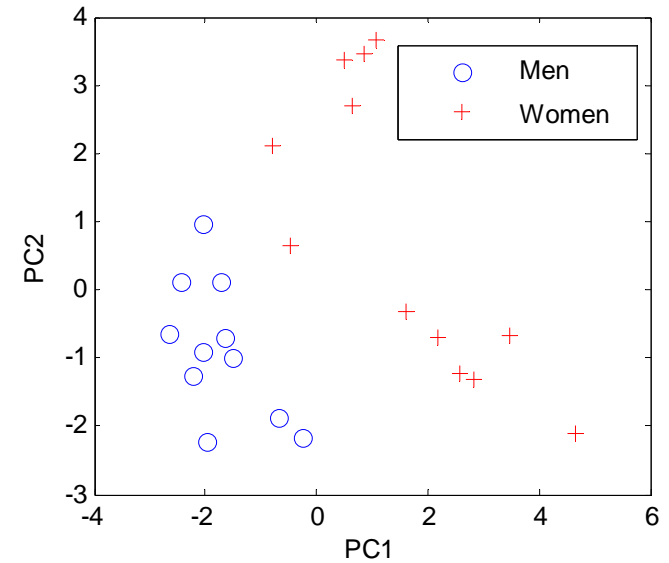
Universiteit van Amsterdam



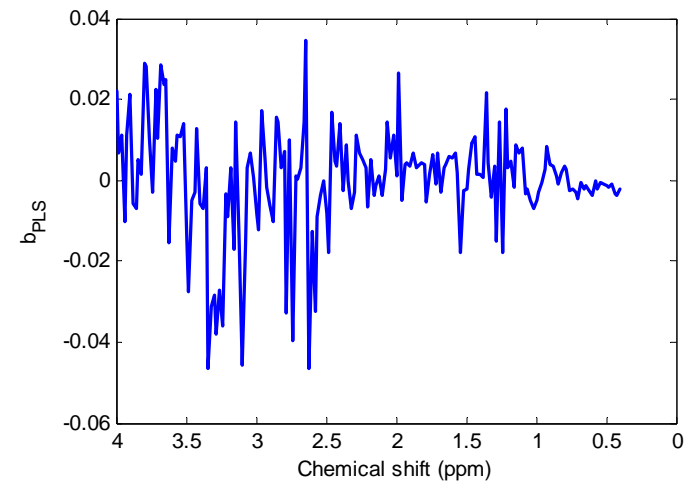
General PLSDA application



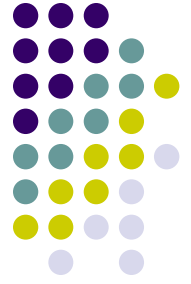
$$Q^2 = 0.27$$



- Is there really a difference between the groups ?
- What is the meaning of Q^2 ?
- Which are the most important peaks for discrimination ?

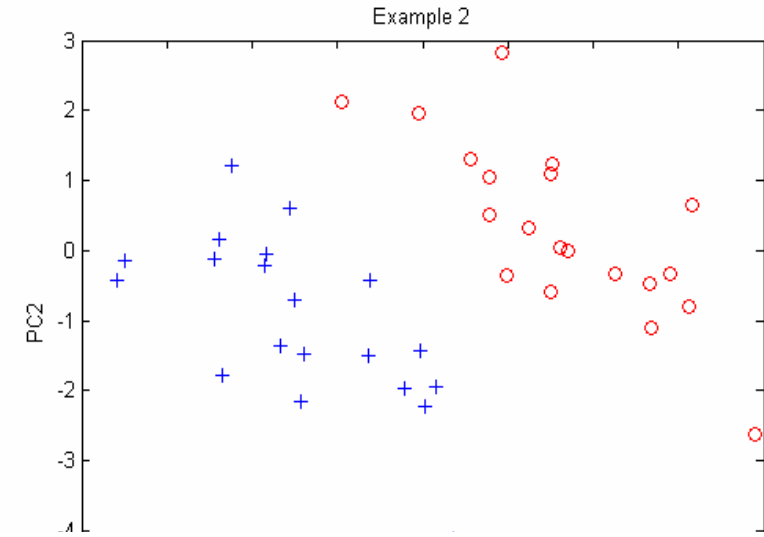
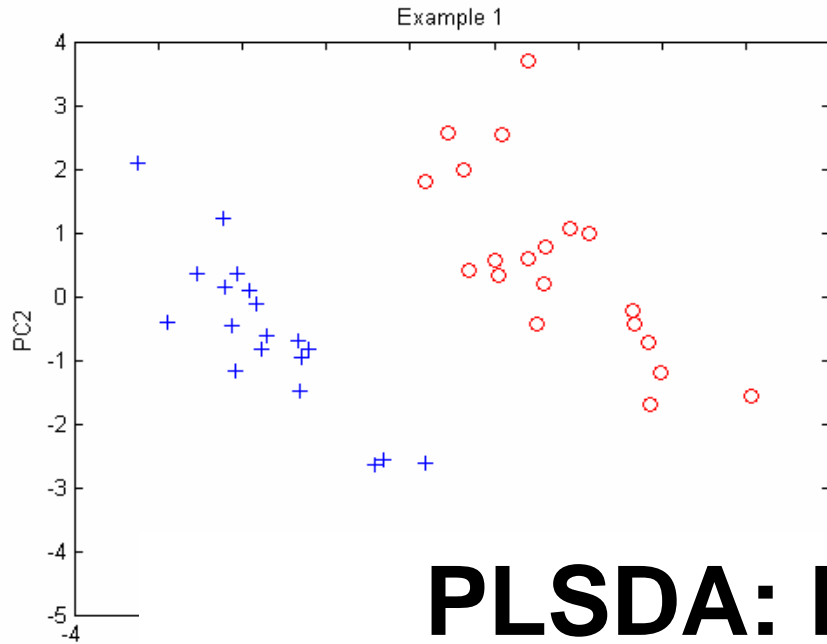


