Statistics for bioinformatics

Supervised classification

Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université (AMU), France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

Supervised classification - Introduction

- In a previous lesson, we presented the problem of ***clustering***, which consists in grouping objects without any a priori definition of the groups.
- The groups emerge from the clustering itself (***class discovery***). Clustering is thus ***unsupervised***.

- In some cases, one would like to focus on some pre-defined classes :
  - classifying tissues as cancer or non-cancer;
  - classifying tissues between different cancer types;
  - classifying genes according to pre-defined functional classes (e.g. metabolic pathway, different phases of the cell cycle, ...);
  - …

- In such cases, we know a priori some elements of each of the envisaged classes. We can use these elements as ***training set*** to build a classifier.
- We can reserve a subset of the known elements to build a ***testing set***, in order to evaluate the accuracy of the trained classifier.
- After training and testing, the classifier can be used later to ***assign*** new objects to the prior classes.
- This whole process is called ***supervised classification***.

Supervised classification methods

- There are many alternative methods for supervised classification
  - Discriminant analysis (linear or quadratic)
  - Bayesian classifiers
  - K-nearest neighbours (**KNN**)
  - Support Vector Machines (**SVM**)
  - Neural networks
  - ...
- Some methods rely on strong assumptions.
  - Discriminant analysis is based on an assumption of multivariate normality.
  - In addition, linear discriminant analysis assumes that all the classes have the same variance.
- The choice of the method thus depends on the structure and on the size of the data sets.
- A recurrent problem for all methods is the ***risk of over-fitting***: when the variable space contains more dimensions than the number of training objects, the classifier can be "fooled". This problem can be reduced by selecting a relevant subset of the variables (***feature selection***).

Global versus local classifiers

- As we saw for regression, classifiers can be global or local.
  - Global classifiers use the same classification rule in the whole data space. The rule is built on the whole training set.
    - Example: discriminant analysis
  - For local classifiers, a rule is made in the different sub-spaces on the basis of the neighbouring training points.
    - Example: KNN

Study case 1: discriminating acute leukemia samples: ALL versus AML

(data from Golub et al., 1999)

Cancer types (Golub, 1999)

- Golub et al. (1999) compared the profiles of expression of ~7000 human genes in patients suffering from two different cancer types: ALL or AML, respectively.
- Selected the 50 genes most correlated with the cancer type.

- The article by Golub et al. (1999) was motivated by the need to develop efficient diagnostics to predict the cancer type from blood samples of patients.
- They proposed a “molecular signature” of cancer type, allowing to discriminate ALM from ALL.
- This first “historical” study relied on somewhat arbitrary criteria to select the genes composing this signature, and the way to apply them to classify new patients.
- We will present here the classical methods used in statistics to classify “objects” (patients, genes) in pre-defined classes.

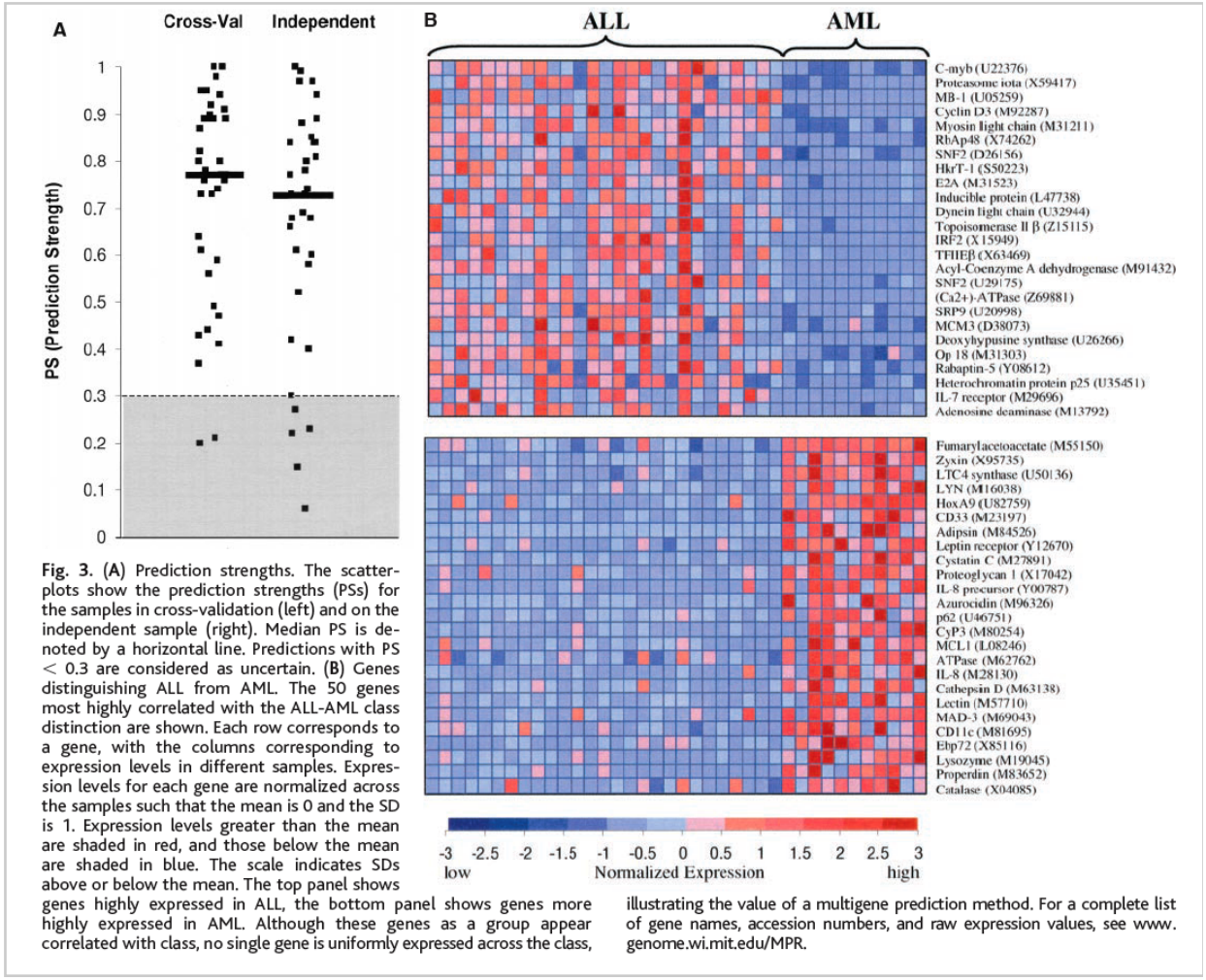
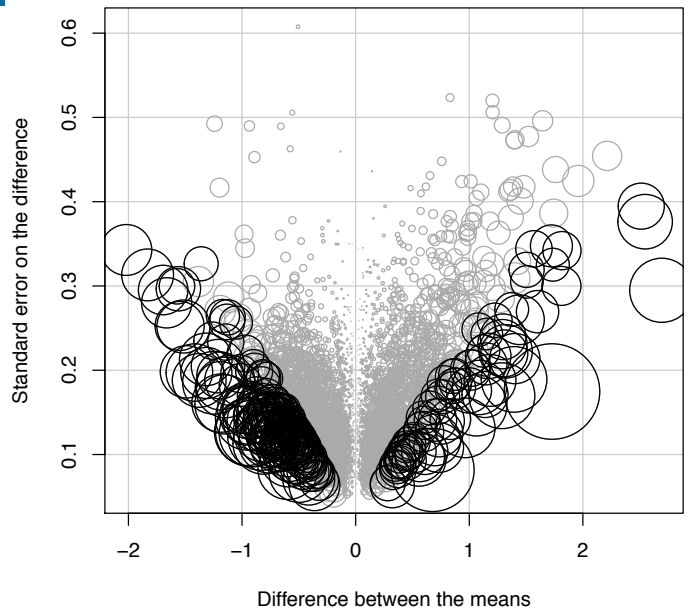


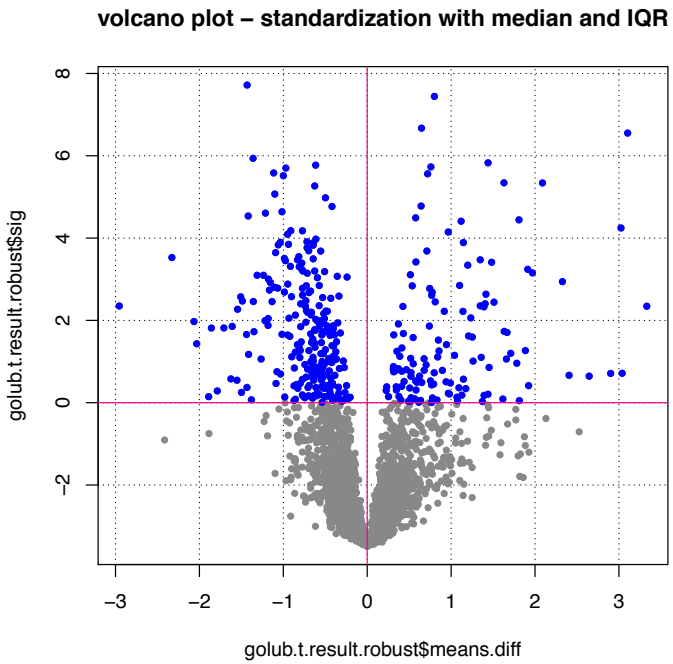
Fig. 3. (**A**) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. (**B**) Genes distinguishing ALL from AML. The 50 genes most highly correlated with the ALL-AML class distinction are shown. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. Expression levels for each gene are normalized across the samples such that the mean is 0 and the SD is 1. Expression levels greater than the mean are shaded in red, and those below the mean are shaded in blue. The scale indicates SDs above or below the mean. The top panel shows genes highly expressed in ALL, the bottom panel shows genes more highly expressed in AML. Although these genes as a group appear correlated with class, no single gene is uniformly expressed across the class, illustrating the value of a multigene prediction method. For a complete list of gene names, accession numbers, and raw expression values, see www.genome.wi.mit.edu/MPR.

- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-7.

Golub et al (1999)

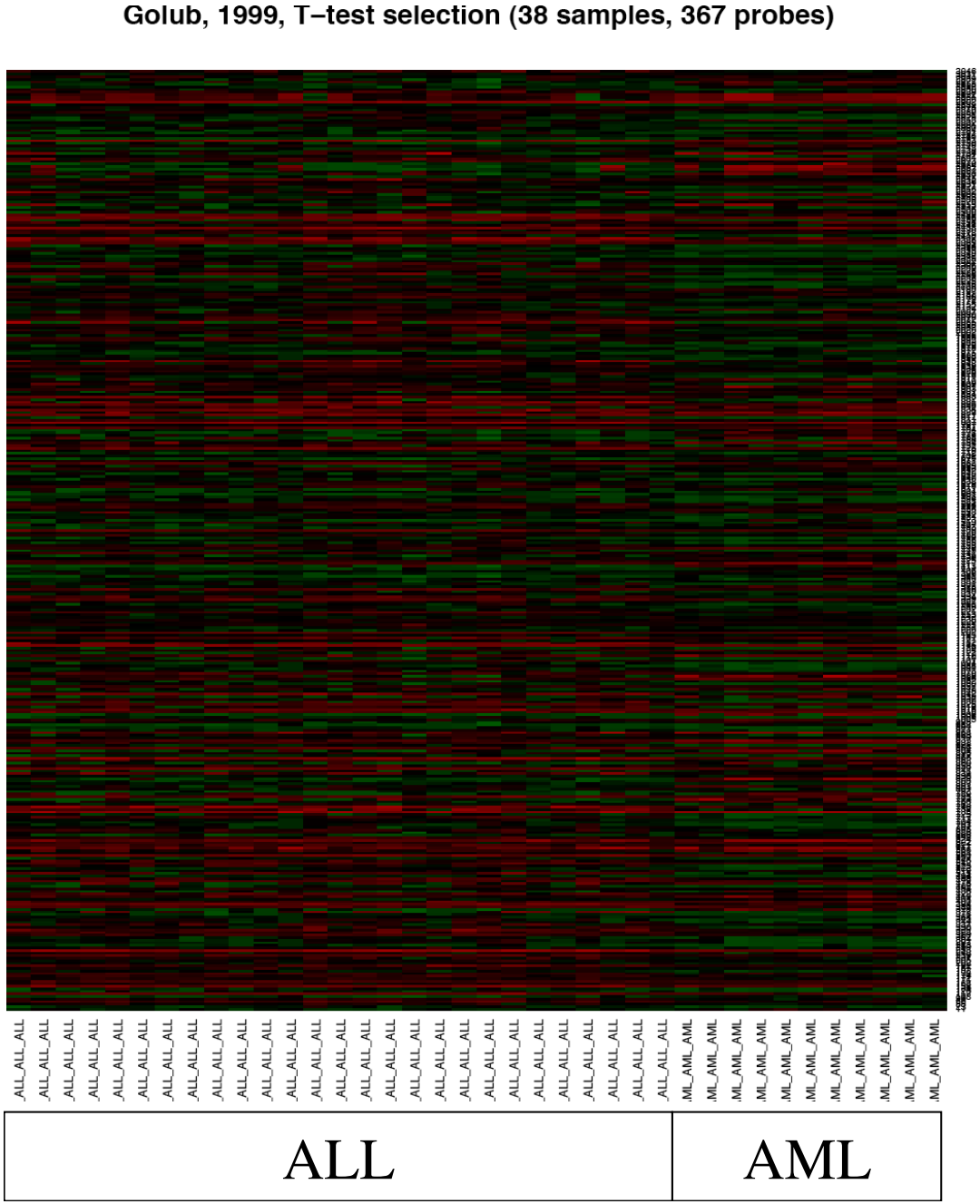
- Data source: Golub et al (1999). First historical publication searching for molecular signatures of cancer type.
- Training set
  - 38 samples from 2 types of leukemia
    - 27 Acute lymphoblastic leukemia (note: 2 subtypes: ALL-T and ALL-B)
    - 11 Acute myeloid leukemia
  - Original data set contains ~7000 genes
  - Filtering out poorly expressed genes retains 3051 genes
- We re-analyze the data using different methods.
- Selection of differentially expressed genes (***DEG***)
  - Welch t-test with robust estimators (median, IQR) retains 367differentially expressed genes with E-value <= 1.
  - Top plot: circle radius indicates T-test significance.
  - Bottom plot (volcano plot):
    - sig = -log10(E-value) >= 0





- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531-7.

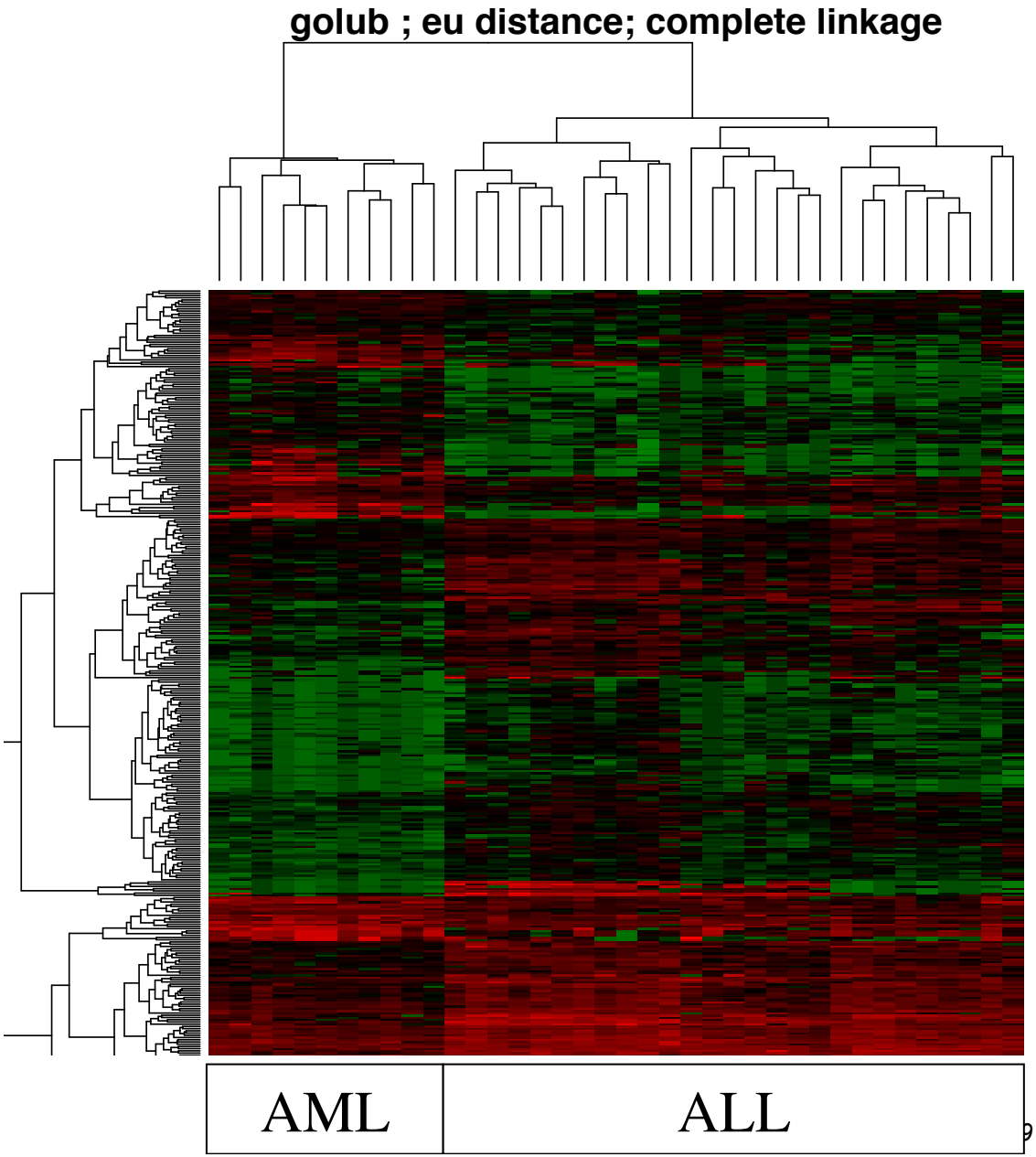
Golub 1999 - Profiles of selected genes



- The 367 gene selected by the T-test have apparently different profiles.
  - Some genes seem greener for the ALL patients (27 leftmost samples).
  - Some genes seem greener for the AML patients (11 rightmost samples).
- This image is however hard to interpret, because genes are not sorted by any specific criterion.

Golub – hierarchical vlustering of DEG genes / profiles

- Hierarchical clustering perfectly separates the two cancer types (AML versus ALL).
- This perfect separation is observed for various metrics (Euclidian, correlation, dot product) and agglomeration rules (complete, average, Ward).
- Sample clustering further reveals subgroups of ALL.
- Gene clustering reveals 4 groups of profiles:
  - AML red, ALL green
  - AML green, ALL red
  - Overall green, stronger in AML
  - Overall red, stronger in ALL

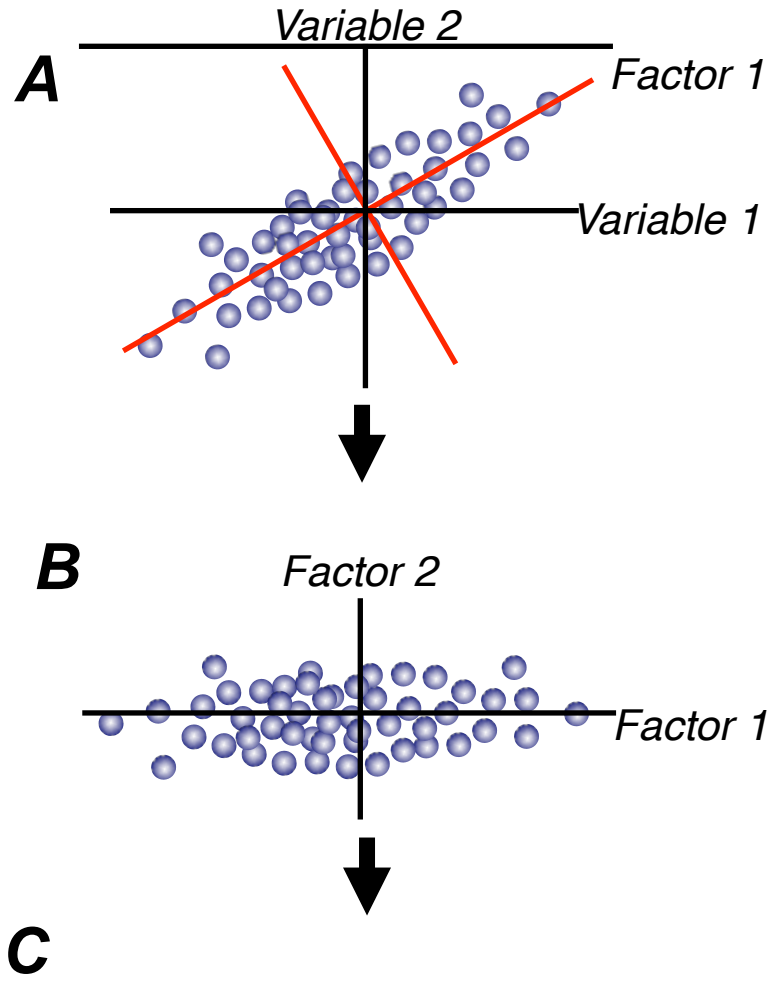


Principal component analysis – principle of the method

A. Multidimensional data
   - n objects, p variables (in this case p=2)

B. Principal components
   - n objects, p factors
   - Each factor is a linear combination of variables

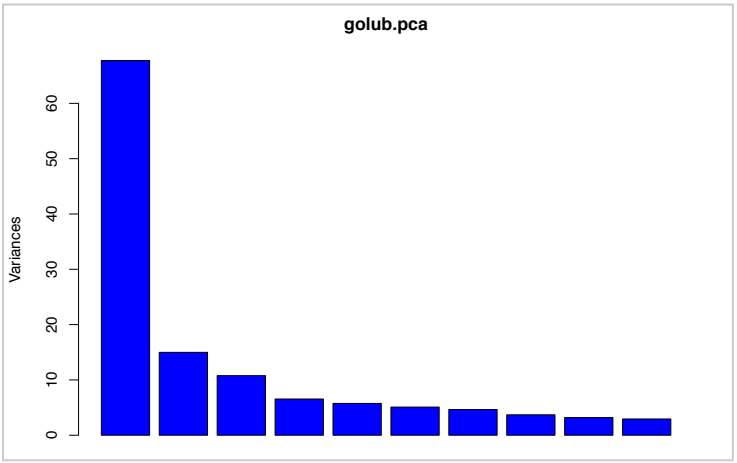
C. Reduction in dimensions
   - Selection of a subset of principal components
   - q factors, with q < p (in this case, q=1)

