Multivariate analysis of 'omics data

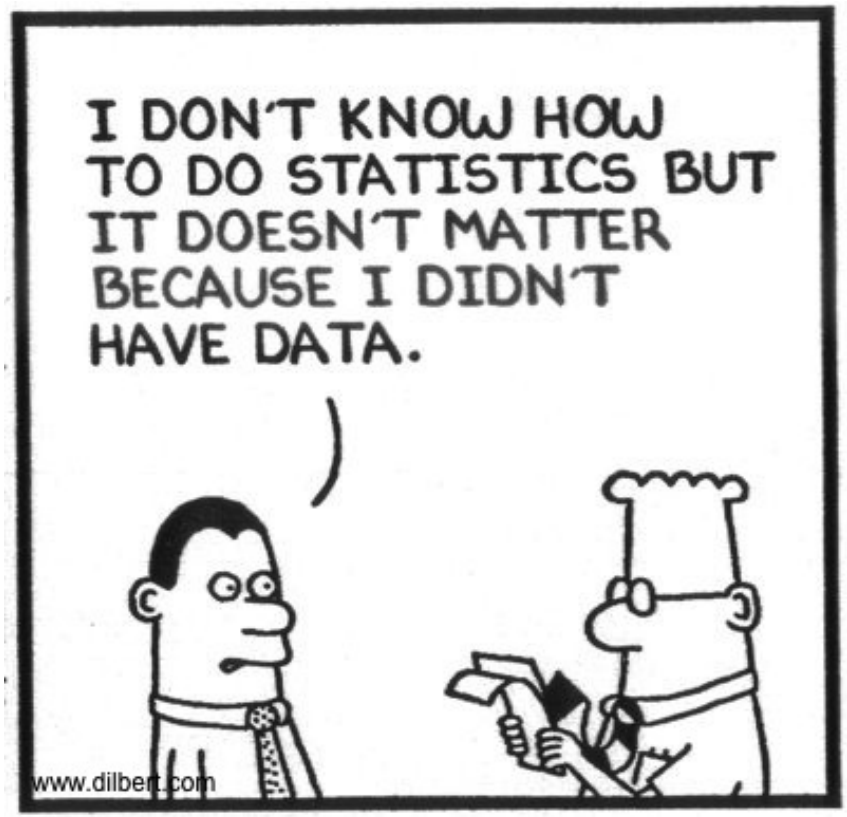
Multivariate discriminant analysis

A/Prof. Kim-Anh Lê Cao

Melbourne Integrative Genomics
School of Mathematics & Statistics
University of Melbourne

kimanh.lecao@unimelb.edu.au

Learning objectives for this short course



Learning objectives for this short course

Theory

- Understand the main concepts of multivariate dimension reduction methods
- Choose the 'right' method for the 'right' biological question
- Be aware of the benefits and limitations of all methods presented
- Interpretation of the graphical outputs

Practice

- Ability to use provided `R` code on own data
- Perform several types of multivariate analyses ranging from data exploration to biomarker selection using `mixOmics`
- Be critical of the results obtained

Supervised methods aim to model a relationship between the data and a measurable outcome

If the outcome is a discrete variable (e.g. type of treatment)
$\rightarrow$ classification.

If the outcome is a continuous variable (e.g. BMI)
$\rightarrow$ regression

Supervised analysis is different from unsupervised analysis where no explicit outcome was given!

Here we will focus on classification using multivariate methods.

Classification analysis aims

- **Descriptive** aim: weight the variables in an *optimal* manner so that their combination best separates the k classes of samples,
  $\rightarrow$ according to a statistical criterion

- **Predictive** aim: predicting the class of a new samples given its variables values
  $\rightarrow$ construction of a classifier (= set of rules)
  $\rightarrow$ diagnostic/prognostic measures w.r.t sensitivity and specificity (ROC, AUC)

