DUBii - Module - Statistics with R

Machine learning

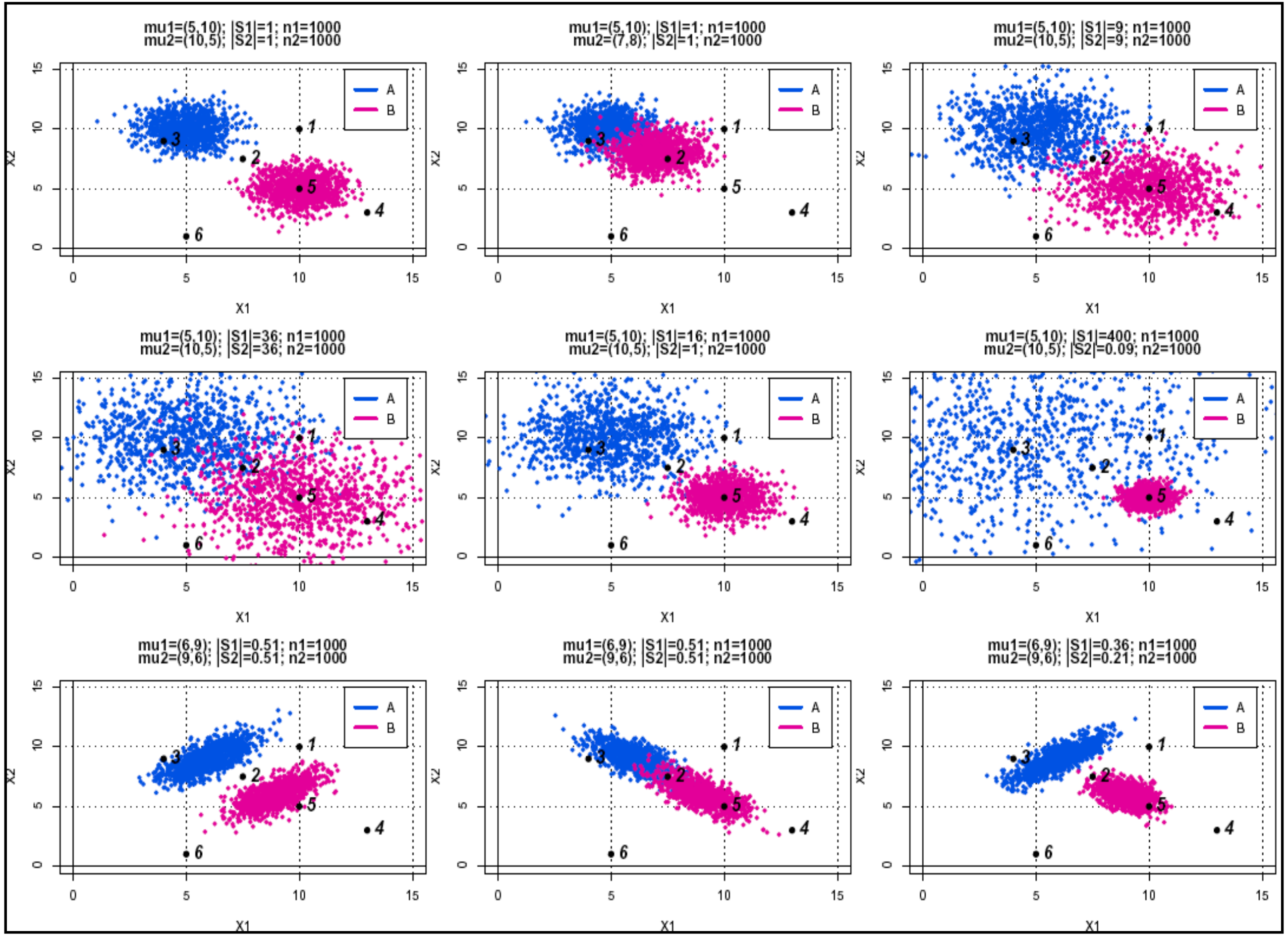
Jacques van Helden
ORCID 0000-0002-8799-8584

Institut Français de Bioinformatique (**IFB**)
French node of the European **ELIXIR** bioinformatics infrastructure

Aix-Marseille Université (AMU)
Lab. Theory and Approaches of Genomic Complexity (**TAGC**)

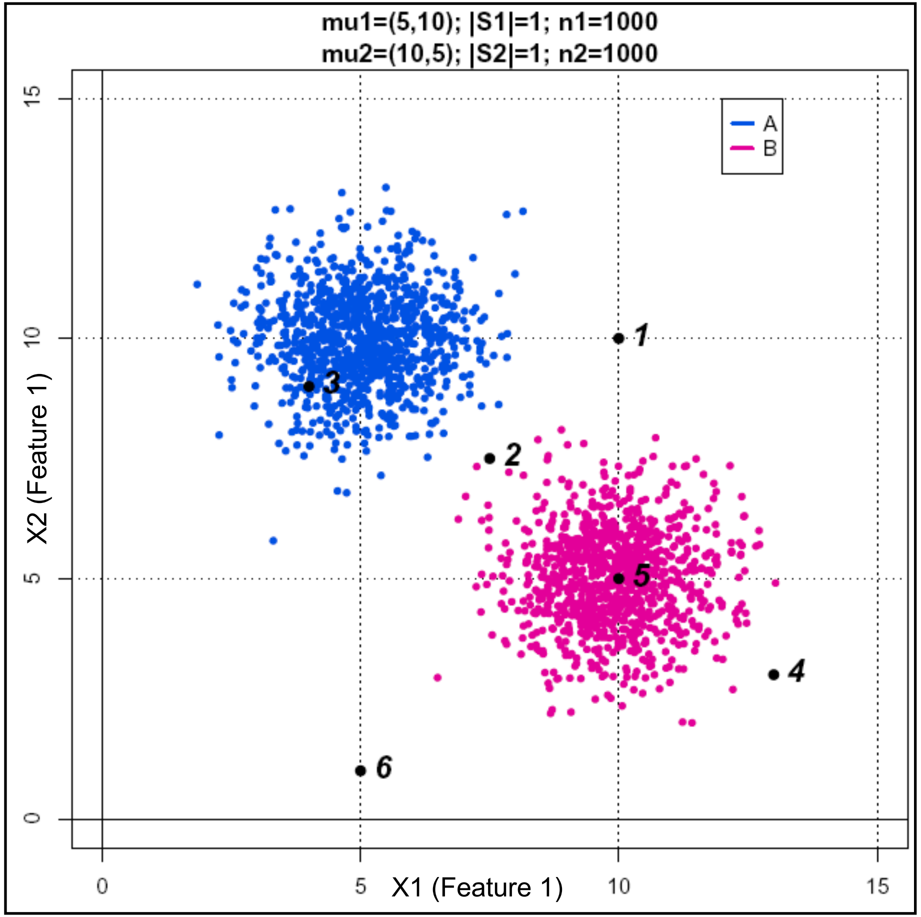
Brain-learning exercise : assign individuals to groups based on their features

Conceptual illustration with two predictor variables



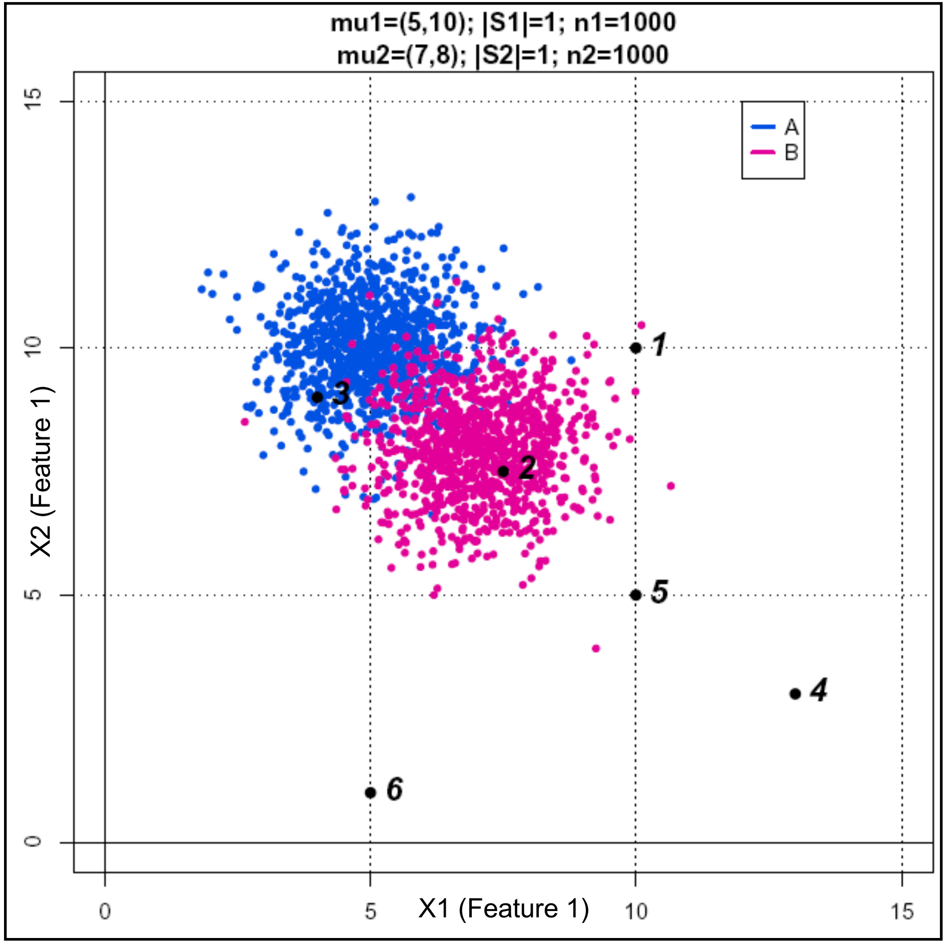
- In the next slides, we will provide you with a higher-resolution of the plots, which represent represent a study case.
- Exercise: assign intuitively each individual (black dot) to one of the two groups (A, B).
  - At each step, ask yourself the following questions.
  - Which criterion did you use to assign an individual to a group?
  - 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?

Conceptual illustration with two variables – Study case 1



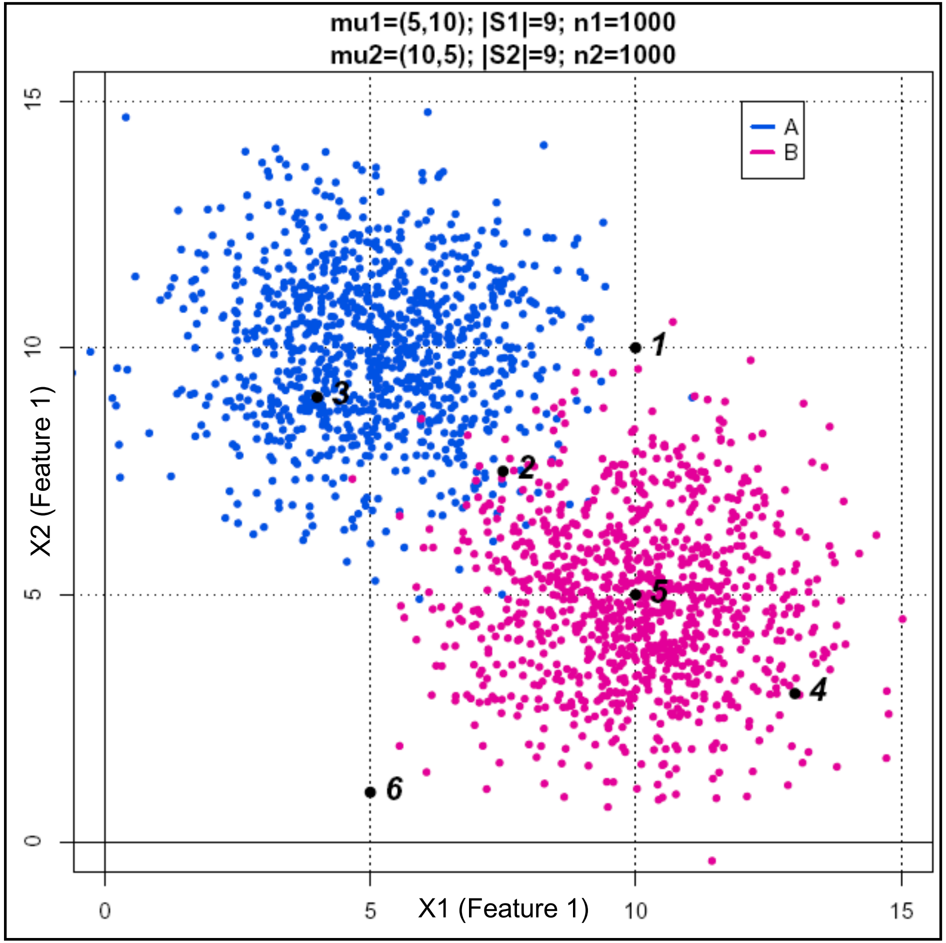
- Inspect the distribution of points for the two groups of individuals (pink, blue) on the 2-dimensional feature space.

Conceptual illustration with two variables – Study case 2



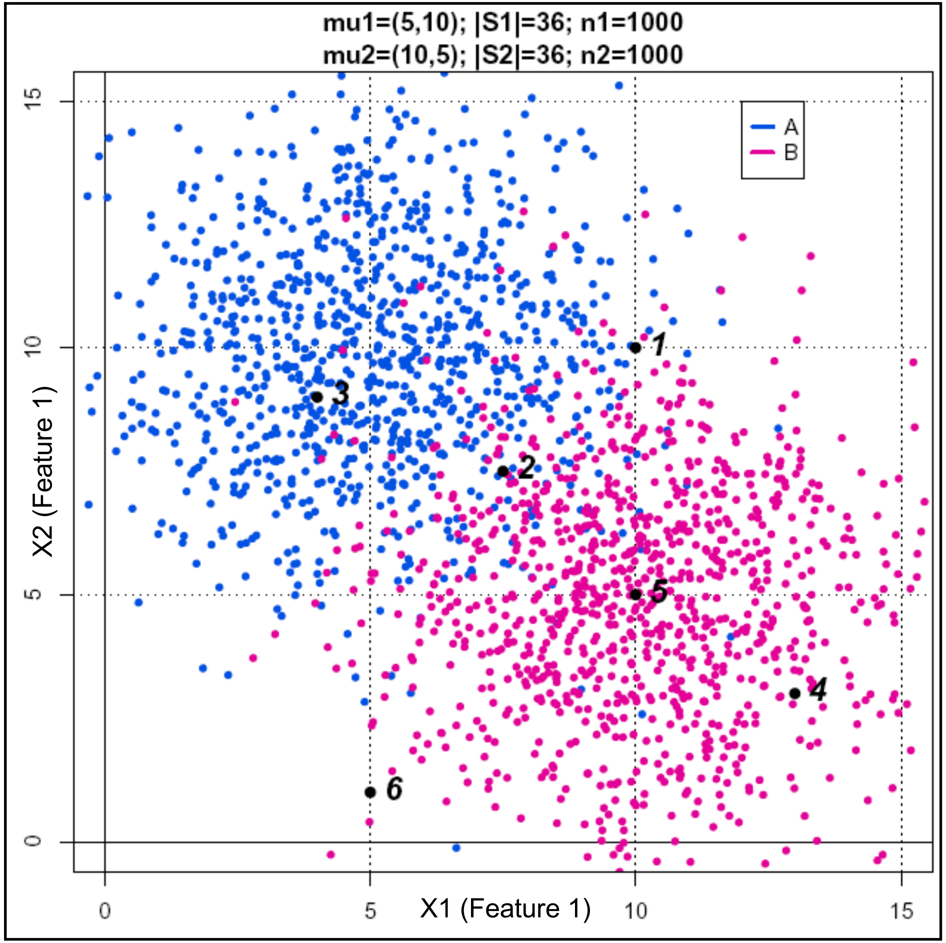
- Effect of the group ***centre location***.

Conceptual illustration with two variables – Study case 3



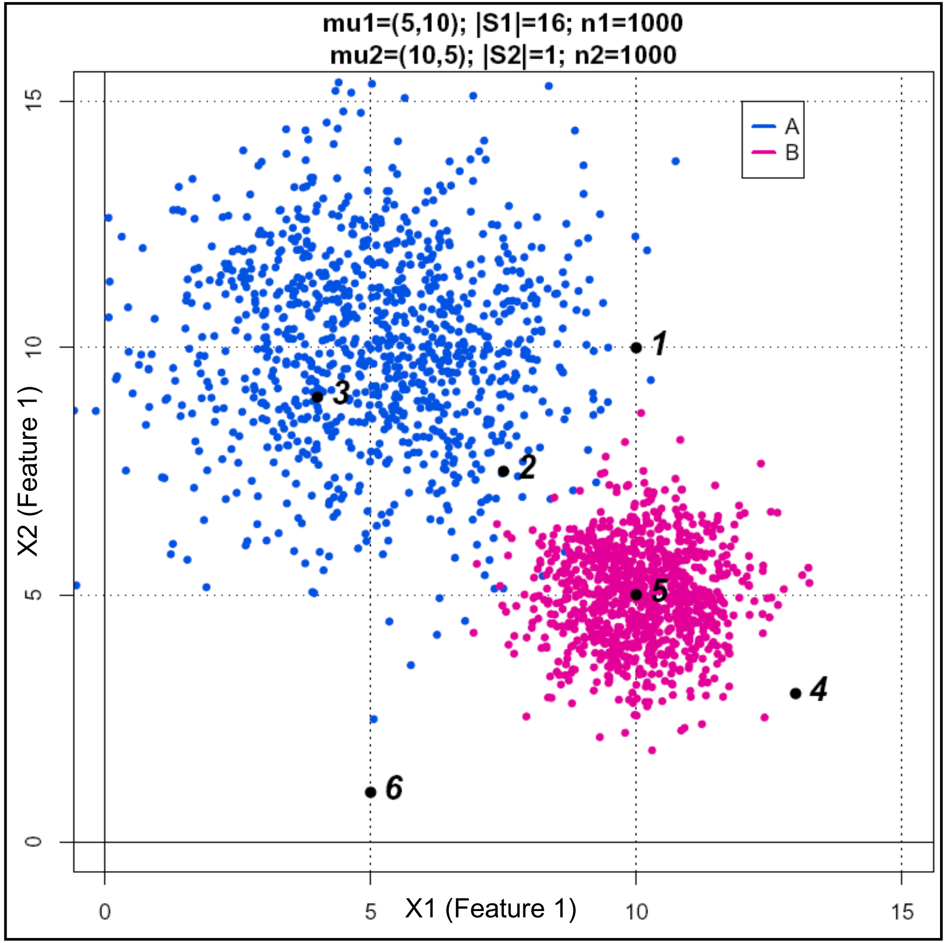
- Effect of the group ***variance***.