First test using random data

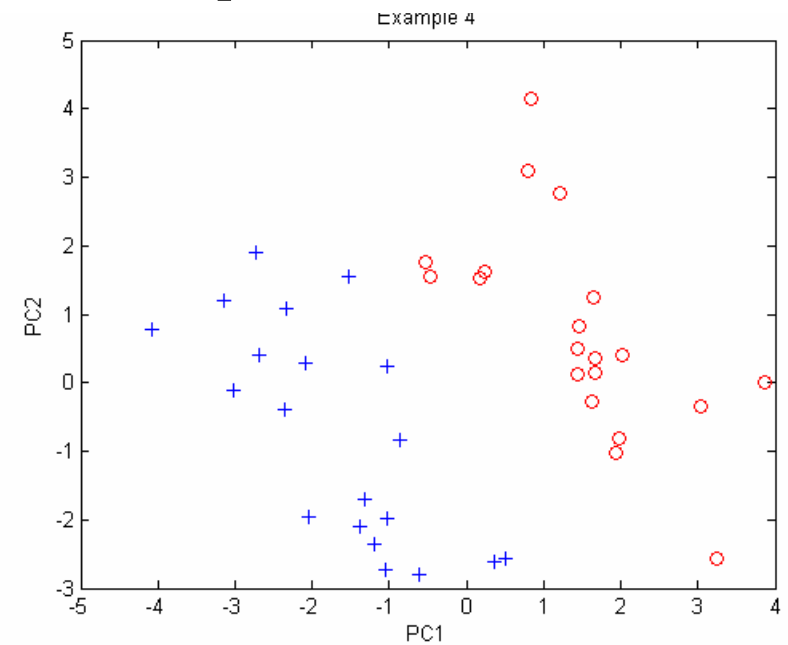
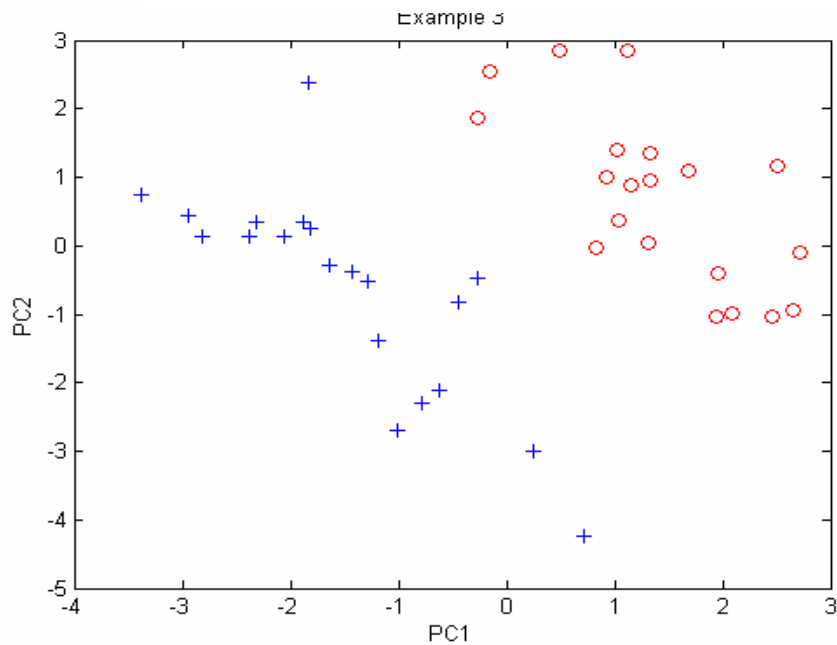


- No difference expected !

Simulations results



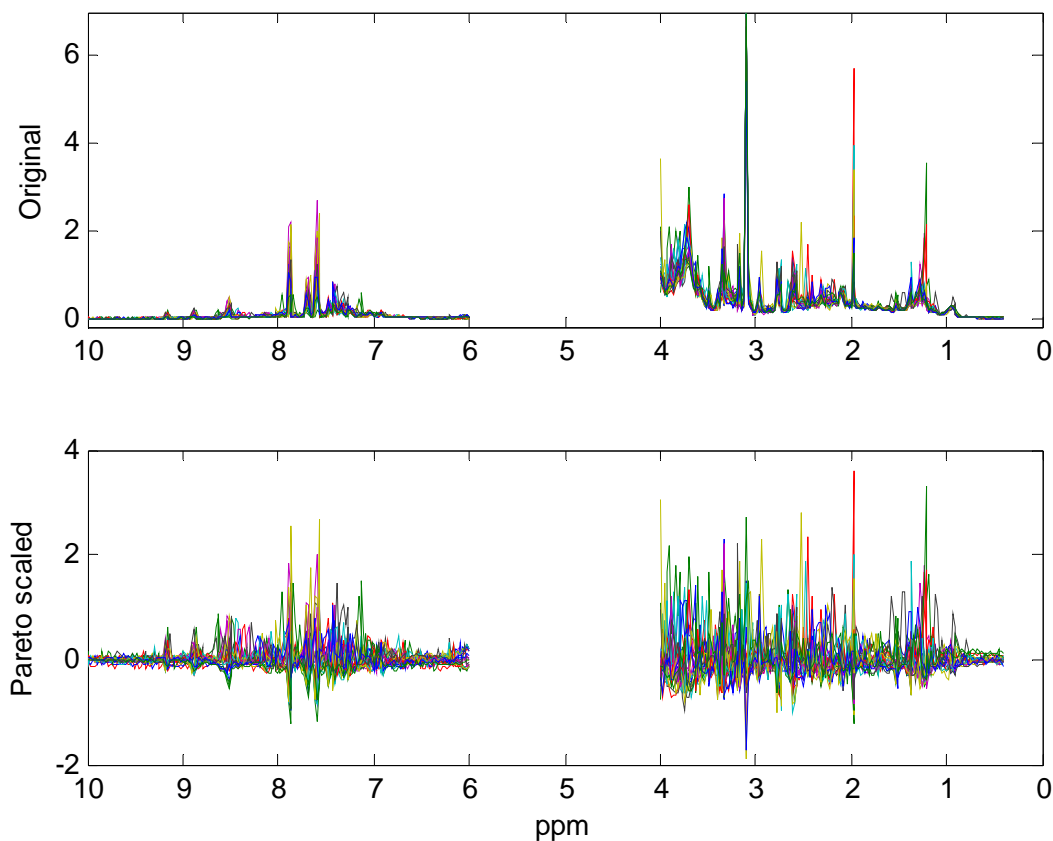
PLSDA: Eager to please



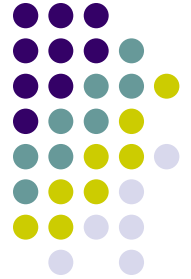
Discrimination by sex with PLSDA



NMR spectra



- Urine samples of healthy volunteers: women (12) and men (11).
- Spectra (0.4-4, 6-10 ppm) were binned (0.02 ppm).
- Each bin normalized to account for volume differences
- Pareto scaling (meancenter + scaling by $s^{1/2}$) before analysis