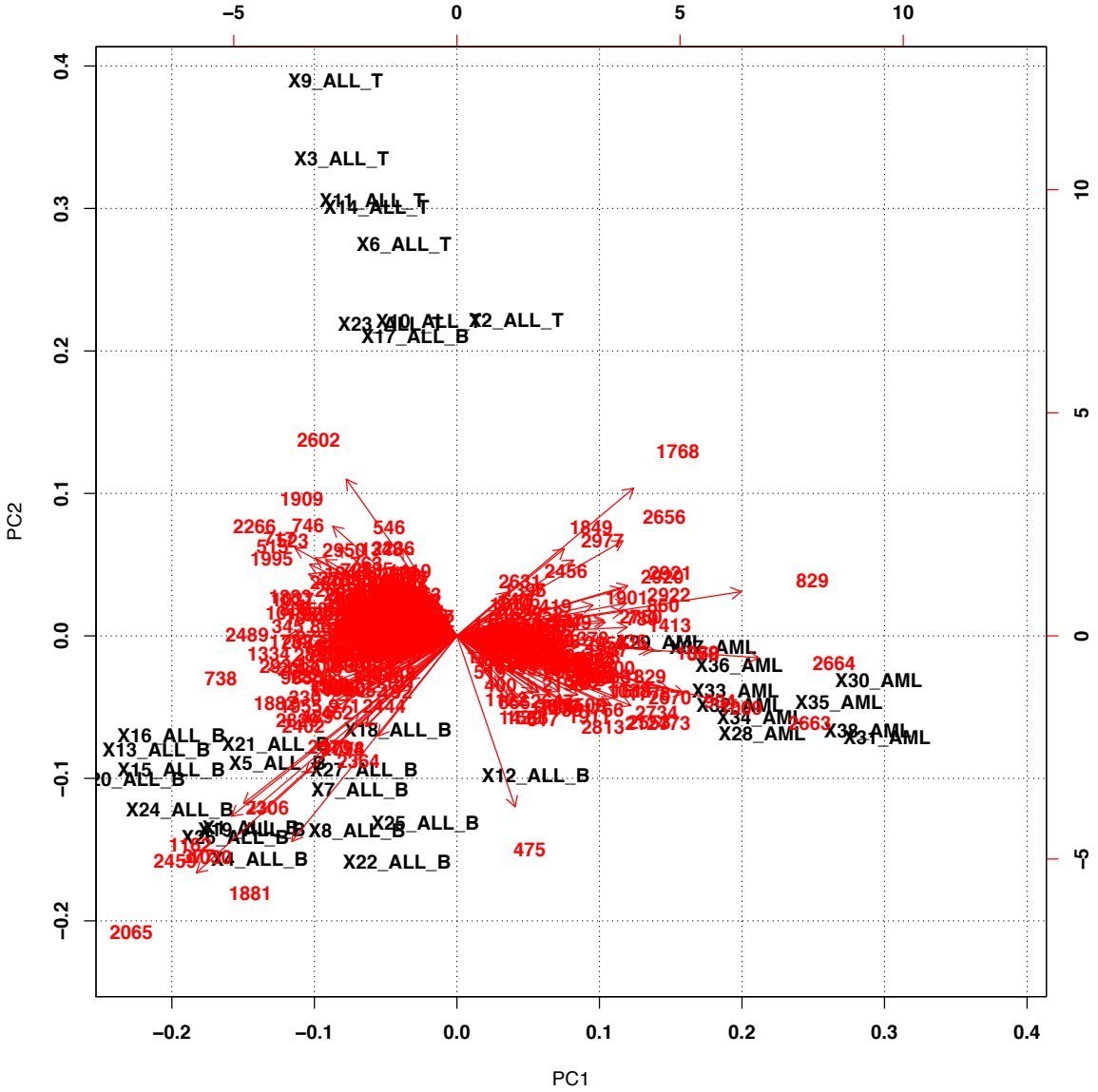


- Gilbert, D., Schroeder, M. & van Helden, J. (2000). Trends in Biotechnology 18) 487-495.

Principal component analysis (PCA)

- Principal component analysis (PCA) relies on a transformation of a multivariate table into a multi-dimensional table of “components”.
- With Golub dataset,
  - Most variance is captured by the first component.
  - The first component (Y axis) clearly separates ALL from AML patients.
  - The second component splits the AML set into two well-separated groups, which correspond almost perfectly to T-cells and B-cells, resp.





Study case 2: classifying ALL subtypes (data from Den Boer et al., 2009)

Den Boer et al., 2009 : procedure

- Den Boer et al (2009) use Affymetrix microarrays to characterize the transcriptome of 190 Acute Lymphoblastic Leukemia of different types.
- They use these profiles to select “transcriptome signatures” that will serve for diagnostics purposes: assigning new samples to one of the cancer types.
- They apply an elaborate procedure relying on an inner and an outer loop of cross-validation.

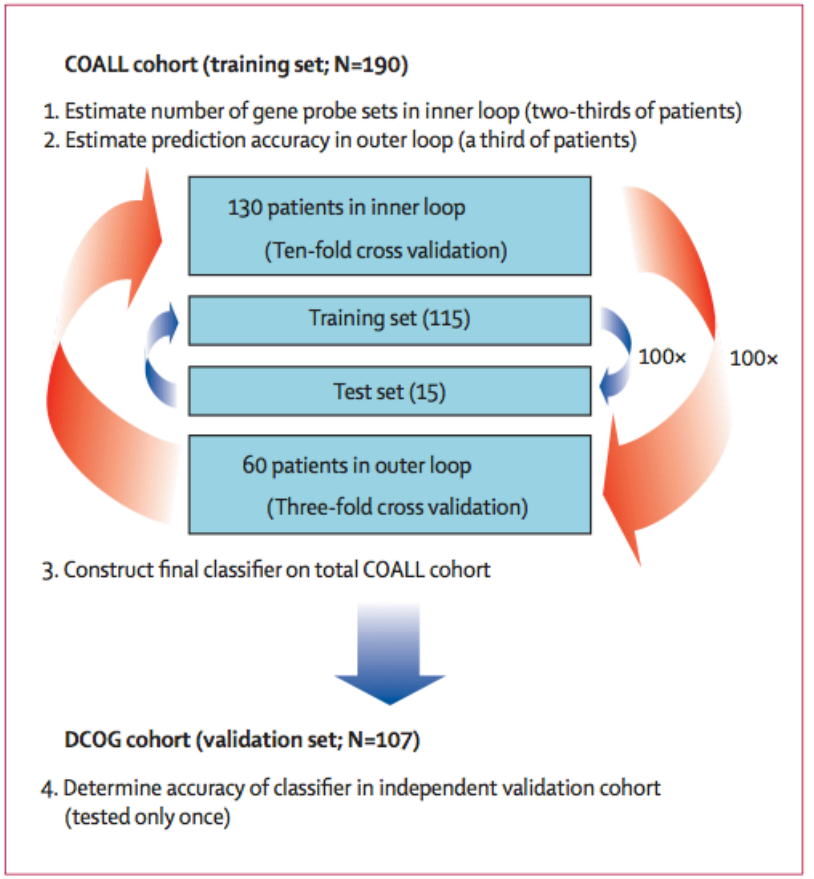


Figure 1: **Identification of a gene-expression signature enabling classification of paediatric ALL**

hyperdiploid	44
pre-B ALL	44
TEL-AML1	43
T-ALL	36
E2A-rearranged (EP)	8
BCR-ABL	4
E2A-rearranged (E-sub)	4
MLL	4
BCR-ABL + hyperdiploidy	1
E2A-rearranged (E)	1
TEL-AML1 + hyperdiploidy	1

- Data source: Den Boer et al. 2009. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol 10(2): 125-134.

Den Boer 2009 - The transcriptomic signature

- The training procedure was used to select 100 genes whose combined expression levels can be used to assign samples to cancer subtypes.
- The heatmaps show that the selected genes are differentially expressed
  - between subtypes of the training set (left);
  - between subtypes of an independent testing set (right).

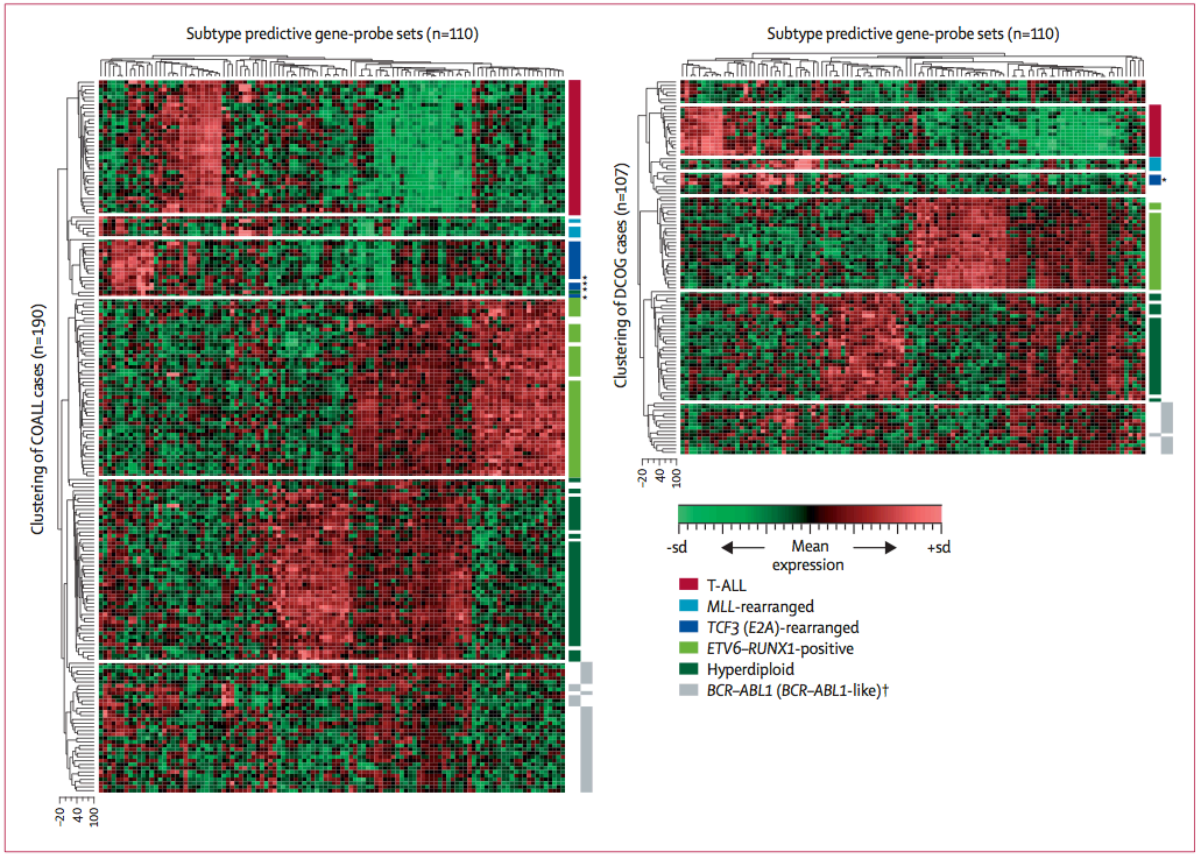


Figure 2: **Clustering of ALL subtypes by gene-expression profiles**
Hierarchical clustering of patients from the COALL (left) and DCOG (right) studies with 110 gene-probe sets selected to classify paediatric ALL. Heat map shows which gene-probe sets are overexpressed (in red) and which gene probe sets are underexpressed (in green) relative to mean expression of all gene-probe sets (see scale bar). *Patients with *E2A*-rearranged subclone (15–26% positive cells). †Right column of grey bar denotes *BCR–ABL1*-like cases.

- Den Boer et al. 2009. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol 10(2): 125-134.

Den Boer 2009 - Accuracy of the classifier

- The signature has an excellent diagnostic value: for the well-represented cancer types, the sensitivity and specificity are >90%.
- **Note:** "accuracy" is misleading some subtypes have 98% accuracy with 0% sensitivity, because some classes are represented by a very small number of samples.

hyperdiploid	44
pre-B ALL	44
TEL-AML1	43
T-ALL	36
E2A-rearranged (EP)	8
BCR-ABL	4
E2A-rearranged (E-sub)	4
MLL	4
BCR-ABL + hyperdiploidy	1
E2A-rearranged (E)	1
TEL-AML1 + hyperdiploidy	1

Sn = TP / (TP + FN)

Sp = TN / (TN + FP)

PPV = TP / (VP + FP)

	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	Accuracy (%)
T-lineage ALL	100 (100–100)	100 (100–100)	100 (100–100)	100 (100–100)	100 (100–100)
ETV6-RUNX1-positive	100 (100–100)	97·8 (95·7–97·8)	93·3 (87·5–93·3)	100 (100–100)	98·3 (96·7–98·3)
Hyperdiploid	100 (92·9–100)	97·8 (95·7–97·8)	92·6 (86·7–93·3)	100 (97·8–100)	96·7 (95·0–98·3)
E2A-rearranged	100 (75·0–100)	100 (98·2–100)	100 (80·0–100)	100 (98·2–100)	98·3 (98·3–100)
BCR–ABL1-positive	0 (0–0)	100 (100–100)	0 (0–0)	98·3 (98·3–98·3)	98·3 (98·3–98·3)
MLL-rearranged	0 (0–0)	100 (100–100)	0 (0–0)	98·3 (98·3–98·3)	98·3 (98·3–98·3)
Overall values	93·5 (93·5–95·7)	78·6 (78·6–85·7)	93·6 (93·2–95·6)	80·0 (76·4–84·6)	90·0 (88·3–91·7)

Data from the COALL study. Data are median (25th–75th percentile). Accuracy is for 100 iterations that include 130 cases to build the classifier and 60 other patients to determine the diagnostic test values in each interation (three-fold cross validation). Overall values based on the classification of all cases, including the B-other group.

Table 1: **Diagnostic test values for the classification of acute lymphoblastic leukaemia by three-fold cross-validation approach**

	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy
T-lineage ALL	15/15 (100%)	92/92 (100%)	15/15 (100%)	92/92 (100%)	107/107 (100%)
ETV6–RUNX1-positive	24/24 (100%)	81/83 (97·6%)	24/26 (92·3%)	81/81 (100%)	105/107 (98·1%)
Hyperdiploid	28/28 (100%)	74/79 (93·7%)	28/33 (84·8%)	74/74 (100%)	102/107 (95·3%)
E2A-rearranged	2/2 (100%)	104/105 (99·0%)	2/3 (66·7%)	104/104 (100%)	106/107 (99·1%)
BCR–ABL1-positive	0/1 (0%)	106/106 (100%)	0/0	106/107 (99·1%)	106/107 (99·1%)
MLL-rearranged	0/4 (0%)	103/103 (100%)	0/0	103/107 (96·3%)	103/107 (96·3%)
Overall values	69/74 (93·2%)	25/33 (75·8%)	69/77 (89·6%)	25/30 (83·3%)	94/107 (87·9%)

Data are number of predicted cases/total per subtype (%). DCOG cohort (107 patients) used to validate independently the predictive value of classification by gene expression signature (tested only once). Overall values based on the classification of all cases, including the B-other group. The specificity, positive predictive value, and accuracy are 100% for *E2A*-rearranged cases if the B-other case with an *E2A*-rearranged subclone (21% positive cells) is included as true positive case (webappendix).

Table 2: **Diagnostic test values for independent validation group**

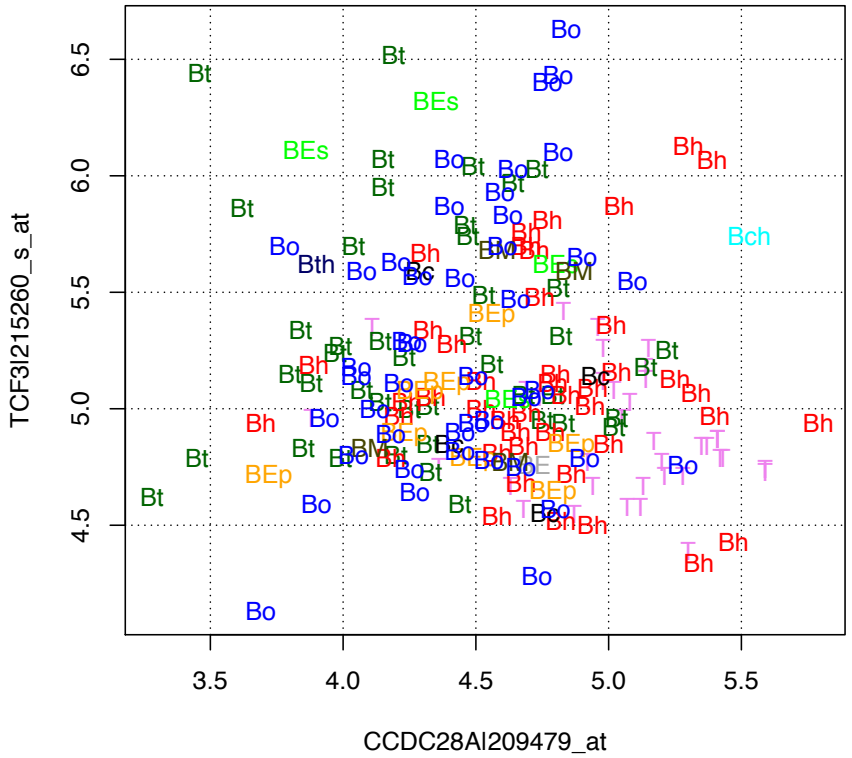
- Den Boer et al. 2009. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol 10(2): 125-134.

Den Boer 2009 - Exploring some profiles

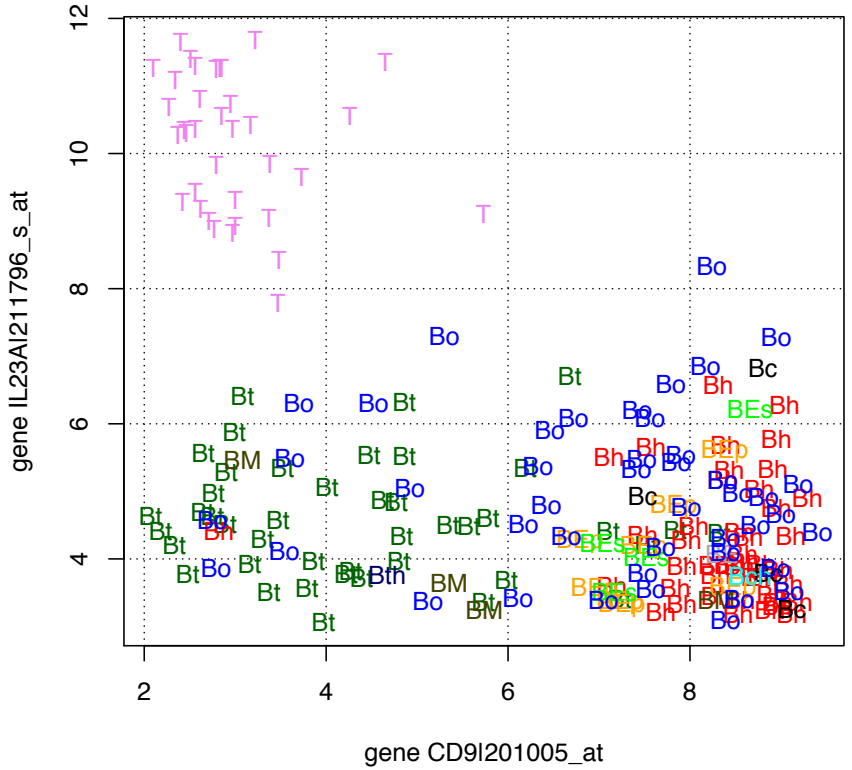
- Left: expression for 2 genes selected at random. Each symbol represents one sample, coloured by cancer type. All cancer types are intermingled.
- Right: expression of the 2 genes with the highest sample-wise variance. The first gene (CD9) separates cell types T and Bt (low expression) from Bh, Bep, Br (high expression). Bo is dispersed over the whole range.
- Question: how can we identify a combination of genes that discriminate the different subtypes as well as possible ?

Bh	hyperdiploid	44
Bo	pre-B ALL	44
Bt	TEL-AML1	43
T	T-ALL	36
BEp	E2A-rearranged (EP)	8
Bc	BCR-ABL	4
BEs	E2A-rearranged (E-sub)	4
BM	MLL	4
Bch	BCR-ABL + hyperdiploidy	1
BE	E2A-rearranged (E)	1
Bth	TEL-AML1 + hyperdiploidy	1

Den Boer (2009), randomly selected genes



2 genes with the highest variance



Discriminant analysis : methodological principles

Multivariate data with a nominal criterion variable

- One disposes of a set of objects (the **sample**) which have been previously assigned to predefined classes.
- Each object is characterized by a series of quantitative variables (the **predictors**), and its class is indicated in a separated column (the **criterion variable**).