Conceptual illustration with two variables – Study case 4



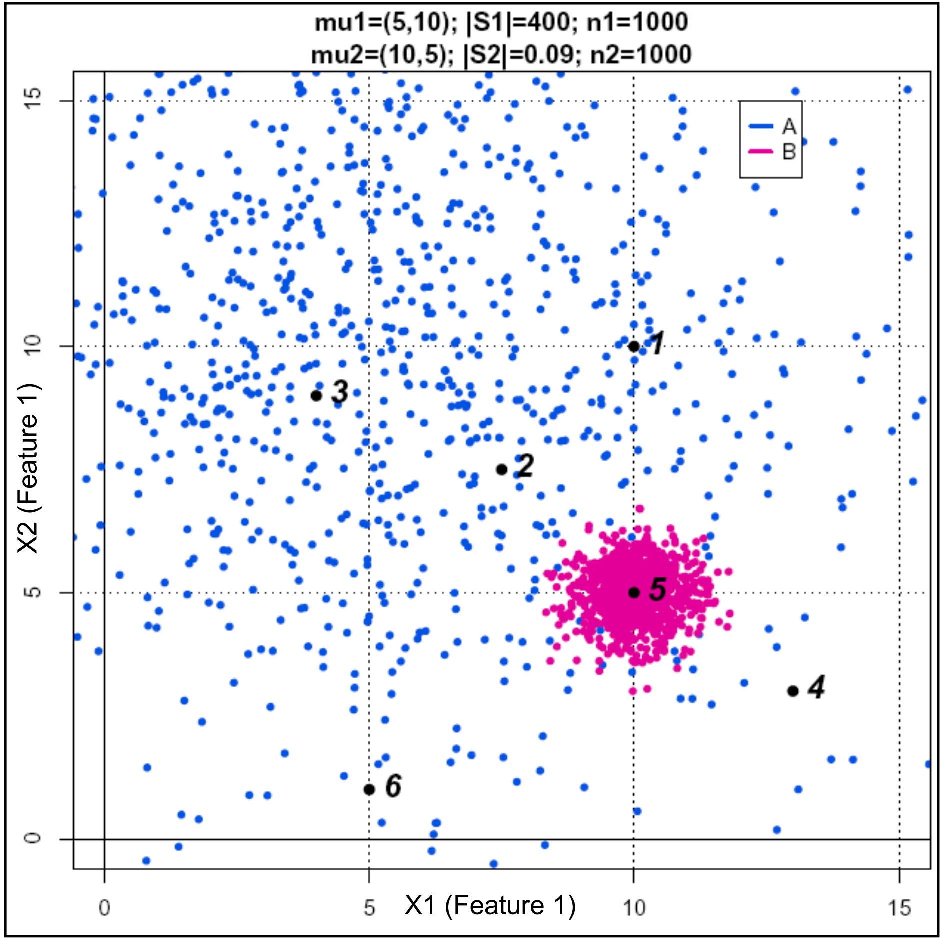
- Effect of the group ***variance***.

Conceptual illustration with two variables – Study case 5



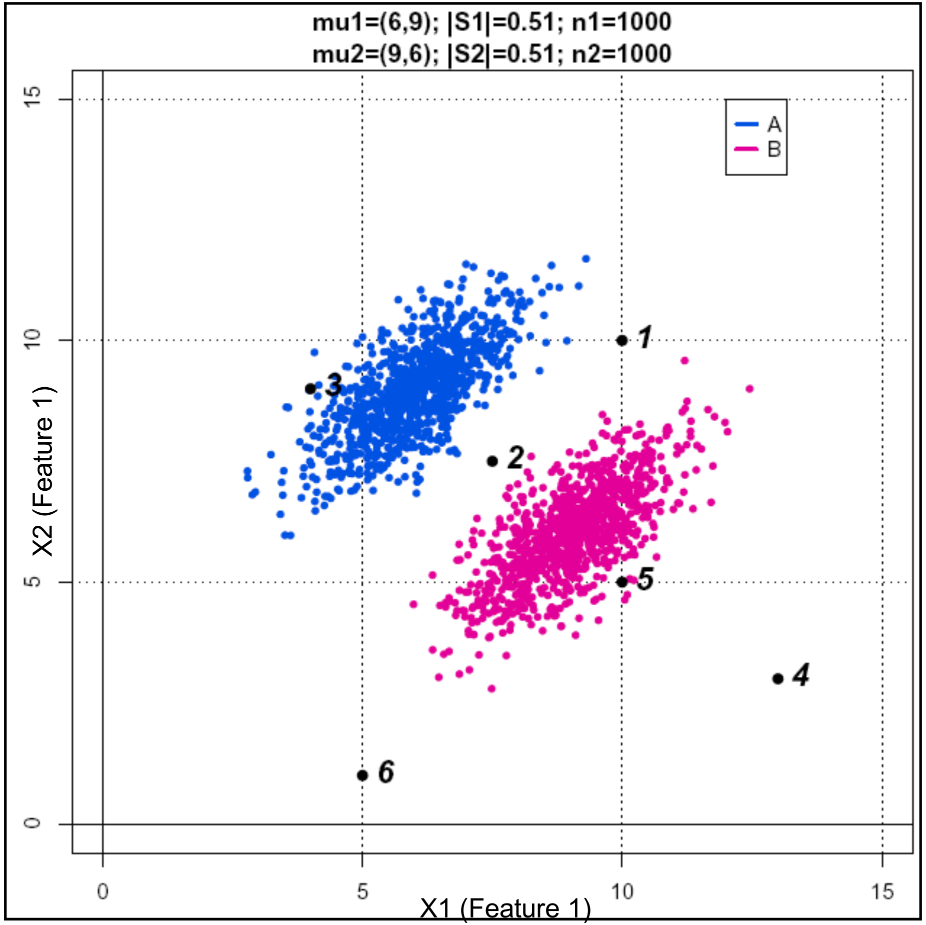
- Impact of the group-specific variances (heteroscedasticity of the data)

Conceptual illustration with two variables – Study case 6



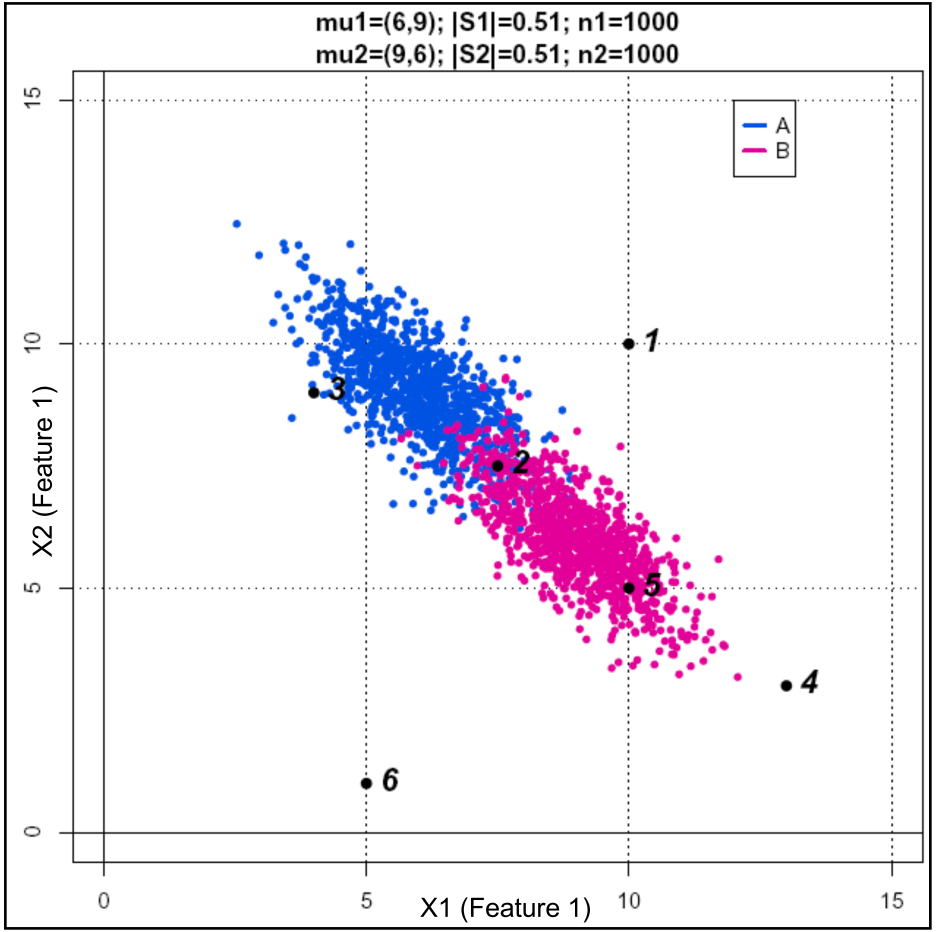
- Impact of the group-specific variances (heteroscedasticity of the data)

Conceptual illustration with two variables – Study case 7



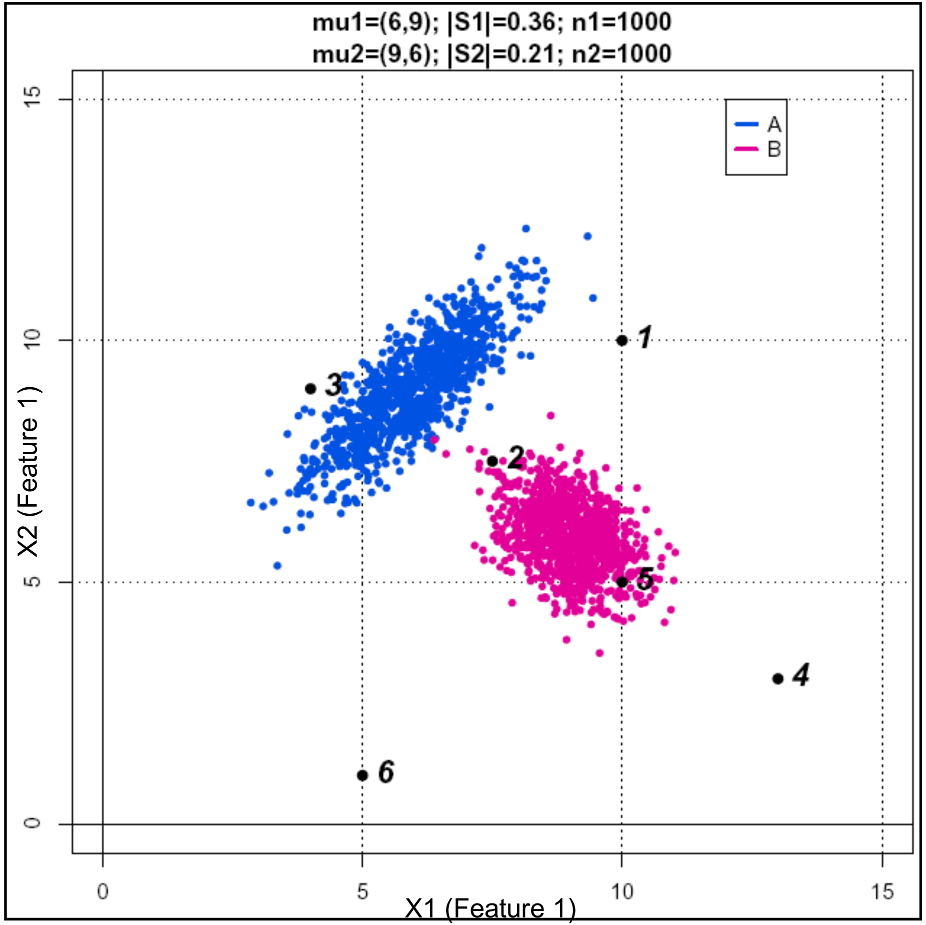
- Effect of the ***covariance*** between features.

Conceptual illustration with two variables – Study case 8



- Effect of the ***covariance*** between features

Conceptual illustration with two variables – Study case 9



- Group-specific ***covariances*** between features.
  - The two groups have different covariance matrices: the clouds of points are elongated in different directions.
  - How does this difference affects group assignments ?

Statistics Applied to Bioinformatics

Multivariate analysis
Introduction

Jacques van Helden
ORCID 0000-0002-8799-8584

Institut Français de Bioinformatique (**IFB**)
French node of the European **ELIXIR** bioinformatics infrastructure

Aix-Marseille Université (AMU)
Lab. Theory and Approaches of Genomic Complexity (**TAGC**)

Multivariate data

- Each row represents one object (also called unit)
- Each column represents one variable

	variable 1	**variable 2**	**…**	**variable p**
individual 1	x_{11}	x_{21}	…	x_{p1}
individual 2	x_{12}	x_{22}	…	x_{p2}
individual 3	x_{13}	x_{23}	…	x_{p3}
individual 4	x_{14}	x_{24}	…	x_{p4}
individual 5	x_{15}	x_{25}	…	x_{p5}
individual 6	x_{16}	x_{26}	…	x_{p6}
individual 7	x_{17}	x_{27}	…	x_{p7}
individual 8	x_{18}	x_{28}	…	x_{p8}
…	…	…	…	…
individual n	x_{1n}	x_{2n}	…	x_{pn}

Multivariate data with an outcome variable

- The outcome variable (also called criterion variable) can be
  - qualitative (nominal) : classes (e.g. cancer type)
  - quantitative (e.g. survival expectation for a cancer patient)

	Predictor variables				**Outcome variable**
	variable 1	**variable 2**	**...**	**variable p**	**variable p+1**
individual 1	x_{11}	x_{21}	…	x_{p1}	y_{1}
individual 2	x_{12}	x_{22}	…	x_{p2}	y_{2}
individual 3	x_{13}	x_{23}	…	x_{p3}	y_{3}
individual 4	x_{14}	x_{24}	…	x_{p4}	y_{4}
individual 5	x_{15}	x_{25}	…	x_{p5}	y_{5}
individual 6	x_{16}	x_{26}	…	x_{p6}	y_{6}
individual 7	x_{17}	x_{27}	…	x_{p7}	y_{7}
individual 8	x_{18}	x_{28}	…	x_{p8}	y_{8}
...	…	…	…	…	…
individual n	x_{1n}	x_{2n}	…	x_{pn}	y_{n}

Predictive approaches - Training set

- The training set is used to build a predictive function
- This function is used to predict the value of the outcome variable for new objects

Training set

	Predictor variables				**Outcome variable**
	variable 1	**variable 2**	**...**	**variable p**	**variable p+1**
individual 1	x_{11}	x_{21}	…	x_{p1}	y_1
individual 2	x_{12}	x_{22}	…	x_{p2}	y_2
individual 3	x_{13}	x_{23}	…	x_{p3}	y_3
...	…	…	…	…	…
individual N_train	x_{1n}	x_{2n}	…	x_{pn}	y_n

Set to predict

	Predictor variables				**Outcome variable**
	variable 1	**variable 2**	**...**	**variable p**	**variable p+1**
individual 1	x_{11}	x_{21}	…	x_{p1}	?
individual 2	x_{12}	x_{22}	…	x_{p2}	?
individual 3	x_{13}	x_{23}	…	x_{p3}	?
...	…	…	…	…	…
individual N_pred	x_{1n}	x_{2n}	…	x_{pn}	?

Training set

	Predictor variables				Outcome variable
	variable 1	variable 2	...	variable p	variable p+1
individual 1	x_{11}	x_{12}	…	x_{1p}	y_1
individual 2	x_{21}	x_{22}	…	x_{2p}	y_2
individual 3	x_{31}	x_{32}	…	x_{3p}	y_3
...	…	…	…	…	…
individual ntrain	x_{n1}	x_{n2}	…	x_{np}	y_n

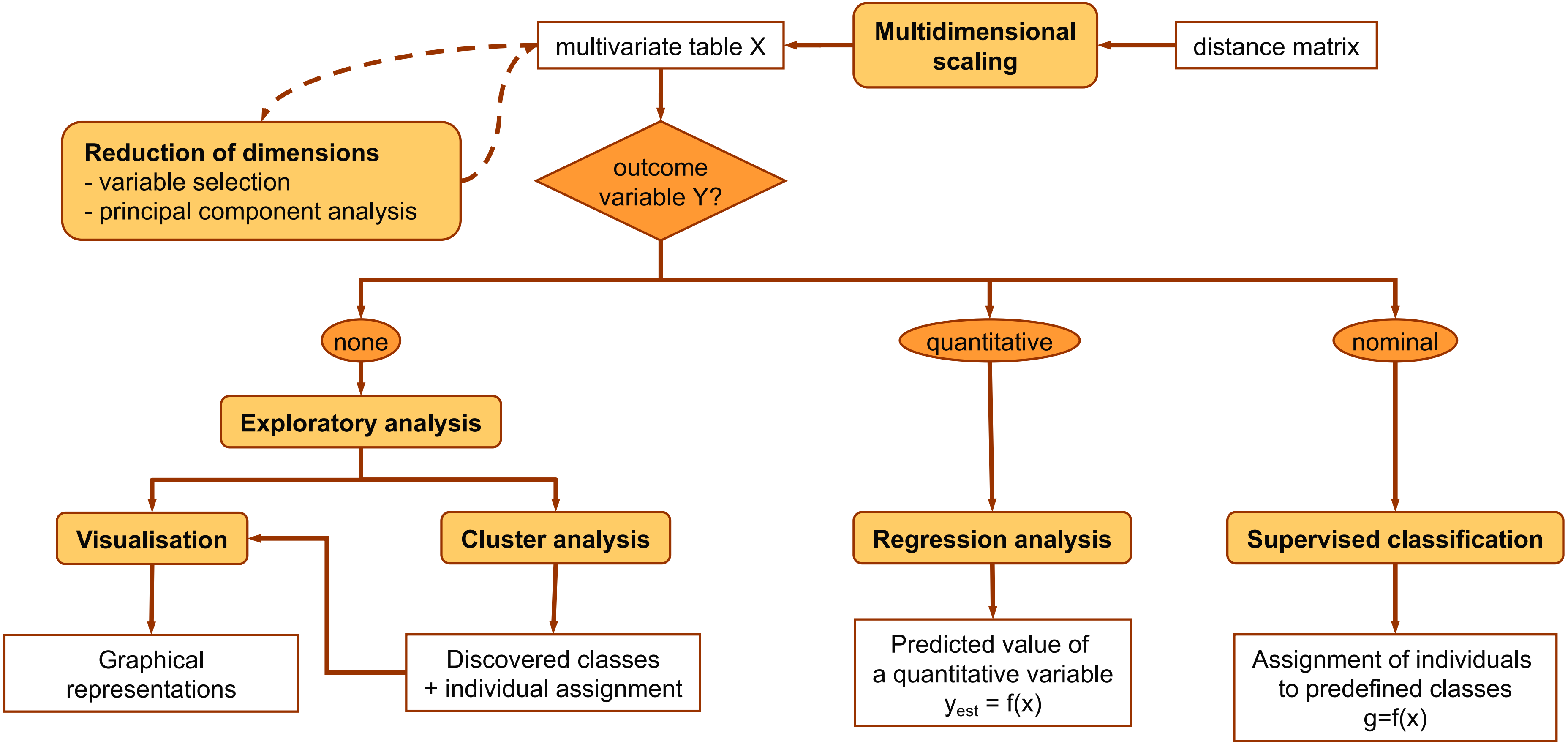
Testing set

	Predictor variables				Outcome variable	
	variable 1	variable 2	...	variable p	variable p+1 (known value)	variable p+1 (predicted)
individual 1	x_{11}	x_{12}	…	x_{1p}	y_1	y'_1
individual 2	x_{21}	x_{22}	…	x_{2p}	y_2	y'_2
individual 3	x_{31}	x_{32}	…	x_{3p}	y_3	y'_3
...	…	…	…	…	…	…
individual ntest	x_{n1}	x_{n2}	…	x_{np}	y_{ntest}	y'_{ntest}

et to predict

	Predictor variables				Outcome variable
	variable 1	variable 2	...	variable p	variable p+1
individual 1	x_{11}	x_{12}	…	x_{1p}	?
individual 2	x_{21}	x_{22}	…	x_{2p}	?
individual 3	x_{31}	x_{32}	…	x_{3p}	?
...	…	…	…	…	…

Flowchart of the approaches in multivariate analysis



multivariate table X
Multidimensional scaling
distance matrix
Reduction of dimensions
- variable selection
- principal component analysis
outcome variable Y?
none
quantitative
nominal
Exploratory analysis
Visualisation
Cluster analysis
Regression analysis
Supervised classification
Graphical representations
Discovered classes + individual assignment
Predicted value of a quantitative variable $y_{est} = f(x)$
Assignment of individuals to predefined classes g=f(x)

Quizz

Check your understanding of the concepts presented in the previous slides by applying them to your own data.