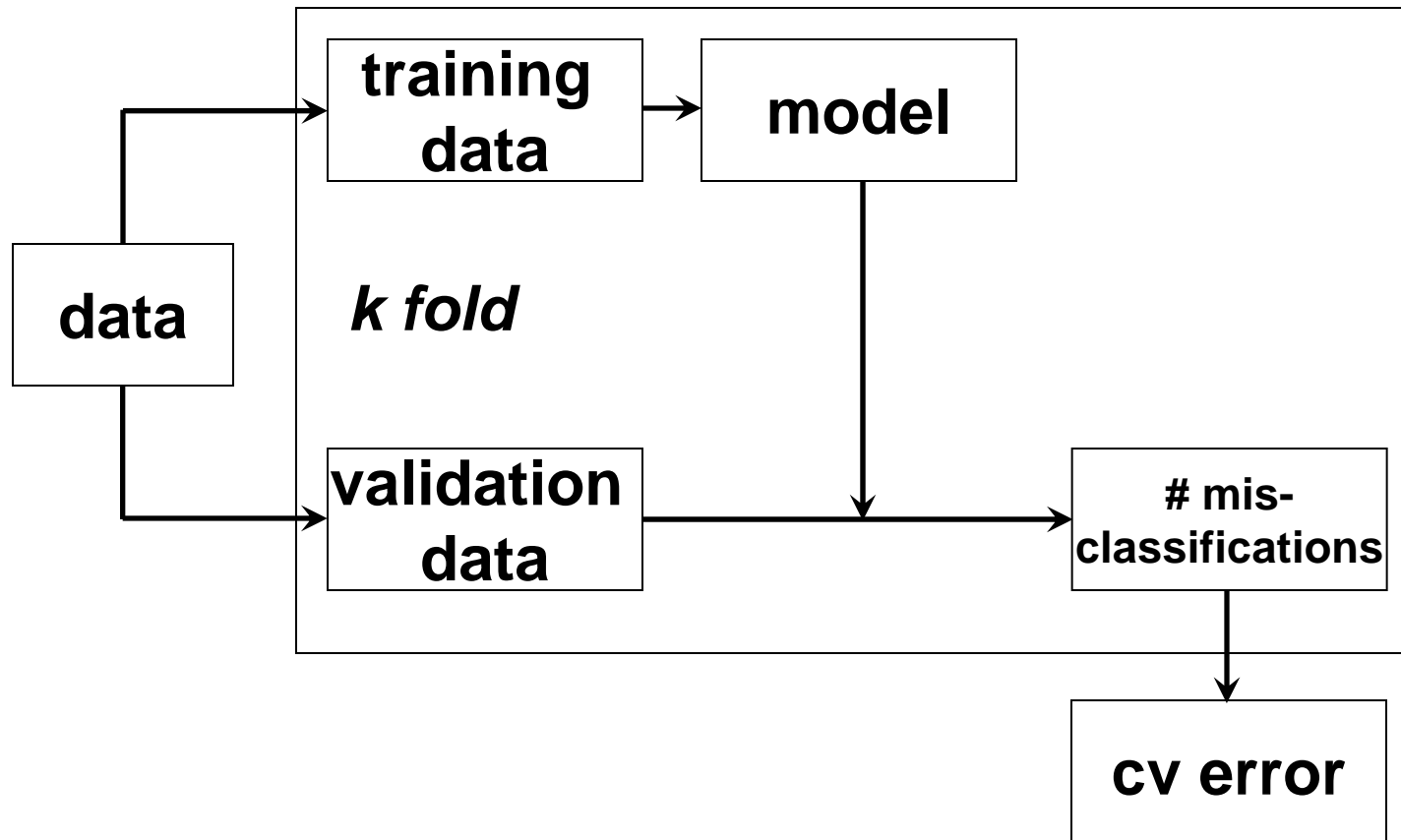
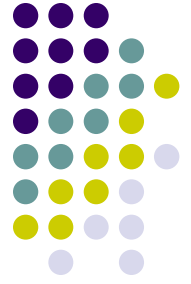


om/naturebiotechnology
Summary recommendations for
standardization and reporting of
metabolic analyses

