

Pattern discovery

Summary

Finding combinations of genes that are involved in some cellular process -- biological pathways.

- Gene clustering:
Groups of genes with similar expression patterns.
- Data Visualization:
Finding patterns such as gene clusters, outliers, other non-trivial patterns
- Two way clustering:
Clustering genes and subjects at the same time.

Cluster Analysis:

Group the genes (or samples) into k distinct natural groups.

Hierarchical clustering: Build a hierarchical tree

Inter point distance is normally the Euclidean distance (some times we may use Manhattan distance).

Inter cluster distance:

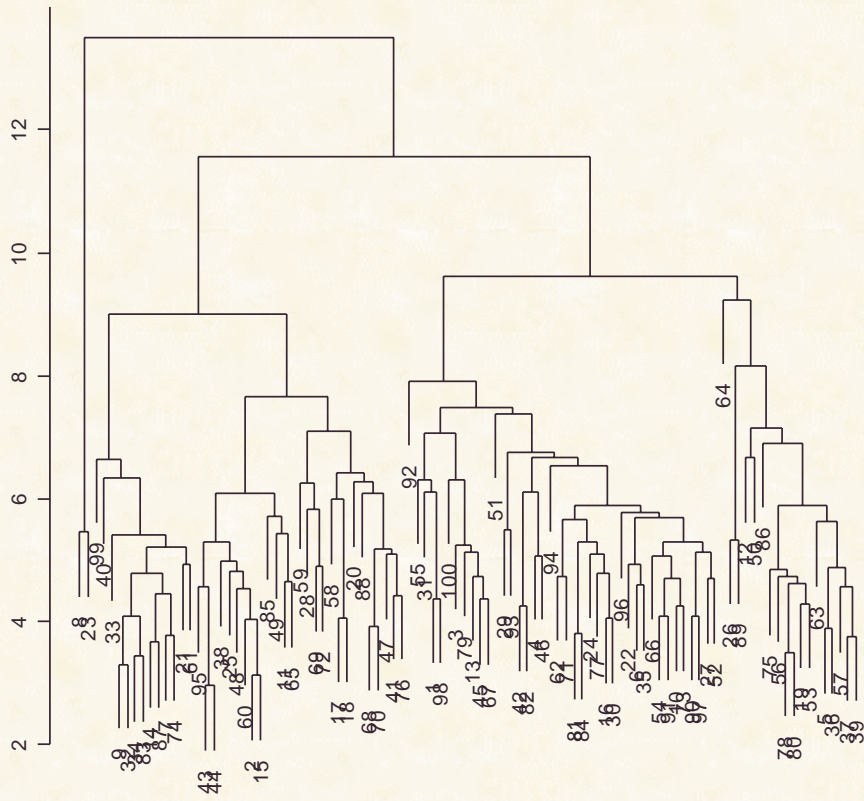
- Single Linkage: distance between the closes two points
- Complete Linkage: distance between the furthest two points
- Average Linkage: Average distance between every pair of points
- Ward: R^2 change.

Build a hierarchical tree:

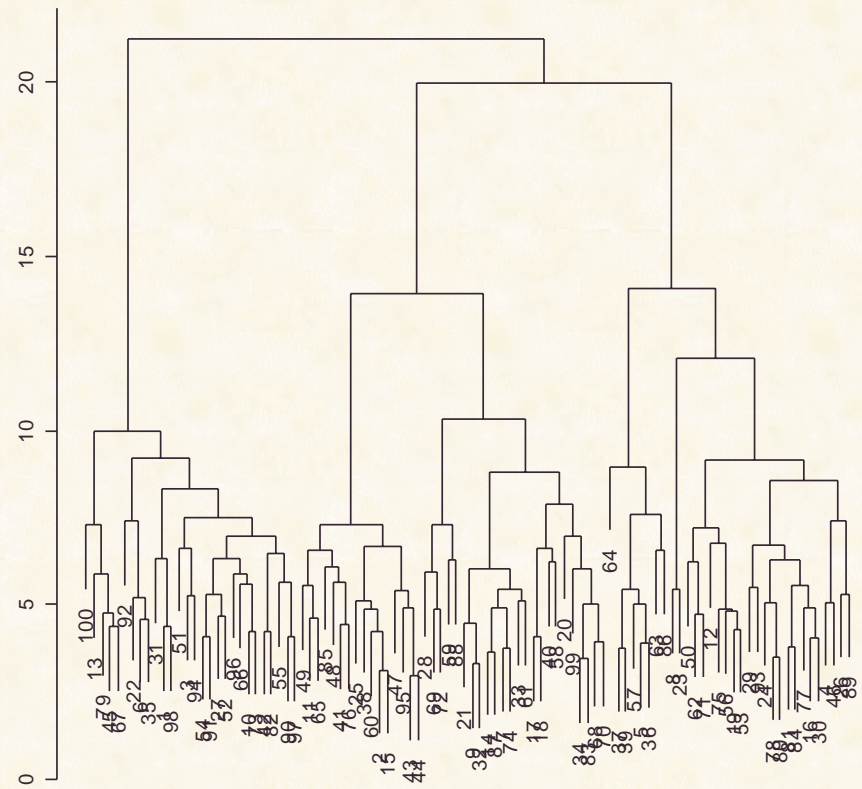
1. Start with a cluster at each sample point
2. At each stage of building the tree the two closest clusters joint to form a new cluster.