1. Describe in one sentence a typical case of multidimensional data that is handled in your domain.
2. Explain how you would organise this dataset into a multivariate structure
   - □ What would correspond to the individuals?
   - □ What would correspond to the variables?
   - □ How many individuals (n) would you have?
   - □ How many variables (p) would you have?
   - □ Do you dispose of one or several outcome variable(s)?
   - □ If so, are these quantitative, qualitative or both?
3. Based on the conceptual framework defined above, which kind of approaches would be you envisage to extract which kind of relevant information from this data? Note that several approaches can be combined to address different questions.

Historical (vintage) examples

Historical example of clustering heat map

- Spellman et al. (1998).
- Systematic detection of genes regulated in a periodic way during the cell cycle.
- Several experiments were regrouped, with various ways of synchronization (elutriation, cdc mutants, …)
- ~800 genes showing a periodic patterns of expression were selected (by Fourier analysis)

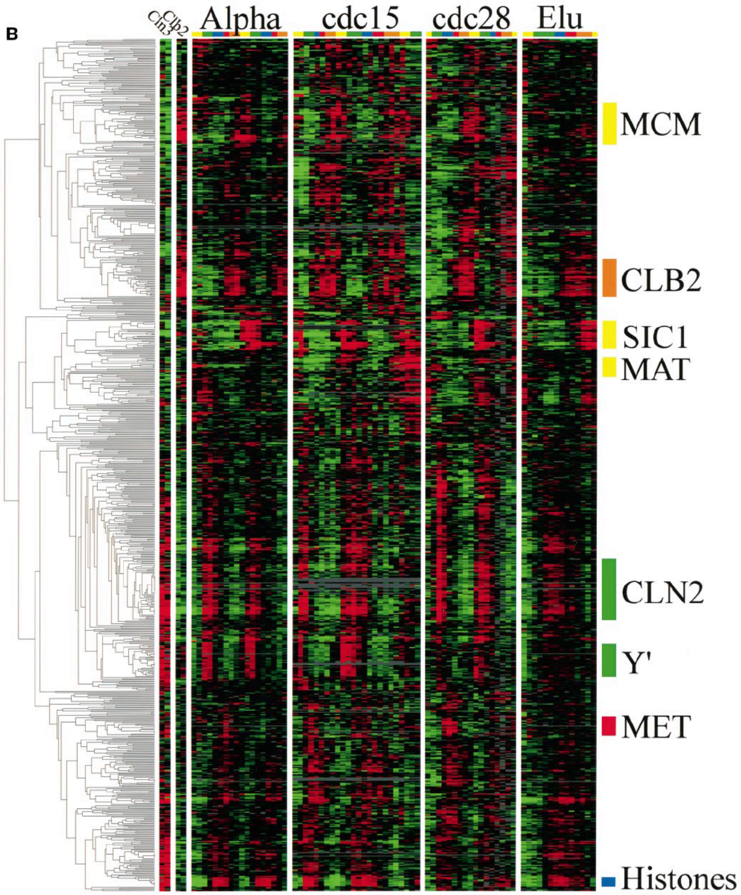
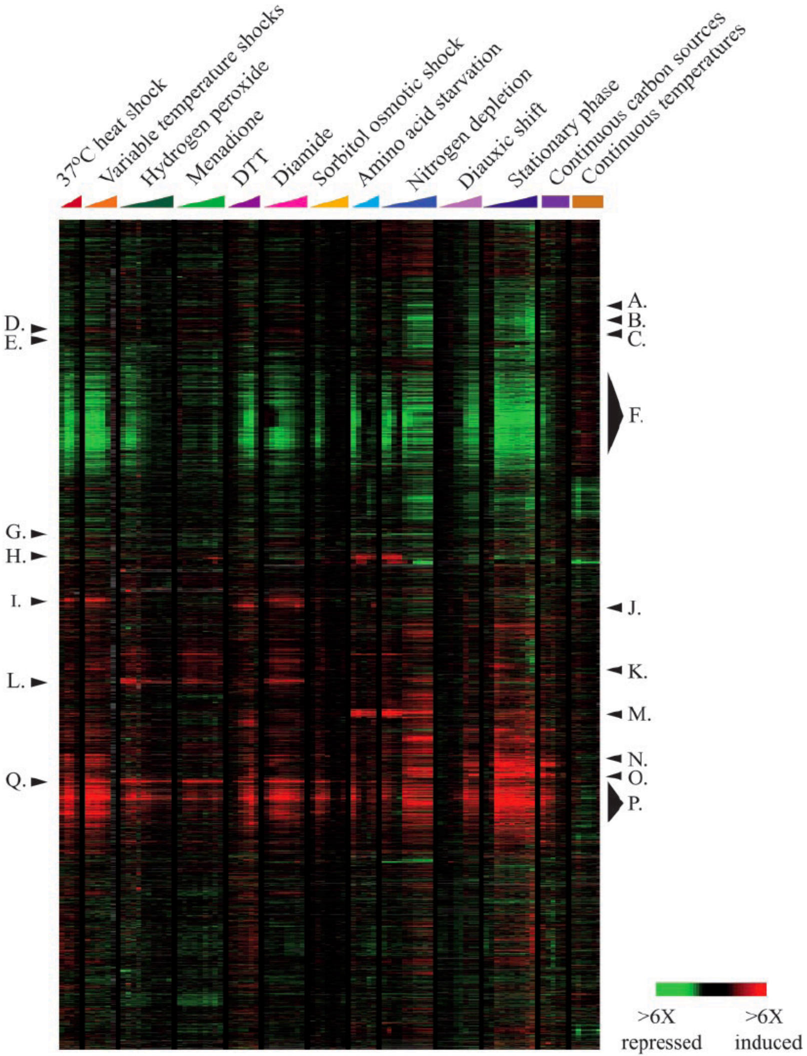


Figure 1. (cont).

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell 9, 3273-97.Time profiles of yeast cells followed during cell cycle.

Stress response in yeast

- Gasch et al. (2000) tested the transcriptional response of yeast genome to
  - Various stress conditions (heat shock, osmotic shock, …)
  - Drugs
  - Alternative carbon sources
  - …
- The heatmap shows clusters of genes having similar profiles of responses to the different types of stress.



Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. & Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11, 4241-57.

Cancer types (Golub, 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.
- Goal: use these genes as molecular signatures for the diagnostic of new patients.

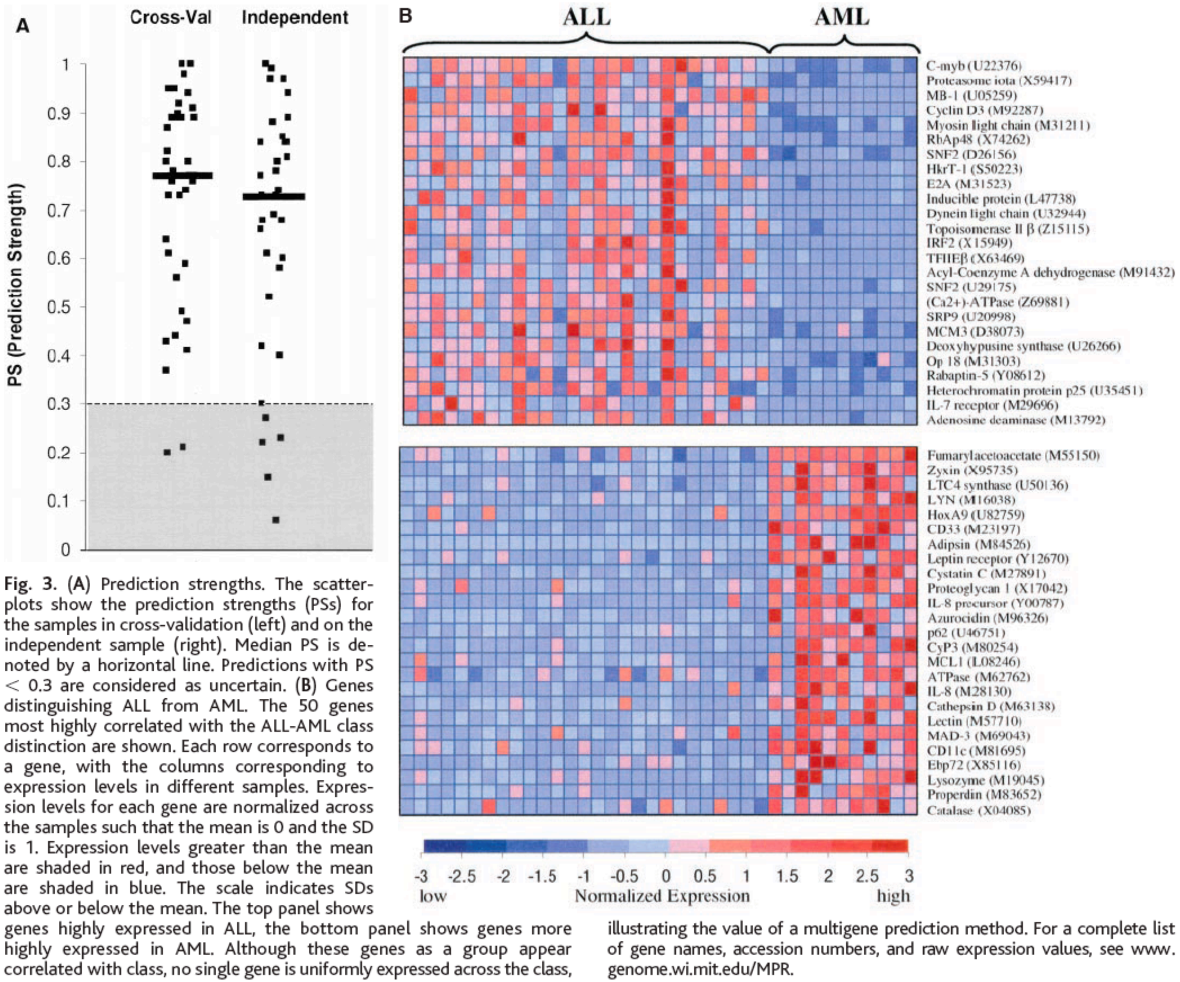


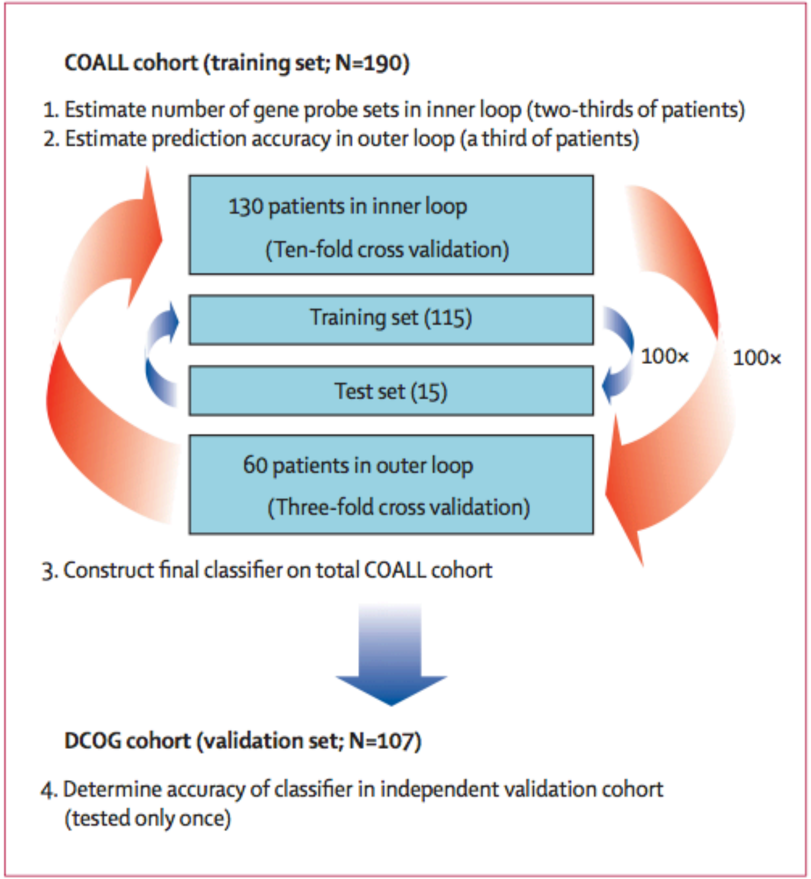
Fig. 3. (**A**) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. (**B**) Genes distinguishing ALL 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.

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.
- edure relying on an ss-validation.

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



COALL cohort (training set; N=190)

1. Estimate number of gene probe sets in inner loop (two-thirds of patients)
2. Estimate prediction accuracy in outer loop (a third of patients)

130 patients in inner loop (Ten-fold cross validation)

Training set (115)

Test set (15)

60 patients in outer loop (Three-fold cross validation)

3. Construct final classifier on total COALL cohort

DCOG cohort (validation set; N=107)

4. Determine accuracy of classifier in independent validation cohort (tested only once)

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

- 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 selects 100 gens 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 the testing set (right).
- The heatmap is bi-clustered, in order to identify simultaneously the groups of patients (rows), and groups of genes (columns) based on the similarity between expression profiles.

- 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.

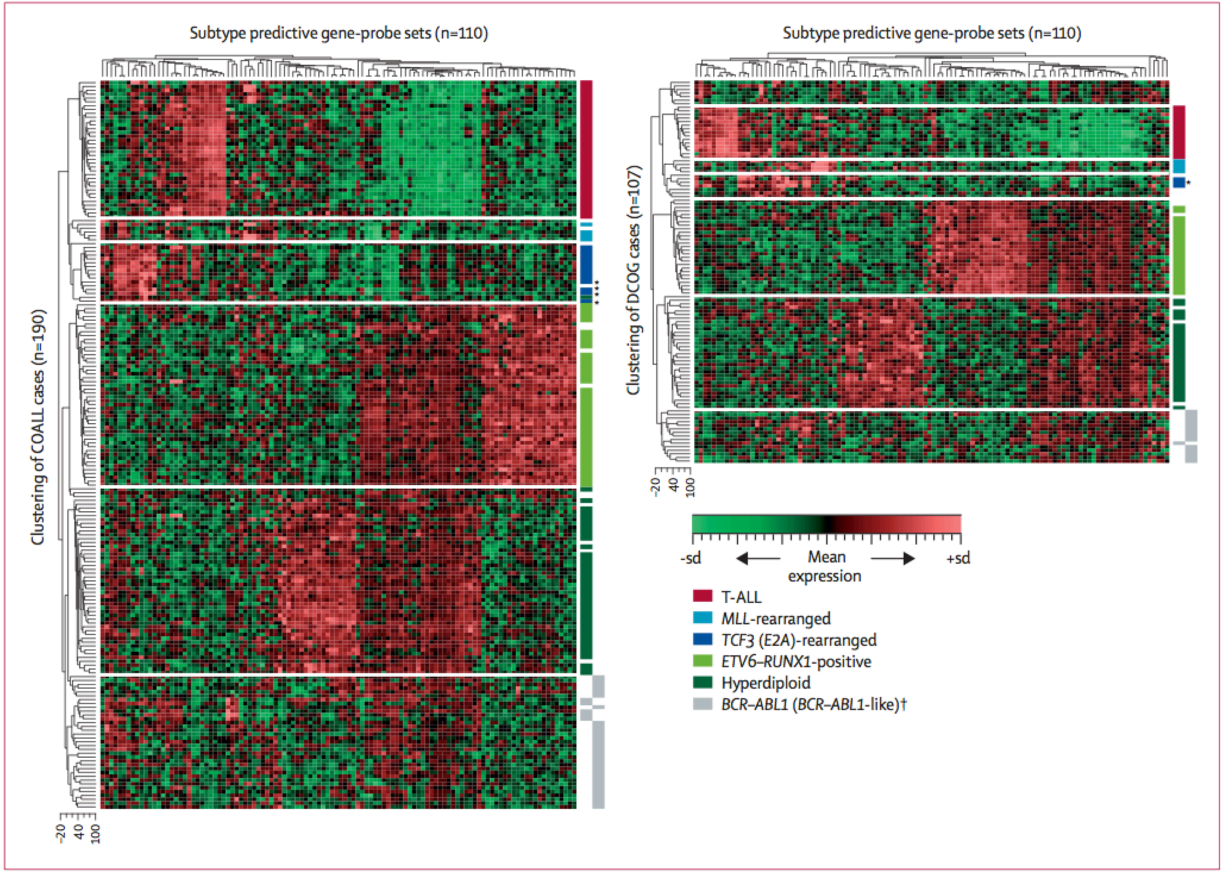


***Figure 2:* Clustering of ALL subtypes by gene-expression profiles**

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.

Statistics for bioinformatics

Supervised classification

Jacques van Helden

Aix-Marseille Université (AMU)
Lab. Theory and Approaches of Genomic Complexity (**TAGC**)

Institut Français de Bioinformatique (**IFB**)
French node of the European **ELIXIR** bioinformatics infrastructure

https://orcid.org/0000-0002-8799-8584

Supervised classification - Introduction

- In the previous chapter, we presented the problem of ***clustering***, which consists in grouping objects *without any a priori definition of the groups*. The group definition 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, ...)
- The classifier can be built with a **training set**, and used later for classifying new objects. This is called **supervised classification**.

Supervised classification methods

- There are many alternative methods for supervised classification
  - Discriminant analysis (linear: **LDA** or quadratic: **QDA**)
  - Bayesian classifier
  - K-nearest neighbours (**KNN**)
  - Support Vector Machine (**SVM**)
  - Decision tree
  - Random Forest (**RF**)
  - Neural network (**NN**)
  - ...
- Questions
  - Which method should we choose?
  - How should we tune its parameters?
  - How to evaluate the respective performances of the methods and parametric choices?

Supervised classification methods

Choosing the best method is not trivial

- Some methods rely on strong assumptions.
  - LDA and QDA: multivariate normality.
  - LDA: all the classes have the same covariance matrix.
- Some methods implicitly rely on Euclidian distance (e.g. KNN)
- Some methods require a large training set, to avoid over-fitting.
- Global vs local classifiers.
  - Global classifiers (e.g. LDA, QDA): same classification rule in the whole data space. The rule is built on the whole training set.
  - Local classifiers (e.g. KNN): rules are made in different sub-spaces on the basis of the neighbouring training points.
- The choice of the method thus depends on the structure and on the size of the data sets.

Choosing the best parameters is not trivial

- KNN: number of neighbours
- LDA, QDA: prior/posterior probabilities
- SVM: kernel
- Decision trees
- RF: number of iterations
- …

Study case 1: ALL versus AML (data from Golub et al., 1999)

Cancer types (Golub, 1999)

- A founding paper: 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.
- Goal: use these genes as molecular signatures for the diagnostic of new patients.