	Predictor variables				**Criterion variable**
	variable 1	**variable 2**	**…**	**variable p**	**class**
object 1	$x_{1,1}$	$x_{2,1}$	…	$x_{p,1}$	**A**
object 2	$x_{1,2}$	$x_{2,2}$	…	$x_{p,2}$	**A**
object 3	$x_{1,3}$	$x_{2,3}$	…	$x_{p,3}$	**A**
…	…	…	…	…	**…**
object i	$x_{1,i}$	$x_{2,i}$	…	$x_{p,i}$	**B**
object i+1	$x_{1,i+1}$	$x_{2,i+1}$	…	$x_{p,i+1}$	**B**
object i+2	$x_{1,i+2}$	$x_{2,i+2}$	…	$x_{p,i+2}$	**B**
…	…	…			
object n-1	$x_{1,n-1}$	$x_{2,n-1}$	…	$x_{p,n-1}$	**K**
object n	$x_{1,n}$	$x_{2,n}$	…	$x_{p,n}$	**K**

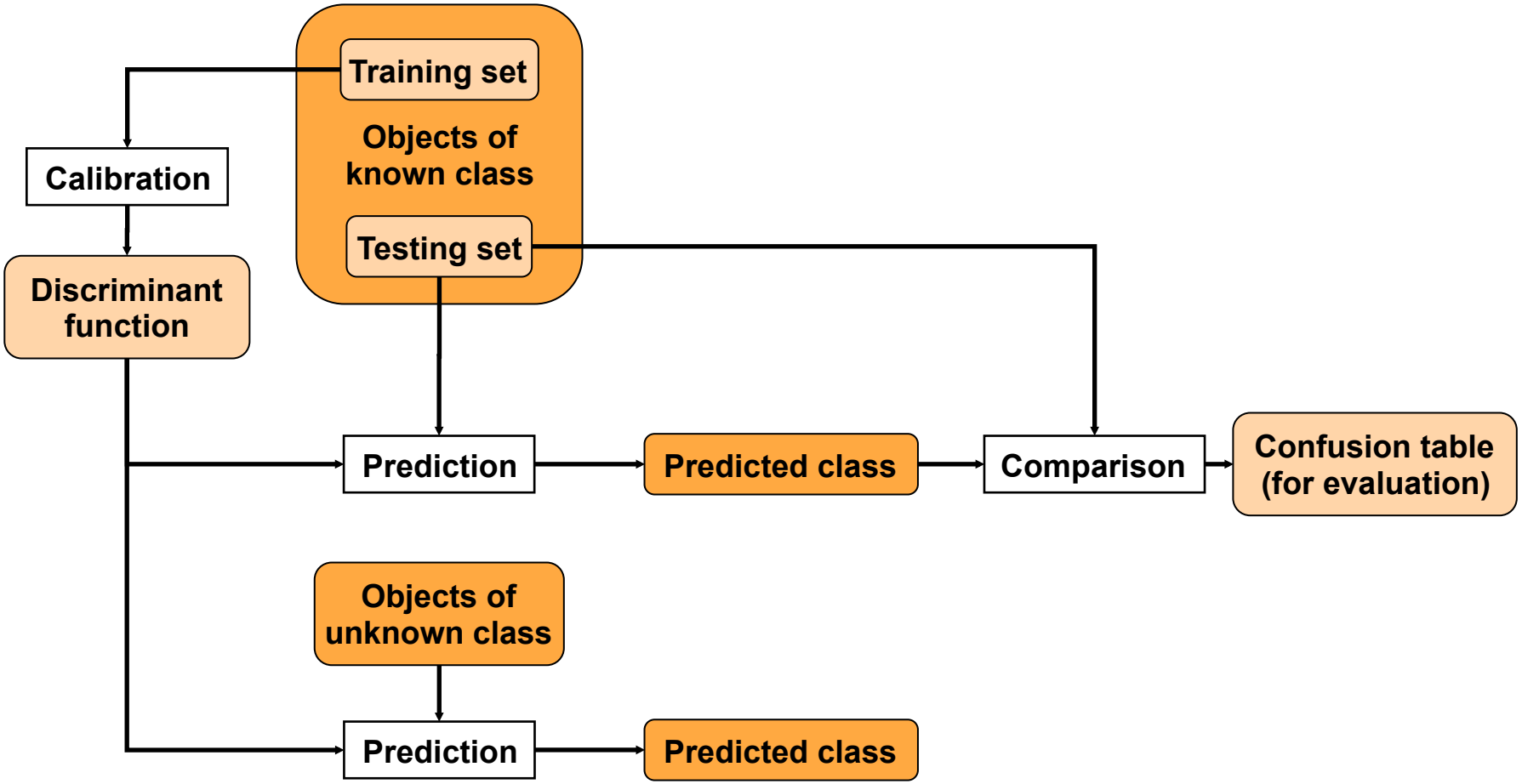
Discriminant analysis - training and prediction

- **Training phase**
  - The sample is used to build a discriminant function.
- **Testing phase**
  - The quality of the discriminant function is evaluated on an independent data set.
- **Prediction phase**
  - The discriminant function is used to predict the value of the criterion variable for new objects.

	Predictor variables				Criterion variable
	variable 1	**variable 2**	**…**	**variable p**	**class**
object 1	x_{11}	x_{21}	…	x_{p1}	**A**
object 2	x_{12}	x_{22}	…	x_{p2}	**A**
object 3	x_{13}	x_{23}	…	x_{p3}	**B**
…	…	…	…	…	**…**
object n_{train}	x_{1n}	x_{2n}	…	x_{pn}	**K**

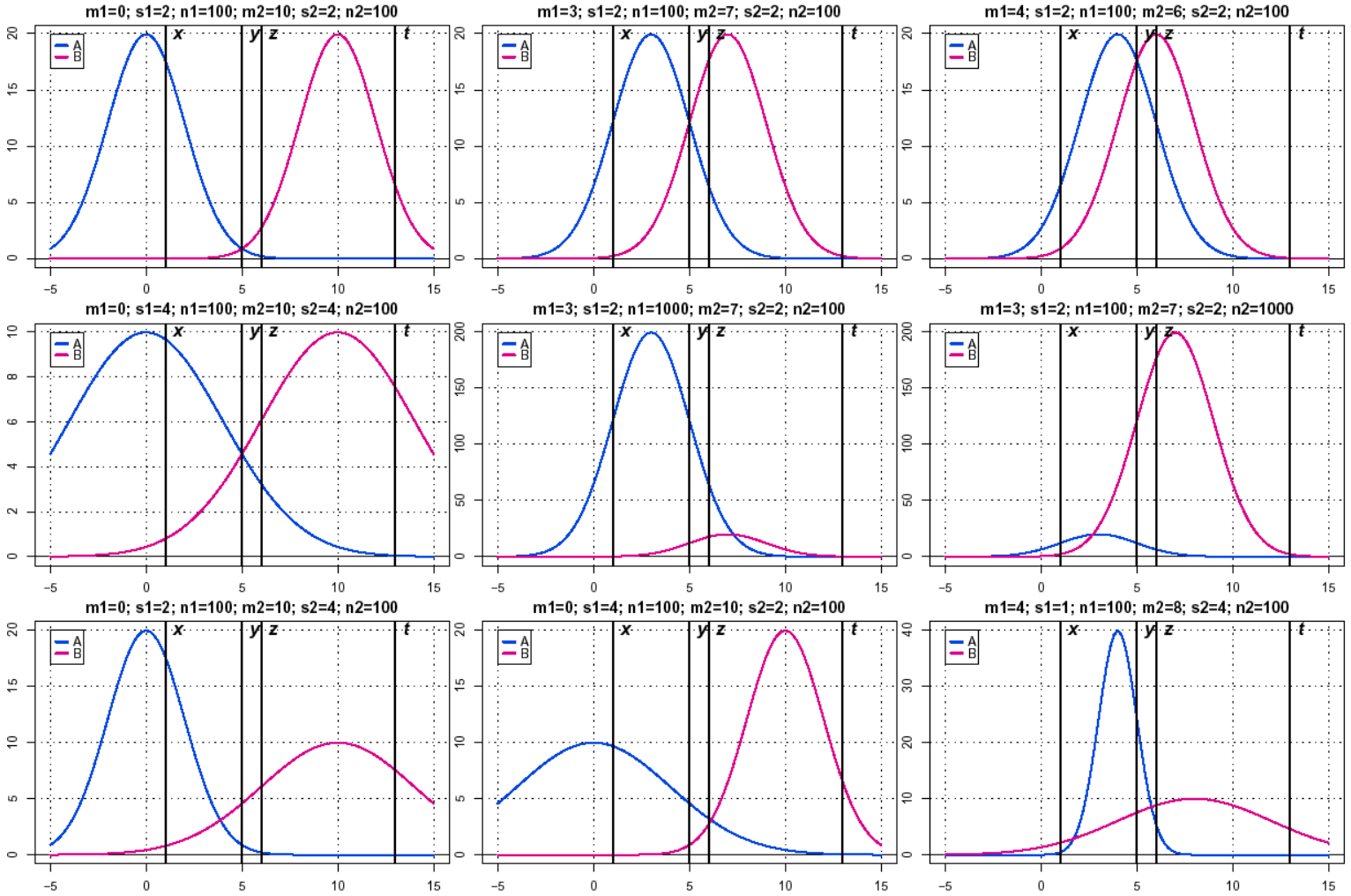
	Predictor variables				Criterion variable
	variable 1	**variable 2**	**…**	**variable p**	**class**
object 1	x_{11}	x_{21}	…	x_{p1}	**?**
object 2	x_{12}	x_{22}	…	x_{p2}	**?**
object 3	x_{13}	x_{23}	…	x_{p3}	**?**
…	…	…	…	…	**…**
object n_{pred}	x_{1n}	x_{2n}	…	x_{pn}	**?**

Discriminant analysis



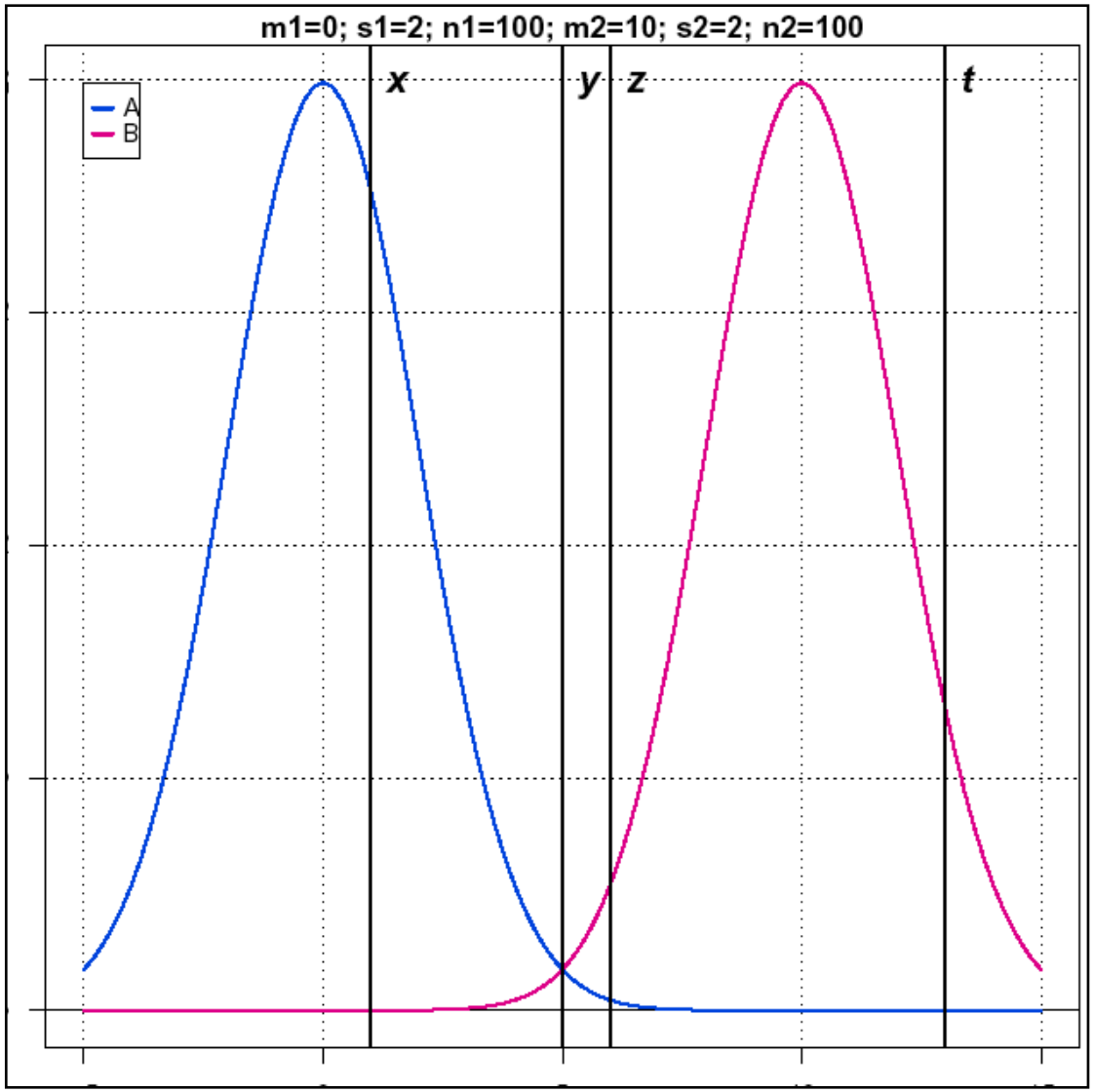
Conceptual illustration with a single predictor variable

- Given two predefined classes (A and B), try intuitively to assign a class to each new object (X positions denoted by vertical black bars).
- How confident do you feel for each of your predictions ?



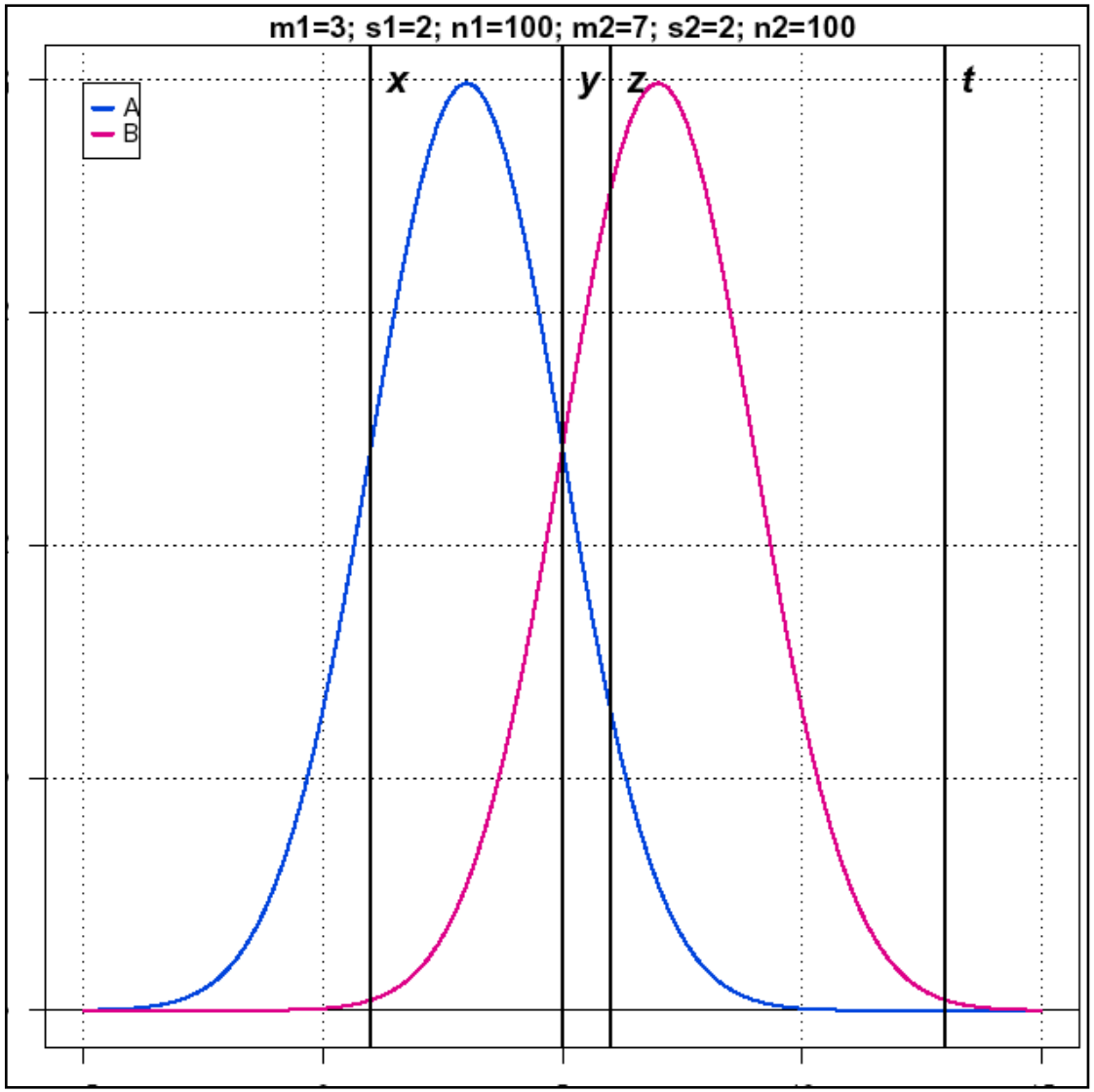
- What is the effect of the respective means ?
- What is the effect of the respective standard deviations ?
- What is the effect of the population sizes ?

Conceptual illustration with a single variable



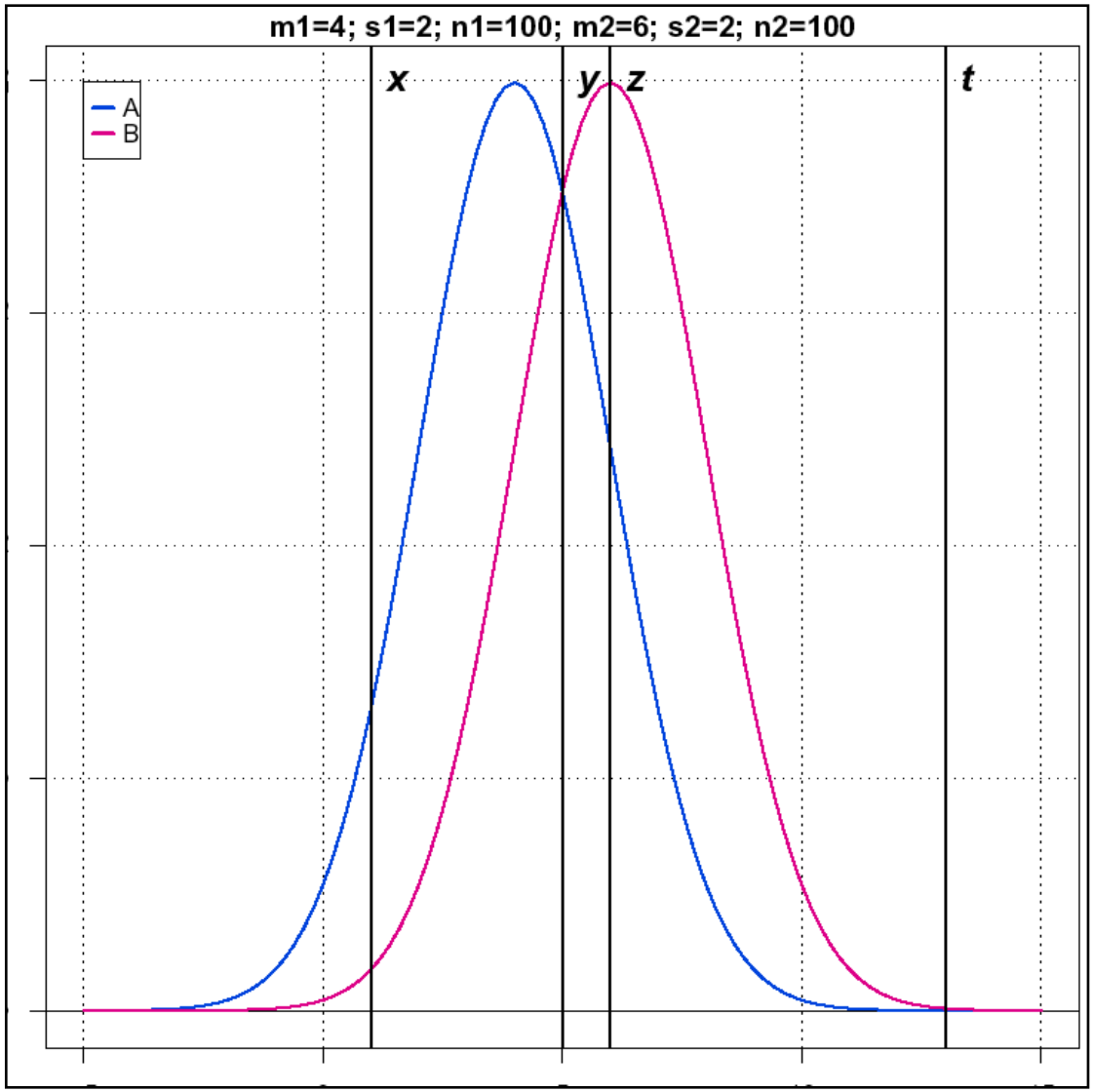
- In this conceptual example, the two populations have the same mean and variance.
- To which group (A or B) would you assign the points at coordinate x, y, z, t, respectively ?

Conceptual illustration with a single variable



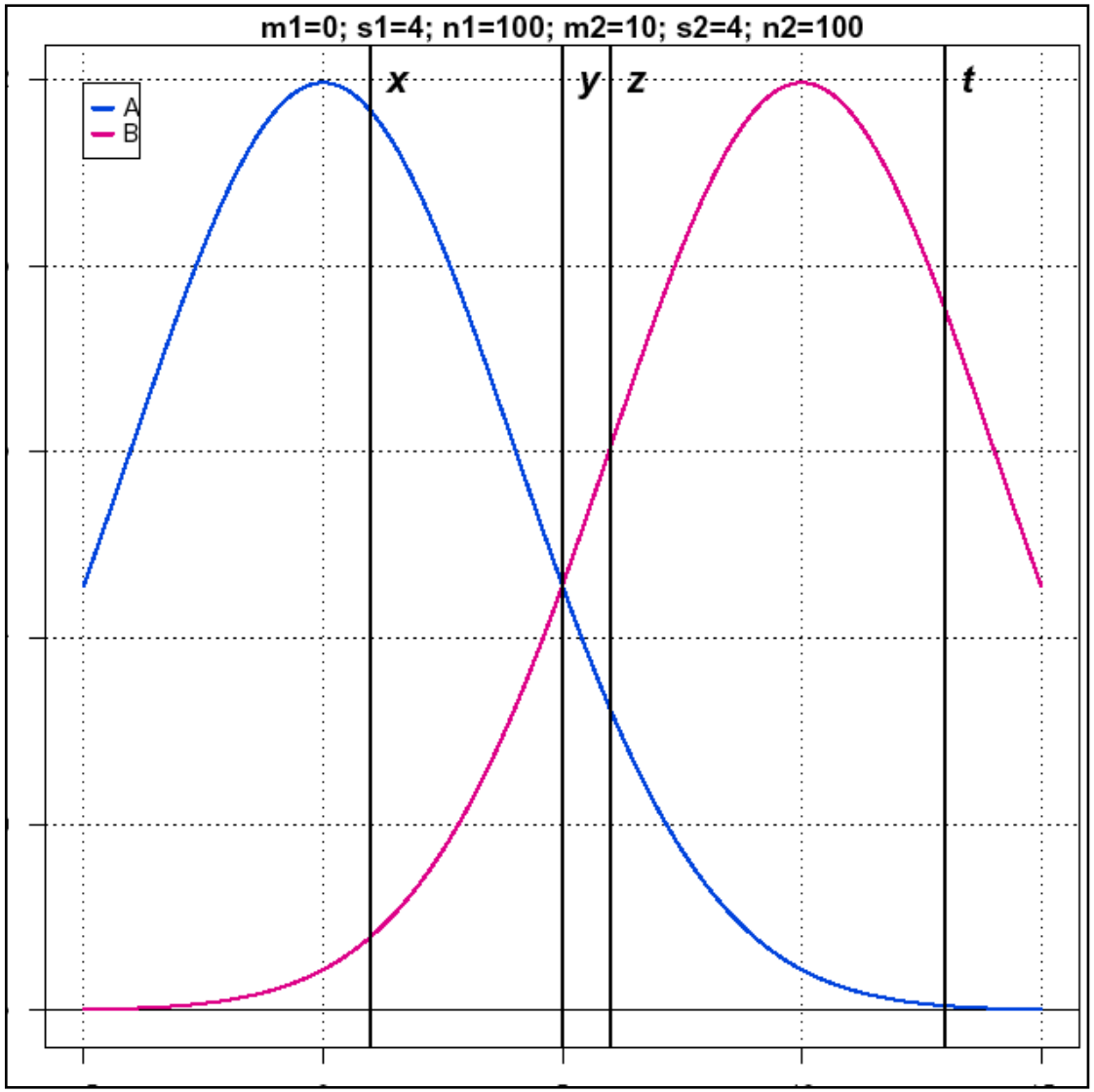
- Same exercise.
- This example shows that the assignation is affected by the position of the group centres.