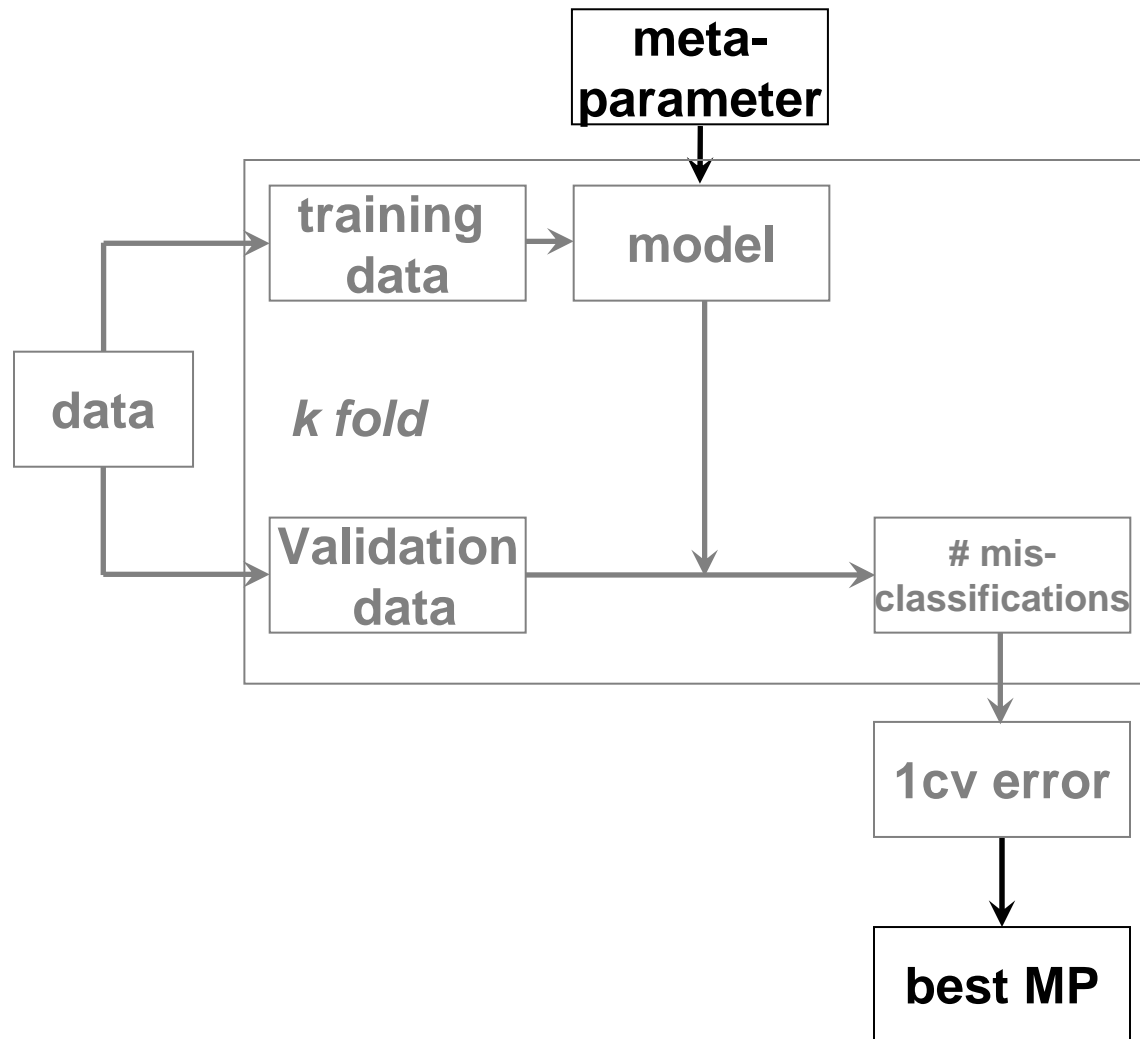
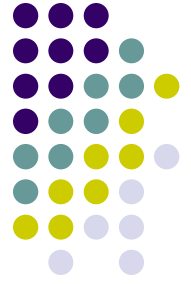
Cross validation:

Each observation must be left out in turn and the prediction error calculated in each case. These may then be averaged to give an overall model predictability.

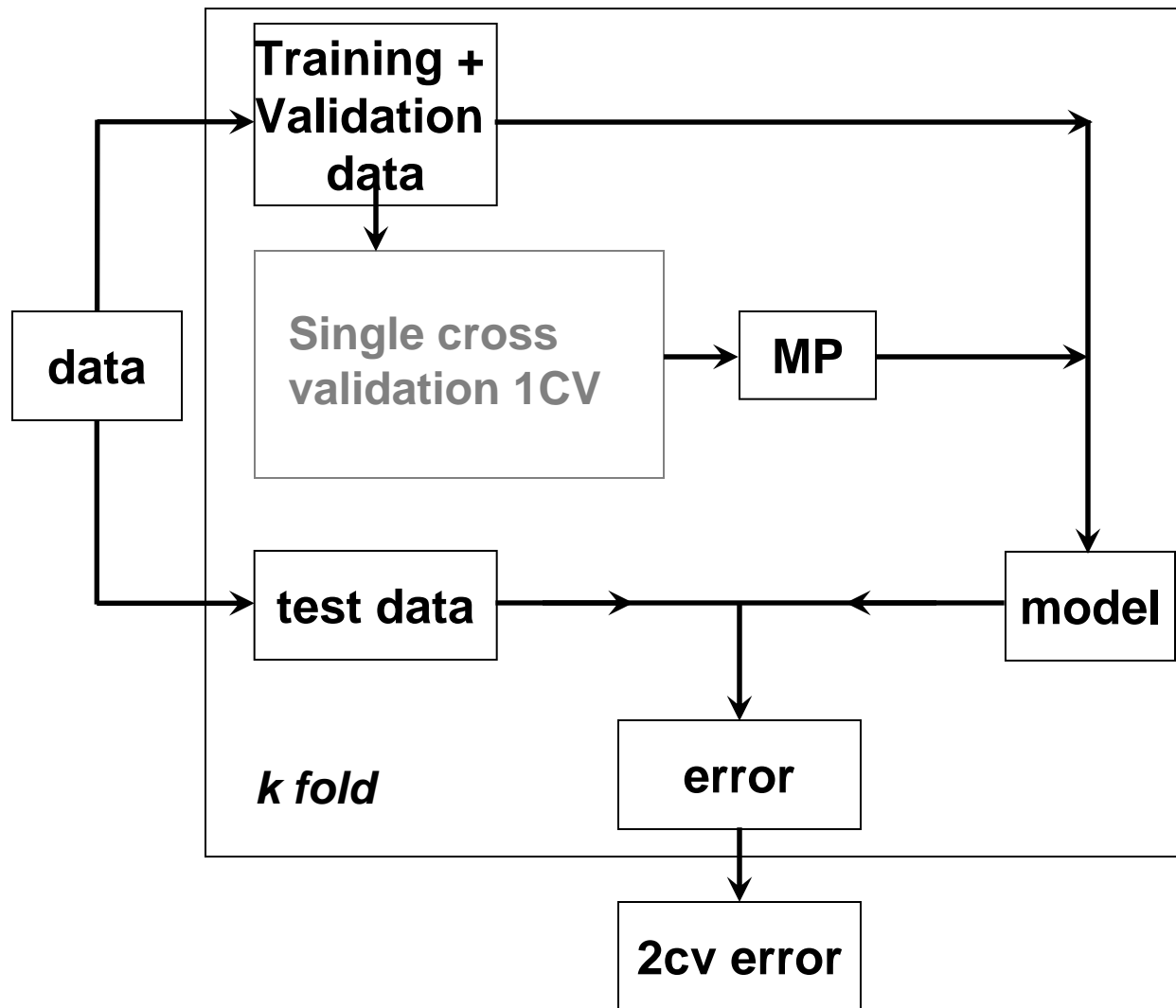
Single Cross validation (1CV)