Feature selection

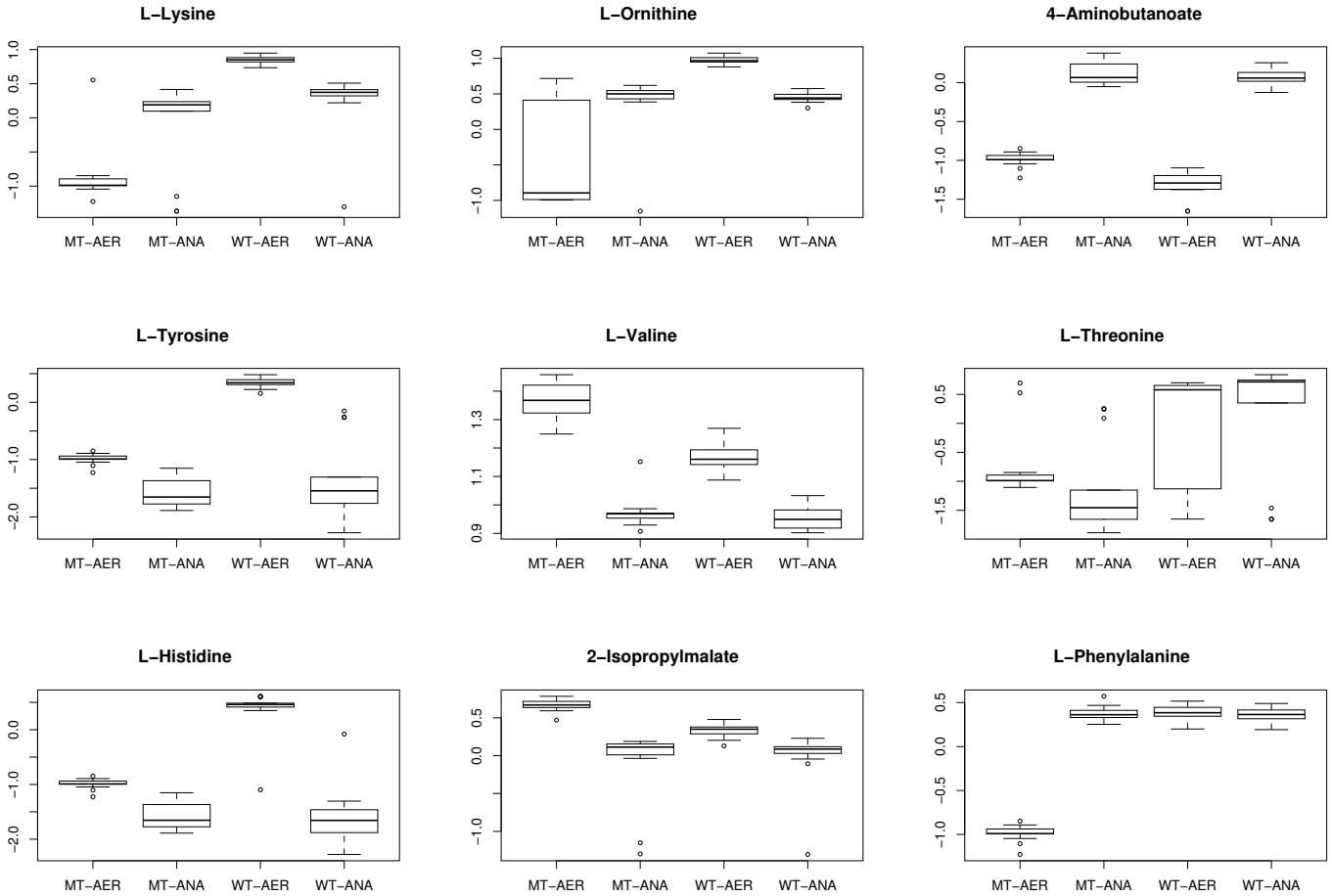
Classification rule built on

- All variables (e.g. genes) or
- a small subset of variables

$\rightsquigarrow$ In molecular biology, a *biomarker panel or molecular signatures* = subset of molecular features with high discriminative power.

$\rightsquigarrow$ Multivariate variable selection often represent a diverse biomarker signature that can not be obtained using univariate statistical methods.

Example of molecular signature



Yeast metabolite data: multivariate biomarker signature w.r.t groups

Multiclass classification

Classification task may involve to separate

- Two groups (.e.g cases vs control groups)
  $\rightarrow$ binary classification
- More than two groups (e.g. several tumour subtypes)
  $\rightarrow$ multiclass classification

Some classification methods are designed for binary classification only (e.g. Support Vector Machine) and apply *one-vs-one*, *one-vs-all* classification for multiclass problems.

Outline of Part 2

Multivariate classification methods

Aim: Seek for a linear combination of features to characterise or separate two or more classes of samples.

Result of a linear multivariate classifier:

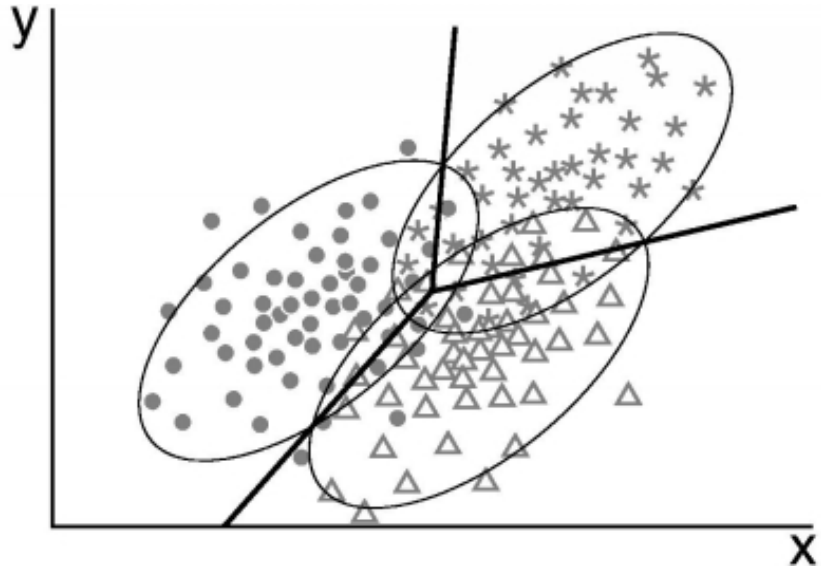
- Dimensionality reduction prior to classification.
- A classifier capable of predicting the class of a new sample based on a linear combination of features.

Multivariate classification approaches:

- Fisher's Linear Discriminant Analysis (LDA)
- Partial Least Squares Discriminant Analysis (PLS-DA)

To oversimplify, you can see this family of approaches as a 'supervised PCA'.

Multivariate classification methods



http://pvermees.andropov.org/

- Maximise the between group variability **and**
- Minimise the within group variability.