Conceptual illustration with a single variable



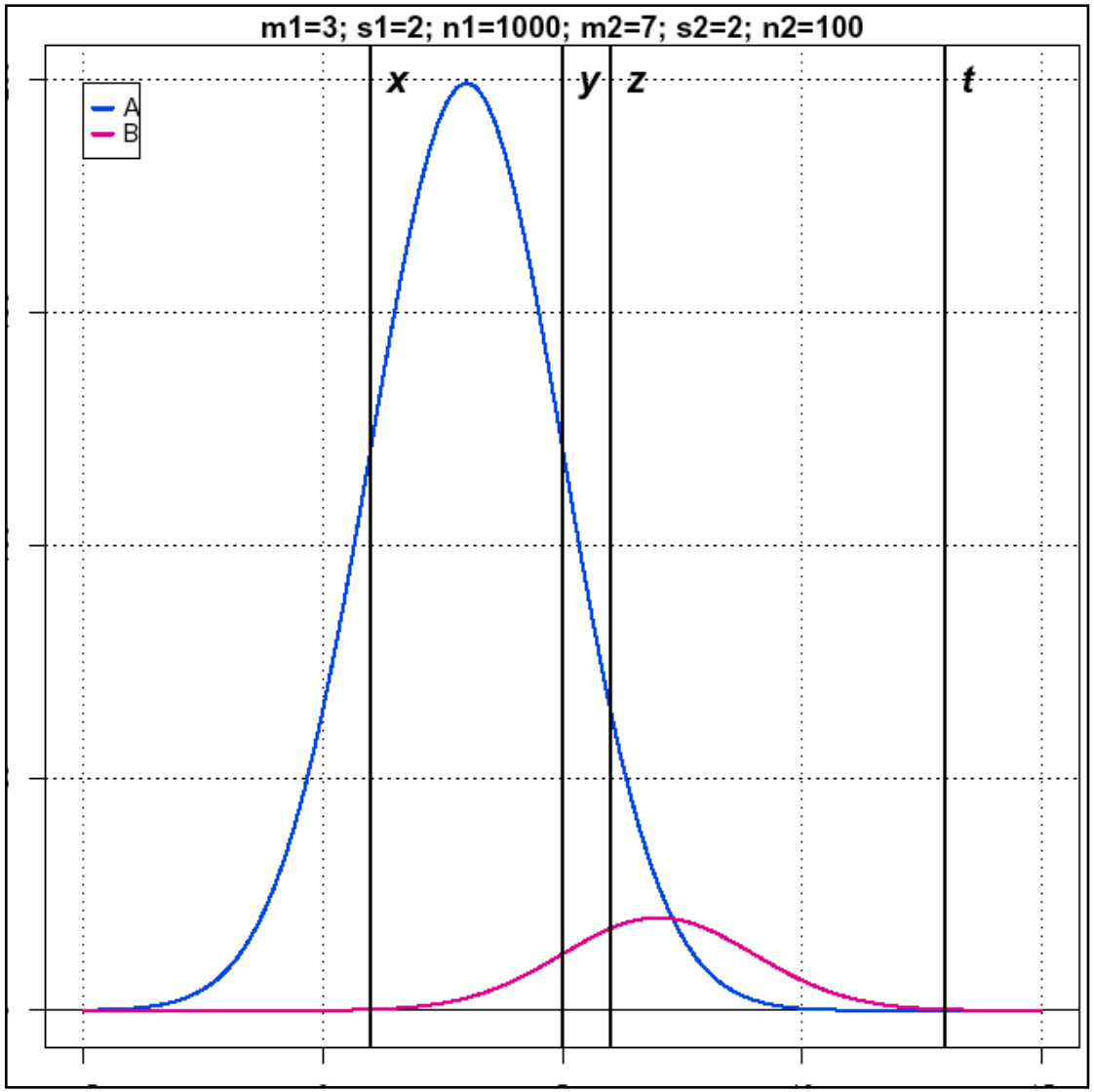
- Same exercise.
- When the centres become too close, some uncertainty is attached to some points (y, but also partly z).
- There is thus an effect of group distance.

Conceptual illustration with a single variable



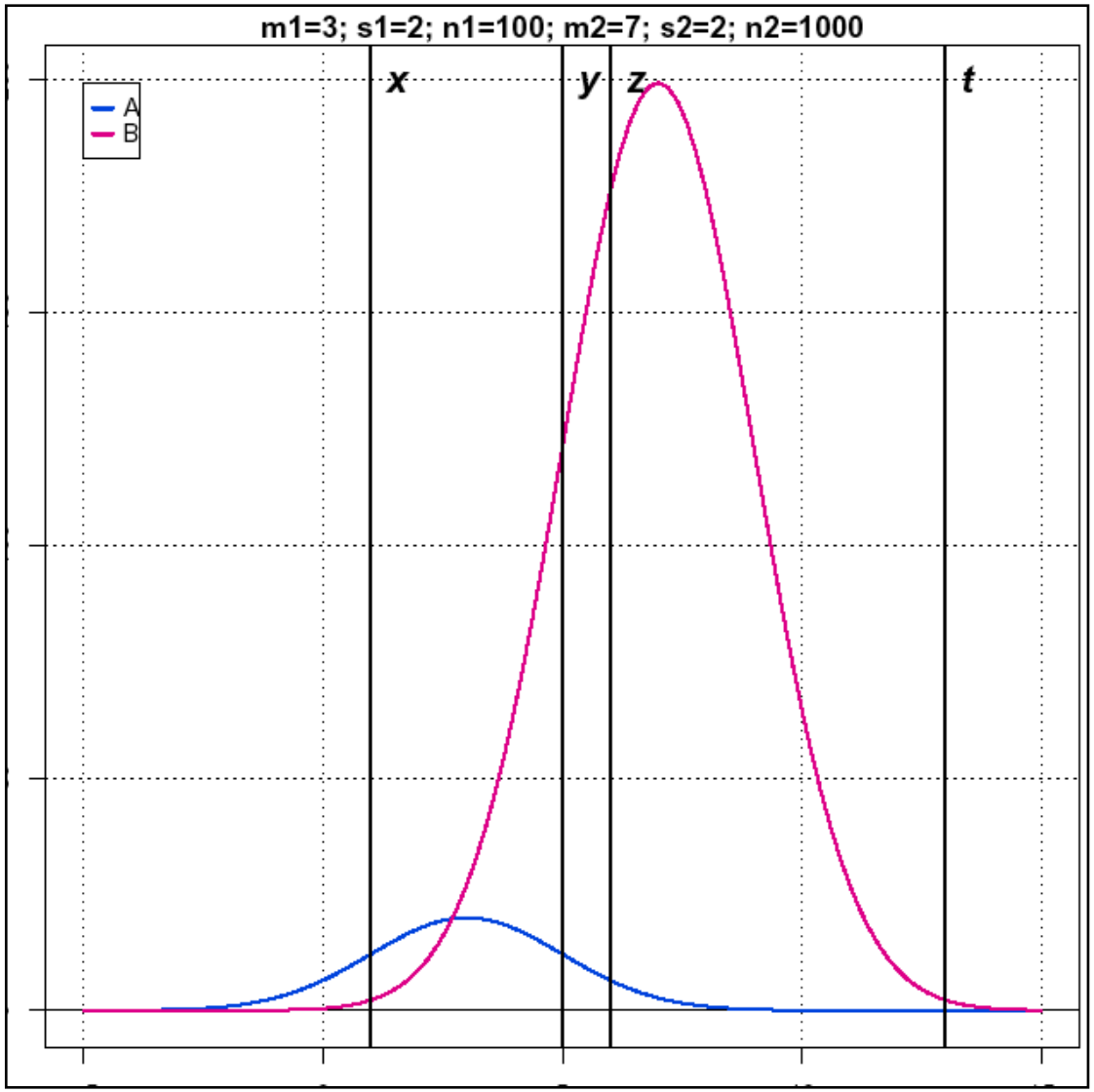
- Same exercise.
- The centres are in the same position as in the first example, but the variance is larger.
- This affects the level of separation of the groups, and raises some uncertainty about the group membership of z.
- The group variance thus affects the assignation.

Conceptual illustration with a single variable



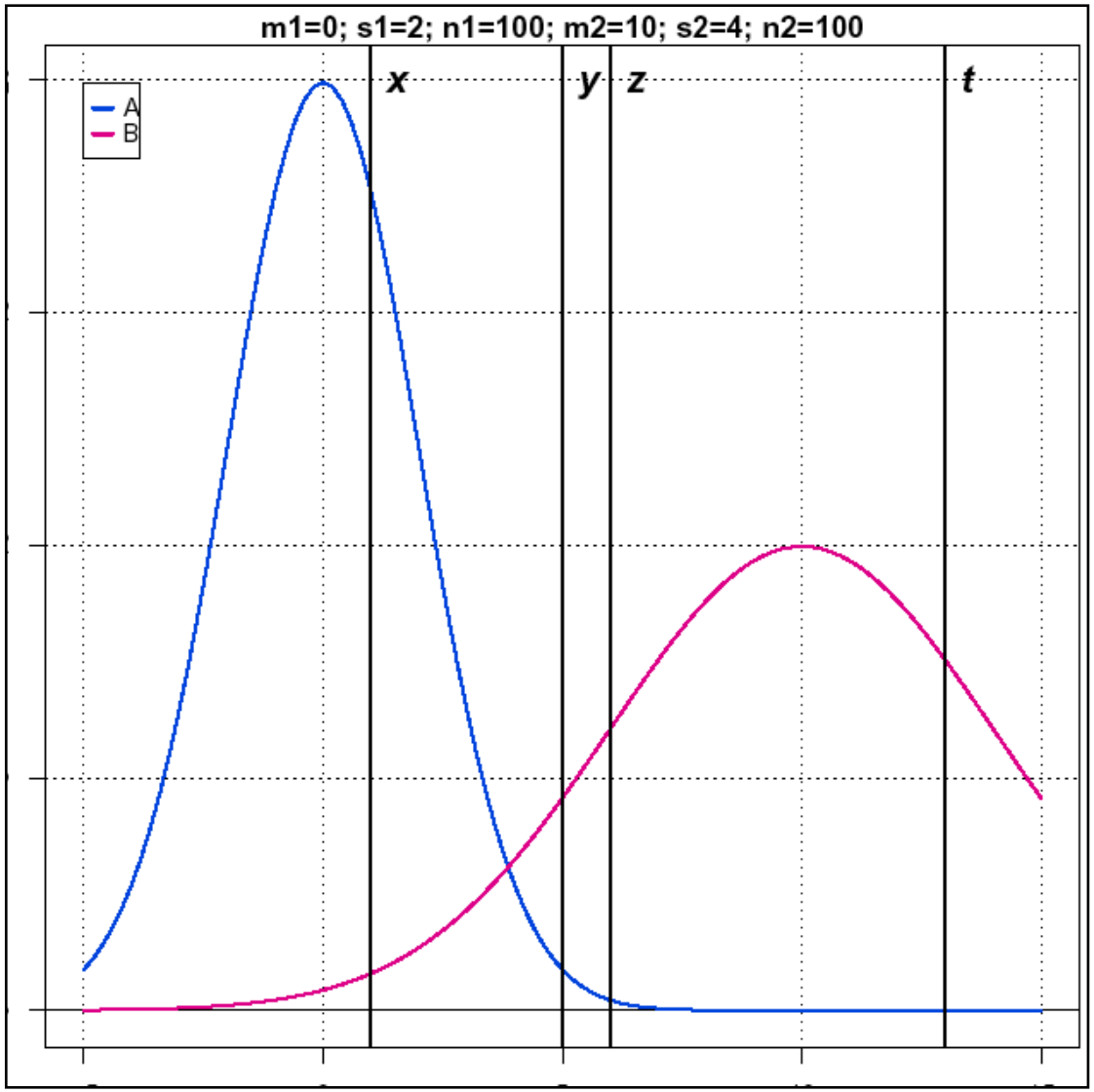
- Same exercise.
- This illustrates the effect of the sample size: if a sample has a much larger size than another one, it will increase the likelihood that some observations were issued from this group.

Conceptual illustration with a single variable



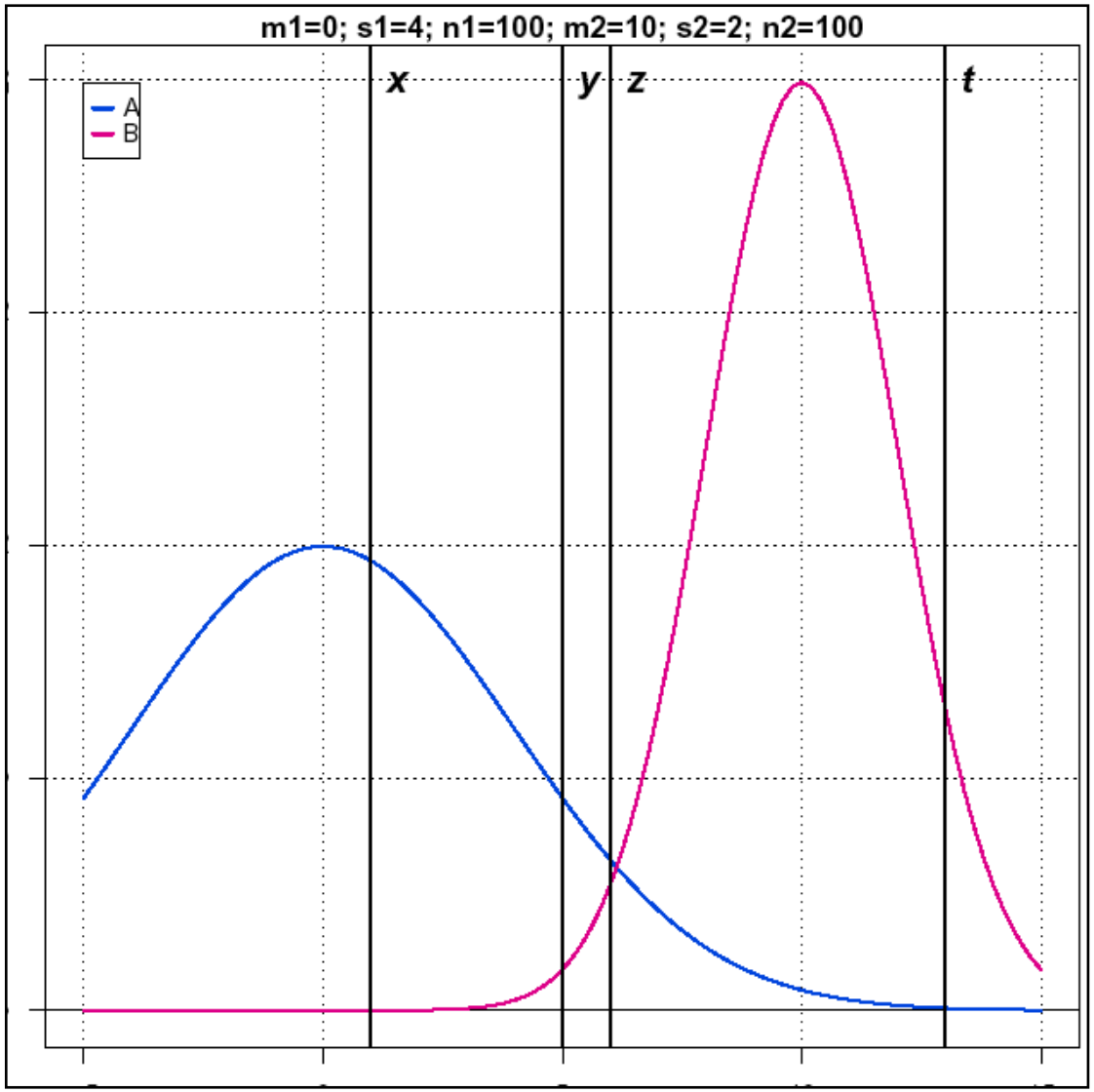
- Same exercise.
- This is the symmetric situation of the preceding figure.
- Although the group centres and variances are identical, the change of sample sizes completely modifies the group assignations.
- This is an effect of ***prior probability***.

Conceptual illustration with a single variable



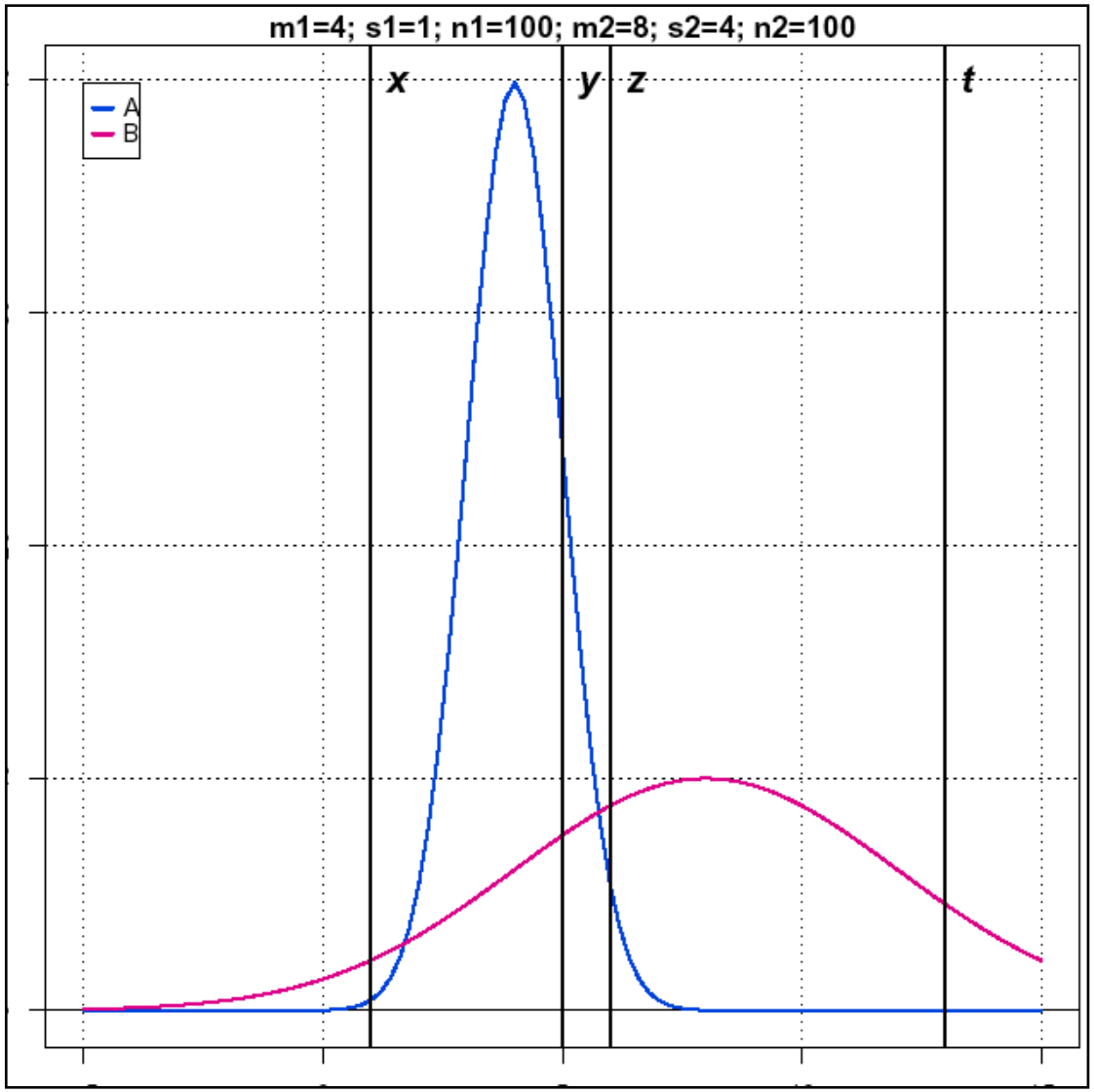
- Same exercise.
- If the two groups have different dispersions, it will affect their likelihood to be the originators of some observations.
- The ***relative dispersion*** of the groups affects the assignation.

Conceptual illustration with a single variable



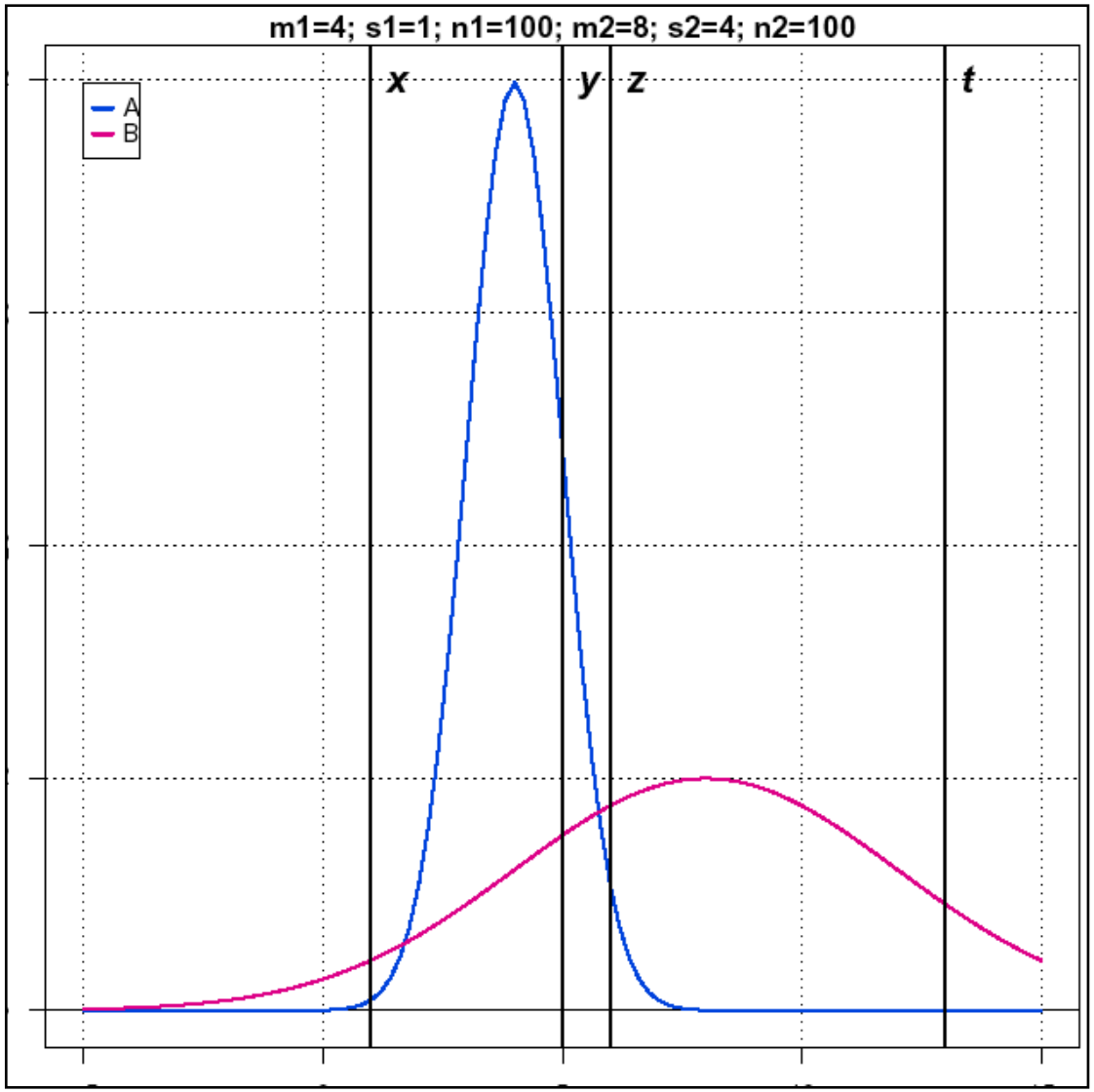
- Same exercise.
- Symmetrical situation of the preceding one: same centres, same sample sizes, but the relative variances vary in the opposite way.
- The ***relative dispersion*** of the groups affects the assignation.