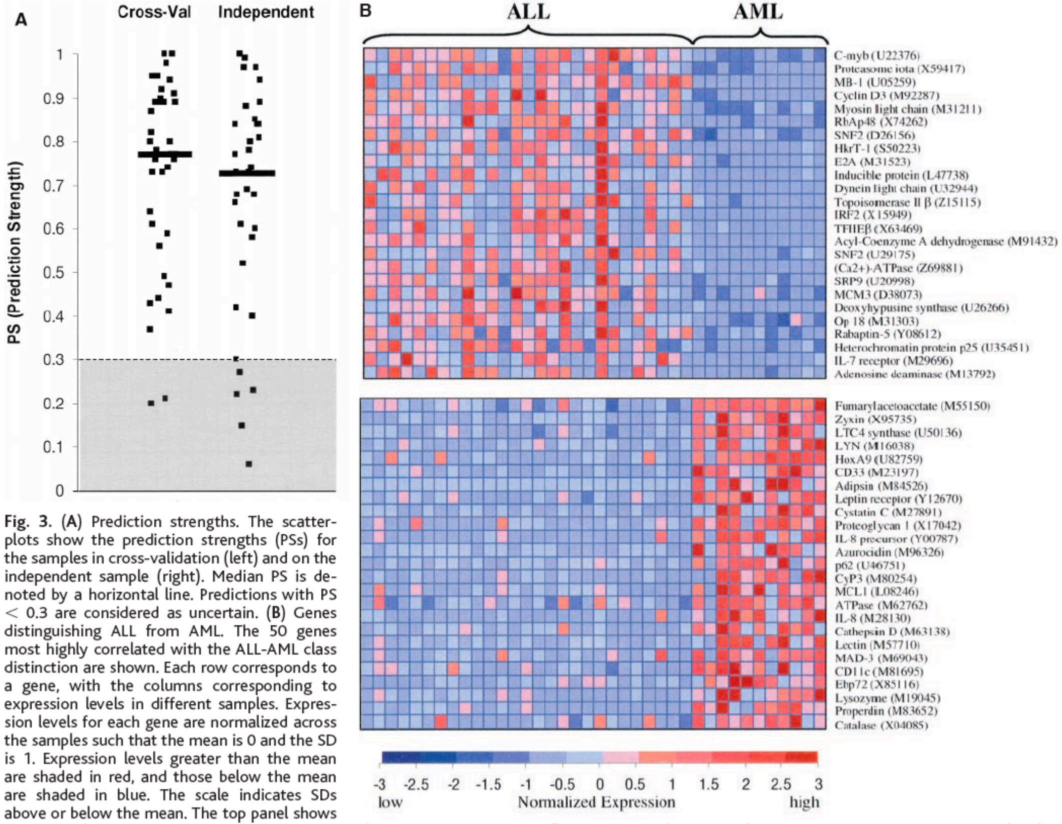


Fig. 3. (**A**) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. (**B**) Genes distinguishing ALL 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.

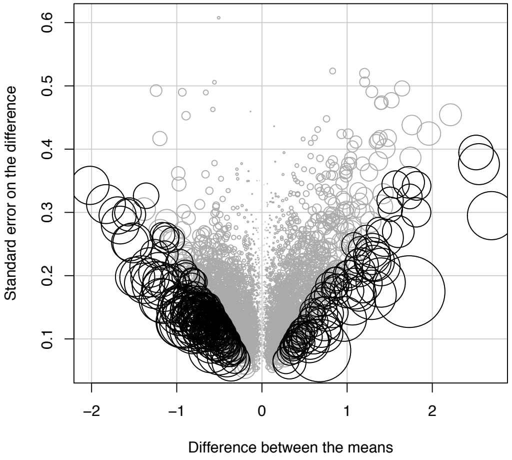
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286,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

Motivation

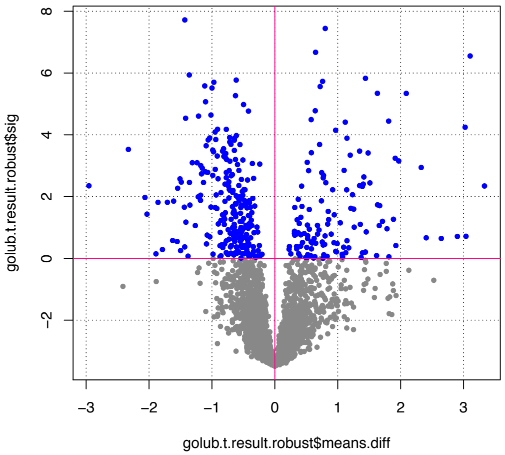
- The article by Golub et al. (1999) was motivated by the need to develop efficient diagnostics to predict the cancer type from blood samples of patients.
- They proposed a “molecular signature” of cancer type, allowing to discriminate ALM from ALL.
- This first “historical” study relied on somewhat 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.

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



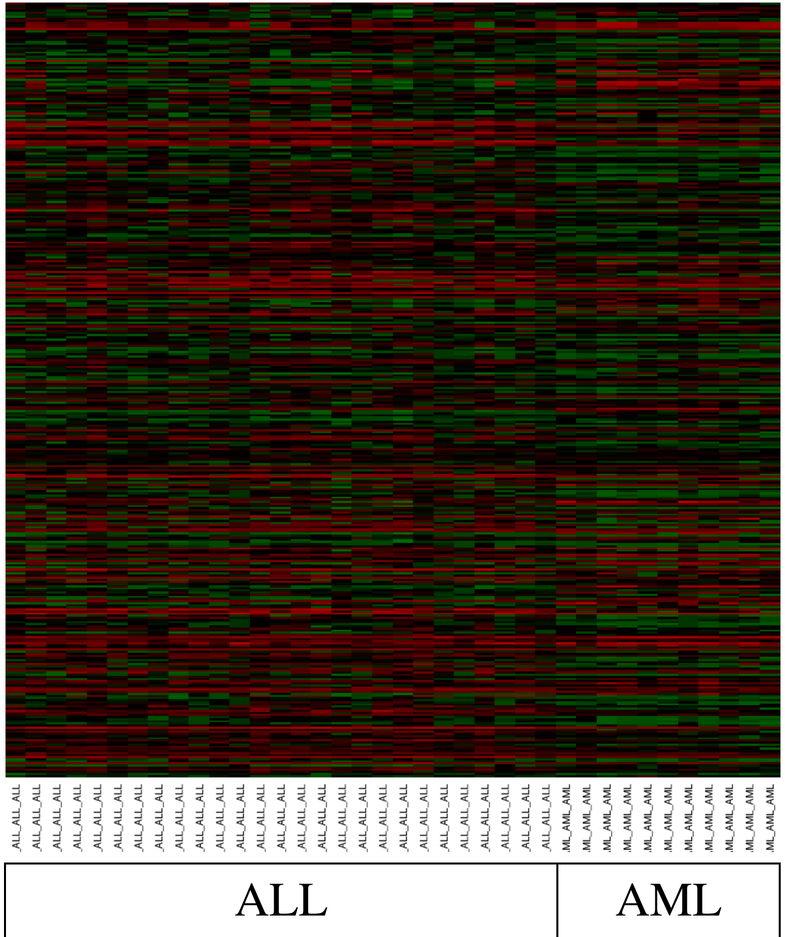
volcano plot – standardization with median and IQR



- 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

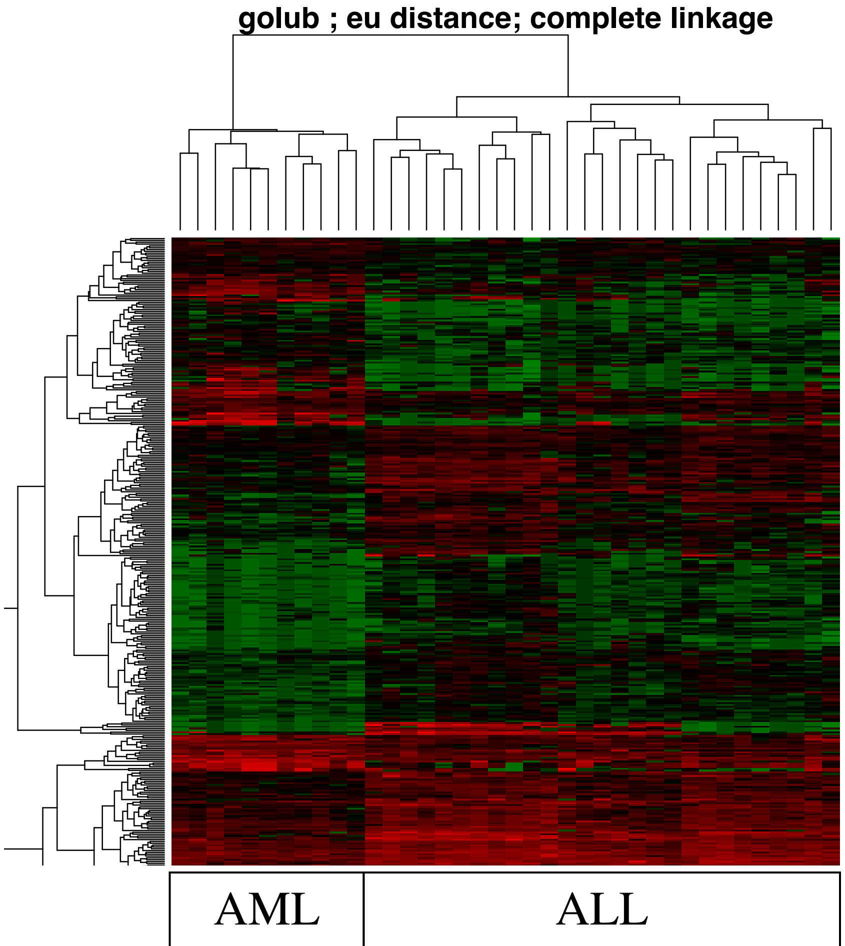
Golub, 1999, T–test selection (38 samples, 367 probes)



- 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)

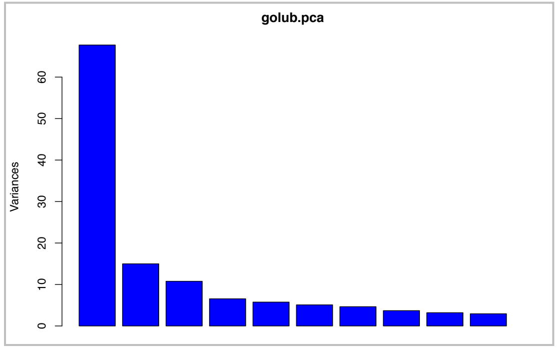
Golub – hierarchical vlustering of DEG genes / profiles

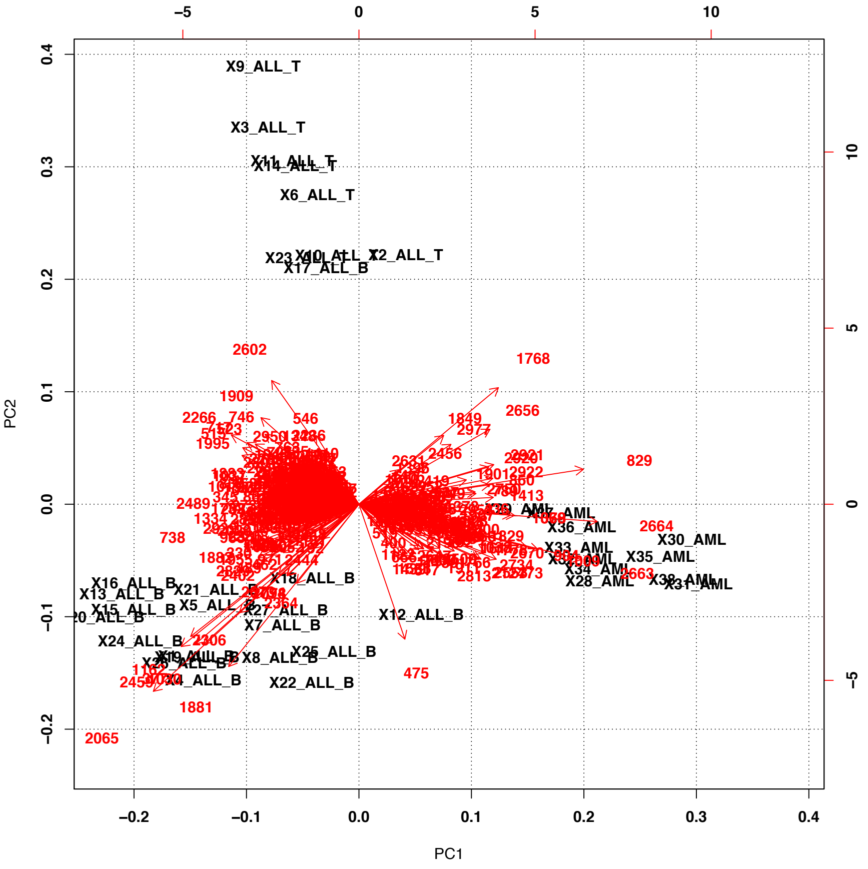
- Hierarchical clustering perfectly separates the two cancer types (AML versus ALL).
- This perfect separation is observed for various metrics (Euclidian, correlation, dot product) and agglomeration rules (complete, average, Ward).
- Sample clustering further reveals subgroups of ALL.
- Gene clustering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



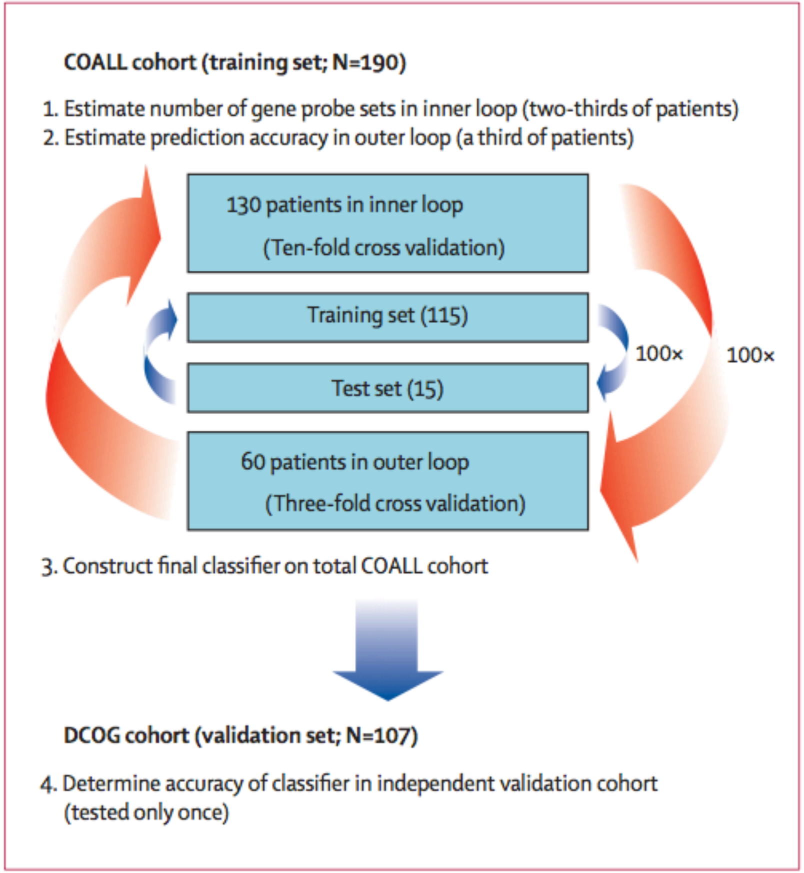


Study case 2: ALL subtypes
(data from Den Boer et al., 2009)

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.

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



COALL cohort (training set; N=190)

1. Estimate number of gene probe sets in inner loop (two-thirds of patients)
2. Estimate prediction accuracy in outer loop (a third of patients)

130 patients in inner loop (Ten-fold cross validation)

Training set (115)

Test set (15)

60 patients in outer loop (Three-fold cross validation)

3. Construct final classifier on total COALL cohort

DCOG cohort (validation set; N=107)

4. Determine accuracy of classifier in independent validation cohort (tested only once)

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

- 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 selects 100 gens 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 the testing set (right).

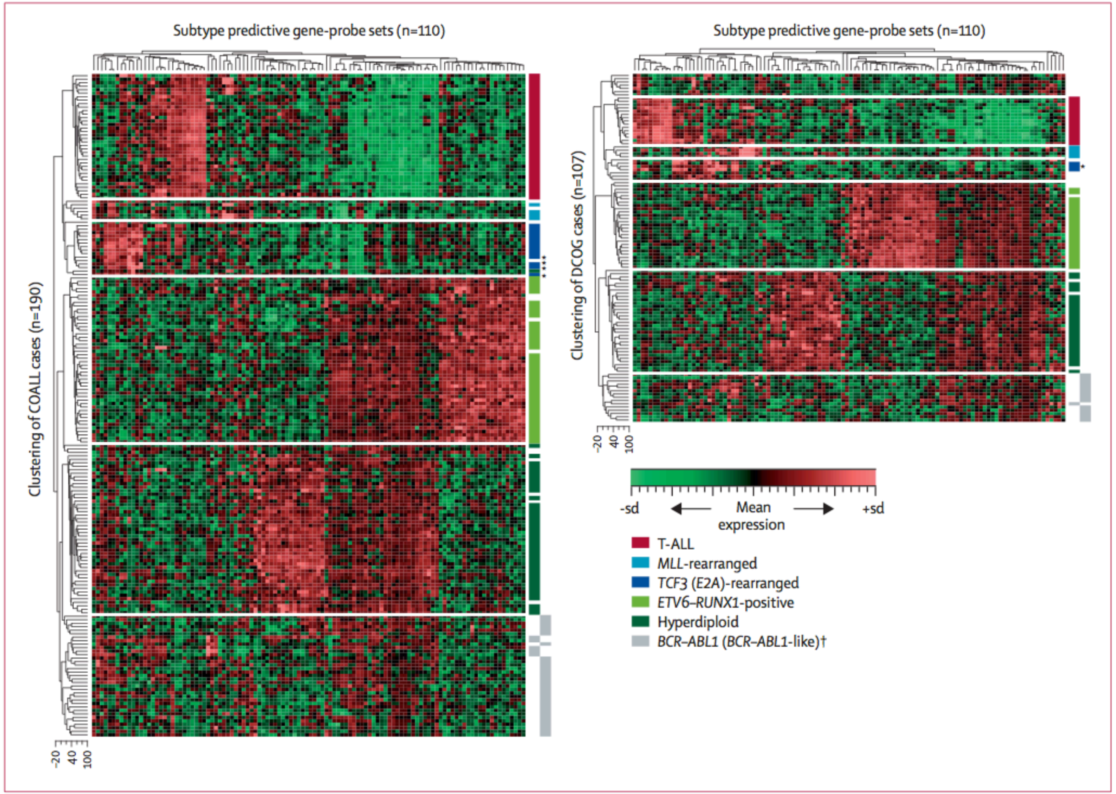


***Figure* 2: Clustering of ALL subtypes by gene-expression profiles**

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

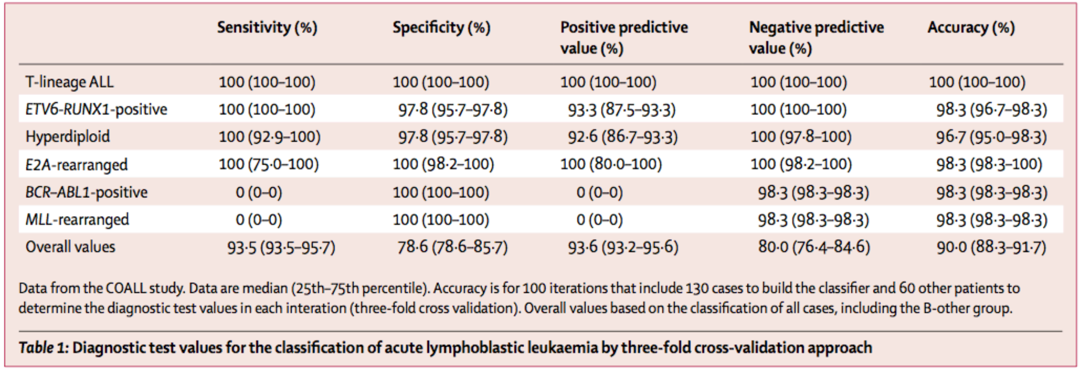
- The signature has an excellent diagnostic value: for the well-represented cancer types, the sensitivity and specificity are >90%.
- **Note:** accuracy is misleading some subtypes have 98% accuracy with 0% sensitivity.