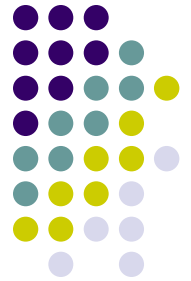
Single Cross validation (1CV)



Double cross validation (2CV)



Validation methods

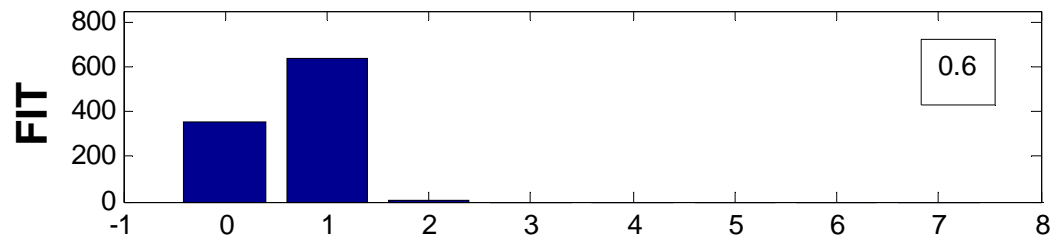


Method	Predicted sample used to calculate model coefficients	Predicted sample used to calculate # components
FIT	yes	yes
1CV	no	yes
2CV	no	no

Misclassifications

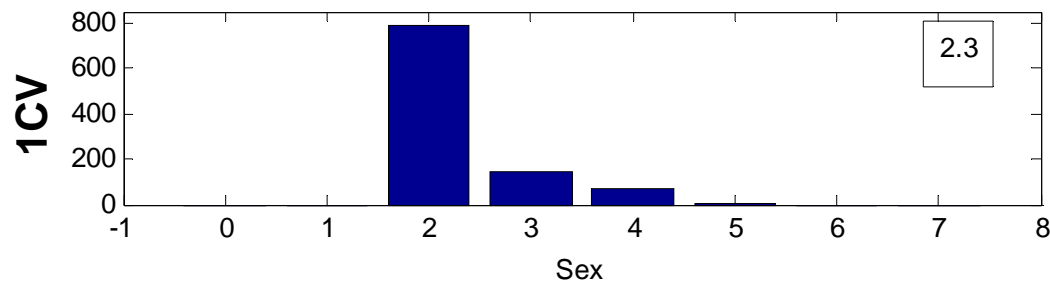


7-fold Cross validation 1000 times



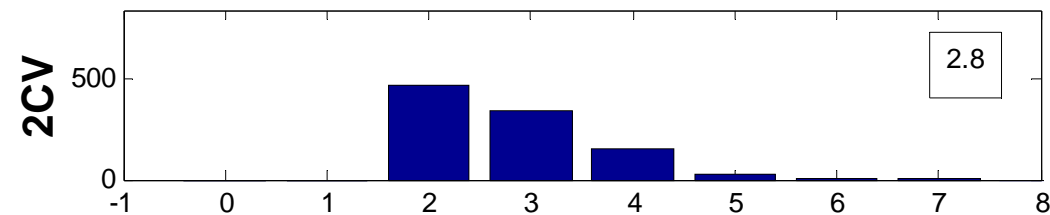
Average # misclassifications

FIT: 0.6



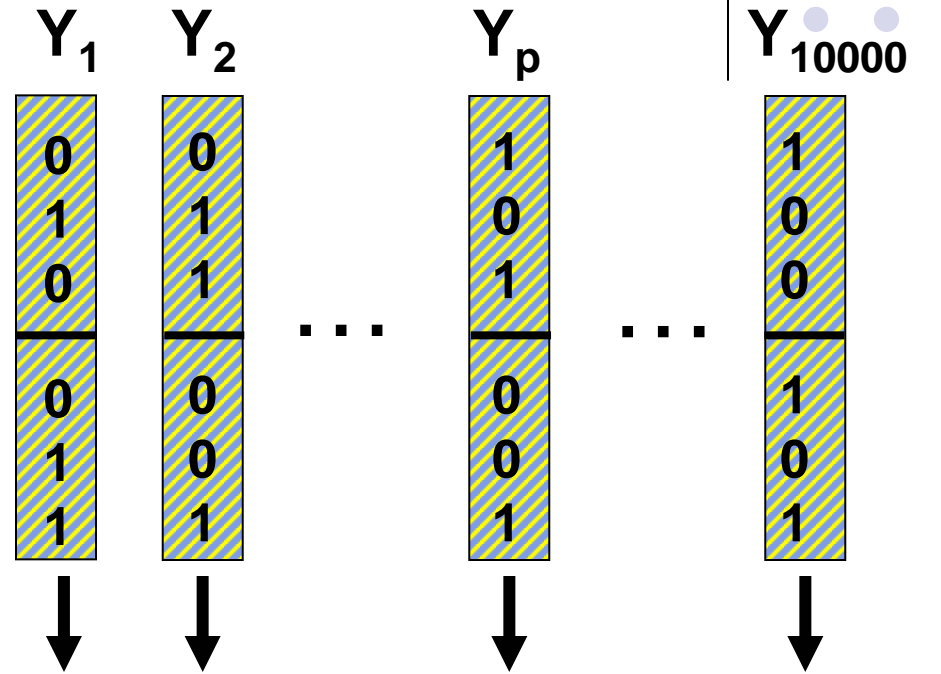
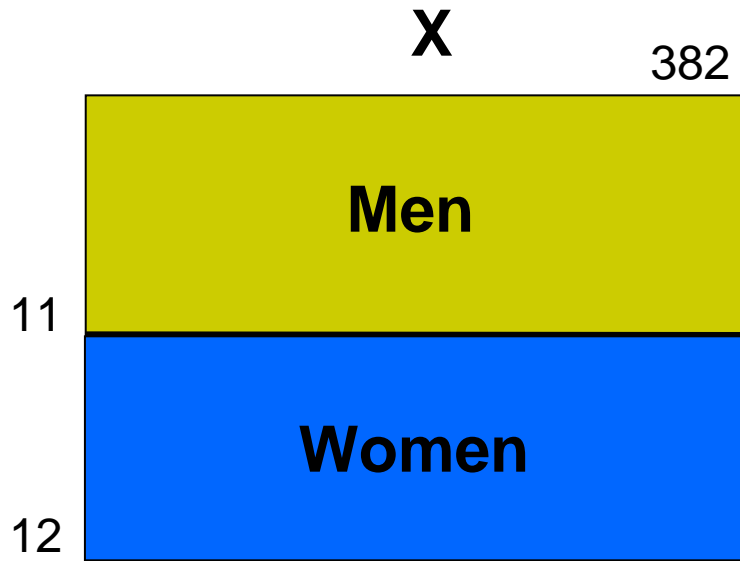
1CV: 2.3

2CV: 2.8



Which value is meaningful?