Tree dendrograms

(i) Average Linkage

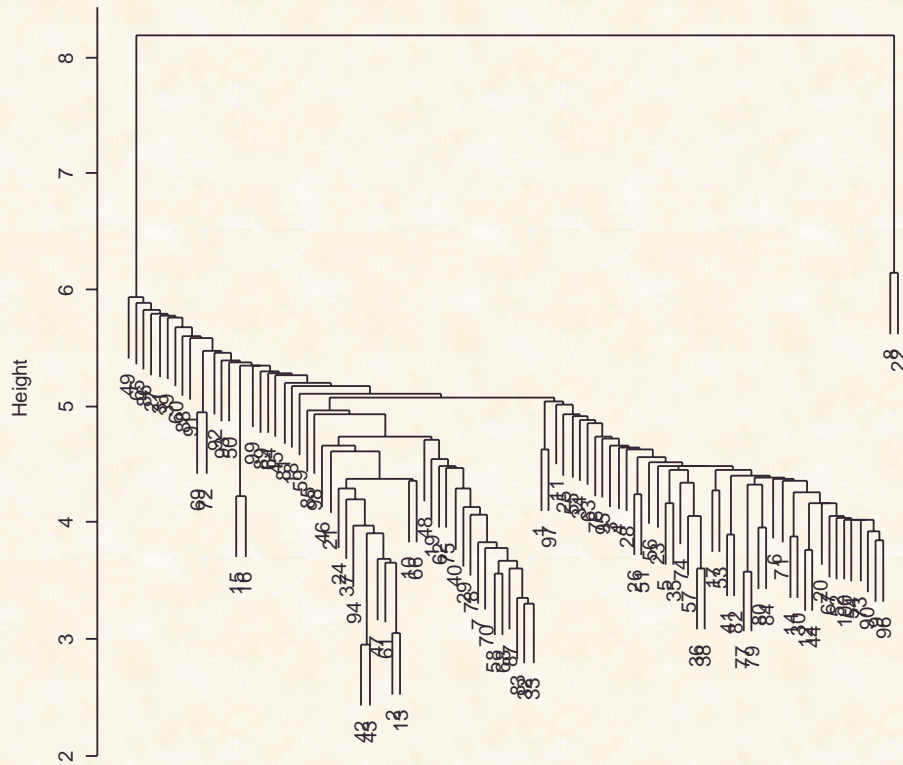


(ii) Complete Linkage

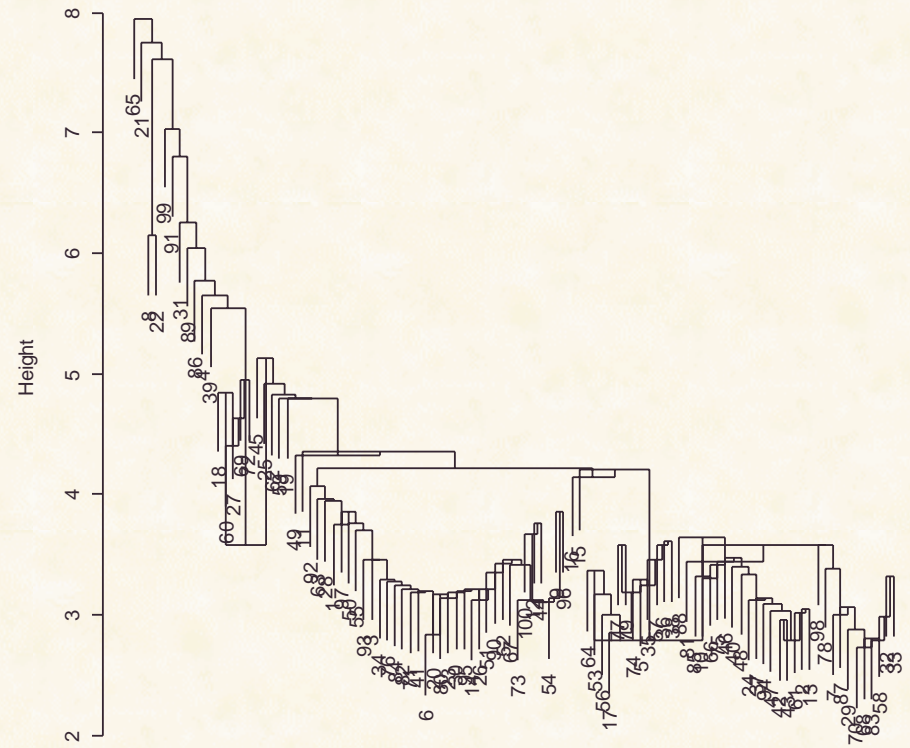


Tree dendrograms

(iii) Single Linkage



(iv) Centroid Linkage

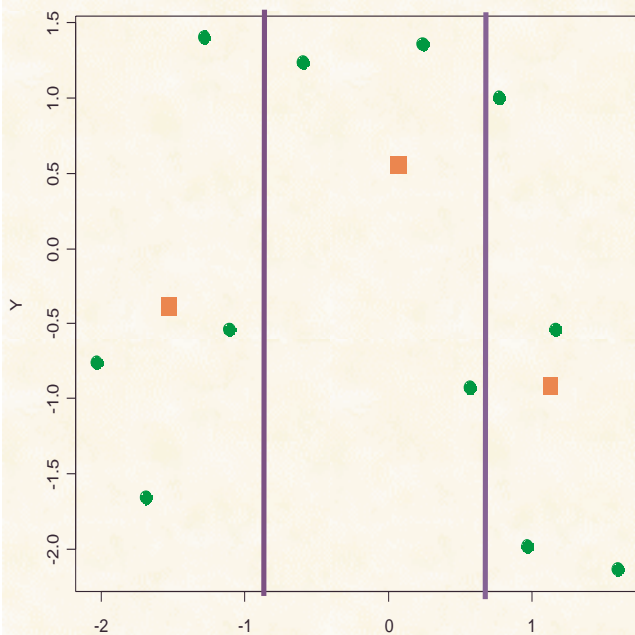


Ward's method

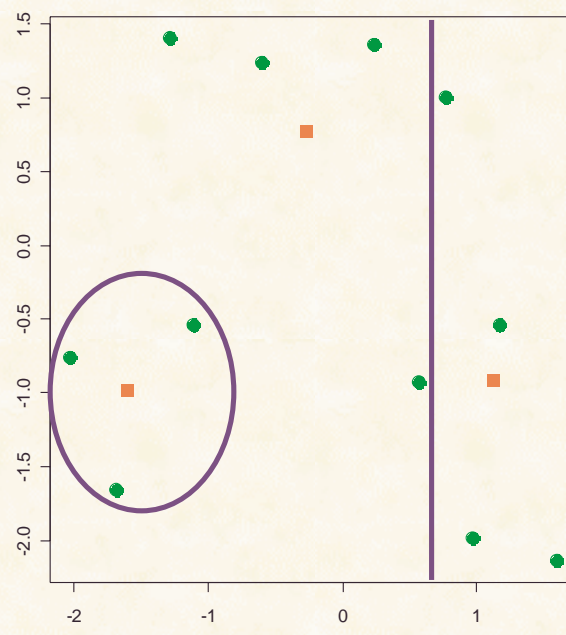


Centroid methods: K-means algorithm.