Multivariate classification methods

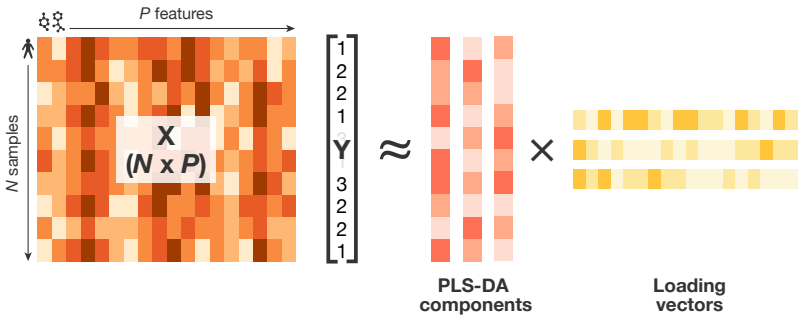
Parameters to choose:

- Number of components (dimensions)
  $\rightarrow$ often K-1, or K max. (K is # classes)

Prediction of a new observation (sample):

- Each component score is a linear combination of variables where variables weights are defined.
- A new sample score can be calculated which predicts the class membership

PLS-DA includes sample group information



- covariance is an unstandardized version of the correlation (Appendix A)
- decomposition of the data matrix X in relation with the outcome Y with a set of components and loading vectors for dimension reduction

The problem to solve is:

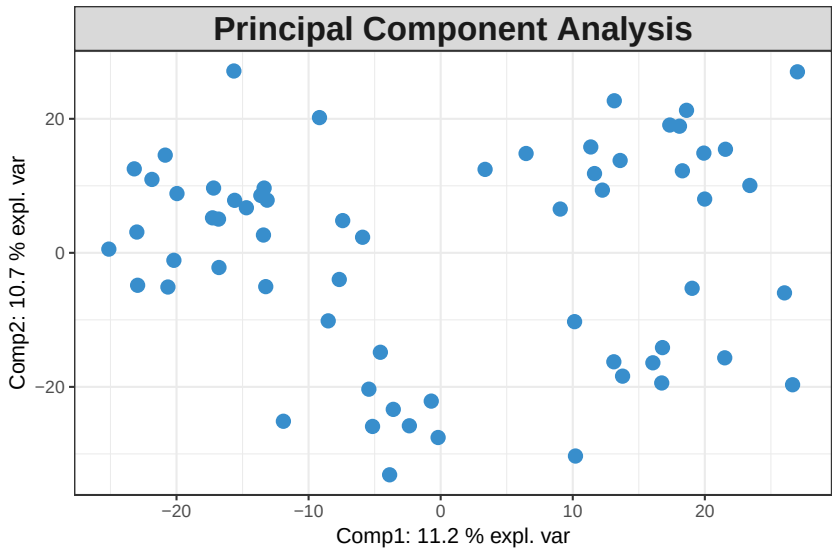
$$\max_{||\mathbf{a}||=1,||\mathbf{b}||=1} cov(X\mathbf{a}, Y\mathbf{b})$$

$\mathbf{t} = X\mathbf{a}$ and $\mathbf{u} = Y\mathbf{b}$ are the PLS-DA components.

Y is coded internally in the function as a dummy matrix with K columns, so $Y\mathbf{b}$ is a linear combination of the outcome categories

Visualisation: data projected into a small subspace spanned by the components

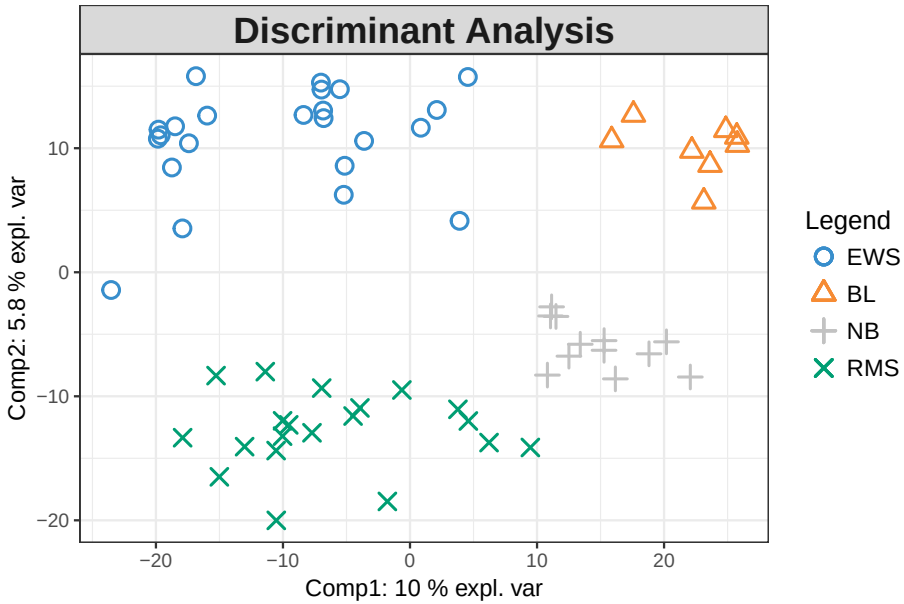
With PCA we would have: visualisation of 63 samples x 2,300 variables (genes)



- 'Similar' samples (based on their variable values) cluster
- Unsupervised exploratory analysis: no information about sample groups included in the model

Visualisation: data projected into a small subspace spanned by the components

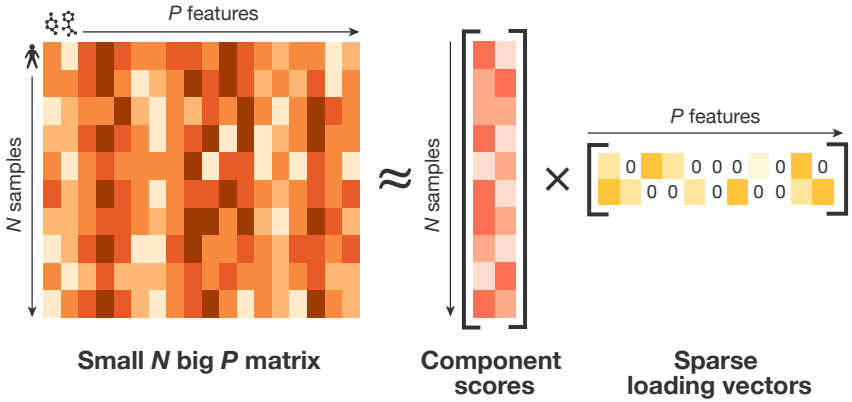
With PLS-DA: visualisation of 63 samples x 2,300 genes according to their tumour subtype