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)

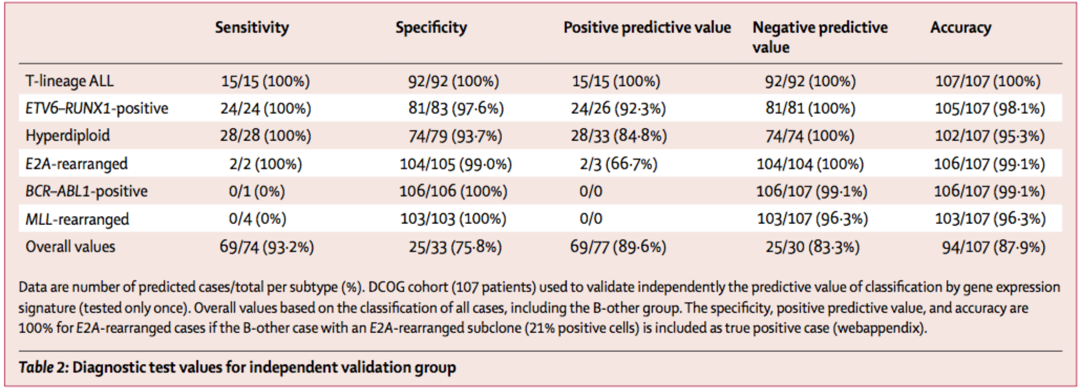
PPV = TP / (VP + FP)



	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).

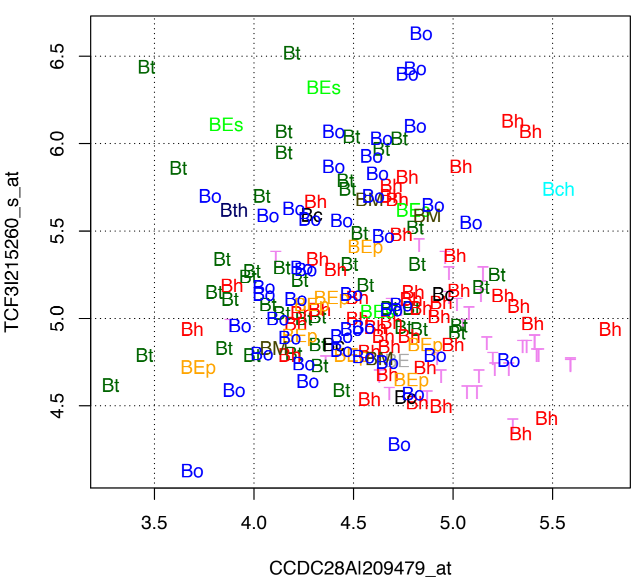
Table 2: **Diagnostic test values for independent validation group**

- 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.

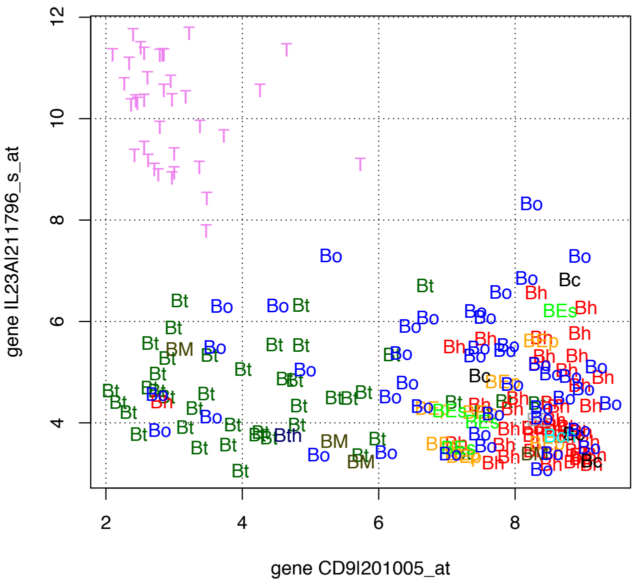
Den Boer 2009 - Exploring some profiles

- 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 ?

Den Boer (2009), randomly selected genes



2 genes with the highest variance



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

Supervised classification: 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**

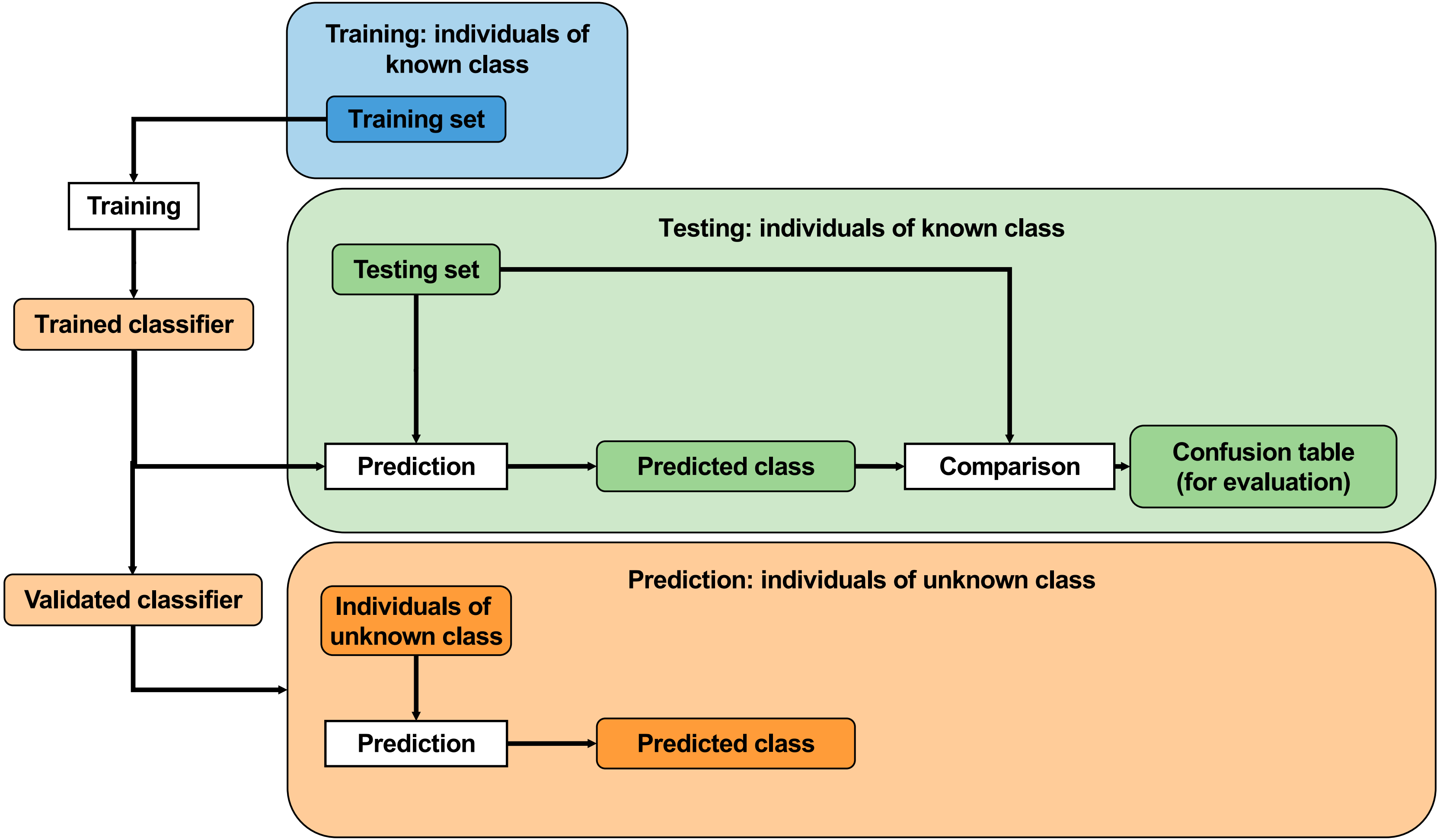
Supervised classification – training and prediction

- **Training phase (training + evaluation)**
  - The sample is used to build a discriminant function
  - The quality of the discriminant function is evaluated
- **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**

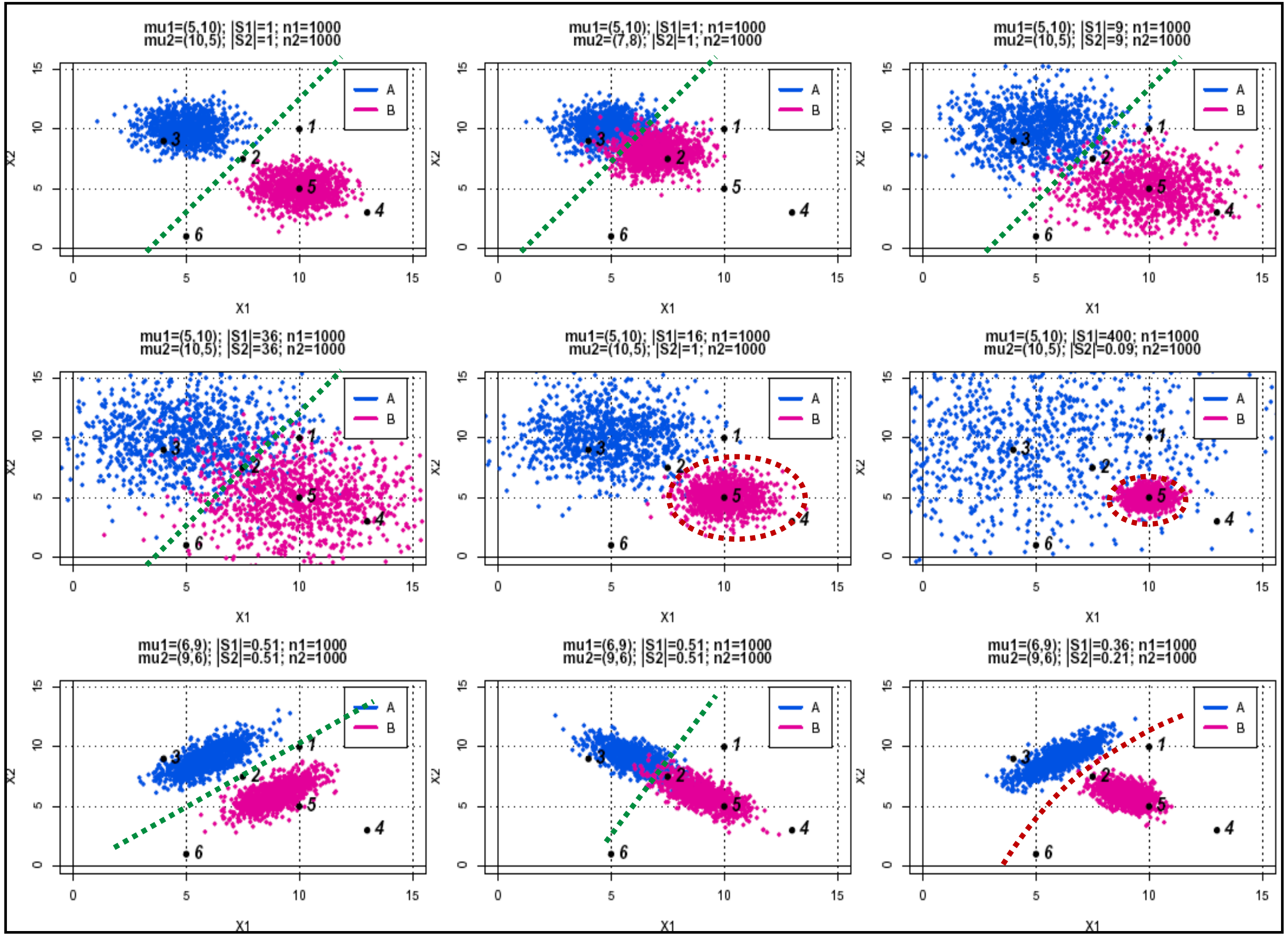
	Predictor variables				Criterion variable
	variable 1	**variable 2**	**…**	**variable p**	**class**
object 1	x_{11}	x_{21}	…	x_{p1}	**?**
object 2	x_{12}	x_{22}	…	x_{p2}	**?**
object 3	x_{13}	x_{23}	…	x_{p3}	**?**
…	…	…	…	…	**…**
object n_{pred}	x_{1n}	x_{2n}	…	x_{pn}	**?**

Supervised classification: steps of the general procedure


Training: individuals of known class
Training set
Training
Trained classifier
Testing: individuals of known class
Testing set
Prediction
Predicted class
Comparison
Confusion table (for evaluation)
Validated classifier
Prediction: individuals of unknown class
Individuals of unknown class
Prediction
Predicted class

Discriminant analysis

Linear or quadratic discriminant analysis (LDA vs QDA)



- Equal covariance matrix between groups?
  - Linear Discriminant Analysis (LDA) is appropriate
  - Green lines on the graph
  - The discrimination rule amounts to draw a straight line between the gravity centers of the training groups
- Different covariant matrices?
  - Quadratic Discriminant Analysis is recommended (red boundaries on the graphs)

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**: based on the density function. Assign unit u to group g if

$$\boxed{f(X \mid g) > f(X \mid g') \qquad for\, g' \neq g}$$

  - **Inverse probability rule**: based on the probability. Assign unit u to group g if

$$\boxed{P(X \mid g) > P(X \mid g') \qquad for\, g' \neq g}$$

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

$$\boxed{P(g \mid X) > P(g' \mid X) \qquad for\, g' \neq g}$$

Where

X is the unit vector

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

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

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

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

Choice of the prior probabilities

- The classes may have different proportions between the sample and the population
- For example, we could decide, based on our knowledge of a problem, that it is likely to have 1% of the individuals that belong to the first group, whereas the training set contains 11% of them.

		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**

Evaluating the performances of a classifier

Concepts

- Evaluation settings
  - Internal evaluation (“training error”) versus external test set (“testing error”)
  - Independent testing set
  - Split out of the training set into training and testing subsets
    - Iterative subsampling
    - K-fold cross-validation
    - Leave-one-out (LOO)
- Evaluation statistics
  - Confusion table
  - Misclassification error rate (MER)
  - Additional metrics for two-groups classification
    - FP, FN, TP, TN
    - Many metrics derived from there: Sn, PPV, FPR, FDR, …

Training – testing settings

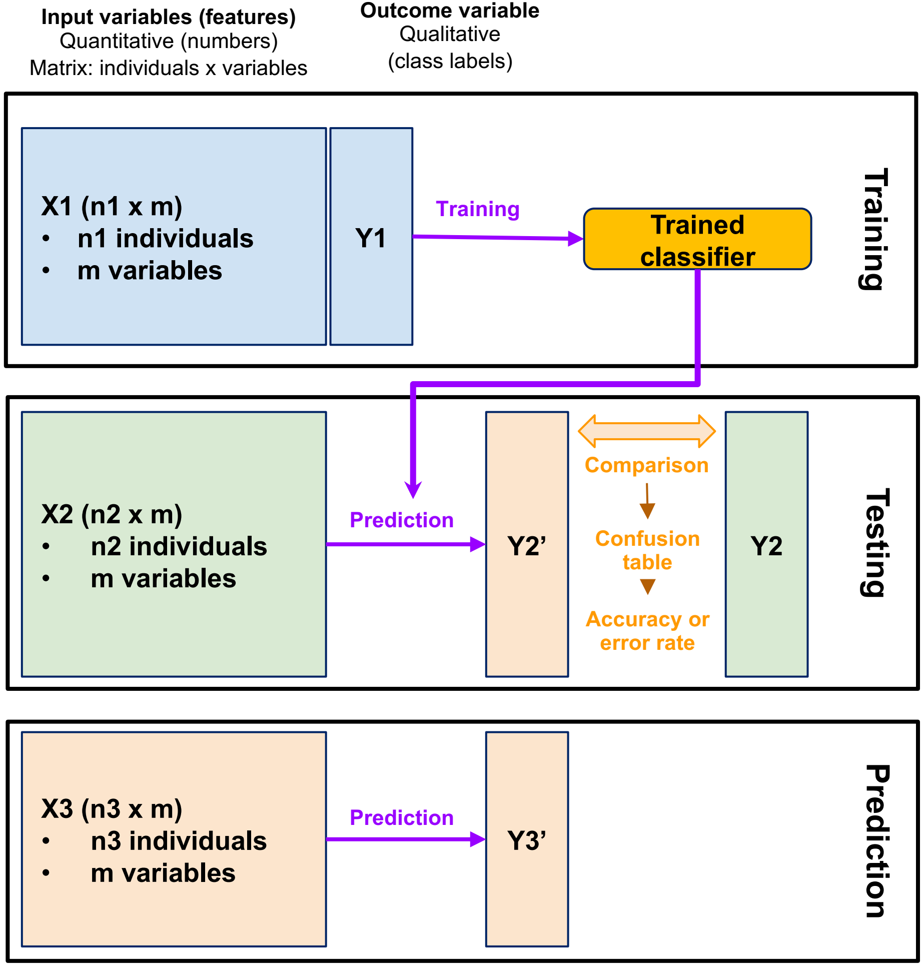
Evaluation of the classifier – predicted and known classes

- The evaluation of a classifier relies on a data set for which we know the class of each individual : the ***testing set***.
- The trained classifier is used to predict the class of each individual of the testing set
- The predicted and known classes are then compared

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

Training, testing and prediction

- Ideally : dispose of an independent testing set
- Alternatives
  - Internal validation (NOT RECOMMENDED)
  - Splitting the training set
    - Iterative subsampling
    - K-fold cross-validation (CV)
    - Leave-one-out (LOO)



Using an independent testing set

- 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 training, and one for testing.
- However, we do not always dispose of two independent sets.
- An alternative setting is to split randomly the samples of known class into two subsets (**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

Training error rate

- One way to evaluate the performances of a classifier is to run it on the training set itself.
- This approach is called **internal analysis**.
- The known and predicted class are then compared for each individual of the training set itself.
- The result is denoted as the ***training error rate*** (the error rate measured on the training set itself).
- **Warning** :
  - This approach is obviously biased, since the training set was used to train the classifier, it is thus optimised for this very specific dataset.
  - The training error rate is thus too optimistic: the performances may be much weaker on an independent set.
  - This approach is not recommended for general purposes.
  - The main interest of this approach is to compare it with an independent testing set (testing error rate) in order to measure the overfitting of the classifier to the particular training set.