Conceptual illustration with a single variable



- Same exercise.
- When the dispersion of one group becomes too high, a simple boundary is not sufficient anymore to separate the two groups.
- In this example, we would classify the leftmost (x) and rightmost (t, and maybe z) objects as B, and the central ones (y) as A. `
- We need thus two boundaries to separate these groups.
- The ***relative dispersion*** of the groups affects the assignation.

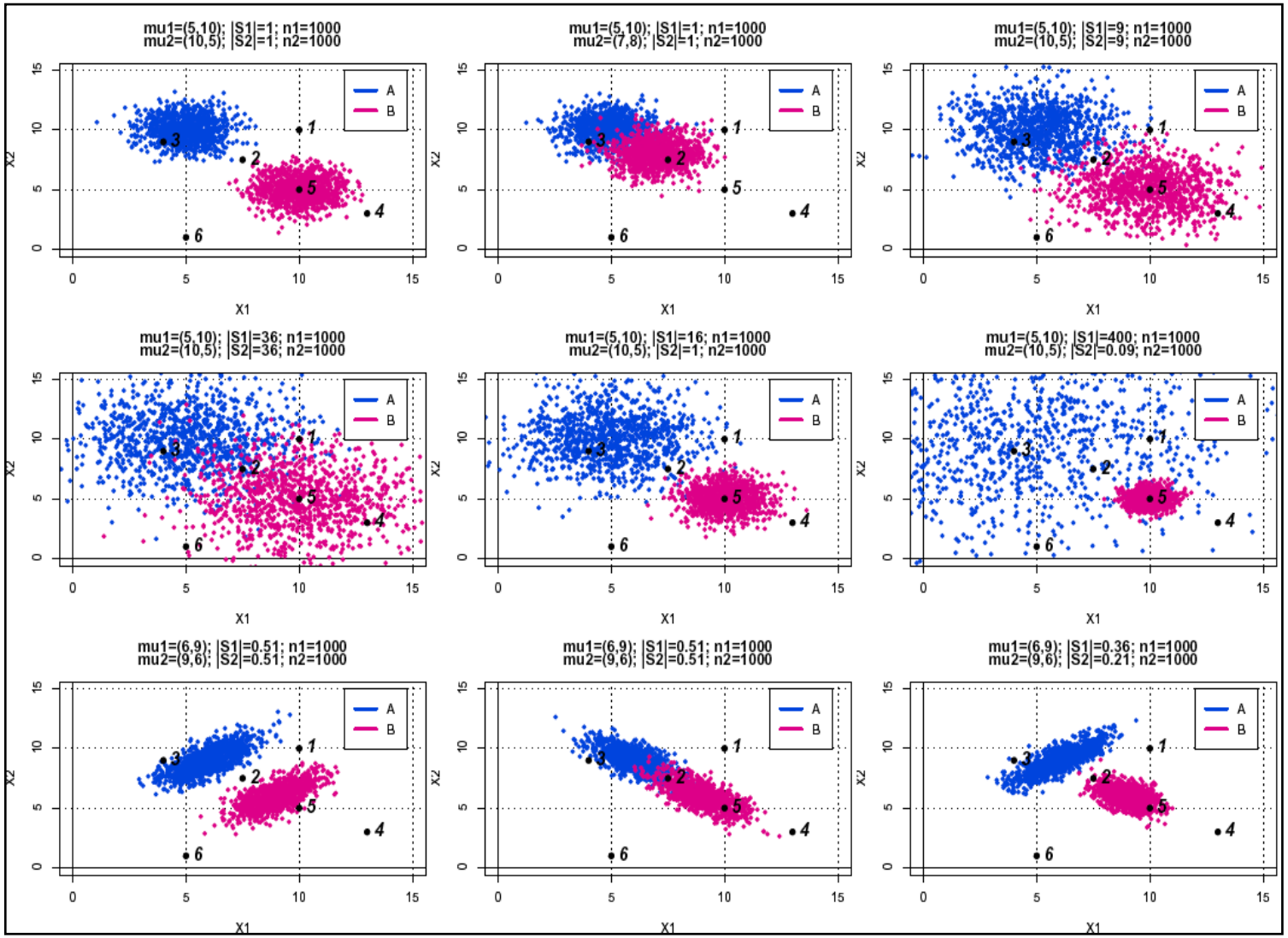
Conceptual illustration with a single variable



- Same exercise.
- Symmetrical situation of the preceding figure.
- The ***relative dispersion*** of the groups affects the assignation.

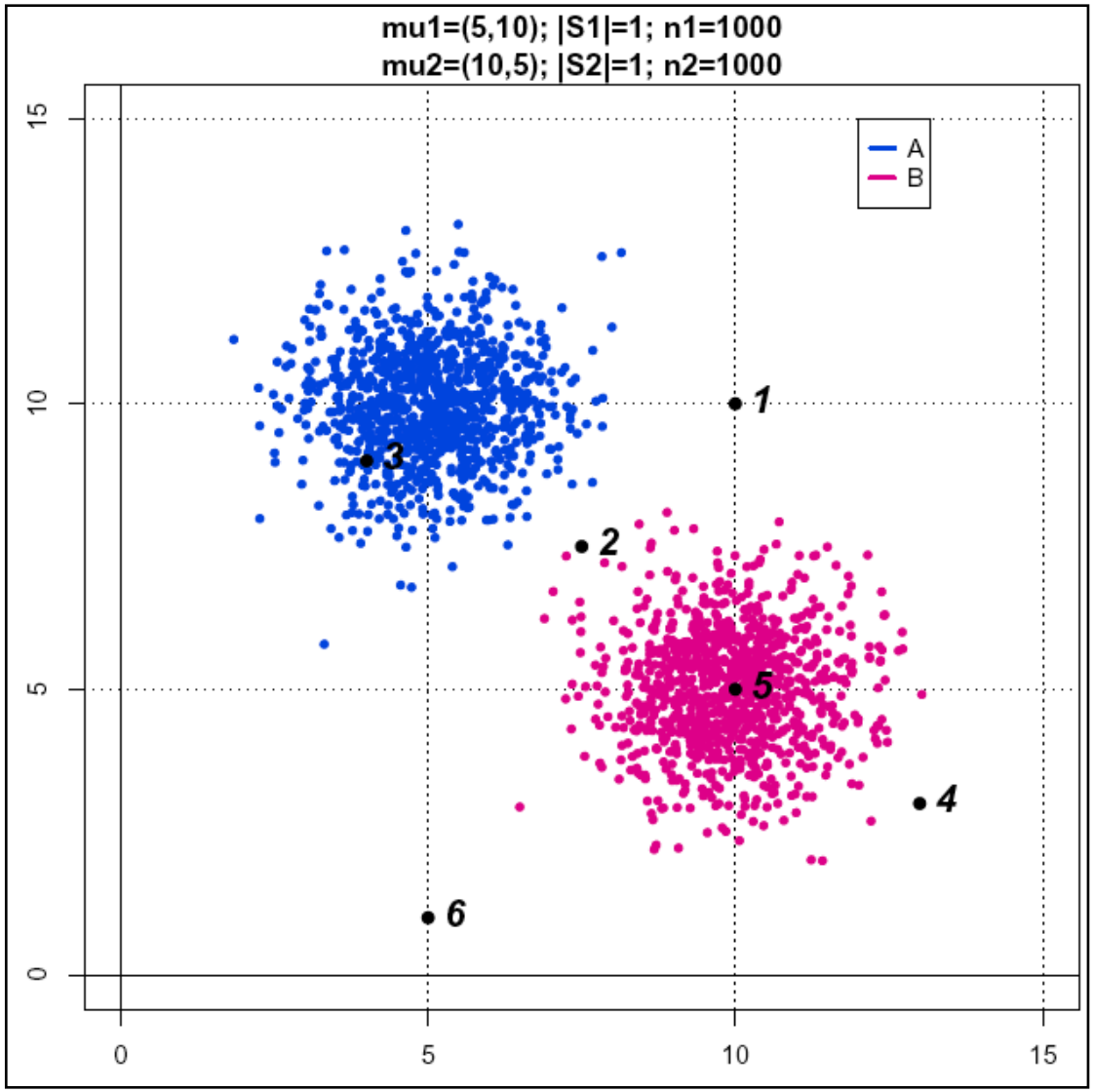
Conceptual illustration with two predictor variables

- Given two predefined classes (A and B), try intuitively to assign a class to each new object (black dots).
- How confident do you feel for each of your predictions ?



- What is the effect of the respective ***means*** ?
- What is the effect of the respective ***standard deviations*** ?
- What is the effect of the ***correlations*** between the two variables ?
- Note that the two population can have ***distinct correlations*** (orientations of the clouds)

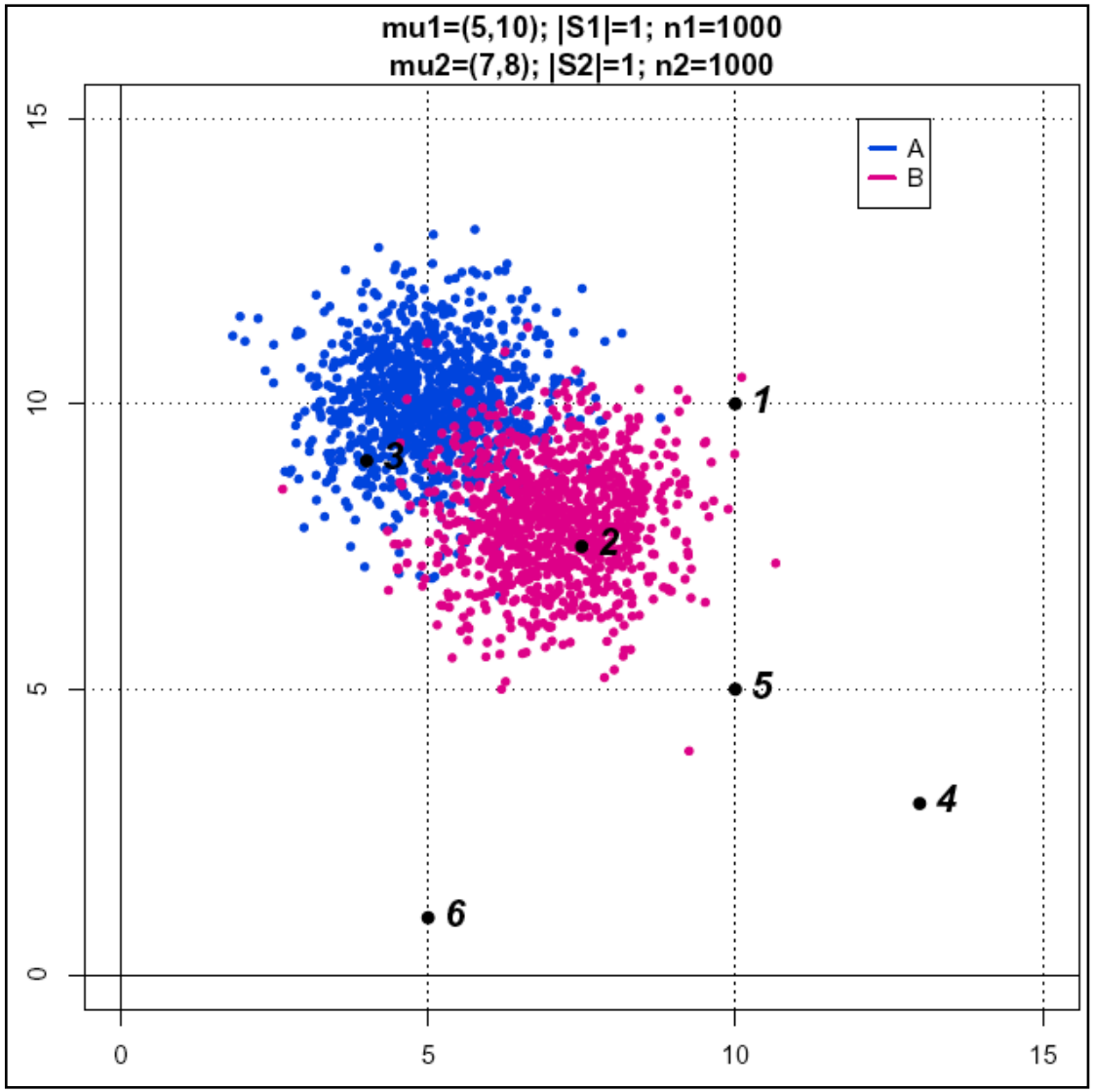
Conceptual illustration with two variable



- The same concepts can be illustrated in a two-dimensional feature space.
- Some additional concepts will appear.
- Try to assign intuitively the points 1 to 6 to either group A or group B.

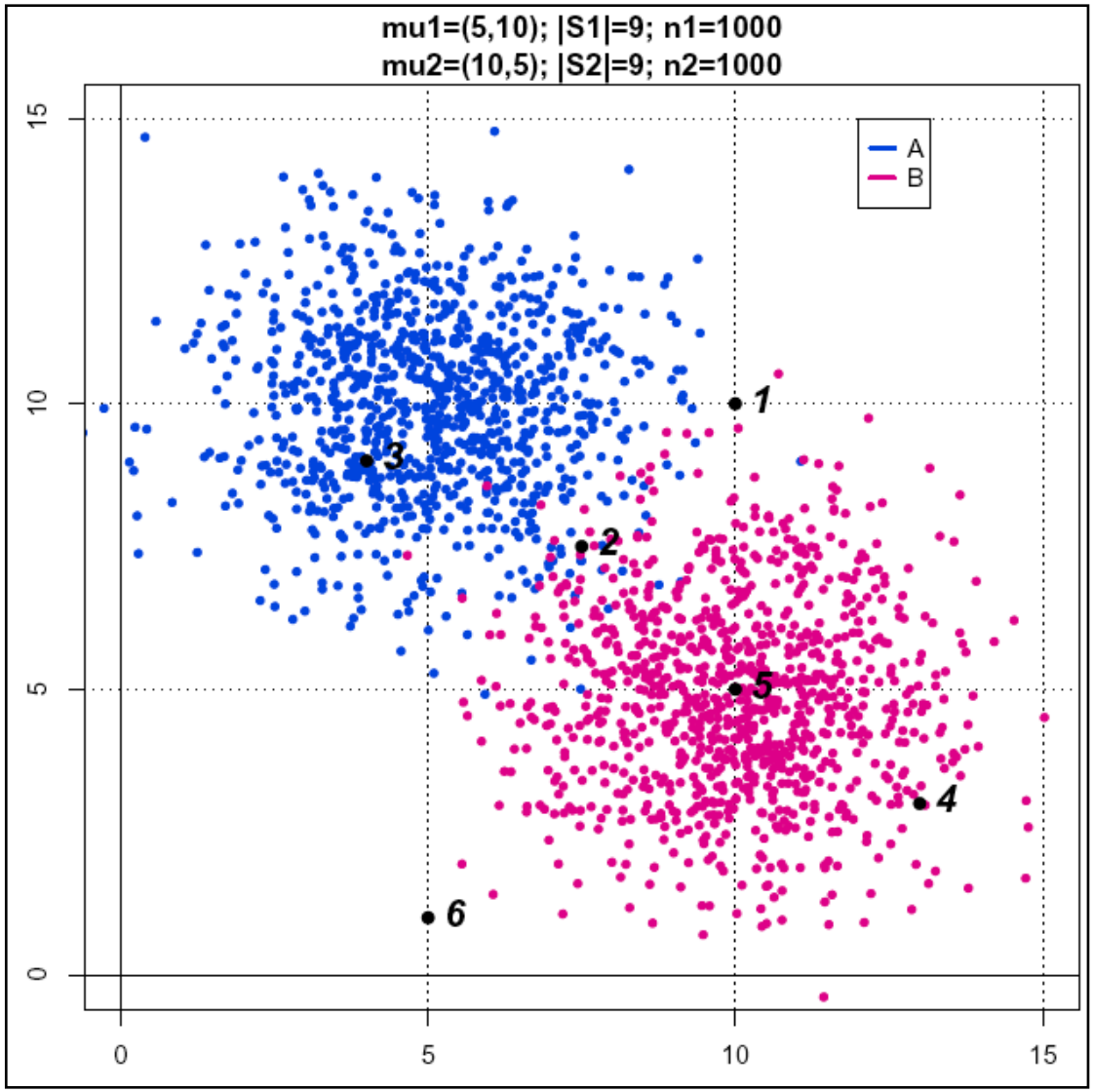
Conceptual illustration with two variable

- Effect of the group ***centre location***.



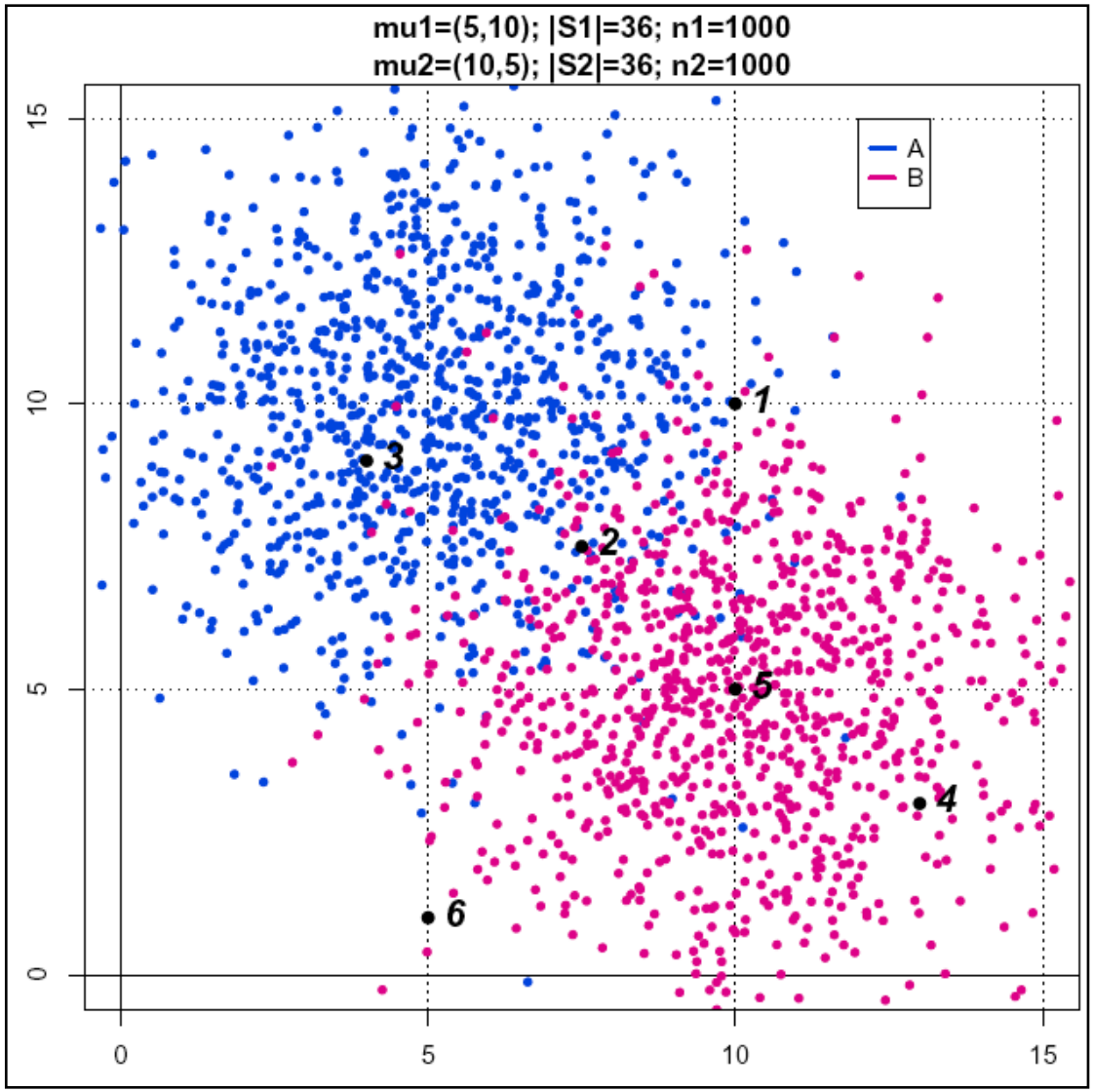
Conceptual illustration with two variable

- Effect of the group ***variance***.