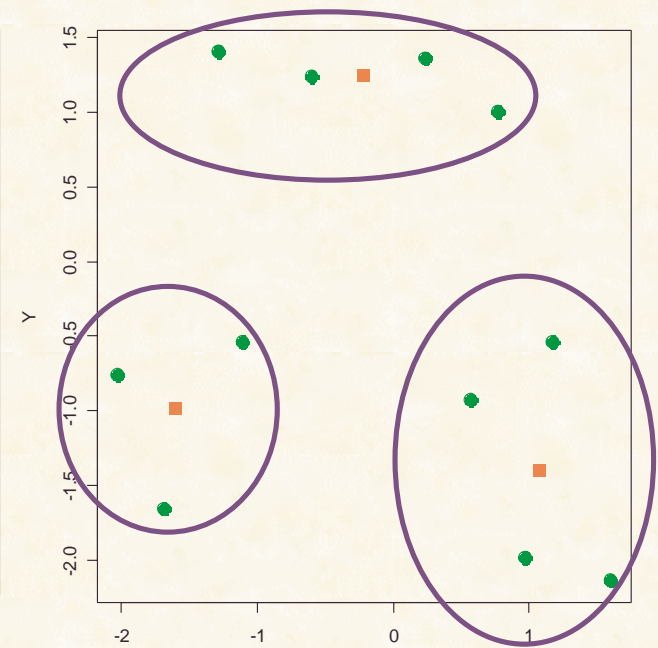
1. K seed points are chosen and the data is distributed among k clusters.
2. At each step we switch a point from one cluster to another if the R^2 is increased.
3. Then the clusters are slowly optimized by switching points until no improvement of the R^2 is possible.