K-fold cross validation

- Split the training set into k parts (e.g. 10-fold cross-validation)
- Iterate for each subset i
    1. Train a classifier with all subsets except subset i
    2. Run the classifier to predict the class of each element of the testing subset (subset i)
- Compare the predicted and known classes for each individual
- Each individual is thus used
    - 1 time for testing
    - k-1 times for training

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.
- Note : LOO is equivalent to perform a N-fold cross validation (where N is the training set size)

Evaluation measures for supervised classification

Evaluation of a classifier – 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.
- When there are more than 2 groups or when the groups are not associated to + and -, the performances are estimated by computing the misclassification error rate (MER)

3-groups confusion table

		Known group			
		A	B	C	SUM
Predicted group	A	8	0	0	8
	B	0	1	1	2
	C	5	18	81	104
	SUM	13	19	82	114

Hits	Diagonal	
Errors	Non-diagonal	
Hit rate	Hits / total	Also named 'accuracy"
MER	Errors / total	Misclassification error rate

Example

		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

Hits	8 + 1 + 81	90
Errors	114 - 90	24
Hit rate	90 / 114	78.95%
MER	24 / 114	21.05%

Evaluation of a classifier – confusion table for 2a-groups classification

- 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 performances of the classifier.
- For 2-groups classification, specific metrics can be applied if one group is considered negative and the other one positive

2-groups classification

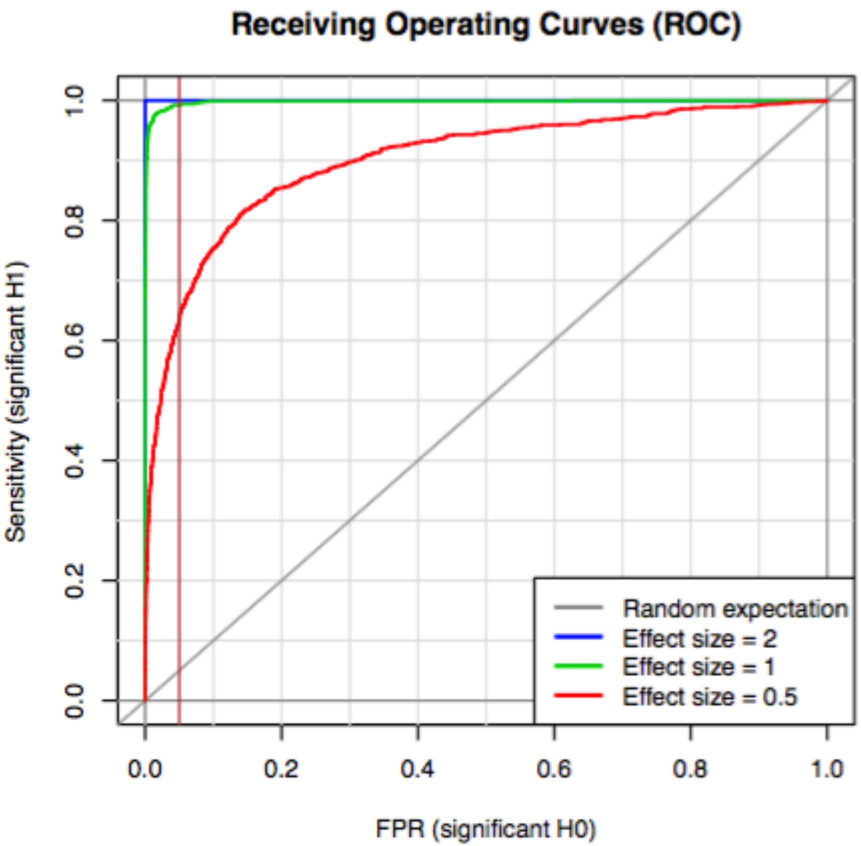
	Known class		
Predicted class	Case	Control	SUM
Case	TP	FP	P
Control	FN	TN	N
SUM	TP+FN	FP+TN	T
Errors	FN+FP		
Correct	TP+TN		
FPR	FP/(FP+TN)		
Sn	TP/(TP+FN)		
FDR	FP/P		

Example

	Known class		
	Case	Control	SUM
Case	99	20	119
Control	1	180	181
SUM	100	200	300
Errors	21	7.00%	
Correct	279	93.00%	
FPR	20/(200)	10.00%	
Sn	99/(100)	99.00%	
FDR	20/119	16.81%	

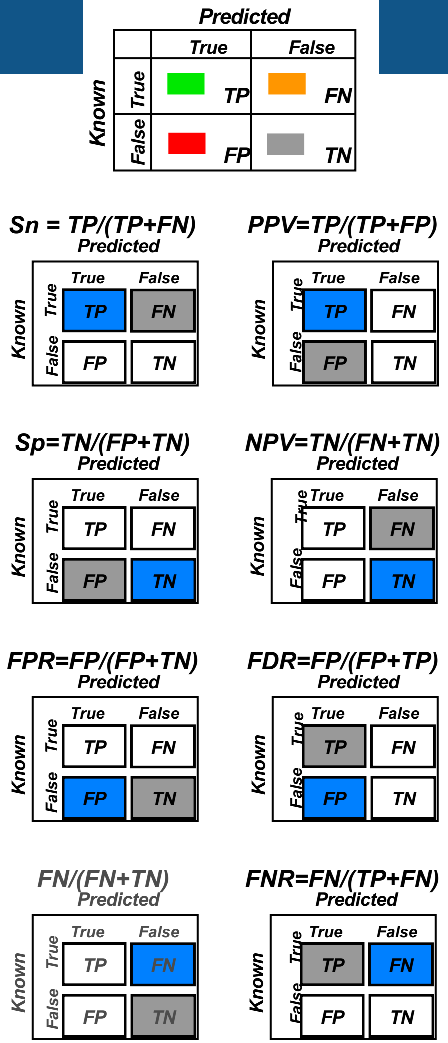
Receiving Operator Characteristics (ROC)

- The Receiving Operator Characteristics (ROC) represents the performances of a classifier as a function of a continuous score (e.g. discriminant function, posterior probability)
- The result is a curve with
  - Abscissa: FPR
  - Ordinate: Sensitivity
- A random classifier will be aligned onto the diagonal
- A perfect classifier achieves FPR=0 and Sn=1 (upper left corner)
- The closer the curves comes to this perfect performance, the better the classifier.
- The Area Under the Curve (AUC) is often used to compare the performances
  - Between classifiers
  - Between different parameter settings for the same classifier



Evaluation statistics for 2-groups classifiers

Abbrev	Name	Formula
TP	True positive	TP
FP	False positive	FP
FN	False negative	FN
TN	True negative	TN
KP	Known Positive	TP+FN
KN	Known Negative	TN+FP
PP	Predicted Positive	TP+FP
PN	Predicted Negative	FN+TN
N	Total	TP + FP + FN + TN
Prev	Prevalence	(TP + FN)/N
ODP	Overall Diagnostic Power	(FP + TN)/N
CCR	Correct Classification Rate	(TP + TN)/N
Sn	**True Positive Rate (Sensitivity)**	**TP/(TP + FN)**
TNR	True Negative Rate (Specificity)	TN/(FP + TN)
FPR	False Positive Rate	FP/(FP + TN)
FNR	False Negative Rate	FN/(TP + FN) = 1-Sn
PPV	**Positive Predictive Value**	**TP/(TP + FP)**
FDR	False Discovery Rate	FP/(FP+TP)
NPV	Negative Predictive Value	TN/(FN + TN)
Mis	Misclassification Rate	(FP + FN)/N
Odds	Odds-ratio	(TP + TN)/(FN + FP)
Kappa	Kappa	((TP + TN) - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))/(N - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))
NMI	NMI n(s)	(1 - -TP*log(TP)-FP*log(FP)-FN*log(FN)-TN*log(TN)+(TP+FP)*log(TP+FP)+(FN+TN)*log(FN+TN))/(N*log(N) - ((TP+FN)*log(TP+FN) + (FP+TN)*log(FP+TN)))
ACP	Average Conditional Probability	0.25*(Sn+ PPV + Sp + NPV)
MCC	Matthews correlation coefficient	(TP*TN - FP*FN) / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))
Acc.a	Arithmetic accuracy	(Sn + PPV)/2
Acc.a2	Accuracy (alternative)	(Sn + Sp)/2
Acc.g	Geometric accuracy	sqrt(Sn*PPV)
Hit.noTN	A sort of hit rate without TN (to avoid the effect of their large number)	TP/(TP+FP+FN)



Various statistics can be derived from the 4 elements of a contingency table (TP, FP, TN, FN).

The arithmetic accuracy may be misleading

Abbrev	Name	Formula
TP	True positive	TP
FP	False positive	FP
FN	False negative	FN
TN	True negative	TN
KP	Known Positive	TP+FN
KN	Known Negative	TN+FP
PP	Predicted Positive	TP+FP
PN	Predicted Negative	FN+TN
N	Total	TP + FP + FN + TN
Prev	Prevalence	(TP + FN)/N
ODP	Overall Diagnostic Power	(FP + TN)/N
CCR	Correct Classification Rate	(TP + TN)/N
Sn	**True Positive Rate (Sensitivity)**	**TP/(TP + FN)**
TNR	True Negative Rate (Specificity)	TN/(FP + TN)
FPR	False Positive Rate	FP/(FP + TN)
FNR	False Negative Rate	FN/(TP + FN) = 1-Sn
PPV	**Positive Predictive Value**	**TP/(TP + FP)**
FDR	False Discovery Rate	FP/(FP+TP)
NPV	Negative Predictive Value	TN/(FN + TN)
Mis	Misclassification Rate	(FP + FN)/N
Odds	Odds-ratio	(TP + TN)/(FN + FP)
Kappa	Kappa	((TP + TN) - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))/(N - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))
NMI	NMI n(s)	(1 - -TP*log(TP)-FP*log(FP)-FN*log(FN)-TN*log(TN)+(TP+FP)*log(TP+FP)+(FN+TN)*log(FN+TN))/(N*log(N) - ((TP+FN)*log(TP+FN) + (FP+TN)*log(FP+TN)))
ACP	Average Conditional Probability	0.25*(Sn+ PPV + Sp + NPV)
MCC	Matthews correlation coefficient	(TP*TN - FP*FN) / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))
Acc.a	Arithmetic accuracy	(Sn + PPV)/2
Acc.a2	Accuracy (alternative)	(Sn + Sp)/2
Acc.g	**Geometric accuracy**	**sqrt(Sn*PPV)**
Hit.noTN	A sort of hit rate without TN (to avoid the effect of their large number)	TP/(TP+FP+FN)

		Predicted	
		True	False
Known	True	🟩 TP	🟧 FN
	False	🟥 FP	▨ TN

Sn = TP/(TP+FN)

		Predicted	
		True	False
Known	True	🟦 TP	▨ FN
	False	FP	TN

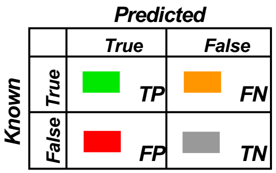
PPV=TP/(TP+FP)

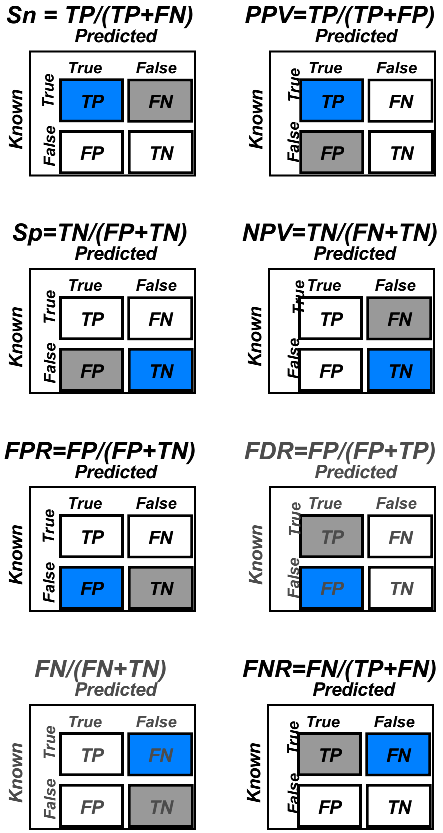
		Predicted	
		True	False
Known	True	🟦 TP	FN
	False	▨ FP	TN

- ***Acc_a = (Sn + PPV)/2***
- An easy way to fool the arithmetic accuracy: *predict all features as positive*
  - **-> *Sn* guaranteed to be 100%**
  - → *Acc_a* guaranteed to be >50%
  - Of course, you have a poor PPV, but the accuracy > 0.5 will be misleading
- The geometric accuracy circumvents this problem
  - ***Acc_g = sqrt(Sn*PPV)***
  - Requires for **both** *Sn* and *PPV* to be high.

TN-based statistics may be misleading

Abbrev	Name	Formula
TP	True positive	TP
FP	False positive	FP
FN	False negative	FN
TN	True negative	TN
KP	Known Positive	TP+FN
KN	Known Negative	TN+FP
PP	Predicted Positive	TP+FP
PN	Predicted Negative	FN+TN
N	Total	TP + FP + FN + TN
Prev	Prevalence	(TP + FN)/N
ODP	Overall Diagnostic Power	(FP + TN)/N
CCR	Correct Classification Rate	(TP + TN)/N
Sn	**True Positive Rate (Sensitivity)**	**TP/(TP + FN)**
TNR	True Negative Rate (Specificity)	TN/(FP + TN)
FPR	False Positive Rate	FP/(FP + TN)
FNR	False Negative Rate	FN/(TP + FN) = 1-Sn
PPV	**Positive Predictive Value**	**TP/(TP + FP)**
FDR	False Discovery Rate	FP/(FP+TP)
NPV	Negative Predictive Value	TN/(FN + TN)
Mis	Misclassification Rate	(FP + FN)/N
Odds	Odds-ratio	(TP + TN)/(FN + FP)
Kappa	Kappa	((TP + TN) - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))/(N - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))
NMI	NMI n(s)	(1 - -TP*log(TP)-FP*log(FP)-FN*log(FN)-TN*log(TN)+(TP+FP)*log(TP+FP)+(FN+TN)*log(FN+TN))/(N*log(N) - ((TP+FN)*log(TP+FN) + (FP+TN)*log(FP+TN)))
ACP	Average Conditional Probability	0.25*(Sn+ PPV + Sp + NPV)
MCC	Matthews correlation coefficient	(TP*TN - FP*FN) / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))
Acc.a	Arithmetic accuracy	(Sn + PPV)/2
Acc.a2	Accuracy (alternative)	(Sn + Sp)/2
Acc.g	Geometric accuracy	sqrt(Sn*PPV)
Hit.noTN	A sort of hit rate without TN (to avoid the effect of their large number)	TP/(TP+FP+FN)




Sn = TP/(TP+FN)

PPV=TP/(TP+FP)

Sp=TN/(FP+TN)

NPV=TN/(FN+TN)

FPR=FP/(FP+TN)

FDR=FP/(FP+TP)

FN/(FN+TN)

FNR=FN/(TP+FN)

- For some types of analyses, TN can represent >99.9%
- Example: predicting transcription factor binding sites in a whole genome.
- -> All the statistics including TN are misleading
- For example, a classifier will have a very high specificity (Sp) and a very low false positive rate (FPR) even though its predictions are mostly wrong.

Machine-learning – classification approaches

K Nearest Neighbours (KNN): principle, pros and cons

Principle

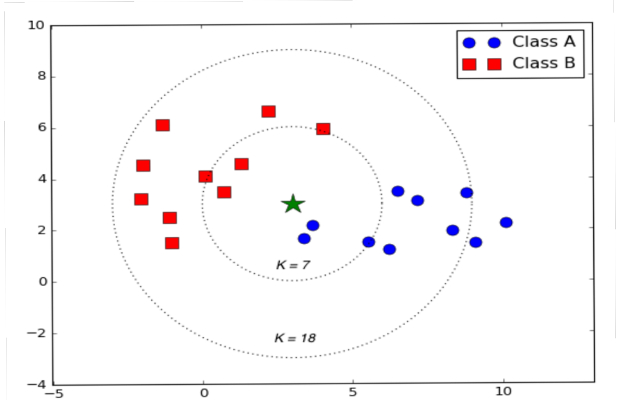
- memorize the positions of individuals training set
- predict the class of an individual based on class labels of its closest neighbours in the training set.

Pros

- Variety of distance criteria to be choose from pretty intuitive and simple.
- no assumptions about data distribution
- No Training Step
- Easy to implement for multi-class problem.

Cons

- *How to choose K ?* No general criterion to choose the optimal number of neighbors.
- Sensitive to the curse of dimensionality (over-fitting)
- Imbalanced data causes problems.
- Outlier sensitivity.
- Slow algorithm
- Missing Value treatment.



Acharya, A. (2017). Comparative Study of Machine Learning Algorithms for Heart Disease Prediction, (April).

Decision trees (DT): principles, pros and cons

Principle

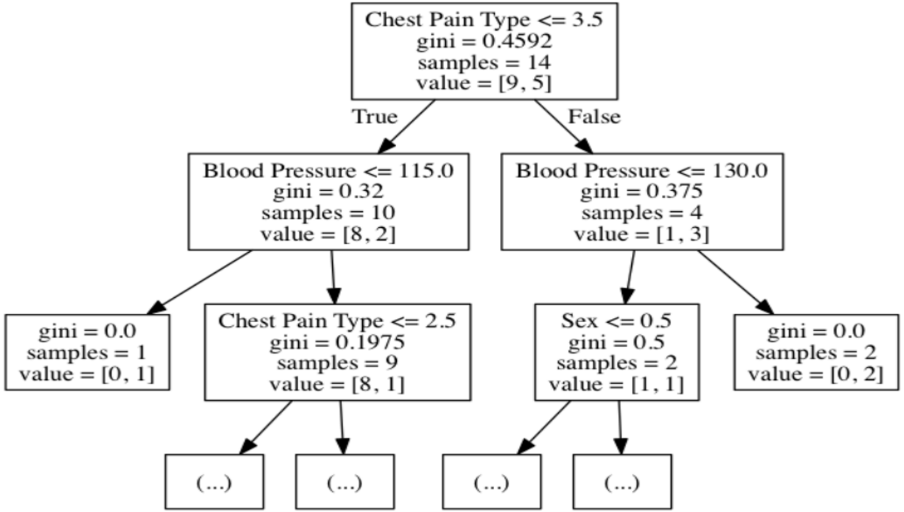
- A **Decision Tree (DT)** builds logical rules (e.g. *if variable i > a threshold, assign to class c*) that progressively lead to assign each sample to a single class.

Pros

- Expressive: one can understand a posteriori which criteria are important for class assignation.

Cons

- Very sensitive to over-fitting
- Lack of generalisation on unseen data.



Acharya, A. (2017). Comparative Study of Machine Learning Algorithms for Heart Disease Prediction, (April).

Random Forest (RF): principle, pros and cons

Principle

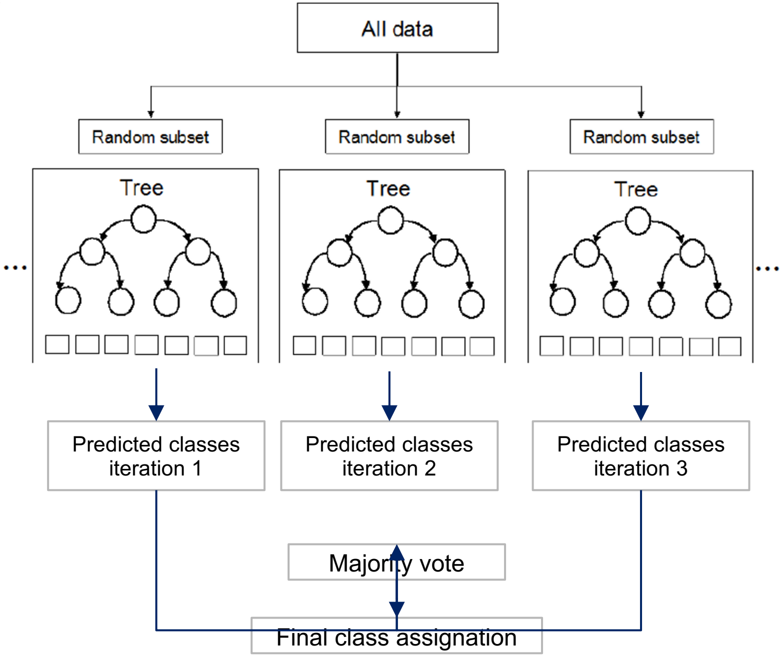
- A **random forest (RF)** is a classifier consisting of a collection of decision trees,
- **Bagging (bootstrapping):** each tree is constructed based on a subset of the training set.
- **Majority vote:** a sample is assigned to the class having the majority of assignations by individual trees.