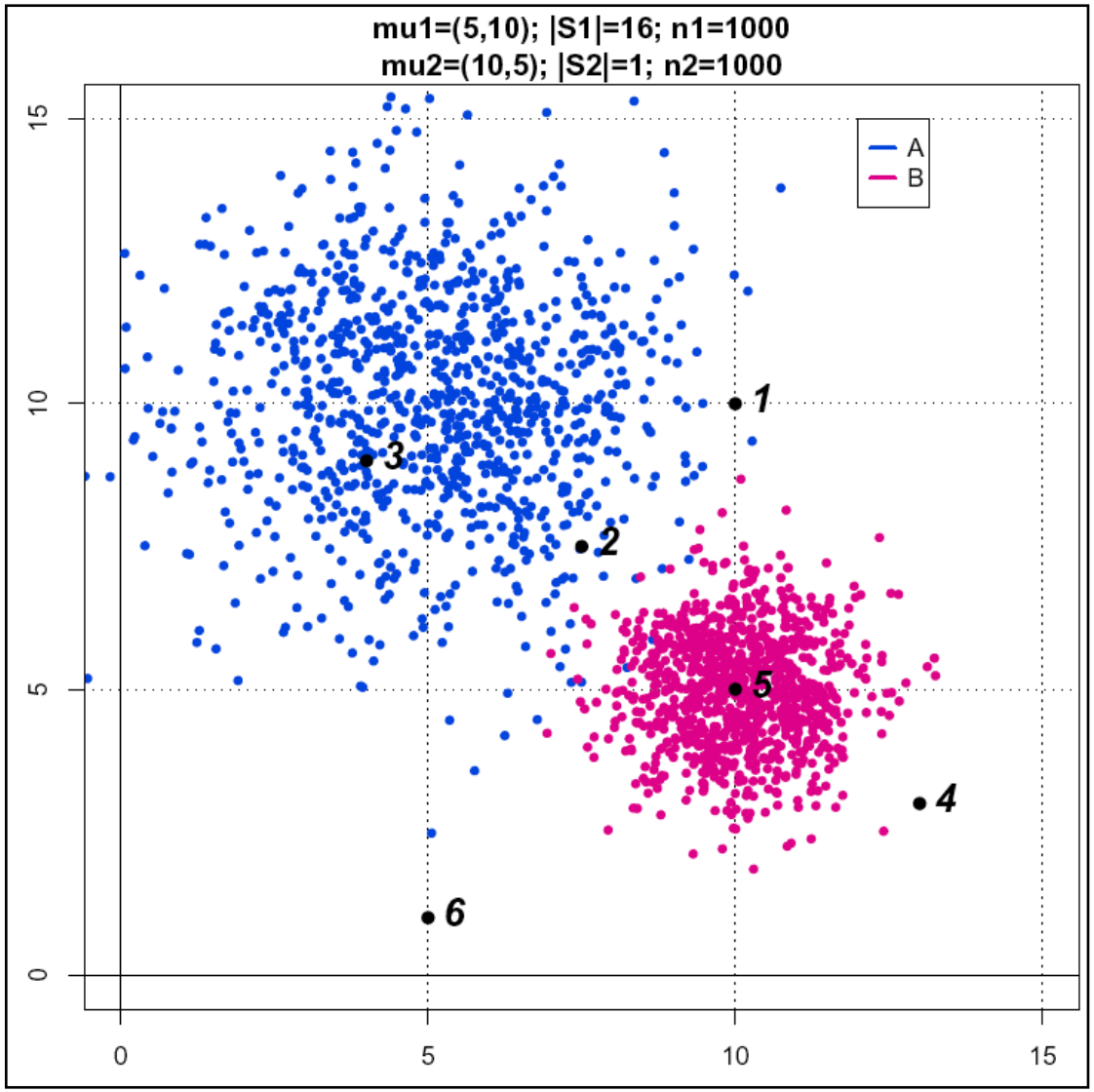
Conceptual illustration with two variable

- Effect of the group ***variance***.



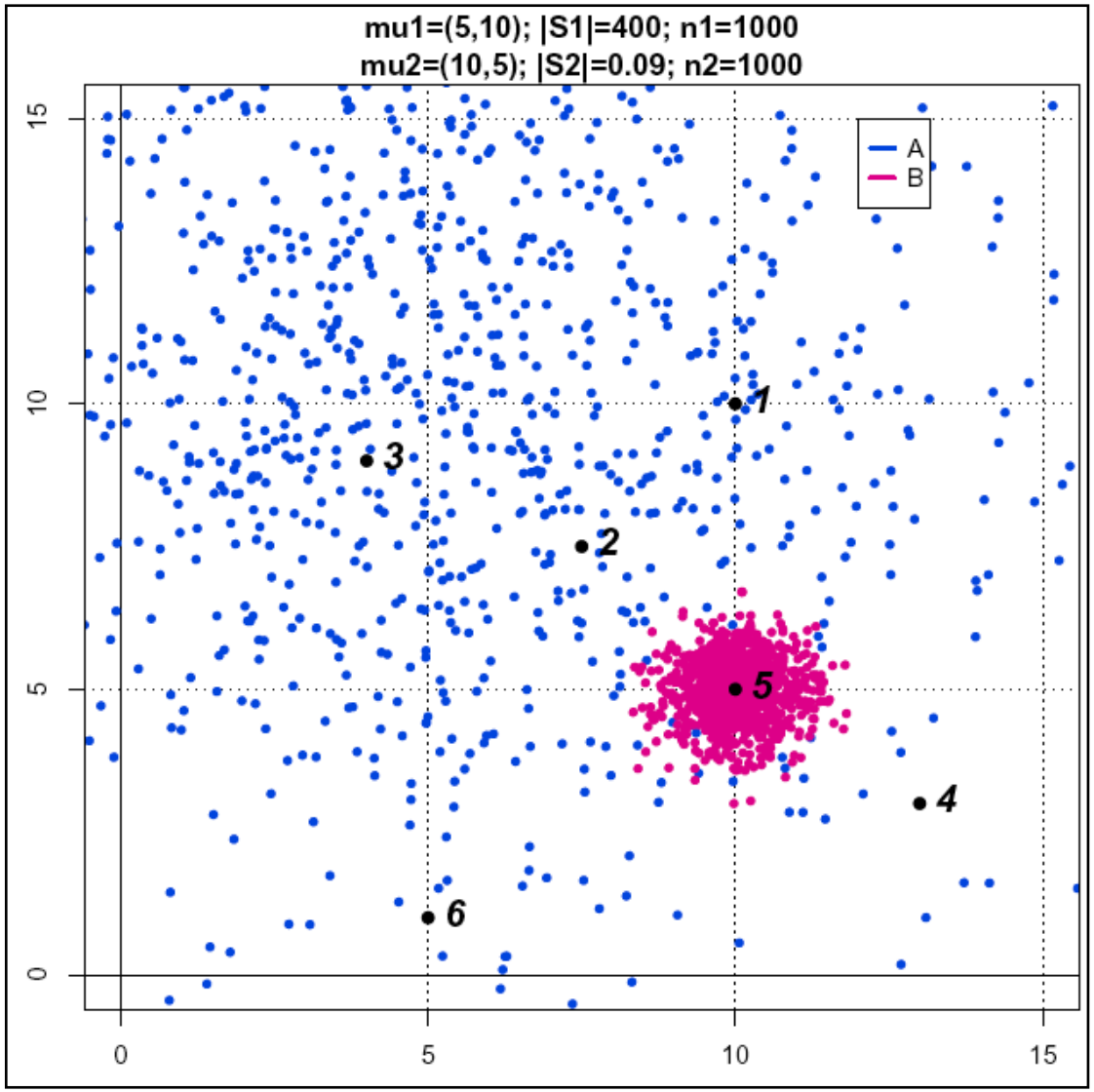
Conceptual illustration with two variable

- Effect of the ***relative variances***.



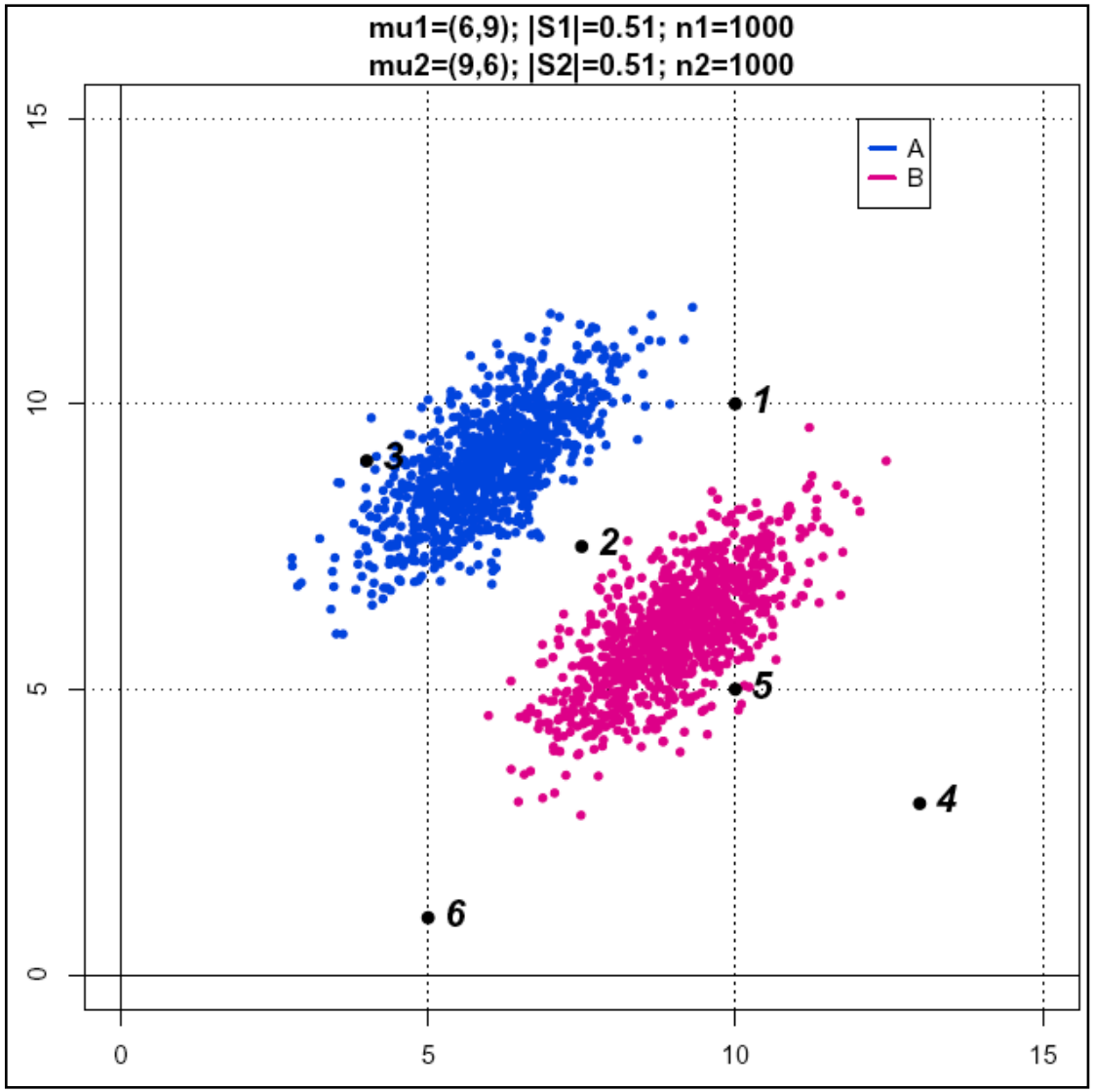
Conceptual illustration with two variable

- Effect of the ***relative variances***.



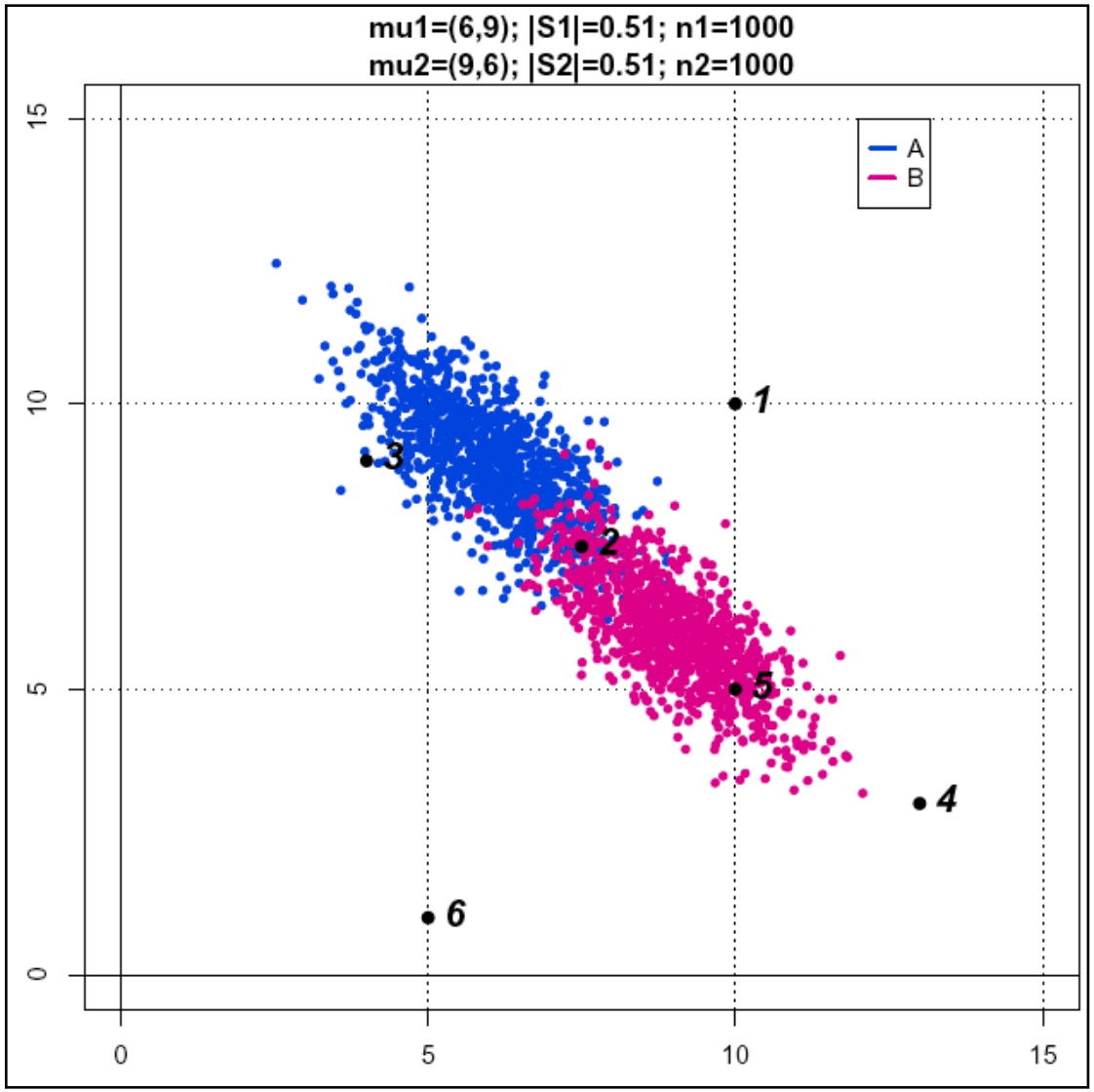
Conceptual illustration with two variable

- Effect of the ***covariance***. between columns of the group.

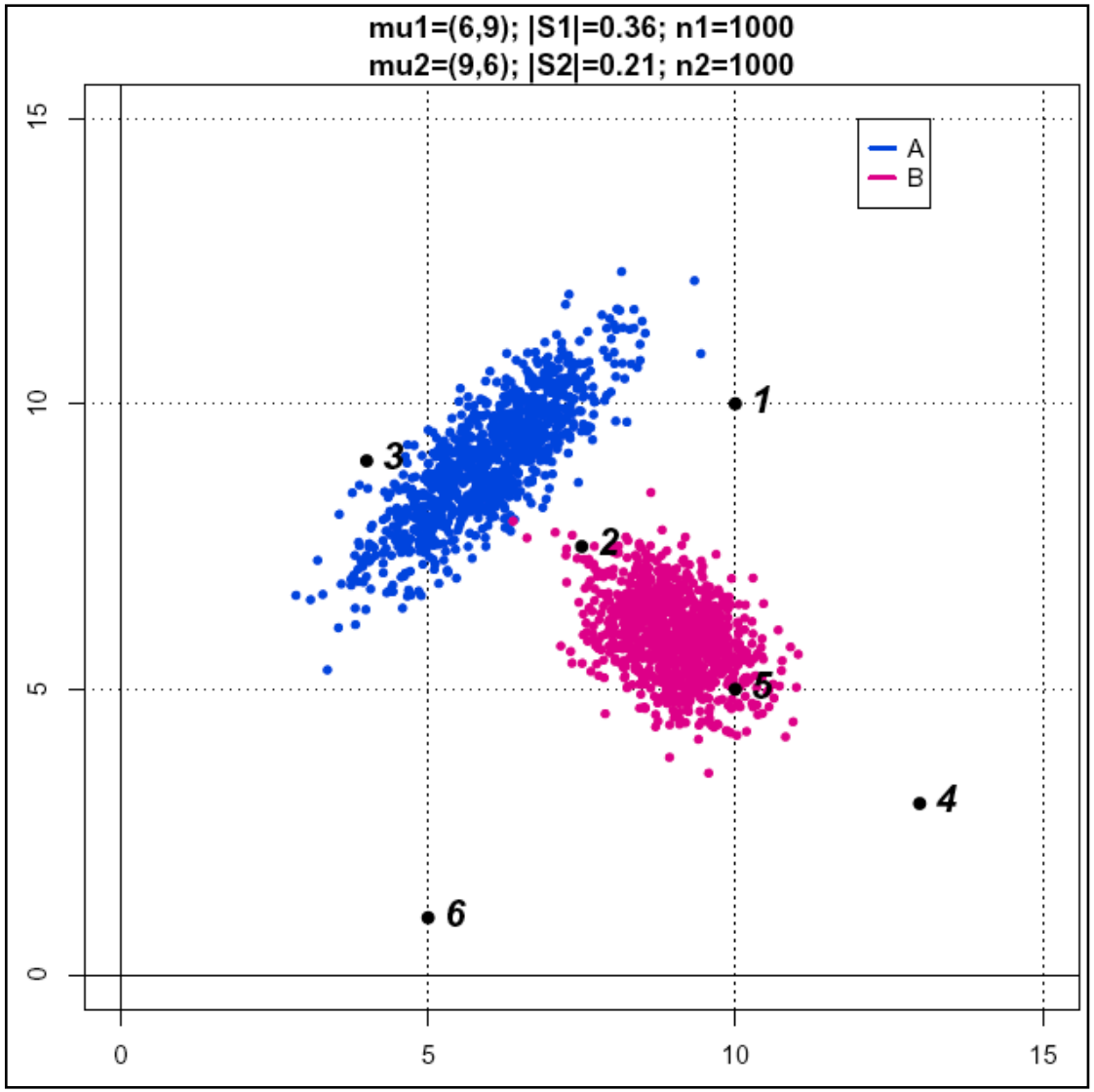


Conceptual illustration with two variable

- Effect of the ***covariance***. between columns of the group.



Conceptual illustration with two variable



- Effect of the ***relative covariance***. between columns of the group.
- The two groups have now different covariance matrices: the clouds are elongated in different directions.
- This affects group assignments (example point 2).

Classification rules

- New units can be classified on the basis of rules based on the calibration sample
- Several alternative rules can be used
  - **Maximum likelihood rule**: assign unit u to group g if

$$f(X \mid g) > f(X \mid g') \qquad for\ g' \neq g$$

  - **Inverse probability rule**: assign unit u to group g if

$$P(X \mid g) > P(X \mid g') \qquad for\ g' \neq g$$

  - **Posterior probability rule**: assign unit u to group g if

$$P(g \mid X) > P(g' \mid X) \qquad for\ g' \neq g$$

Where

X	is the unit ve
g,g'	are two groups
$f(X\|g)$	is the density function of the value X for group g
$P(X\|g)$	is the probability to emit the value X given the group g
$P(g\|X)$	is the probability to belong to group g, given the value X

Posterior probability rule

- The posterior probability can be obtained by application of Bayes' theorem

$$P(g \mid X) = \frac{P(X \mid g)P(g)}{P(X)}$$

$$P(g \mid X) = \frac{P(X \mid g)\pi_g}{\sum_{g'=1}^{k} P(X \mid g')\pi_{g'}}$$

Where

- X is the unit vector
- g is a group
- k is the number of groups
- π_g is the prior probability of group g

Maximum likelihood rule - multivariate normal case

- If the predictor variable is univariate normal

$$f(X \mid g) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_g^2}} e^{-\frac{1}{2}\left(\frac{X-\mu_g}{\sigma_g}\right)^2}$$

- If the predictor variable is multivariate normal

$$f(X \mid g) = \frac{1}{\sqrt{(2\pi)^p}\sqrt{|\Sigma_g|}} e^{\left[-\frac{1}{2}(X-\mu_g)'\Sigma_g^{-1}(X-\mu_g)\right]}$$

Where

- X is the unit vector
- p is the number of variables
- μ_g is the mean vector for group g
- Σ_g is the covariance matrix for group g

Bayesian classification in case of normality

- Each object is assigned to the group which minimizes the function

$$f = P(g)\frac{1}{\sqrt{(2\pi)^p}\sqrt{|\Sigma_g|}}e^{\left[-\frac{1}{2}(X-\mu_g)\Sigma_g^{-1}(X-\mu_g)\right]}$$

Linear versus quadratic classification rule

- There is one covariance matrix per group g.
  - This matrix indicates the covariance between each column (variable) of the data set, for the considered group.
  - The diagonals of this matrix represent the variance (=covariance between a variable and itself)
- When all covariance matrix are assumed to be identical
  - The classification rule can be simplified to obtain a linear function. This is referred to as ***Linear Discriminant Analysis (LDA)***
  - In this case,the boundary between groups will be a plane (2 variables) or a hyper-plane (more than 2 variables).
- If the variances and covariances are expected to differ between groups
  - A specific covariance matrix has to be used for each group.
  - The boundary between two groups is a curve (with two variables) or a hyper-surface (more than 2 variables).
  - This is referred to as ***Quadratic Discriminant Analysis (QDA)***

Evaluation of the discriminant function - confusion table

- One way to evaluate the accuracy of the discriminant function is to apply it to the sample itself. This approach is called **internal analysis**.
- The known and predicted class are then compared for each sample unit.
- **Warning** : internal analysis is too optimistic. This approach is not recommended.