Step 1



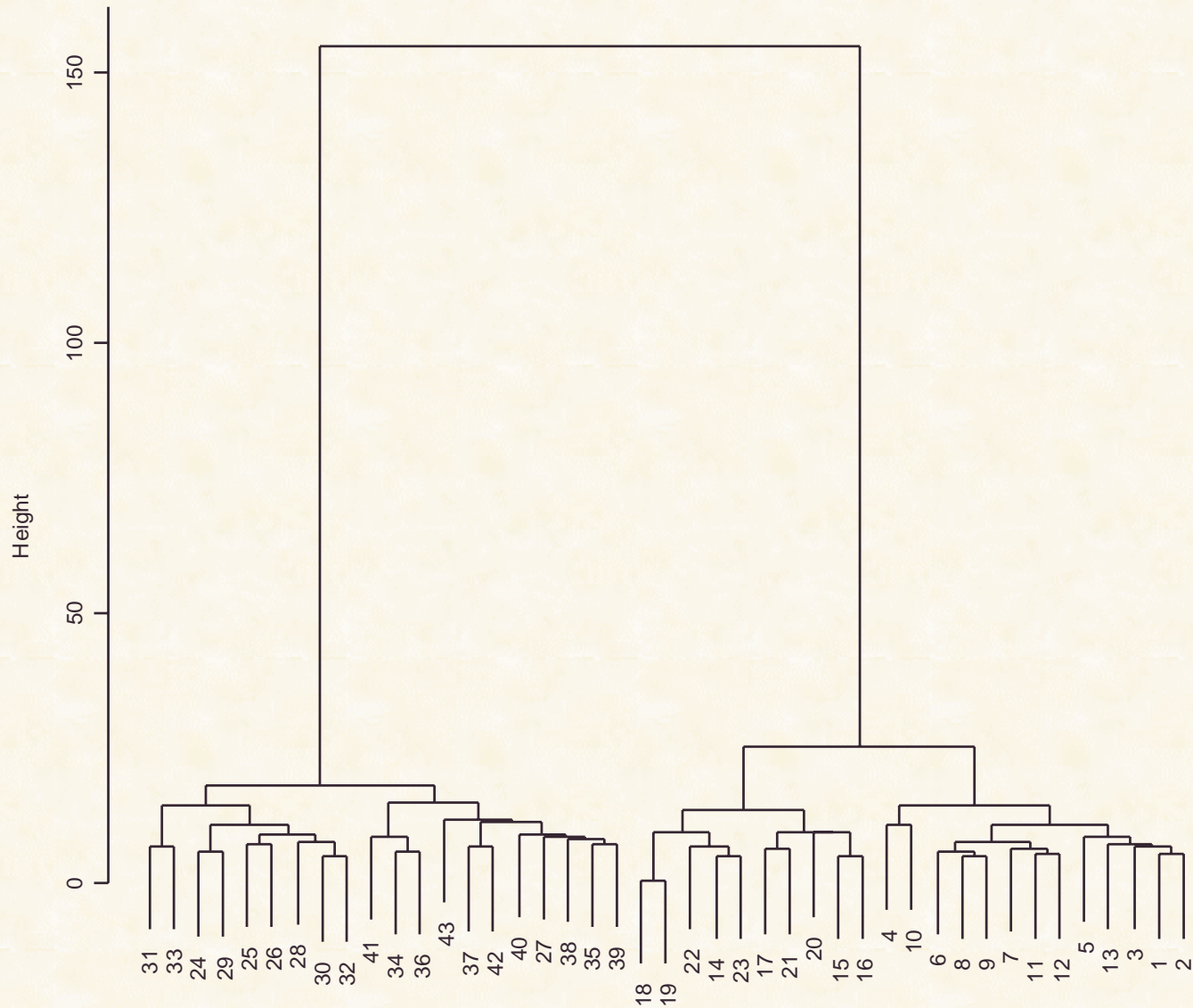
Step 2