Pros

- Reduces the over-fitting problem of the decision trees.

Cons

- Not easy to visually interpret



Adapted from: Anwar Isied and Hashem Tamimi. Using Random Forest (RF) as a transfer learning classifier for detecting Error-Related Potential (ErrP) within the context of P300-Speller.
DOI: 10.12751/nncn.bc2015.0143

Support Vector Machines (SVM): principle, pros and cons

Principle

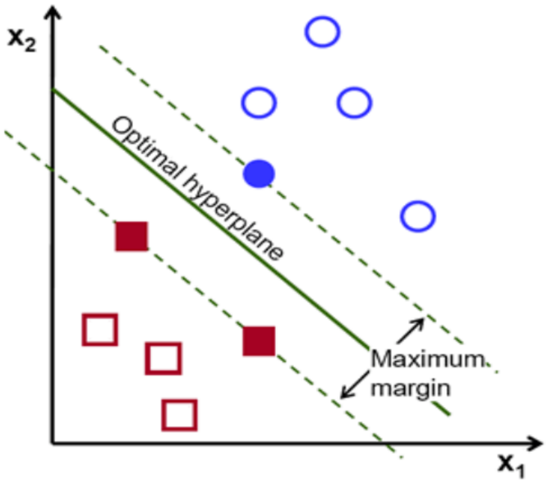
- Separate the various classes by a hyperplane in the feature hyperspace
- SVM is modelled with train data and outputs the hyperplane in the test data.
- The SVM model tries to find the space in the matrix of data where different classes of data can be widely separated and draws a hyperplane.

Pros

- Performs similarly to logistic regression when linear separation
- Performs well with non-linear boundary depending on the kernel used
- Handles well high-dimensional data.

Cons

- Sensitive to overfitting
- Training issues depending on kernel

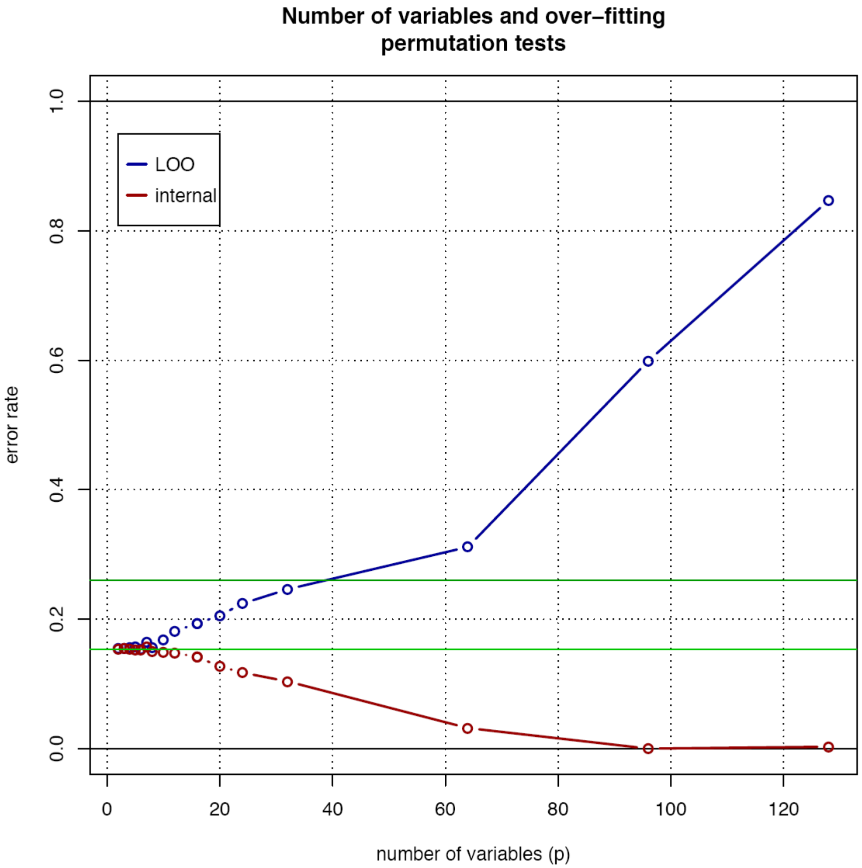


Acharya, A. (2017). Comparative Study of Machine Learning Algorithms for Heart Disease Prediction, (April).

Over-fitting and feature 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.



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.
    - Non-supervised criterion: e.g. sort features by decreasing variance (the relevance is questionable).
    - 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.
    - **All combinations of features**. Generally not tractable: 2^m possibilities
  - **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

Conclusions

Summary – supervised classification

- Setting:
  - a set of quantitative predictor variables (input variables)
  - a single nominal criterion variable (output variable)
- A sample is used to train the classifier training set), which is then evaluated on an independent testing set (testing) before being 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)
    - Iterative subsampling
    - K-fold Cross-validation
    - 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.

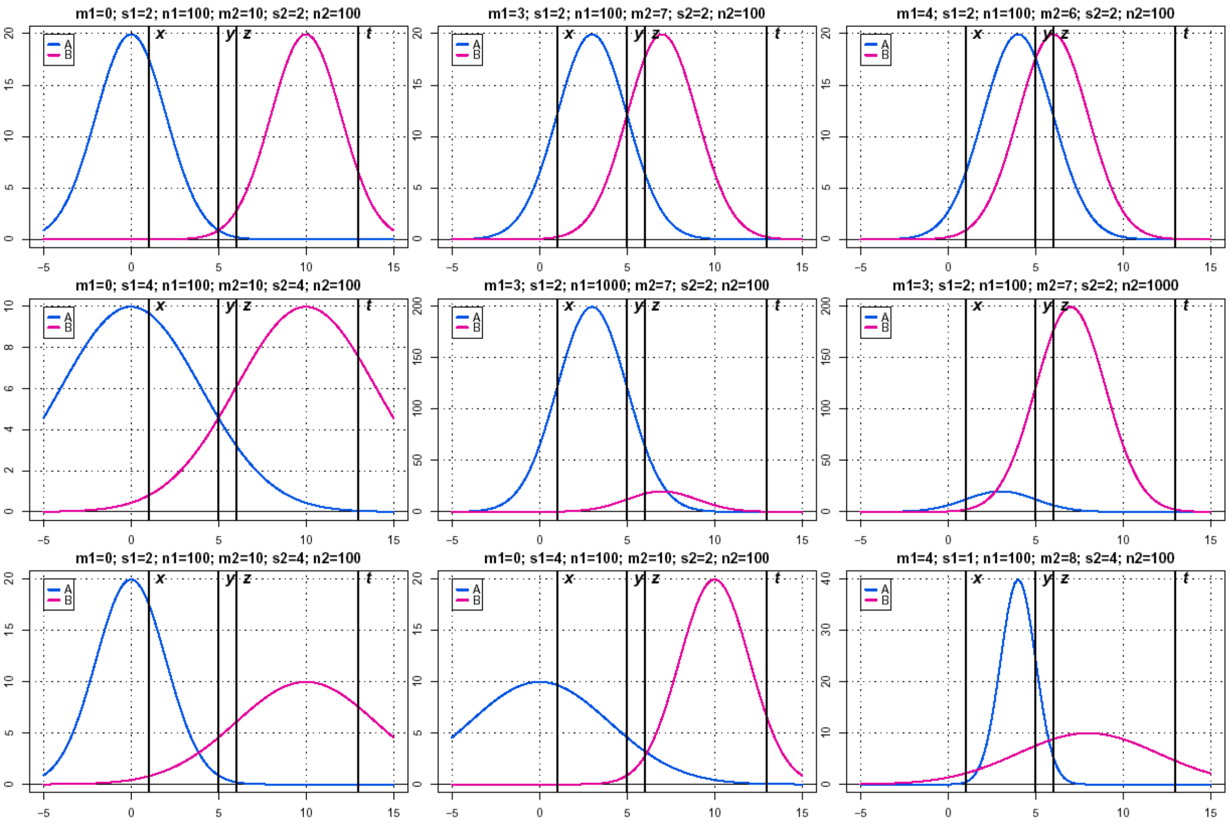
KNN classifiers

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.

Supplementary material

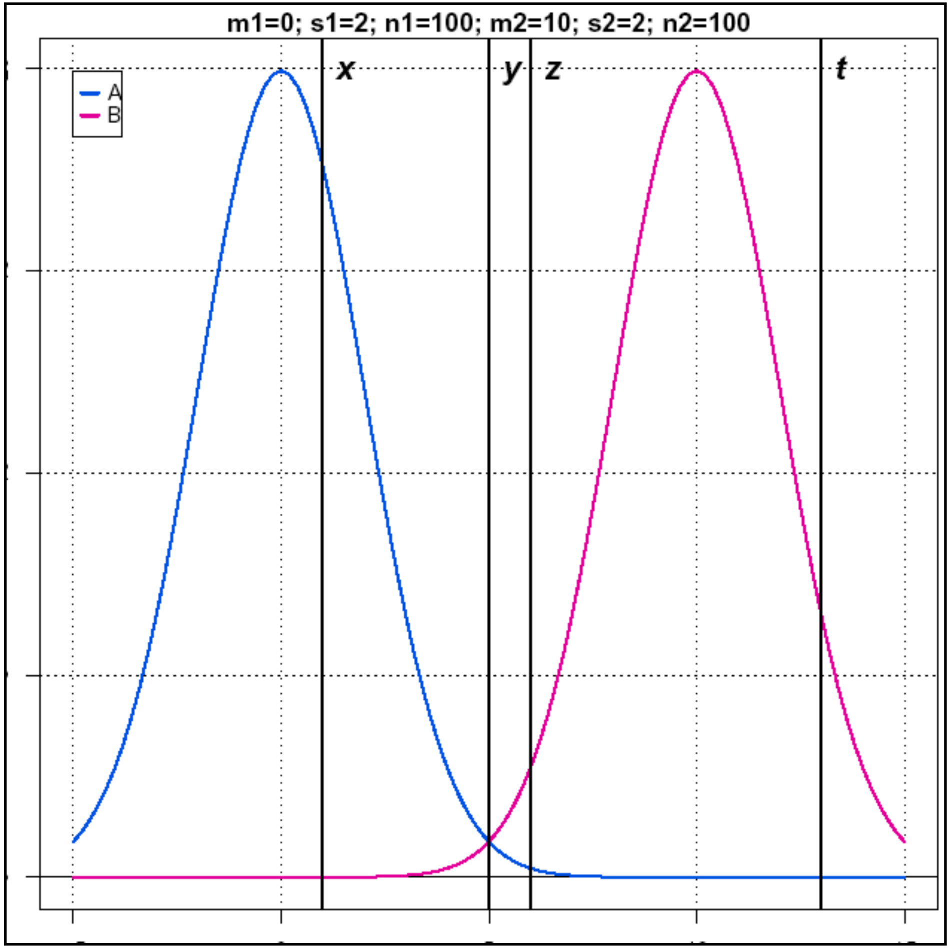
Conceptual illustration with a single predictor variable



Exercise

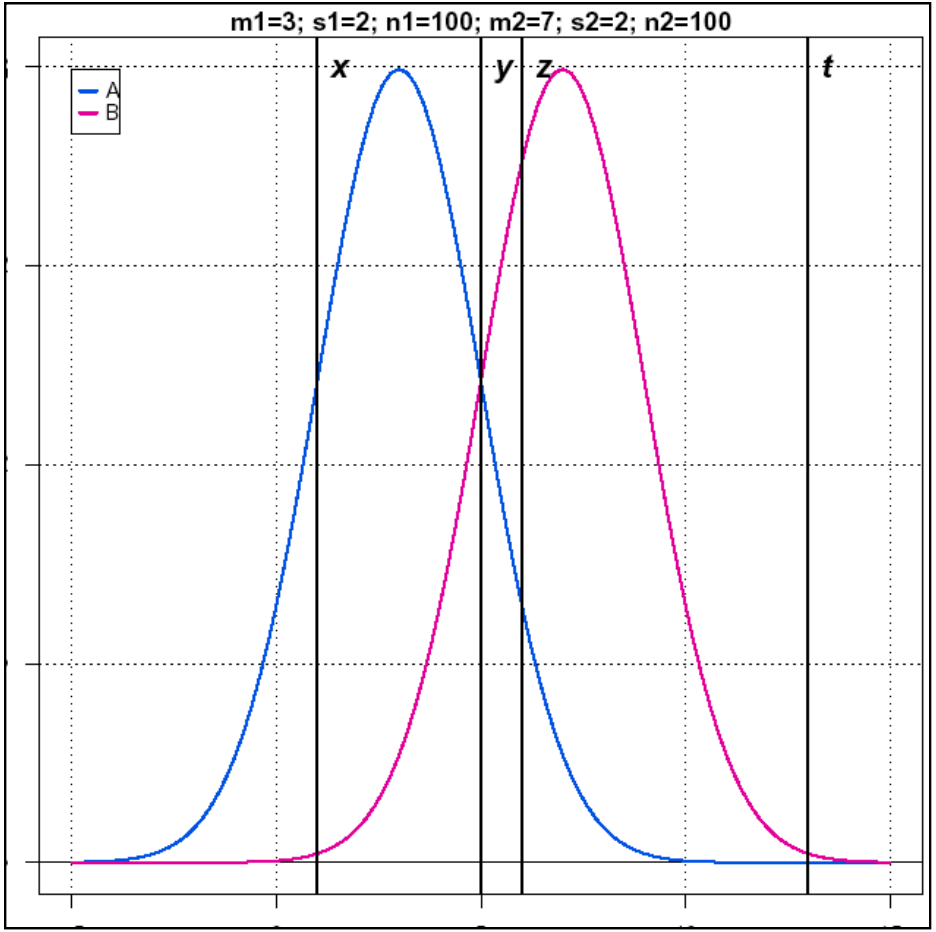
- Given two predefined classes (A and B), try intuitively to assign a class to each new object (X positions denoted by vertical black bars).
- How confident do you feel for each of your predictions ?
- What is the effect of the respective means ?
- What is the effect of the respective standard deviations ?
- What is the effect of the population sizes ?

Conceptual illustration with a single variable



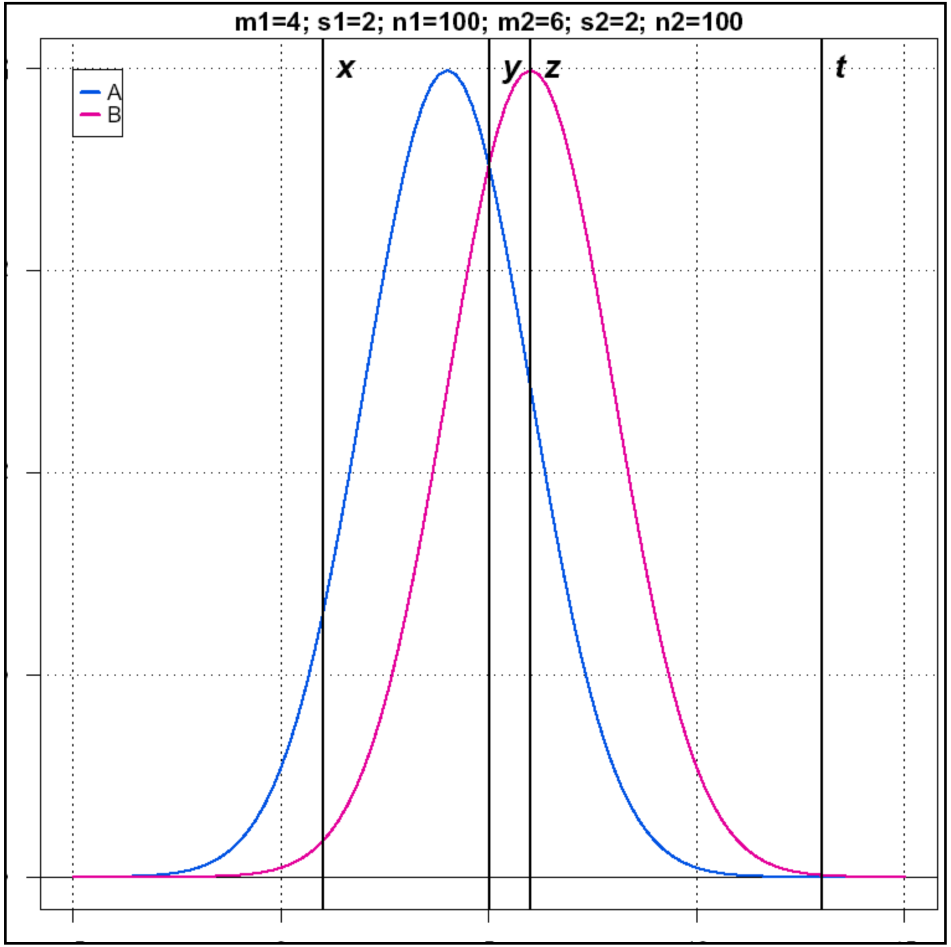
- In this conceptual example, the two populations have equal means and variances.
- To which group (A or B) would you assign the points at coordinate x, y, z, t, respectively ?

Conceptual illustration with a single variable



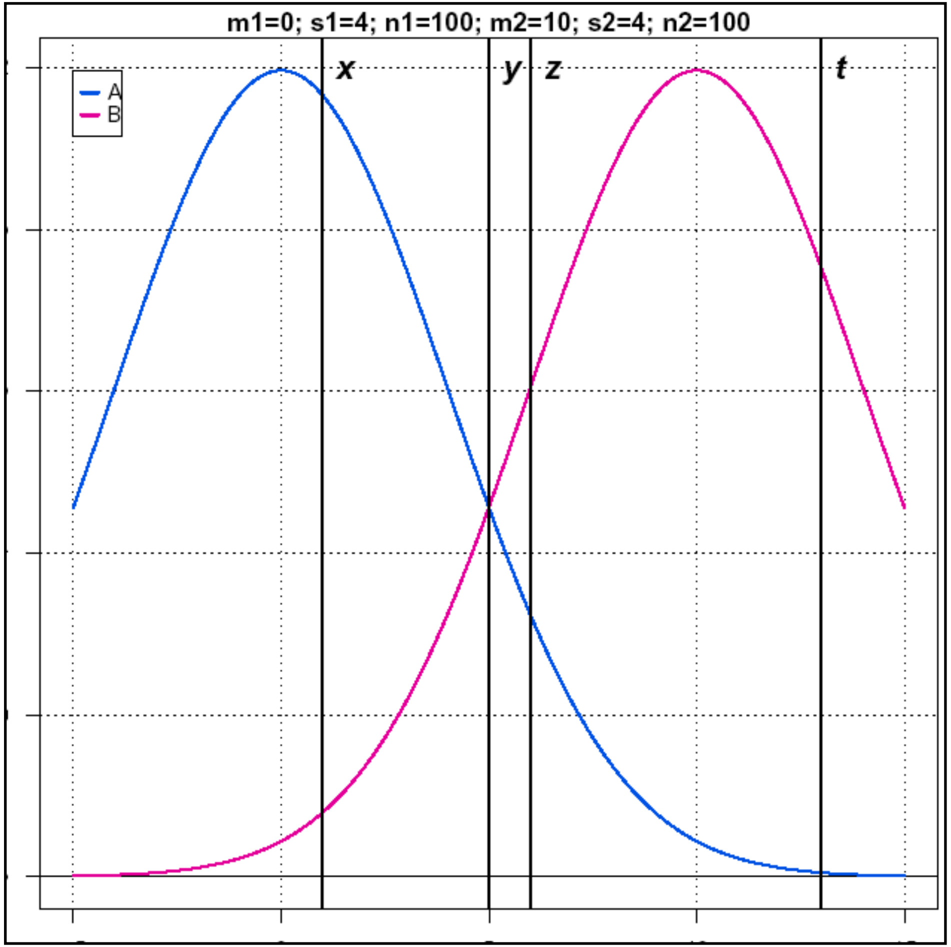
- Same exercise.
- This example shows that the assignation is affected by the position of the group centres.