	Predictor variables				Criterion variable	
	variable 1	**variable 2**	**…**	**variable p**	**known**	**predicted**
object 1	$x_{1,1}$	$x_{2,1}$	…	$x_{p,1}$	**A**	**A**
object 2	$x_{1,2}$	$x_{2,2}$	…	$x_{p,2}$	**A**	**B**
object 3	$x_{1,3}$	$x_{2,3}$	…	$x_{p,3}$	**A**	**A**
…	…	…	…	…	**…**	**…**
object i	$x_{1,i}$	$x_{2,i}$	…	$x_{p,i}$	**B**	**K**
object i+1	$x_{1,i+1}$	$x_{2,i+1}$	…	$x_{p,i+1}$	**B**	**B**
object i+2	$x_{1,i+2}$	$x_{2,i+2}$	…	$x_{p,i+2}$	**B**	**B**
…	…	…				
object n-1	$x_{1,n-1}$	$x_{2,n-1}$	…	$x_{p,n-1}$	**K**	**K**
object n	$x_{1,n}$	$x_{2,n}$	…	$x_{p,n}$	**K**	**K**

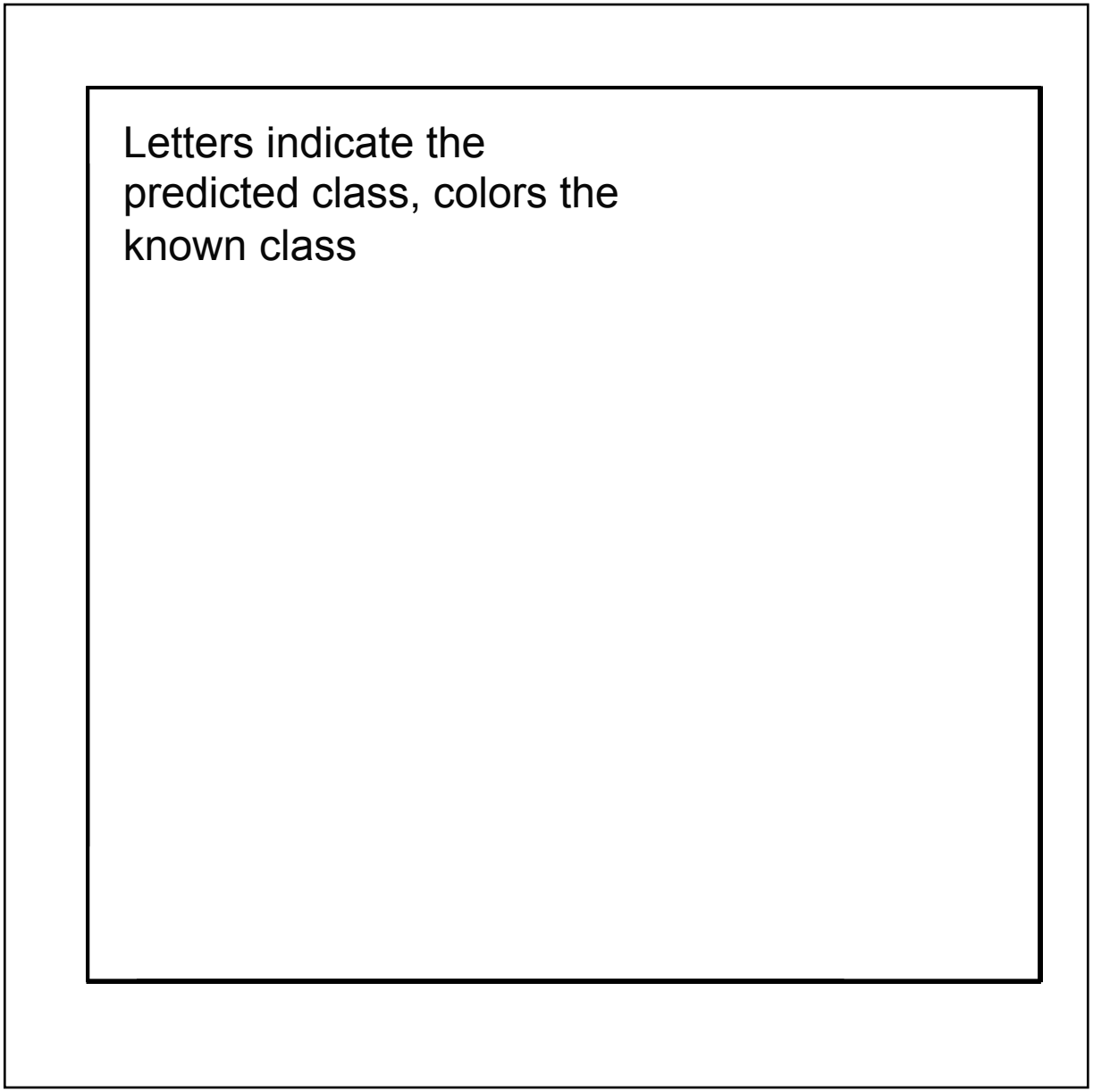
Evaluation of the discriminant function - confusion table

- The results of the evaluation are summarized in a **confusion table**, which contains the count of the predicted/known combinations.
- The confusion table can be used to calculate the accuracy of the predictions.

Confusion table

		Known			
		PHO	MET	CTL	SUM
Predicted	PHO	8	0	0	8
	MET	0	1	1	2
	CTL	5	18	81	104
	SUM	13	19	82	114
	Errors	24	21.05%		
	Correct	90	78.95%		

Evaluation of the discriminant function - plot



- The two first discriminant functions can be used as X and Y axes for plotting the result.
- In the same way as for PCA, X and Y axes represent linear combinations of variables
- However, these combinations are not the same as the first factors obtained by PCA.
  - When comparing with PCA figure, the PHO genes are now all located nearby the X axis.

External analysis

- Using the sample itself for evaluation is problematic, because the evaluation is biased (too optimistic). To obtain an independent evaluation, one needs two separate sets : one for calibration, and one for evaluation. This approach is called **external analysis**.
- The simplest setting is to split randomly the sample into two sets (**holdout approach**) :
  - the **training set** is used to build a discriminant function
  - the **testing set** is used for evaluation

Training set

	Predictor variables				Criterion variable
	variable 1	**variable 2**	**…**	**variable p**	**class**
object 1	x_{11}	x_{21}	…	x_{p1}	A
object 2	x_{12}	x_{22}	…	x_{p2}	A
object 3	x_{13}	x_{23}	…	x_{p3}	B
…	…	…	…	…	…
object n_{train}	x_{1n}	x_{2n}	…	x_{pn}	K

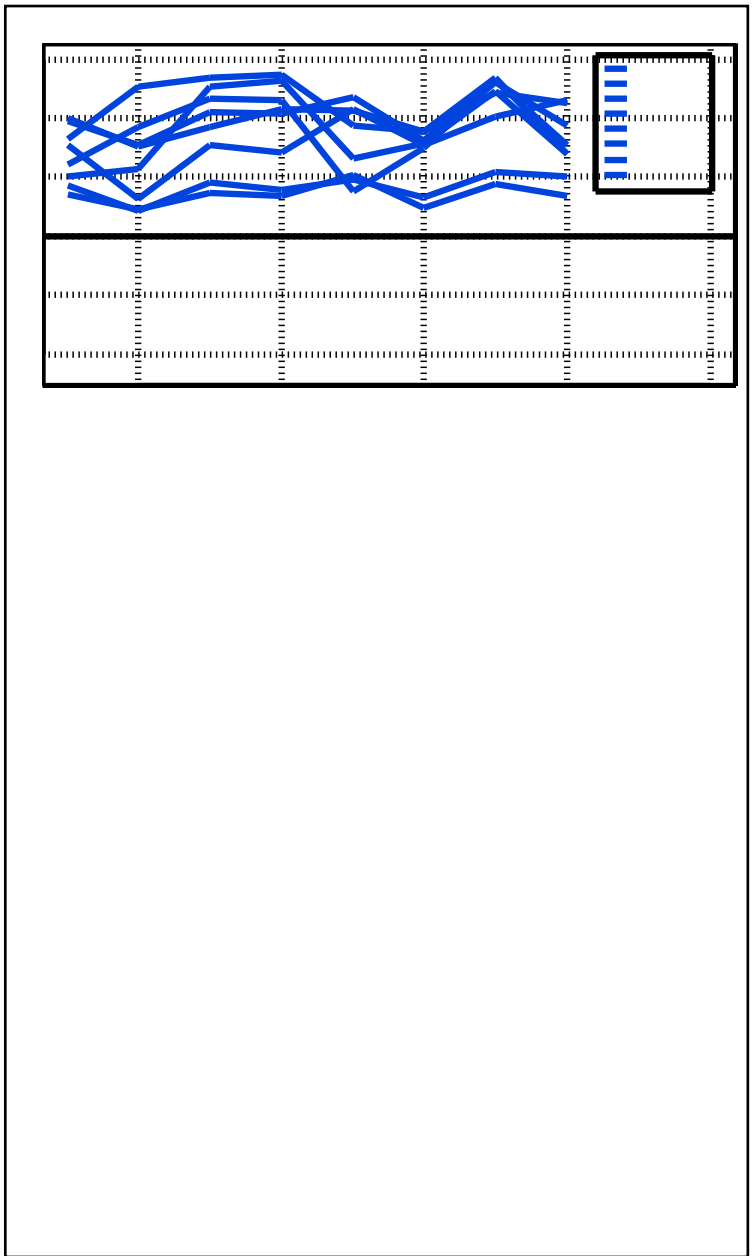
Testing set

	Predictor variables				Criterion variable	
	variable 1	**variable 2**	**…**	**variable p**	**known**	**predicted**
object 1	x_{11}	x_{21}	…	x_{p1}	A	A
object 2	x_{12}	x_{22}	…	x_{p2}	B	A
object 3	x_{13}	x_{23}	…	x_{p3}	B	B
…	…	…	…	…	…	…
object n_{test}	x_{1n}	x_{2n}	…	x_{pn}	K	K

Leave-one-out (LOO) validation

- When the sample is too small, it is problematic to loose half of it for testing.
- In such a case, the **leave-one-out (LOO)** approach is recommended :
  1. Discard a single object from the sample.
  2. With the remaining objects, build a discriminant function.
  3. Use this discriminant function to predict the class of the discarded object.
  4. Compare known and predicted class for the discarded object.
  5. Iterate the above steps with each object of the sample.

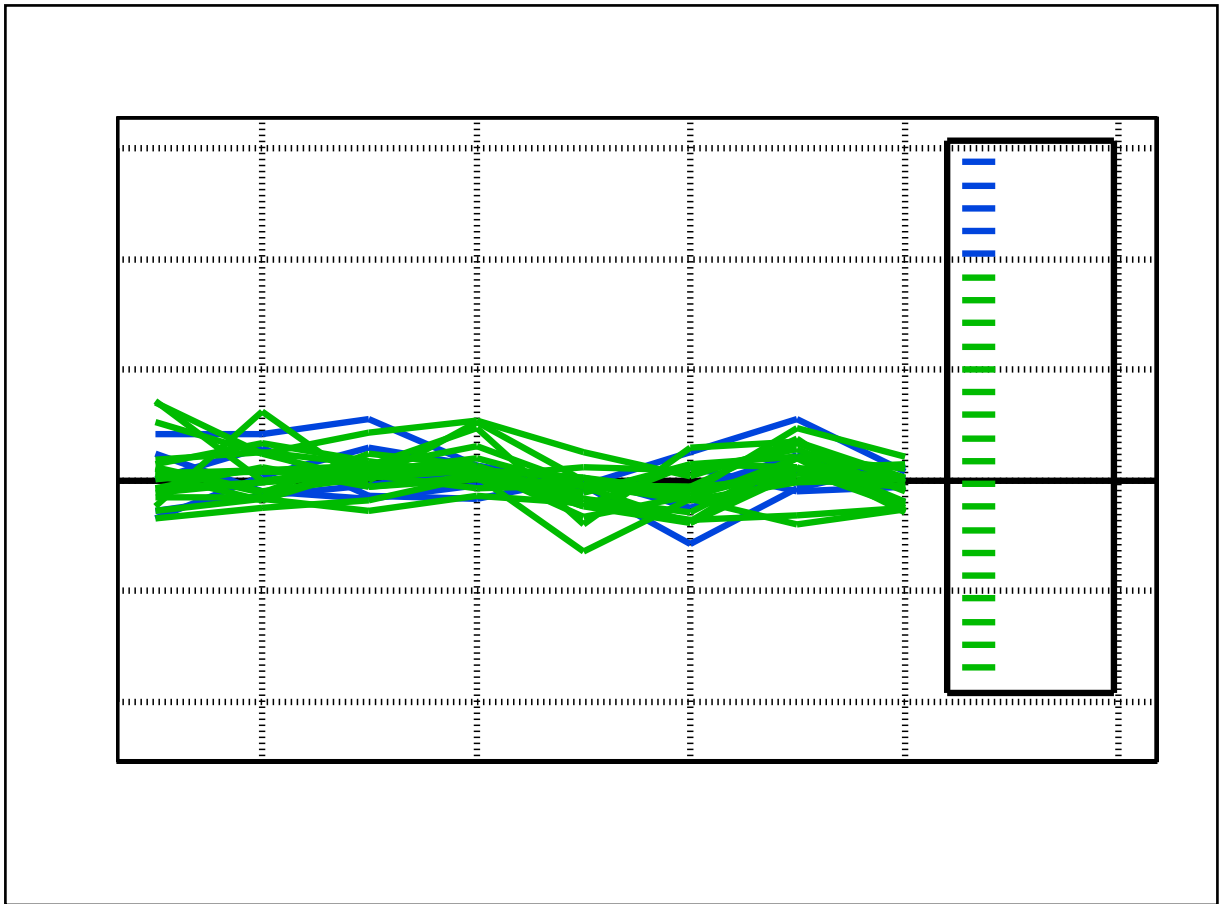
Profiles after prediction



- Example:
  - Gene expression data
  - Linear discriminant analysis
  - Leave-one-out cross-validation.
- Genes predicted as "PHO" have generally high levels of response (but this is not true for all of them)
- A very few genes are predicted as MET.
- Most genes predicted as control have a low levels of regulation.

Analysis of the misclassified units

- The sample might itself contain classification errors. The apparent misclassifications can actually represent corrections of these labelling errors.
- Example : gene expression data - linear discriminant analysis
  All the genes "mis"classified as control have actually a flat expression profile.
  - Most of them are MET genes (indeed, these are not expected to respond to phosphate)
  - the 4 PHO genes (blue) have a flat profile



Evaluation with leave-one-out

- Leave-one-out is more severe for evaluating the accuracy of predictions.

Choice of the prior probabilities

- The classes may have different proportions between the sample and the population
- For example, we could decide, on the basis of our biological knowledge, that it is likely to have 1% rather than 11% of yeast gene responding to phosphate.

		Population	
Class	Sample	Priors from sample	Arbitrary priors
PHO	13 *11%*	659 *11%*	58 *1%*
MET	19 *17%*	964 *17%*	58 *1%*
CTL	82 *72%*	4160 *72%*	5667 *98%*
TOTAL	114	5783	5783

Prediction phase

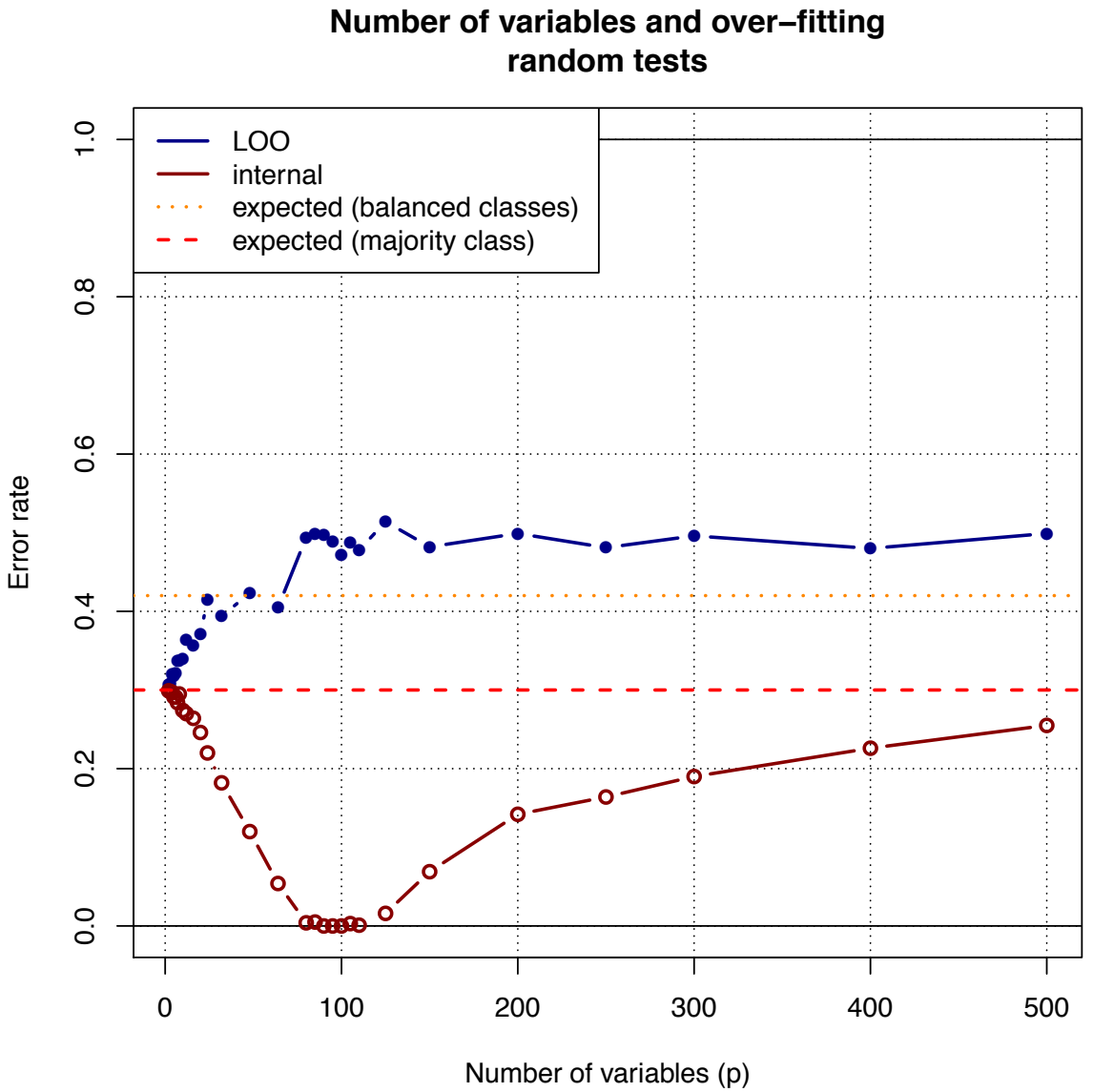
Feature selection (=variable selection)

Feature selection (variable selection)

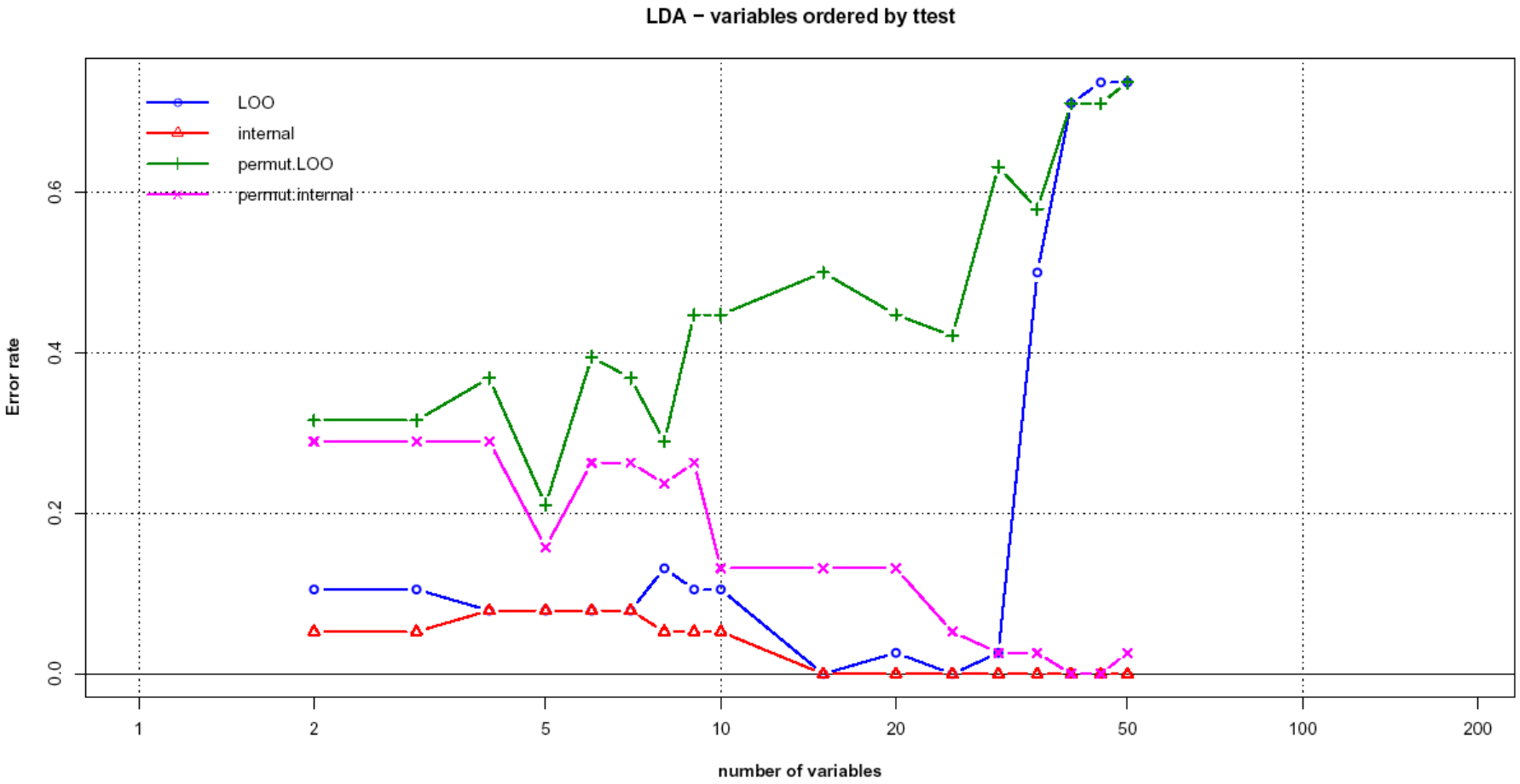
- One approach to circumvent this problem is to select a subset of variables only.
- This subset of variables can be selected according to different rules.
  - **Variable ordering**: variables are ordered according to some criterion, and the topmost variables are retained.
    - Inter-group distances calculated in each variable separately. This inter-group distance can be calculated with the t-test.
    - P-value of the t-test (the P-value is not always linear with the t statistics, since the number of observations can vary from row to row if there are missing values).
  - **Variables combinations**
    - Selection of a subset of variables and estimation of the capability of each subset to classify correctly.
    - The number of possible combinations of variables increases exponentially with the number of variables.
  - **Stepwise selection**
    - Stepwise selection is an heuristics to select a subset of variables in a quadratic time, but they do not guarantee optimality.
      - Forward selection
      - Backward selection
      - Forward-backward selection

Over-fitting

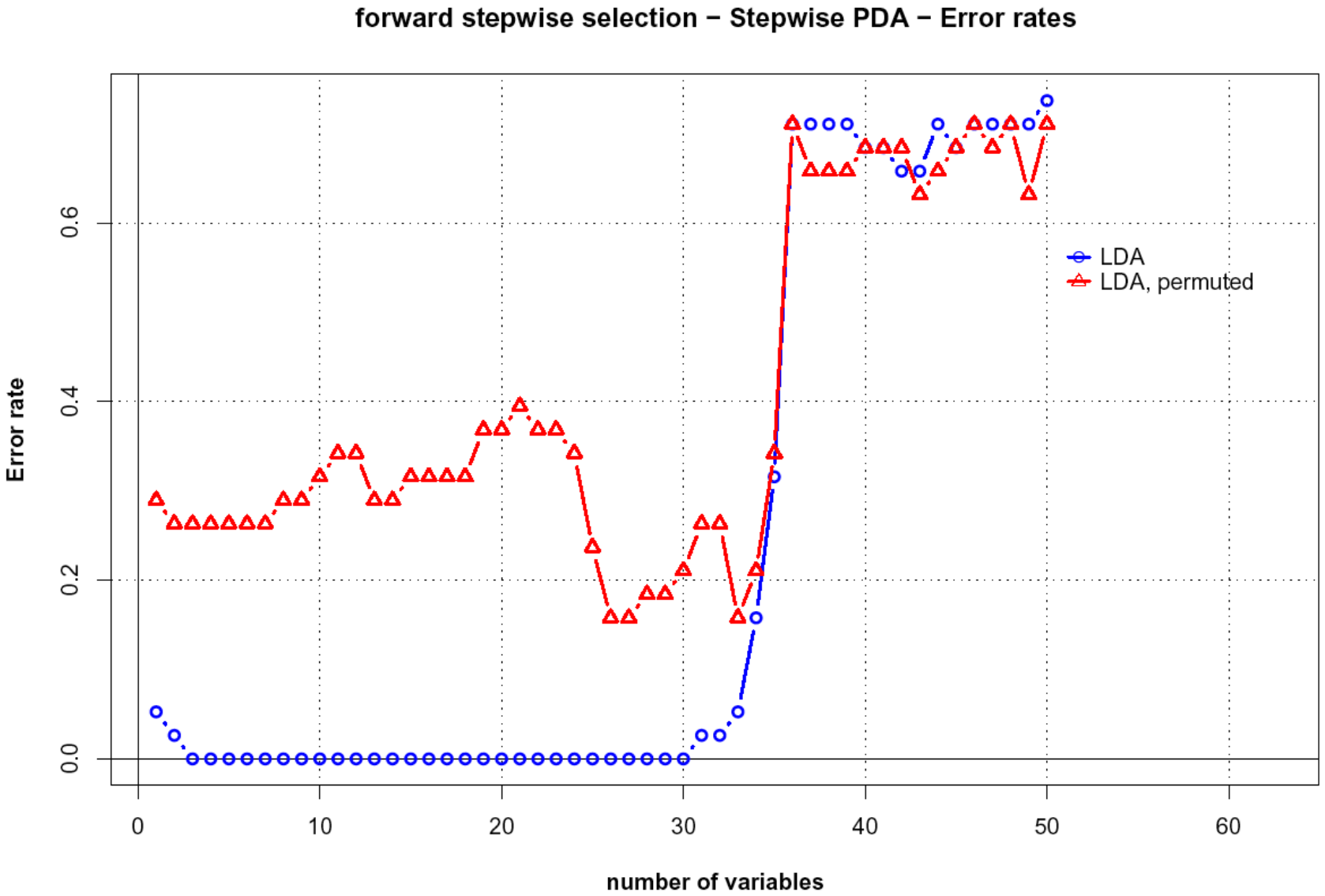
- A typical application of supervised classification is to classify experiments (e.g. patient types) on the basis of the expression profiles.
- In this case, the objects are the experiments, and the variables the genes.
- This raises a problem of over-fitting: the number of variables is much larger than the number of objects in the training set.
- In such situations, the classifier will tend to build a classification rule which perfectly fits the training set, but fails to generalize to other observations.