Permutations



Distribution of statistics for:
Class 0 = Class 1

Make model:
Predict class

} FIT
 } 1CV
 } 2CV
 } Misclassifications
 } Q^2
 } ...

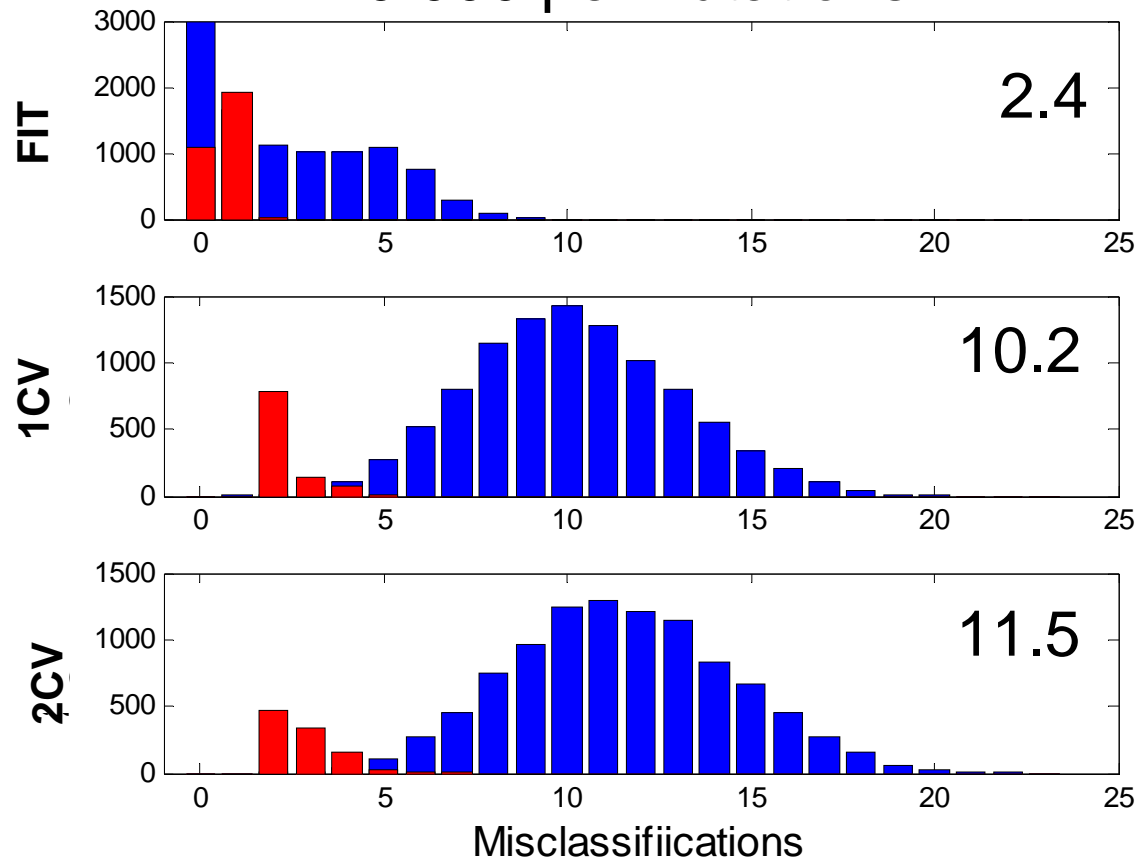


Permutations of Sex classification



Expected on average 11.5 of 23 misclassifications

10.000 permutations

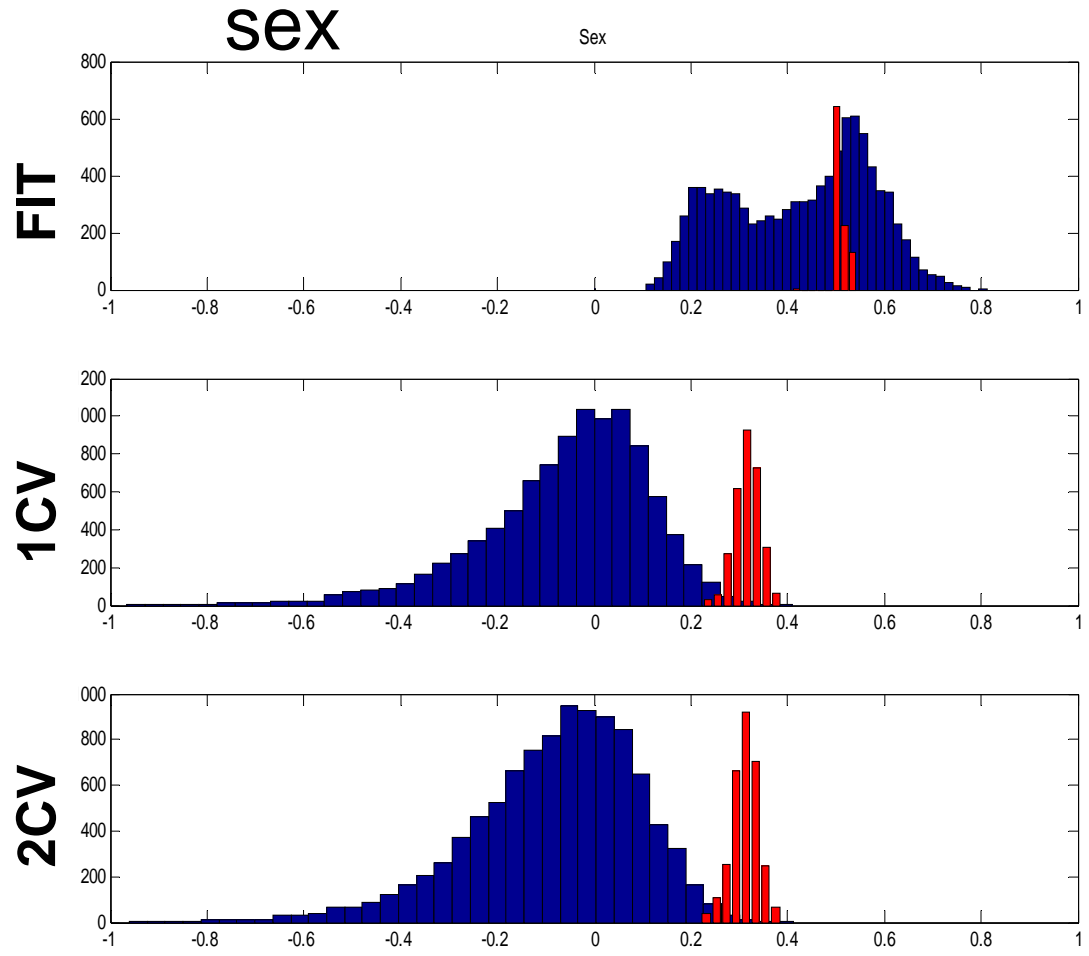


- FIT far too optimistic
- 1CV slightly over-optimistic
- 2CV as expected
- Both 1CV and 2CV give good separation

Validation of Q^2 statistic in PLSDA



$$Q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} = 1 - \frac{PRESS}{SS}$$