Conceptual illustration with a single variable



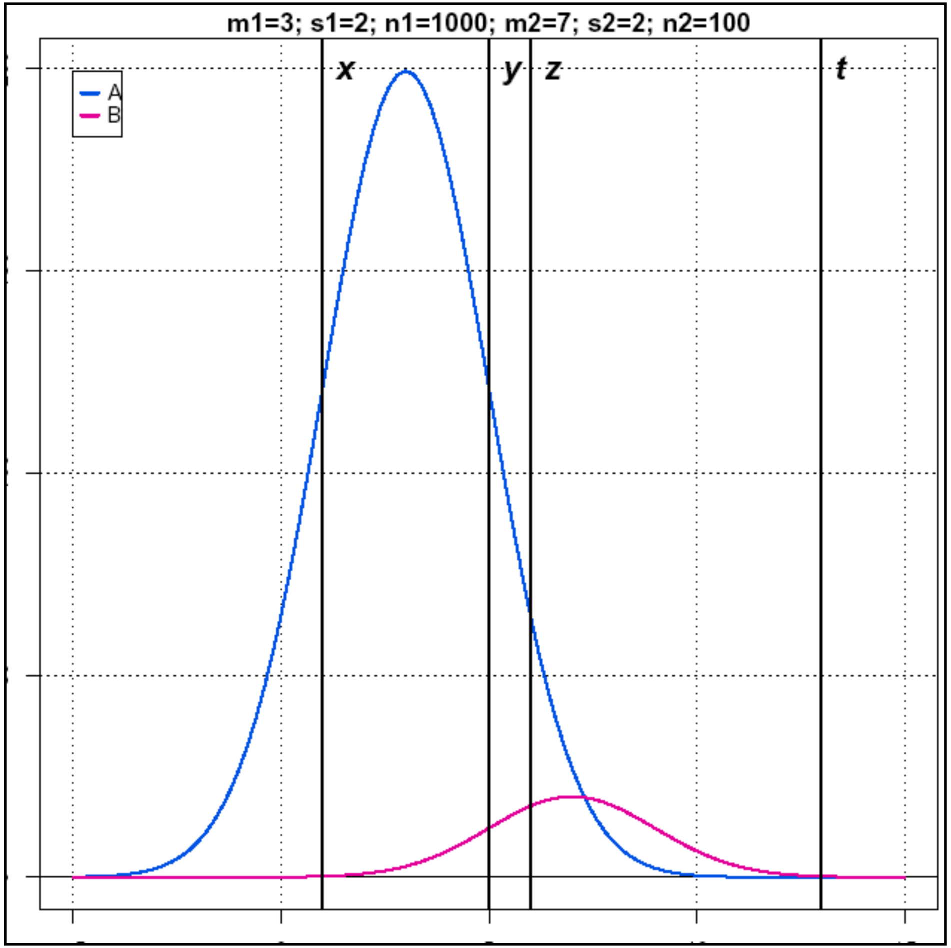
- Same exercise.
- When the centres become too close, some uncertainty is attached to some points (y, but also partly z).
- There is thus an effect of group distance.

Conceptual illustration with a single variable



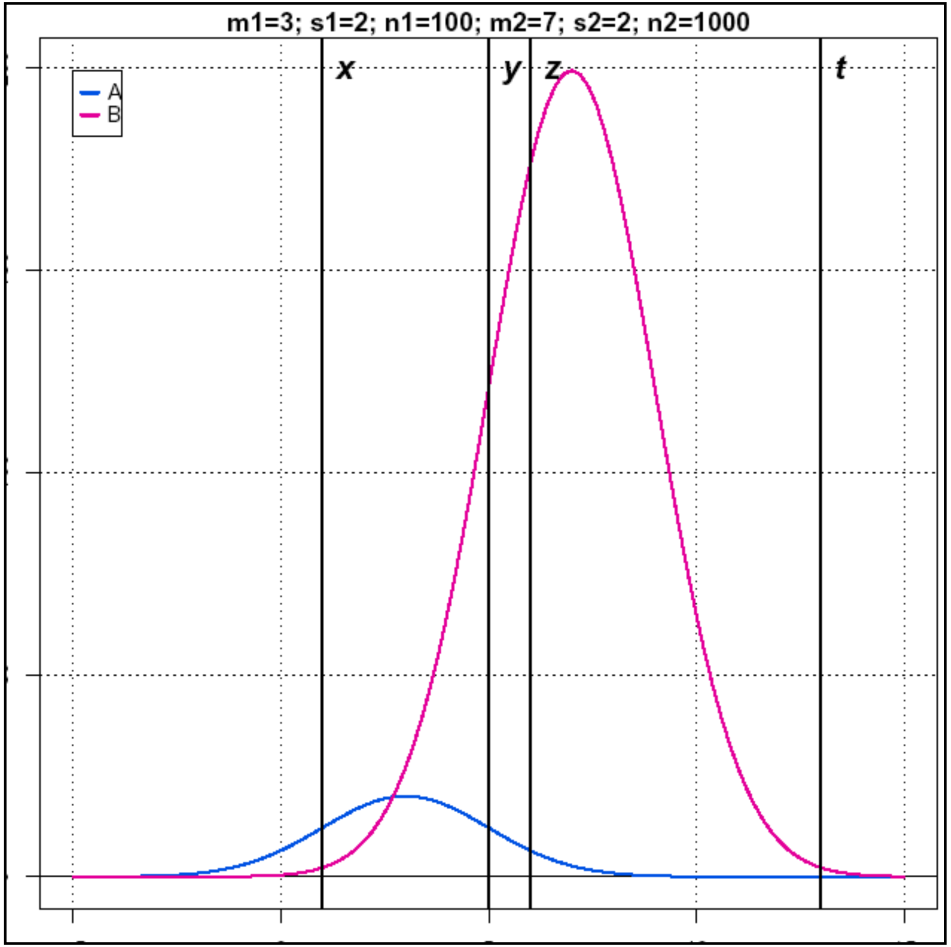
- Same exercise.
- The centres are in the same position as in the first example, but the variance is larger.
- This affects the level of separation of the groups, and raises some uncertainty about the group membership of z.
- The group variance thus affects the assignation.

Conceptual illustration with a single variable



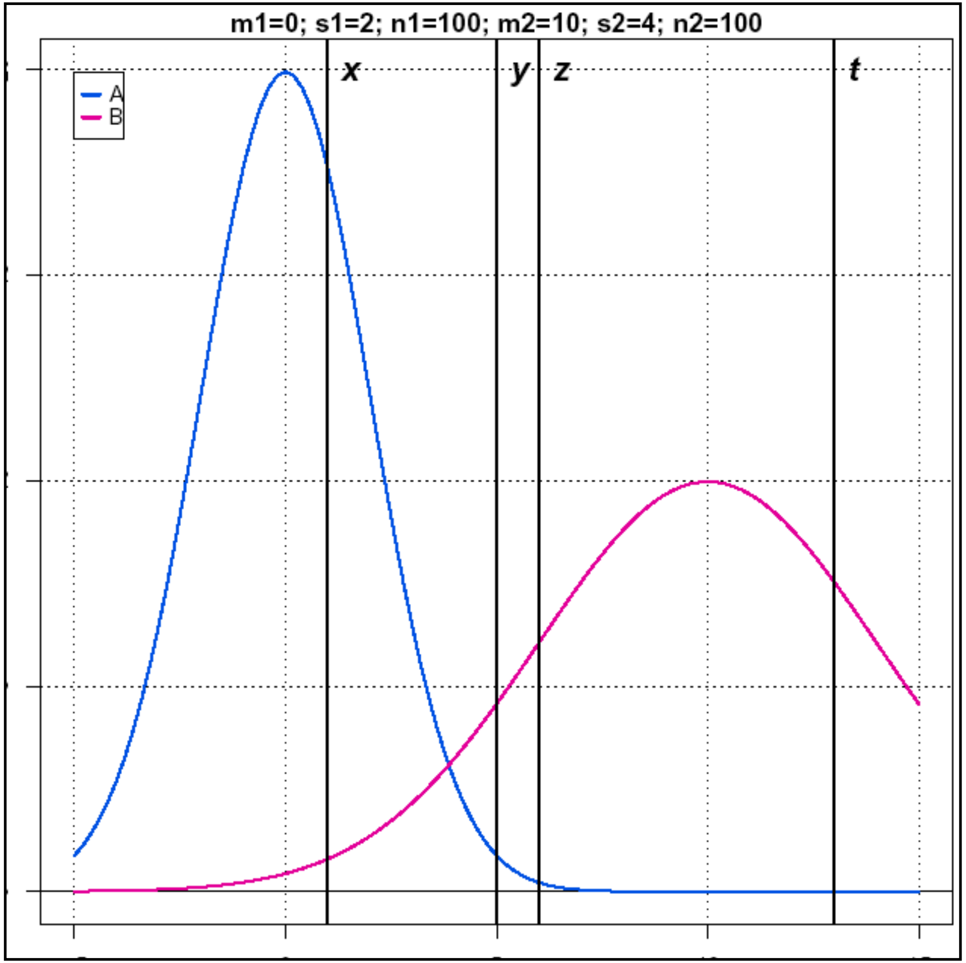
- Same exercise.
- This illustrates the effect of the sample size: if a sample has a much larger size than another one, it will increase the likelihood that some observations were issued from this group.

Conceptual illustration with a single variable



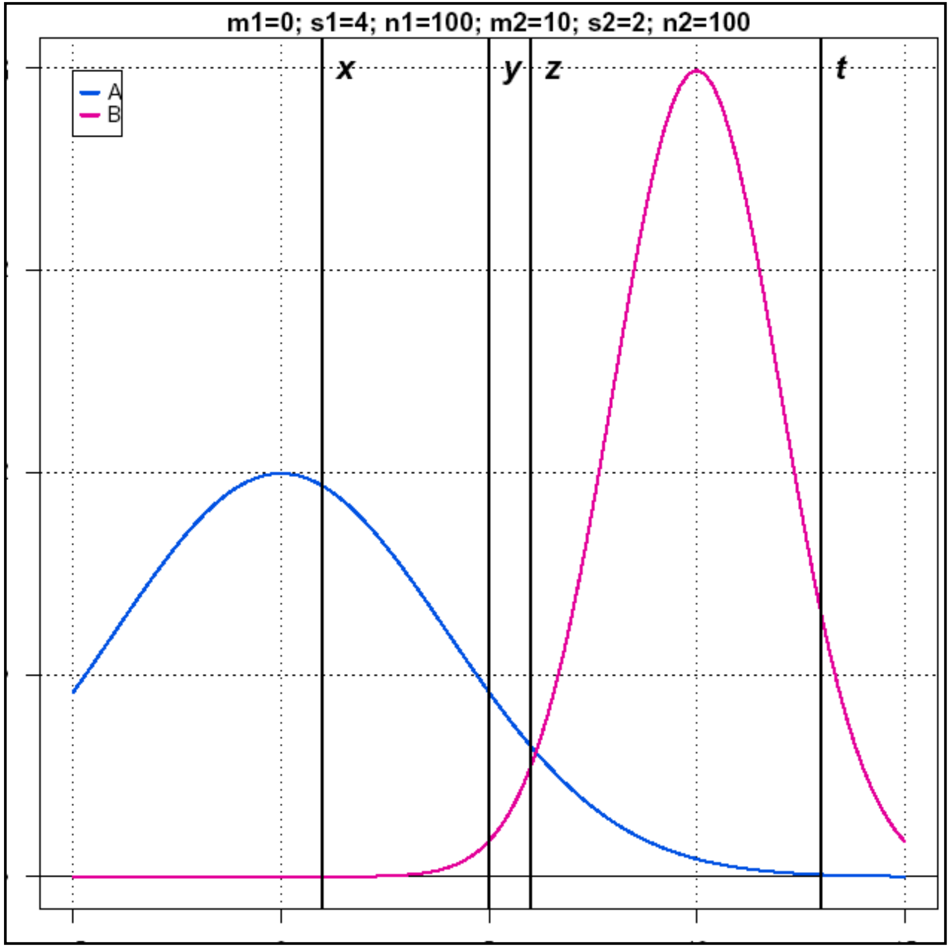
- Same exercise.
- This is the symmetric situation of the preceding figure.
- Although the group centres and variances are identical, the change of sample sizes completely modifies the group assignations.
- This is an effect of ***prior probability***.

Conceptual illustration with a single variable



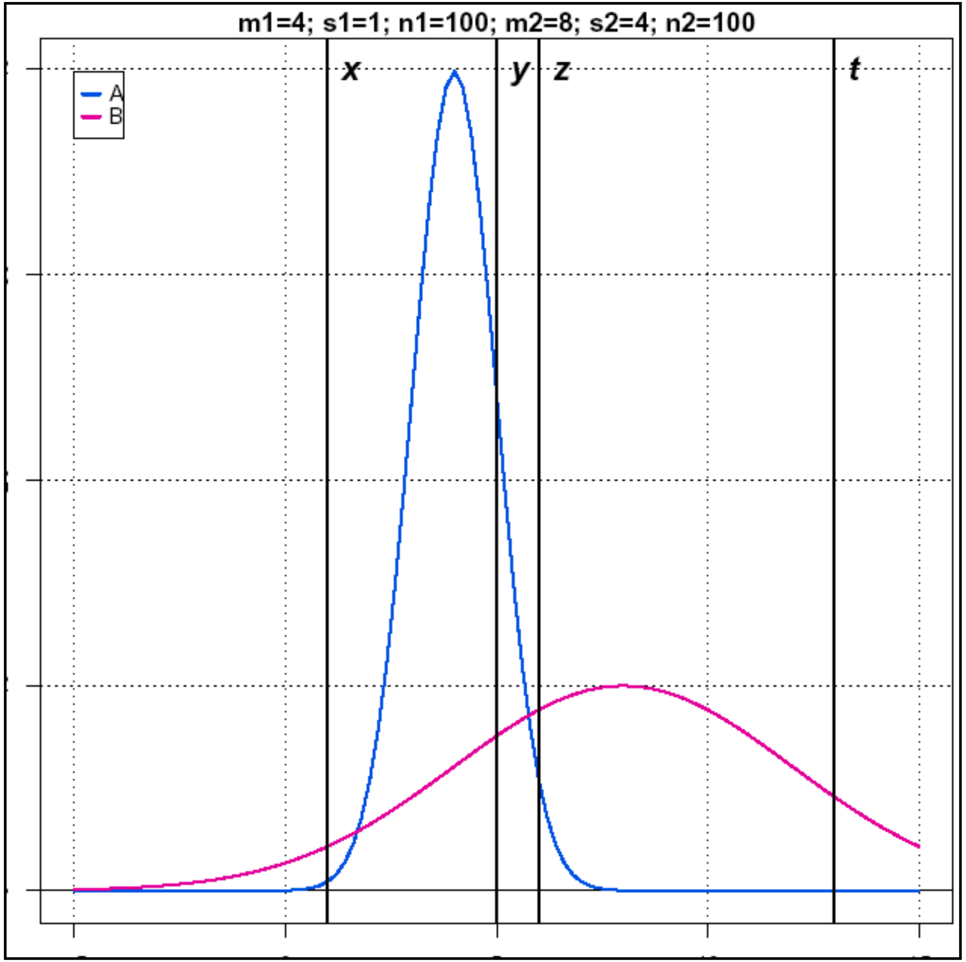
- Same exercise.
- If the two groups have different dispersions, it will affect their likelihood to be the originators of some observations.
- The ***relative dispersion*** of the groups affects the assignation.

Conceptual illustration with a single variable



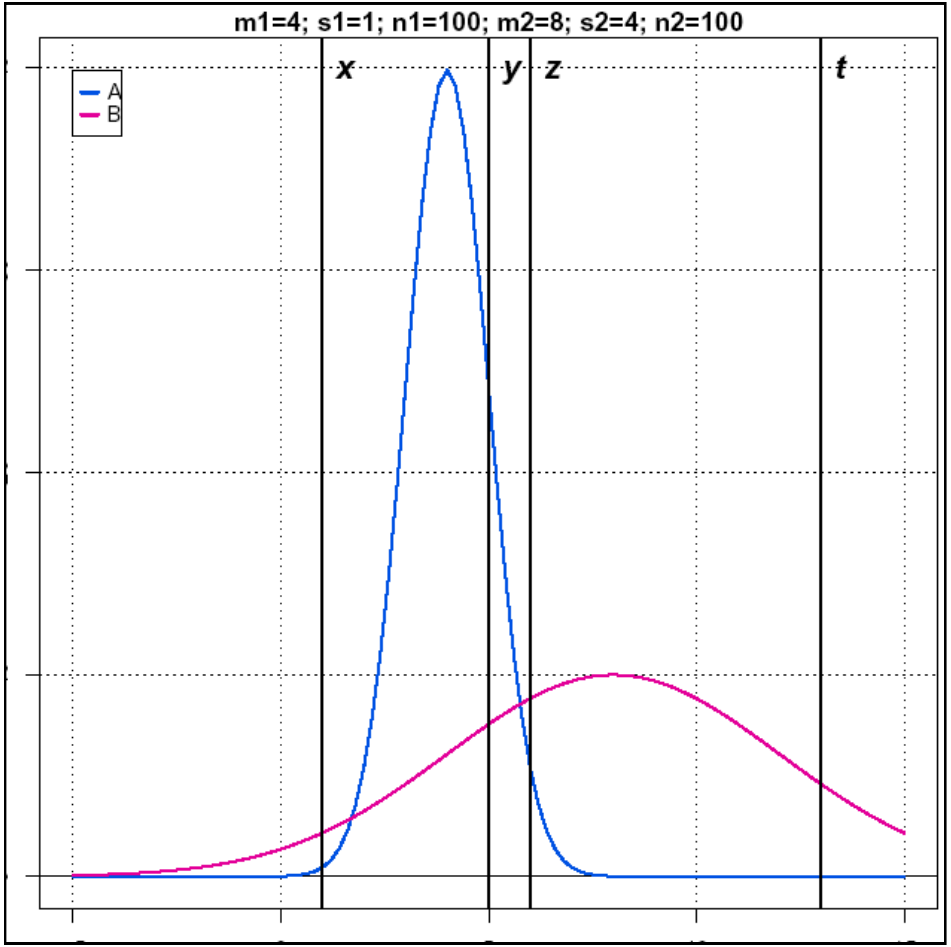
- Same exercise.
- Symmetrical situation of the preceding one: same centres, same sample sizes, but the relative variances vary in the opposite way.
- The ***relative dispersion*** of the groups affects the assignation.

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.

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

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

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

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

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)***