- Samples cluster according to their group
- Supervised analysis: aim is to separate/discriminate sample groups

Further dimension reduction with variable selection

Background: sparse PCA performs variable selection

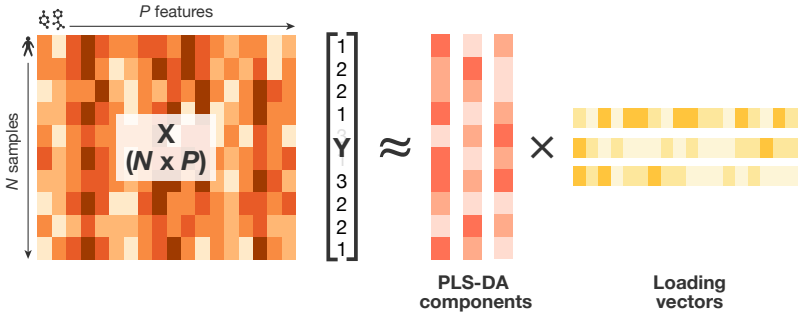


- Shrinks some variable coefficients to zero (variable selection) for each component using an optimal process with LASSO penalisations
- Each component is built on its selected variables only

Shen, H., Huang, J.Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation *J. Multivariate Analysis.*

sparse PLS-DA for variable selection

sparse PLS-DA includes internal variable selection to select the most discriminative variables.

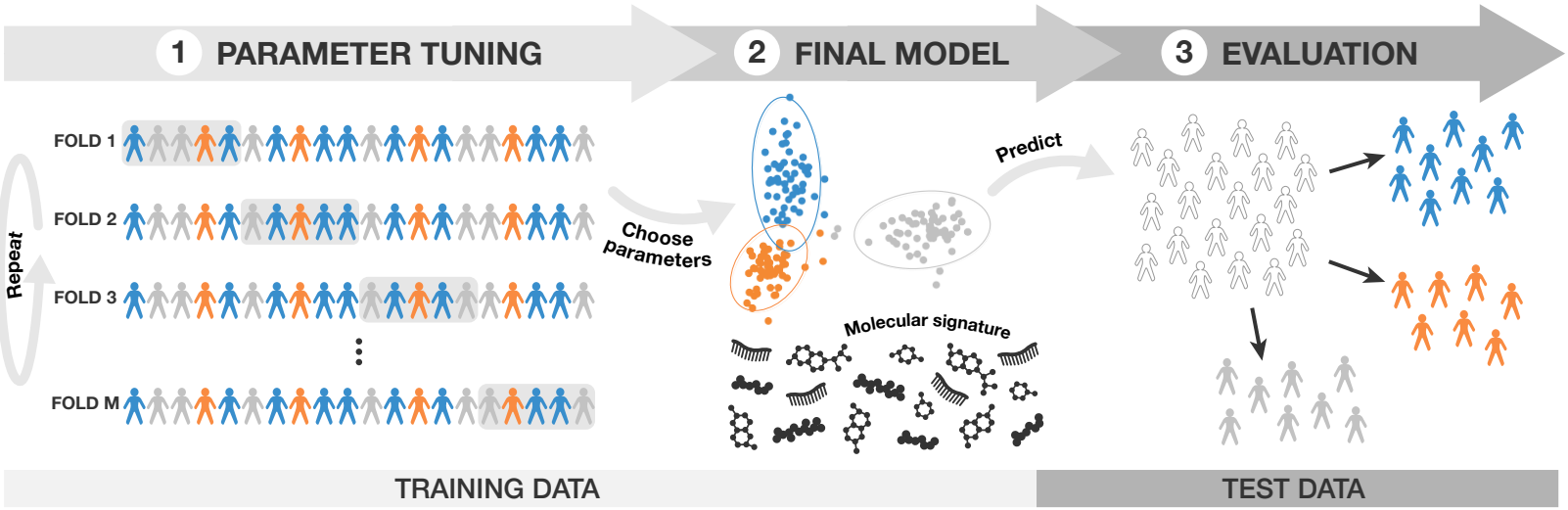


- Shrinks some variable coefficients to zero (variable selection) for each component using LASSO penalisations
- Each component is built on its selected variables only

Lê Cao et al. (2011). Sparse PLS Discriminant Analysis. *BMC Bioinformatics* 12:253.

Outline of Part 2

The PLS-DA process



1. Choose parameters based on cross-validation (CV)
2. Train the model, obtain molecular signature
3. Predict on independent test data, or evaluate performance based on CV

Parameters to choose in sPLS-DA

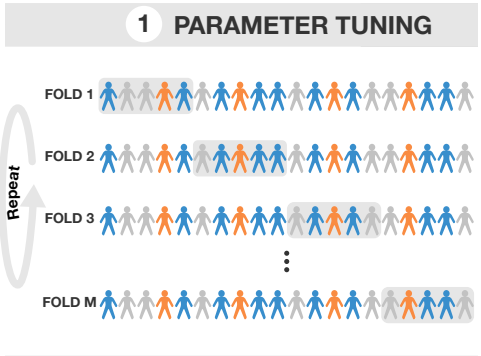
- Number of components: usually $K - 1$ is ok but needs to be checked! (K = number of sample groups / class)
  $\rightsquigarrow$ argument `ncomp`
- Number of variables to select per component
  $\rightsquigarrow$ argument `keepX`

After parameters tuning is performed, we are able to run:

```
splsda.srbct <- splsda(X, Y, ncomp = 3, keepX = c(12, 50, 35))
```

$\rightarrow$ i.e. sPLS-DA on a gene expression dataset X to discriminate 4 tumour subtypes Y with chosen parameters: 3 components, selecting 12, 50 and 35 genes on each component resp.