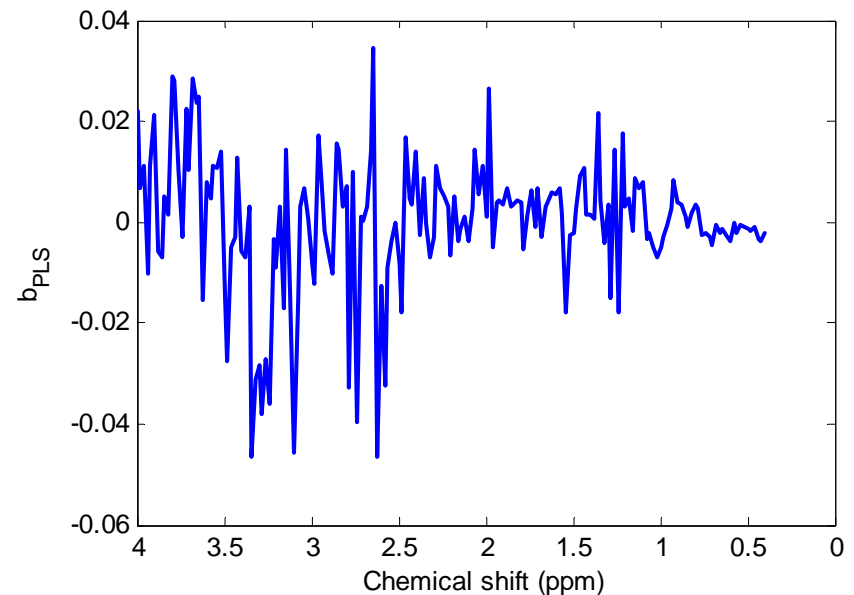


- Q^2 often used to qualify separation.
- Which Q^2 value indicates a good separation?
- Many permutations give $Q^2 > 0$!!
- Both 1CV and 2CV give good separation

Rank products¹, cross validation and permutation to select biomarkers

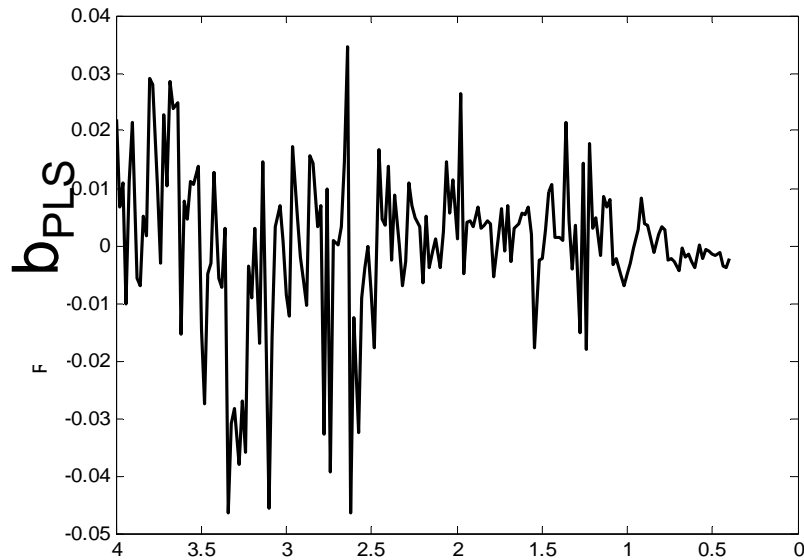
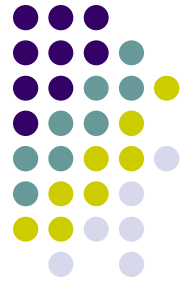


- Rank coefficients according to absolute value. Most important (highest) one gets 1.
- Multiply ranks of 2CV models to find most important variables.
- Calculate rank products of 2CV models of permuted data.
- Compare distributions and select important variables.