Variable ordering with the t-test



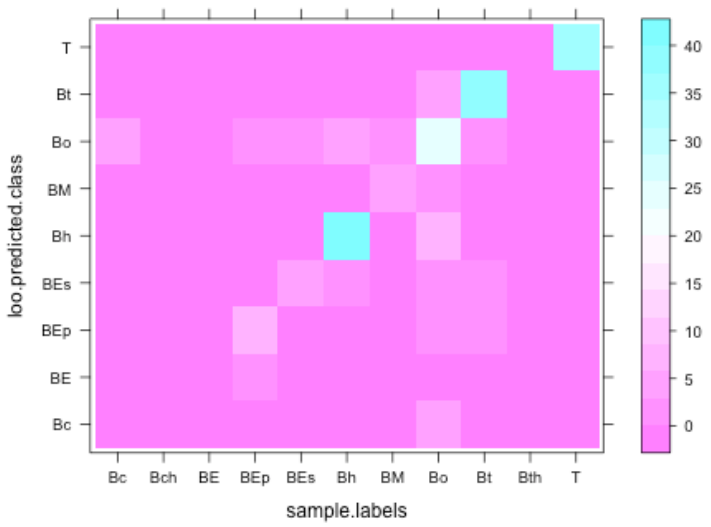
Forward stepwise feature selection



Leave-one-out cross-validation – 20 genes , top-ranking by variance

- Cross-validation of Linear Discriminant Analysis with Den Boer (2009).
- Variables: **20** top-ranking probesets **sorted by decreasing variances**.
- Hit rate: proportion of correct predictions
  - Correct (diagonal): 152
  - Total: 187
  - Hit rate: 81.3%
  - Error rate: 18.7%

	loo.predicted.class								
sample.labels	Bc	BE	BEp	BEs	Bh	BM	Bo	Bt	T
Bc	0	0	0	0	0	0	4	0	0
Bch	0	0	0	0	0	0	0	0	0
BE	0	0	0	0	0	0	0	0	0
BEp	0	1	6	0	0	0	1	0	0
BEs	0	0	0	3	0	0	1	0	0
Bh	0	0	0	1	40	0	3	0	0
BM	0	0	0	0	0	3	1	0	0
Bo	3	0	2	2	7	1	25	4	0
Bt	0	0	2	1	0	0	1	39	0
Bth	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	36

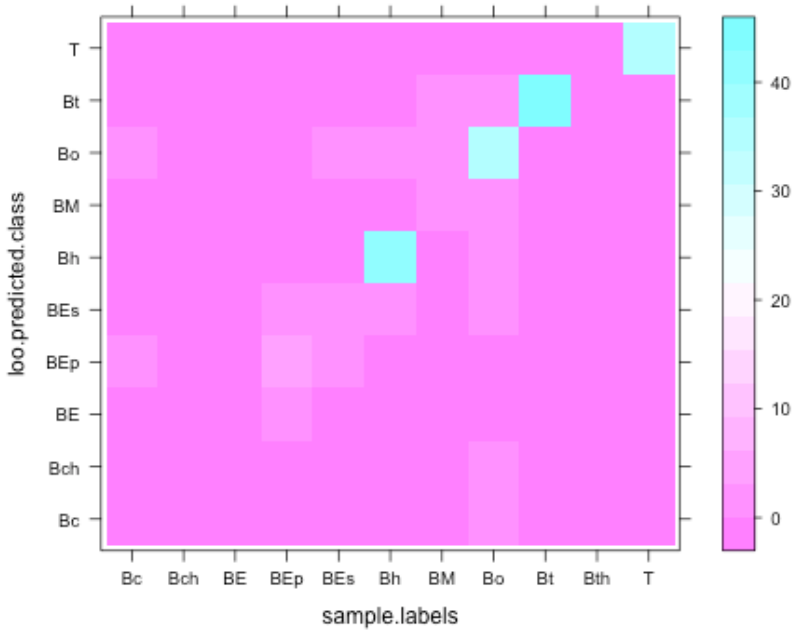


Leave-one-out cross-validation – 20 genes top-ranked by various criteria

- Cross-validation of Linear Discriminant Analysis with Den Boer (2009).
- Variables: 20 top-ranking probesets **sorted by multi-criterion rank** (variance + two-groups Welch tests).
- Hit rate: proportion of correct predictions
  - Correct (diagonal): 164
  - Total: 187
  - Hit rate: **87.7**%
  - Error rate: 12.3%

Predicted class

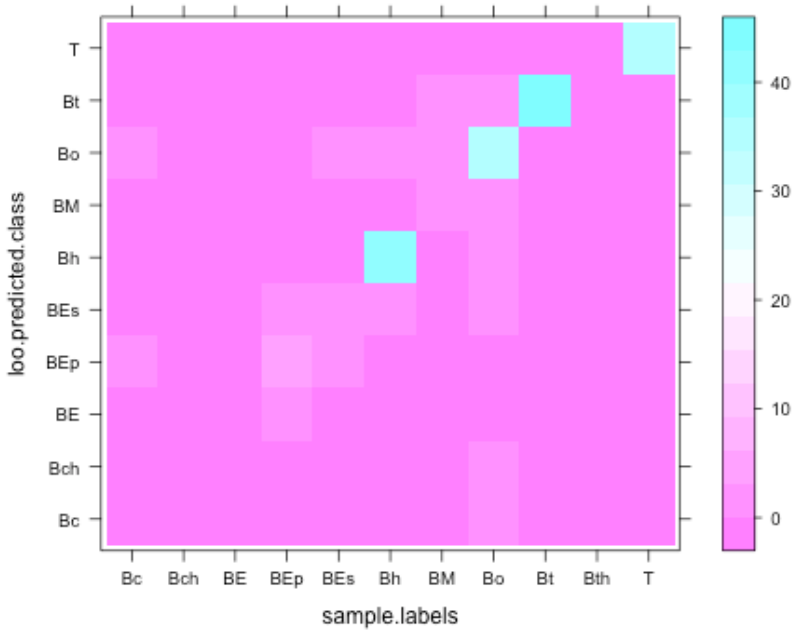
Known class	Bc	Bch	BE	BEp	BEs	Bh	BM	Bo	Bt	Bth	T	Total
Bc	**0**	0	0	1	0	0	0	3	0	0	0	**4**
Bch	0	**0**	0	0	0	0	0	0	0	0	0	**0**
BE	0	0	**0**	0	0	0	0	0	0	0	0	**0**
BEp	0	0	1	**6**	1	0	0	0	0	0	0	**8**
BEs	0	0	0	1	**2**	0	0	1	0	0	0	**4**
Bh	0	0	0	0	1	**41**	0	2	0	0	0	**44**
BM	0	0	0	0	0	0	**2**	1	1	0	0	**4**
Bo	2	1	0	0	1	3	1	**34**	2	0	0	**44**
Bt	0	0	0	0	0	0	0	0	**43**	**0**	0	**43**
Bth	0	0	0	0	0	0	0	0	**0**	**0**	0	**0**
T	0	0	0	0	0	0	0	0	0	0	**36**	**36**
Total	**2**	**1**	**1**	**8**	**5**	**44**	**3**	**41**	**46**	**0**	**36**	**187**



Leave-one-out cross-validation – 100 genes top-ranked by various criteria

- Cross-validation of Linear Discriminant Analysis with Den Boer (2009).
- Variables: **100** top-ranking probesets sorted by multi-criterion rank (variance + two-groups Welch tests).
- Hit rate: proportion of correct predictions
  - Correct (diagonal): 168
  - Total: 187
  - Hit rate: **89.9**%
  - Error rate: 10.2%

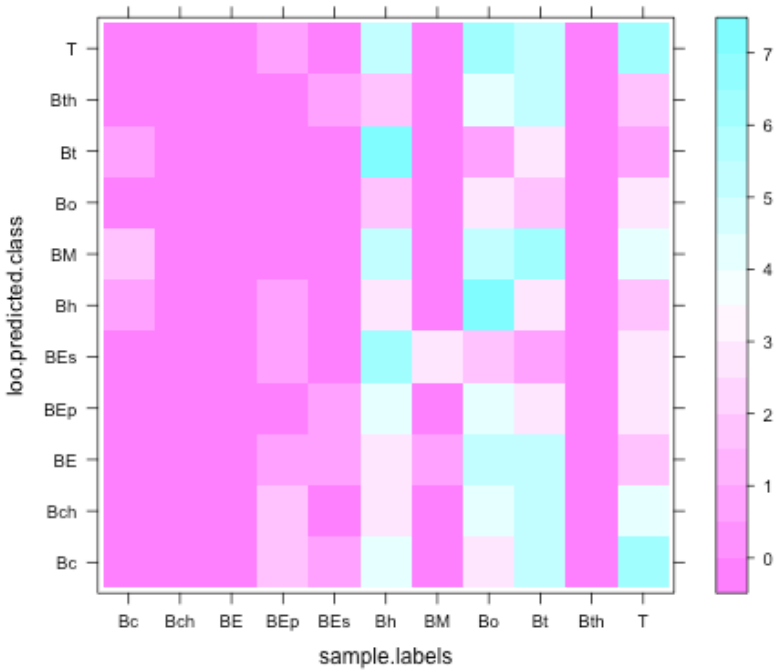
loo.predicted.class								
sample.labels	Bc	BEp	BEs	Bh	BM	Bo	Bt	T
Bc	0	0	0	2	0	2	0	0
Bch	0	0	0	0	0	0	0	0
BE	0	0	0	0	0	0	0	0
BEp	0	8	0	0	0	0	0	0
BEs	0	0	3	0	0	1	0	0
Bh	0	0	1	41	0	2	0	0
BM	0	0	1	0	2	1	0	0
Bo	2	1	0	5	0	35	1	0
Bt	0	0	0	0	0	0	43	0
Bth	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	36



Leave-one-out cross-validation – 200 genes top-ranked by various criteria

- Cross-validation of Linear Discriminant Analysis with Den Boer (2009).
- Variables: **200** top-ranking probesets sorted by multi-criterion rank (variance + two-groups Welch tests).
- Hit rate: proportion of correct predictions
  - Correct (diagonal): 15
  - Total: 187
  - Hit rate: 8%
  - Error rate: **92**%

loo.predicted.class											
sample.labels	Bc	Bch	BE	BEp	BEs	Bh	BM	Bo	Bt	Bth	T
Bc	0	0	0	0	0	1	2	0	1	0	0
Bch	0	0	0	0	0	0	0	0	0	0	0
BE	0	0	0	0	0	0	0	0	0	0	0
BEp	2	2	1	0	1	1	0	0	0	0	1
BEs	1	0	1	1	0	0	0	0	0	1	0
Bh	4	3	3	4	6	3	5	2	7	2	5
BM	0	0	1	0	3	0	0	0	0	0	0
Bo	3	4	5	4	2	7	5	3	1	4	6
Bt	5	5	5	3	1	3	6	2	3	5	5
Bth	0	0	0	0	0	0	0	0	0	0	0
T	6	4	2	3	3	2	4	3	1	2	6



Technical note: approach followed by DenBoer (differs from here)

- Multi-groups discrimination with 6 subtypes only (T-ALL, ETV6–RUNX1-positive, hyperdiploid, E2A- rearranged, BCR–ABL1-positive and MLL-rearranged)
- Training: 190 cases (COALL)
- Inner loop
  - Three-fold cross-validation: 2/3 cases for training, 1/3 for evaluation.
  - 100 iterations
- Variable filtering:
  - for each subtype, selection of the 50 lowest p-values with Wilcoxon's test.
  - For BCR-ABL1 and MLL, used 40 probesets from another source.
- Learning algorithm: radial-kernal support vector machine.
- Selection of the least number of probes by backward selection.

Bh	hyperdiploid	44
Bo	pre-B ALL	44
Bt	TEL-AML1	43
T	T-ALL	36
BEp	E2A-rearranged (EP)	8
Bc	BCR-ABL	4
BEs	E2A-rearranged (E-sub)	4
BM	MLL	4
Bch	BCR-ABL + hyperdiploidy	1
BE	E2A-rearranged (E)	1
Bth	TEL-AML1 + hyperdiploidy	1

Summary - discriminant analysis

- Discriminant analysis is based on a set of quantitative predictor variables, and a single nominal criterion variable.
- A sample is used to build a set of discriminant functions (calibration), which is then used to assign additional units to classes (prediction).
- The discriminant function can be either linear or quadratic. Linear discriminant analysis relies on the assumption that the different classes have similar covariance matrices.
- The accuracy of the discriminant function can be evaluated in different ways.
  - On the whole sample (internal approach)
  - Splitting of the sample into training and testing set (holdout approach)
  - Successively discard each sample unit, build a discriminant function and predict the discarded unit (leave-one-out)
- The efficiency decreases with the p/N ratio. When this ratio is too low, there is a problem of over-fitting.
- Stepwise approaches consist in selecting the subset of variables which raises the highest efficiency.

KNN classifiers

K nearest neighbours

- Discriminant analysis is a global approach to classification: the discriminant rule is established in the same way for the whole data space, on the basis of group centres and covariance matrices. Discriminant analysis is thus a ***global classifier***.
- K nearest neighbour (***KNN***) classifiers takes a very different approach: at each position of the feature space
  - The K closest neighbour points from the training set are identified;
  - A vote is established as a function of the relative proportions of the respective training groups in this set of neighbours.
- KNN is thus a ***local classifier***.
- The choice of K drastically affects group assignments.

Support Vector Machines

Web resources

- Gist
  - Download http://microarray.cpmc.columbia.edu/gist/
  - Web interface http://svm.sdsc.edu/cgi-bin/nph-SVMsubmit.cgi

Old slides

Training set

- There is a subset of objects (in the case below, genes) which can be assigned to predefined classes (e.g. “phosphate”, “methionine” or “control”), on the basis of external information (e.g. biological knowledge).
- These classes will be used as criterion variable.
- Note : the sample class labels might contain some errors (misclassified objects).

Phosphate-responding genes

#	ORF	Gene name	Family
1	YBR093C	PHO5	PHO
2	YDR481C	PHO8	PHO
3	YAR071W	PHO11	PHO
4	YHR215W	PHO12	PHO
5	YOL001W	PHO80	PHO
6	YGR233C	PHO81	PHO
7	YML123C	PHO84	PHO
8	YPL031C	PHO85	PHO
9	YJL117W	PHO86	PHO
10	YCR037C	PHO87	PHO
11	YBR106W	PHO88	PHO
12	YBR296C	PHO89	PHO
13	YHR136C	SPL2	PHO

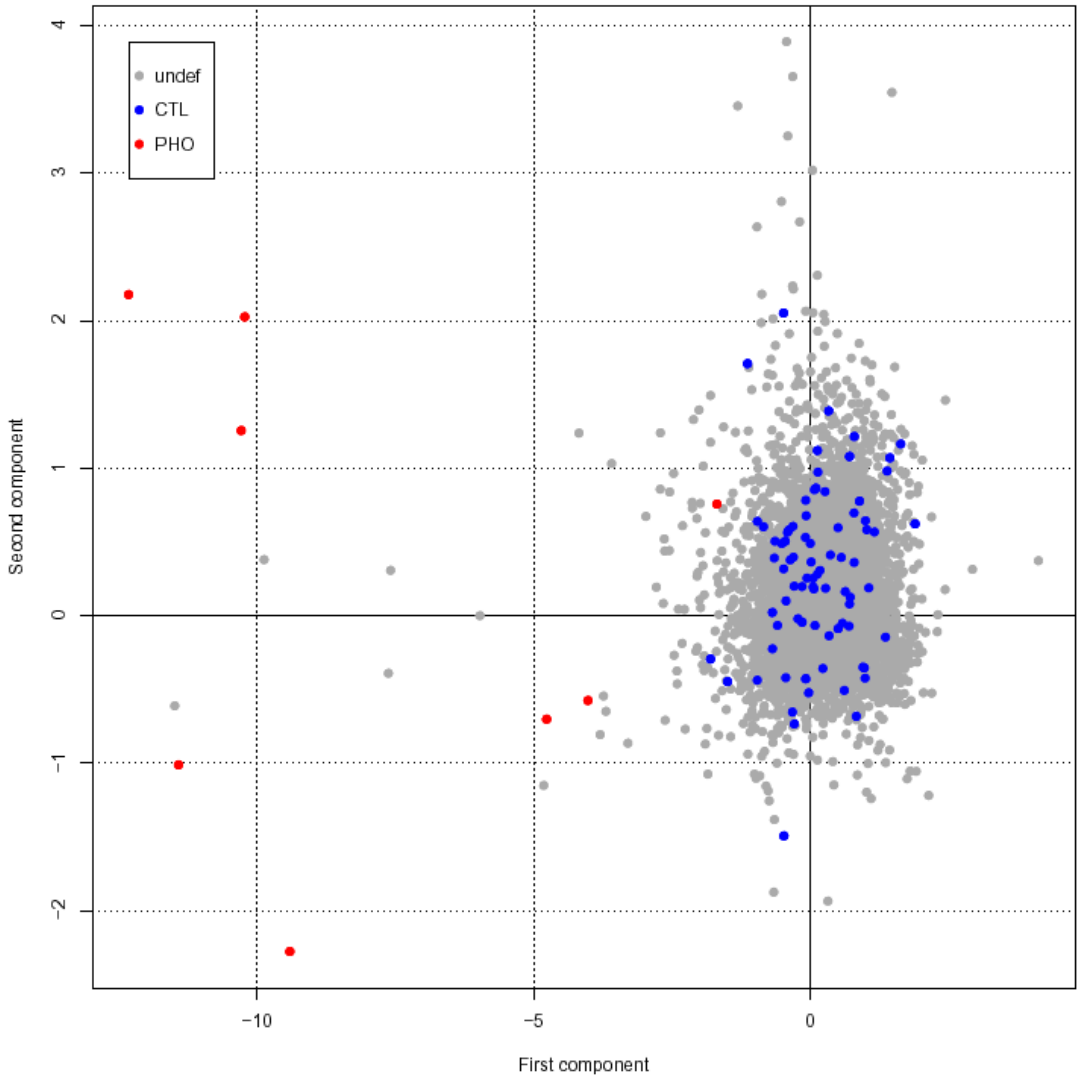
Methionine-responding genes

#	ORF	Gene name	Family
14	YBR213W	MET8	MET
15	YDR253C	MET32	MET
16	YDR502C	SAM2	MET
17	YER091C	MET6	MET
18	YFR030W	MET10	MET
19	YHL036W	MUP3	MET
20	YIL046W	MET30	MET
21	YIR017C	MET28	MET
22	YJR010W	MET3	MET
23	YJR137C	ECM17	MET
24	YKL001C	MET14	MET
25	YKR069W	MET1	MET
26	YLR180W	SAM1	MET
27	YLR303W	MET17	MET
28	YLR396C	VPS33	MET
29	YNL241C	ZWF1	MET
30	YNL277W	MET2	MET
31	YOL064C	MET22	MET
32	YPL038W	MET31	MET

Control genes

#	ORF	Gene name	Family
33	YAL038W	CDC19	CTL
34	YBL005W	PDR3	CTL
35	YBL005W-A	YBL005W-A	CTL
36	YBL005W-B	YBL005W-B	CTL
37	YBL030C	PET9	CTL
38	YBR006W	UGA5	CTL
39	YBR018C	GAL7	CTL
40	YBR020W	GAL1	CTL
41	YBR115C	LYS2	CTL
42	YBR184W	YBR184W	CTL
43	YCL018W	LEU2	CTL
44	YDL131W	LYS21	CTL
45	YDL182W	LYS20	CTL
46	YDL205C	HEM3	CTL
47	YDL210W	UGA4	CTL
48	YDR011W	SNQ2	CTL
49	YDR044W	HEM13	CTL
50	YDR234W	LYS4	CTL
51	YDR285W	ZIP1	CTL
…	…	…	…
112	YPR065W	ROX1	CTL
113	YPR138C	MEP3	CTL
114	YPR145W	ASN1	CTL

2-dimensional visualization of the sample



- If there are many variables, PCA can be used to visualize the sample on the planed formed by the two principal components.
- Example: gene expression data
  - MET genes seem undistinguishable from CTL genes (they are indeed not expected tor espond to phosphate)
  - Most PHO genes are clearly distant from the main cloud of points.
  - Some PHO genes are mixed with the CTL genes.