Parameters tuning based on cross-validation



- We divide our training samples N into M folds of equal size. Each sample is randomly allocated to a fold
- We train a PLS-DA model on $M - 1$ folds and test (predict) the class of the samples form the left-out fold

  Evaluate the performance on the test set, e.g. classification error
- We do this M times (M folds) then repeat the process several times and average the performance across folds and repeats

Class prediction of new samples

The PLSDA model is formulated as:

$$Y = X\beta + E,$$

β is the matrix of the regression coefficients and is unknown and E is the residual matrix. X and Y are matrices of predictors and outcome.

The prediction of a new sample is then

$$Y_{predicted} = X_{new}\hat{\beta},$$

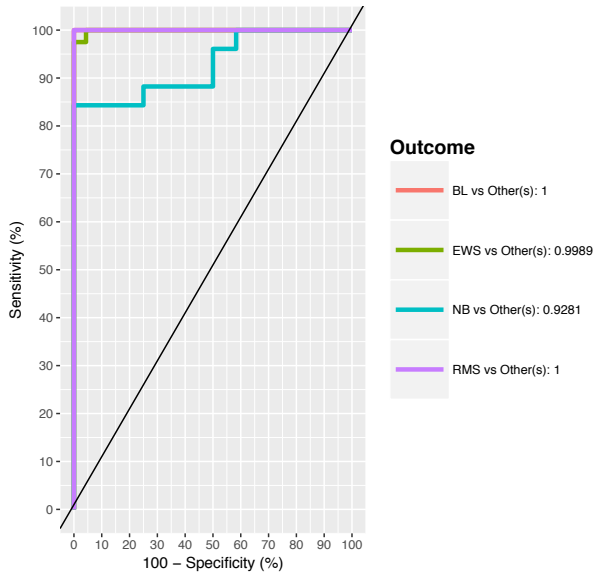
$\hat{\beta}$ estimated regression coefficient matrix, X_{new} data matrix for new samples.

Usually, $Y_{predicted}$ is a continuous numerical value (not a class number!) that is then mapped to a class membership using a prediction distance (see prac).

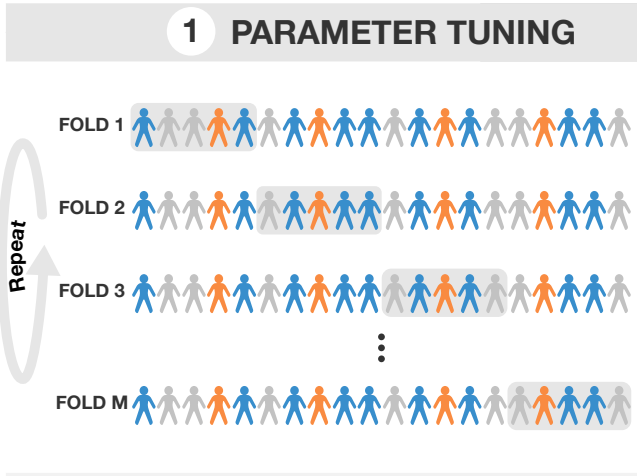
Which performance measure?

- Classification error rate = # misclassified samples/ # samples
- Balanced classification error rate to weight up minority classes
- Sensitivity: proportion of positives that are correctly predicted (e.g cases)
- Specificity: proportion of negatives that are correctly predicted (e.g. controls)

- Area Under the Curve (AUC) from a Receiver Operating Characteristic (ROC)



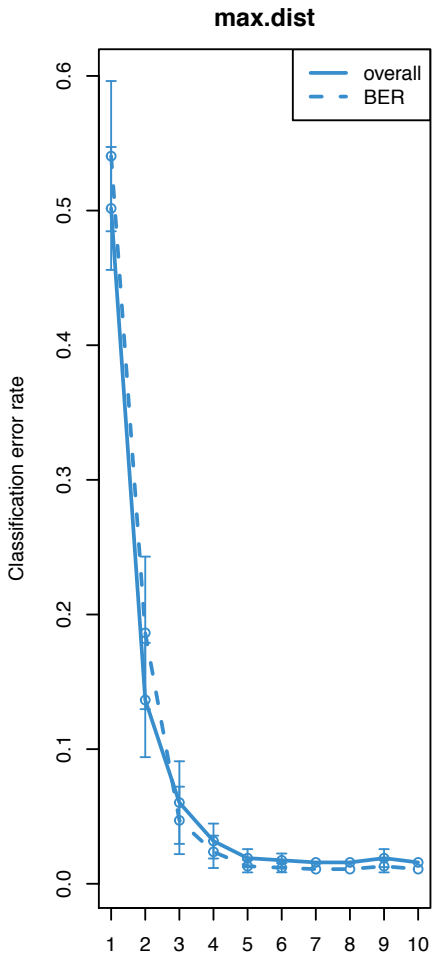
1 - Parameters tuning in practice: `ncomp`



- Set up a grid of parameters values
  e.g. we assess `ncomp` $= 1, 2, ..., 5$.
- Choose the number of folds M so that $N/M \geq 5^*$
- Choose the number of repeats ~ 50
- Look at the performance for the grid of parameters values, and choose final `ncomp` that achieves best performance