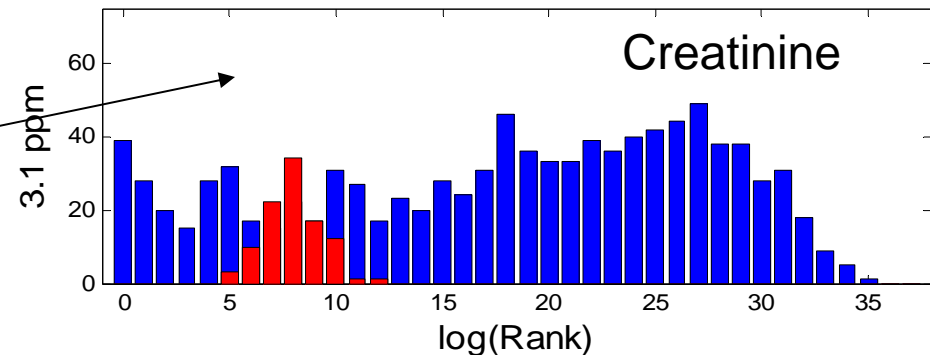
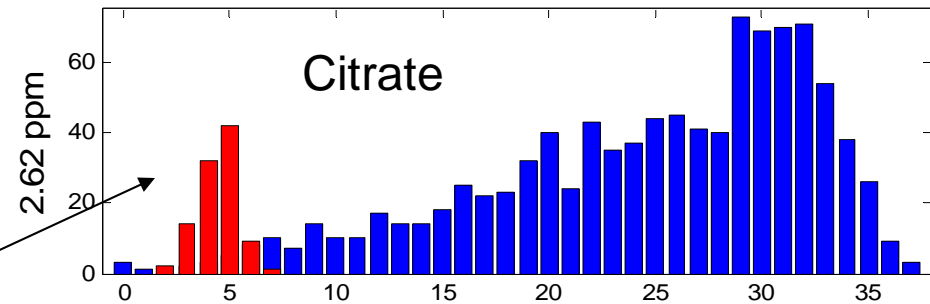
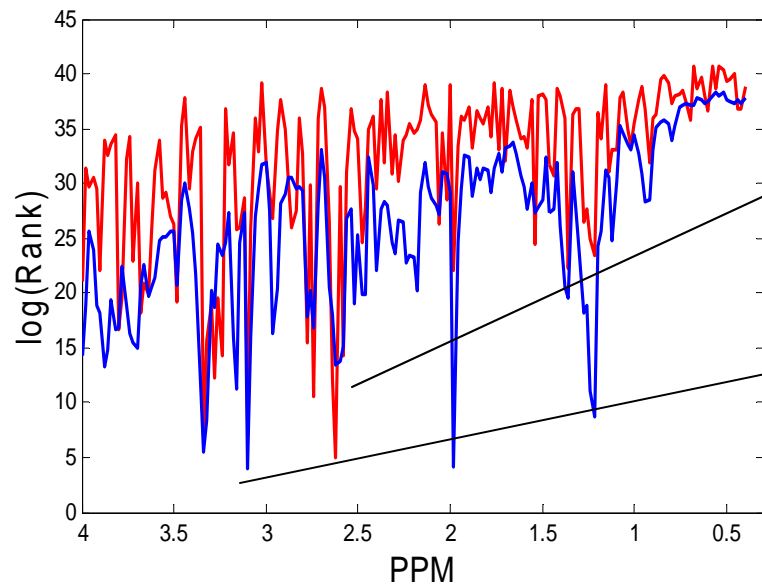


¹Breitling, FEBS lett, 2004, 537, 83-92

Biomarkers by rank products and permutations



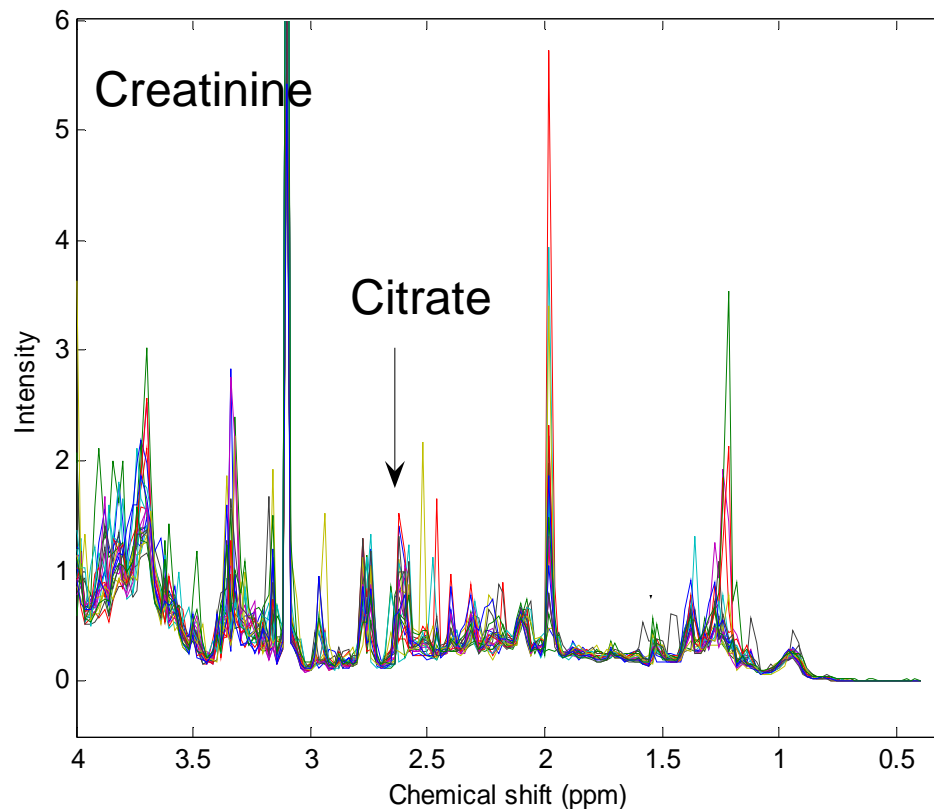
- Rank products indicate which peaks are significant and which not
- **Mean rank crossvalidation** vs **10% quantile of permutations**
- Histograms of rank products of two peaks



Some conclusions



sex



- Use random data as first step to test new data analysis methods.
- Only double cross validation (2CV) truly mimics test set validation.
- Combination of permutations, rank products and many 2CV models is a generally applicable tool for validation of classification and biomarker selection.

Acknowledgements



- Biosystems Data Analysis:
 - Age Smilde, Huub Hoefsloot, Suzanne Smit, Daniel Vis
- Unilever Research Vlaardingen (metabolomics):
 - John van Duynhoven, Ferdi van Dorsten, Ewoud van Velzen