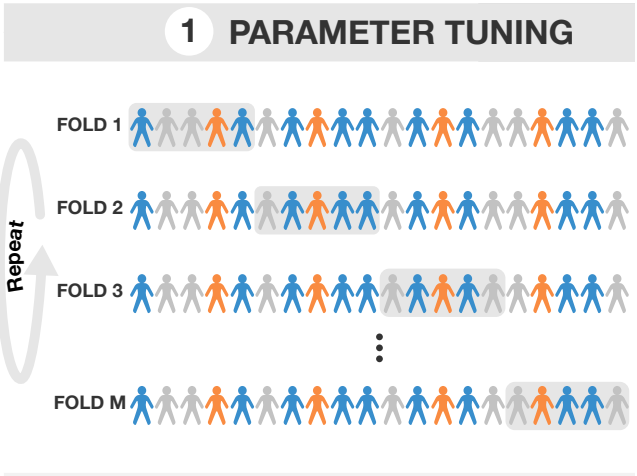
* my rule of thumb. Consider otherwise leave-one-out CV (loo)

1 - Parameters tuning in practice: `ncomp`



- Compare overall vs. balanced error rate (for minority classes)
- Mean error rate and standard error per component
- Could choose 3 or 4 components

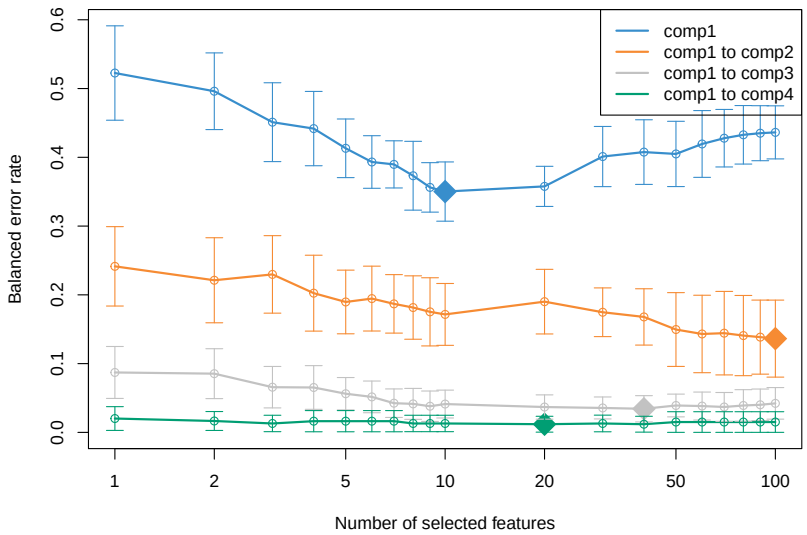
2 - Parameters tuning in practice: keepX



- Set up a grid of parameters values *per* component*
  e.g. evaluate keepX = 5, 10, 25, 50, 100.
- Choose the number of folds M and repeats
- Look at the performance for the grid of parameters values, and choose final keepX *per* component that achieves best performance

* consider going up to ncomp chosen earlier, or +1

2 - Parameters tuning in practice: `keepX`



- Mean error rate and standard error per `keepX` value
- Diamond indicates the minimum `keepX` value per component
- Error rate decreases as we add more components in the model

How many components do we really need?

The (common) pitfall of selection bias in classification

To correctly evaluate the performance of a classifier method during feature selection:

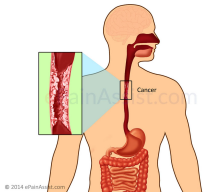
- feature selection and model training has to be evaluated against independent data,
- in other words: the test set should not been used in any way for the inference of the classifier

  Otherwise over optimistic performance otherwise a.k.a overfitting, feature selection bias

→ We use cross-validation when we cannot afford an independent test set.

Outline of Part 2

Oesophageal cancer study



Proteomics assay (129 proteins) of 40 patients, with 20 Barrett's oesophagus benign or 20 oesophageal adenocarcinoma cancer samples.

Aim: develop blood tests for detection and personalised treatment

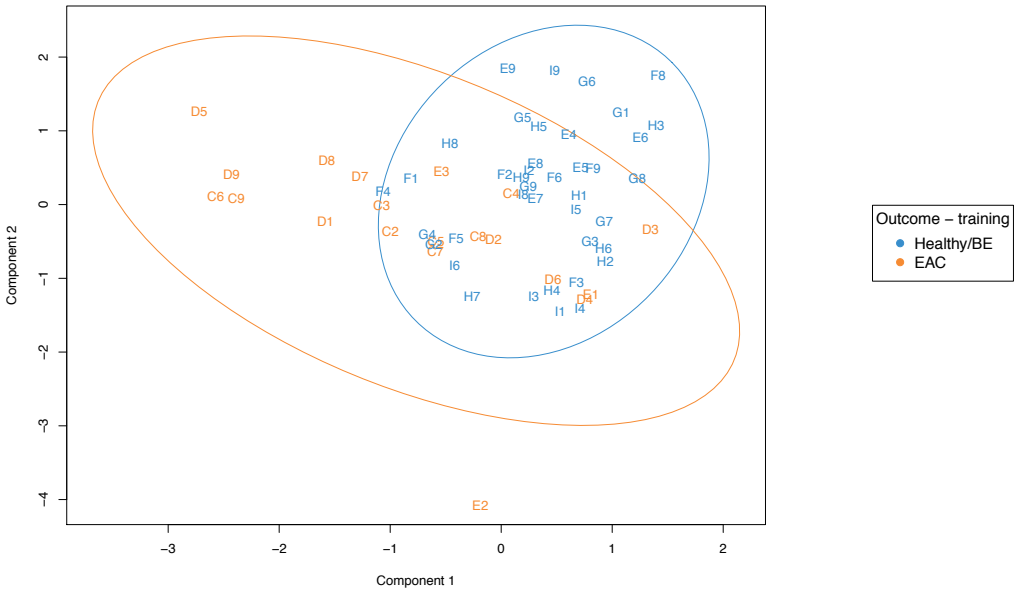
Statistical challenges:

- Small cohort (20 patients per group)
- Data range and variability with proteomics data
- Classical statistical methods fail

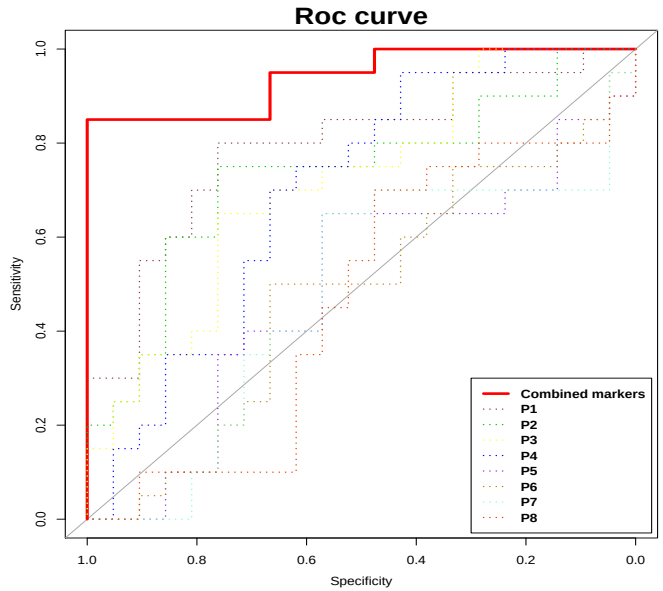
AK Shah, K-A. Lê Cao, B Gautier, MM Hill et al. (2015) Serum glycoprotein biomarker discovery and qualification pipeline reveals novel diagnostic biomarkers for oesophageal adenocarcinoma. *Mol Cell Prot 14(11)*.

Oesophageal cancer study

Training



Sample representation on a multivariate selection of 11 proteins

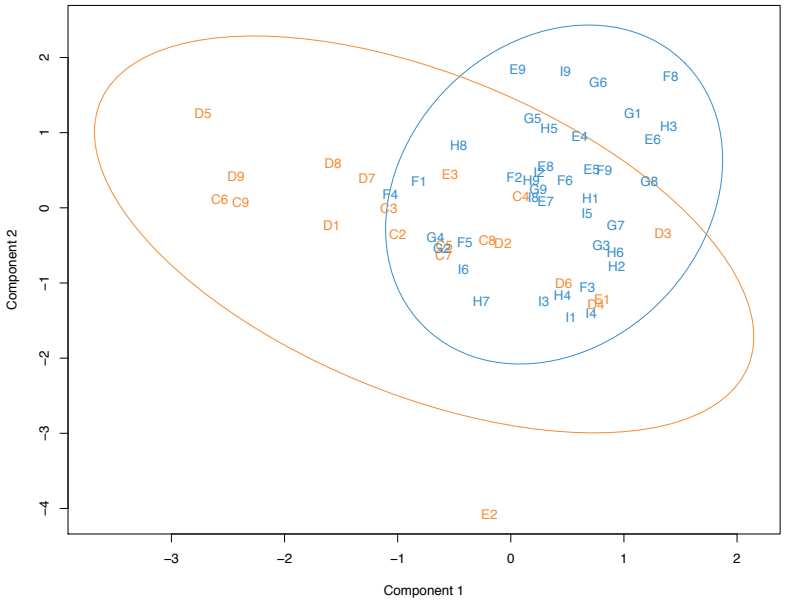


AUC of the combined 11 proteins is $>$ individual proteins

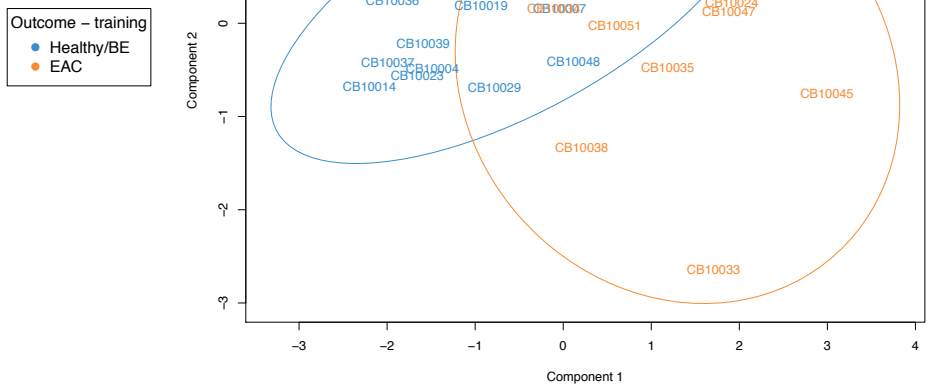
Oesophageal cancer study

Validation

Discovery cohort:
identify 11 biomarkers



Validation cohort:
predict on those same 11 markers

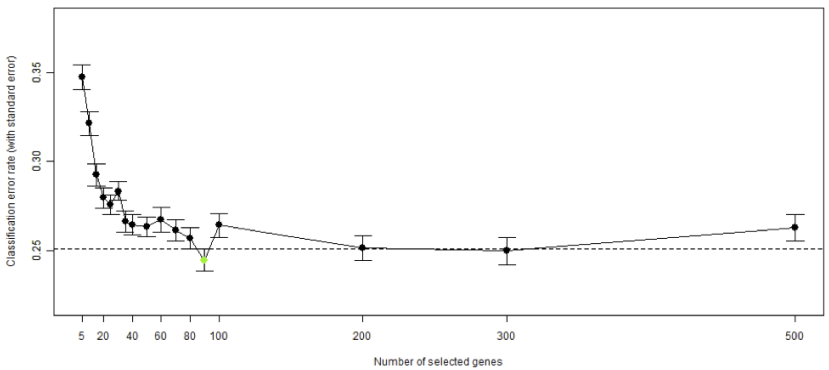


Patent: Hill M, Shah A, Lê Cao K-A (2014). Blood Test for Throat Cancer. WO2016077881A9, Australia.

Kidney transplant study

Genomics assay (p = 27K) of 40 patients with kidney transplant, with acute transplant rejection or no rejection.

Choose `keepX` on component 1 with `tune`:



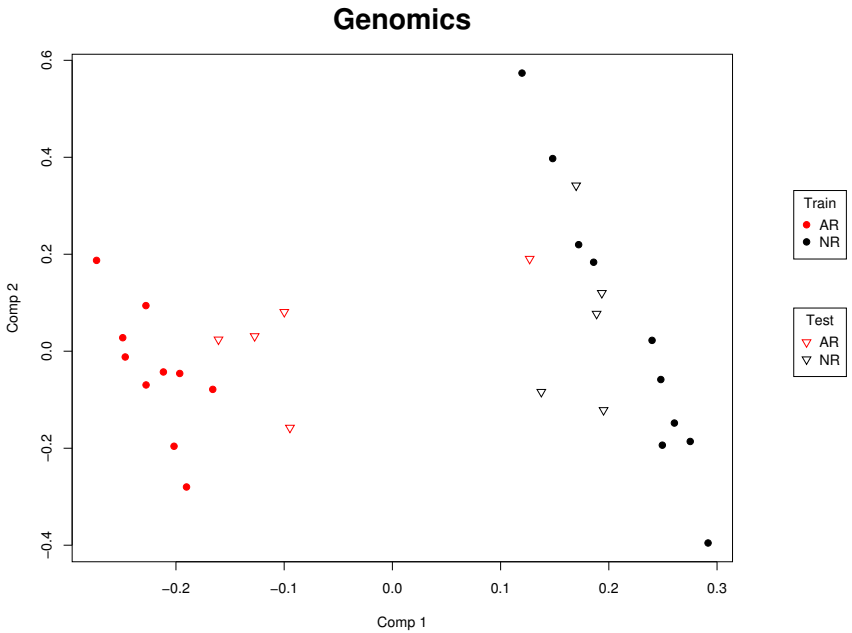
5-fold CV on $n = 26$ samples

$\rightsquigarrow$ Selection of 90 probe sets (genes related to inflammation and innate immune responses)

Günther O. , Lê Cao K-A. et al. (2014) Novel multivariate methods for integration of genomics and proteomics data: Applications in a kidney transplant rejection study, *OMICS: A journal of integrative biology*, 18(11), 682-95.

Kidney transplant study

Testing step



Color represents the true class

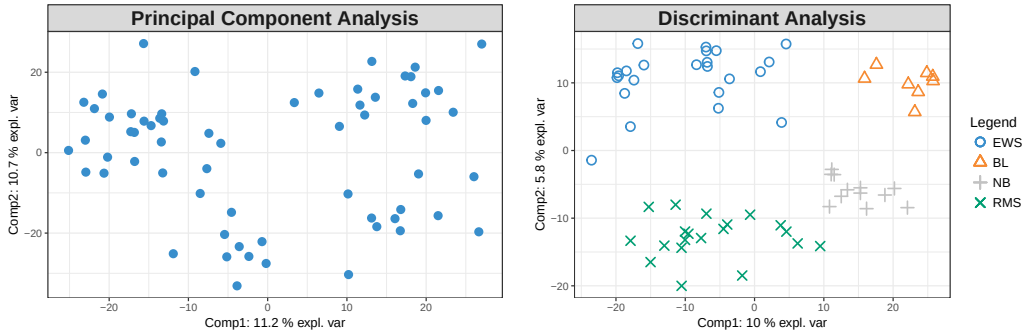
- Test samples ($n_{test} = 14$) overlaid with training sample set
- sPLS-DA fitted model (incl. 90 probe sets) prediction
- Performance assessed on test set

Classifier	**# probes**	**Error rate**	**Sensitivity**	**Specificity**	AUC
sPLS-DA	90	0.14	0.71	1	0.90
PLS-DA	27K (all)	0.21	0.57	1	0.82

Outline of Part 2

sPLS-DA summary I

- PLS-DA and sPLS-DA are supervised methods that require a categorical variable Y as input.
- Both methods aim to discriminate samples based on their group membership.
- Contrary to PCA, PLS-DA aims to find differences between sample groups rather than maximising the variance.



sPLS-DA summary II

- sPLS-DA can be used at different levels: exploratory to biomarker discovery
- May need to use cross-validation for more 'robust' results and parameter tuning
- When performing variable selection, be aware of the problem of overfitting.

Outline of Part 2

Your turn!

`srbct study` to identify a subset of gene markers discriminating types of Small Round Blue Cell Tumors in 63 patients:

- Expression levels of 2,308 genes `srbct$gene` (microarray)
- 4 types of tumours `srbct$class`

Case Study R script is in `Rscripts/casestudy_SRBCT.R`:

- Preliminary PCA to explore the data
- PLS-DA and sPLS-DA
- Optional steps: parameter tuning and prediction (load results from tuning in `RData/`

Answer the 15 Questions to reinforce your learning