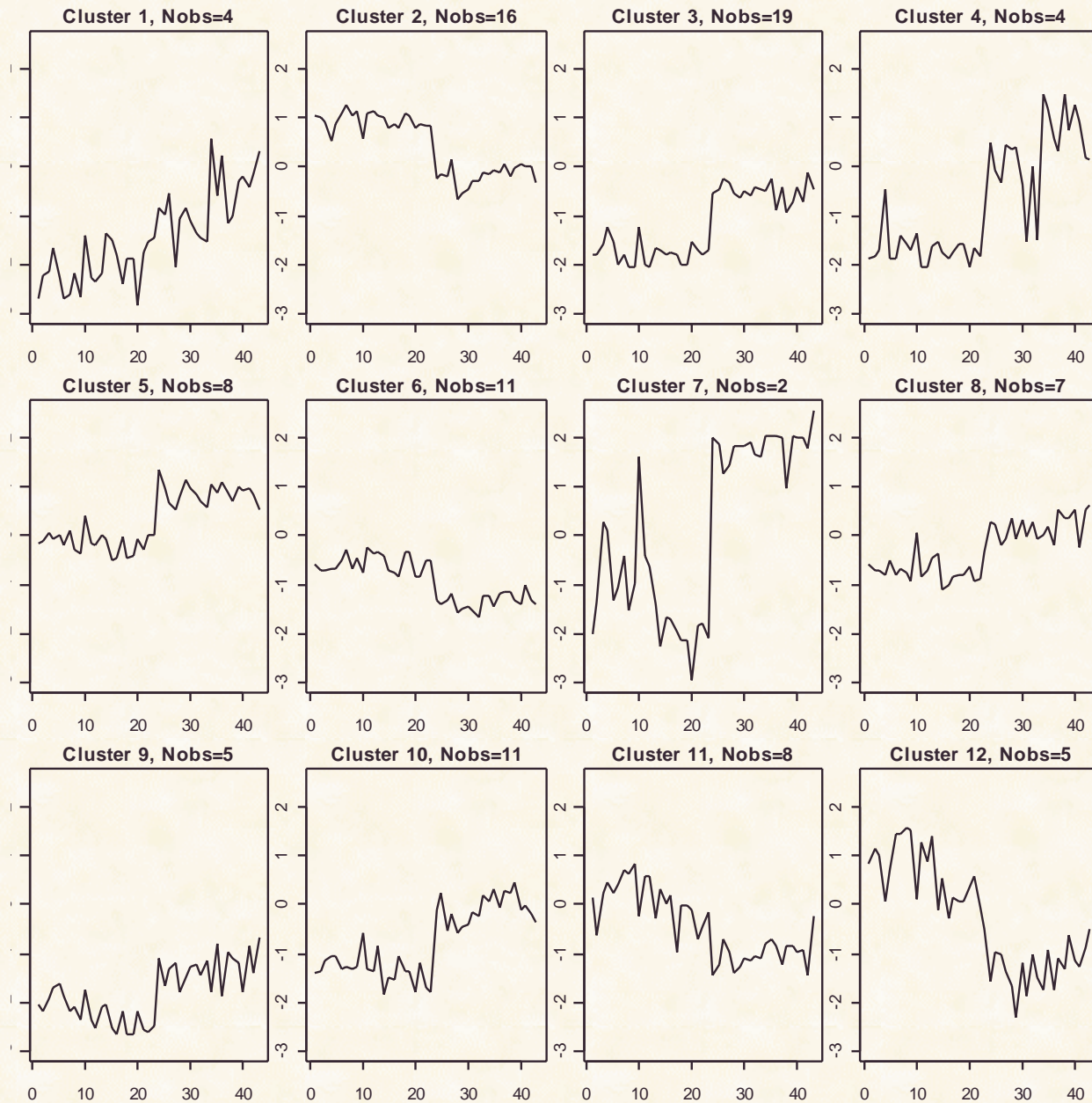
Step n 7

Clustering the observations:

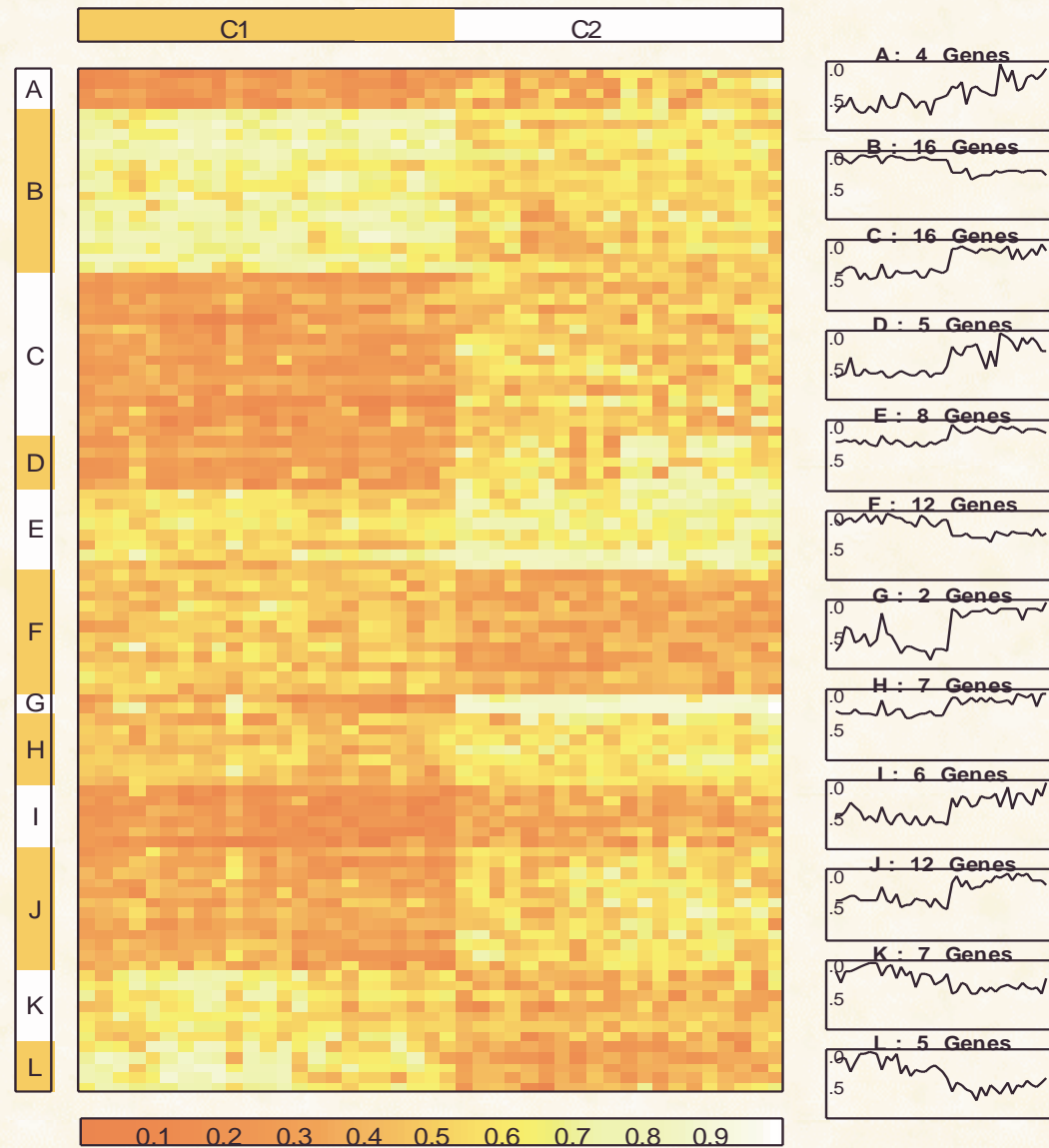
Clustering 43 samples using the 100 genes dataset. Ward's Method.



Cluster Profiles for Ward's method



Microarray graph summarizing the two way clustering



Correlation Matrix

1. Use covariance or correlation matrix?
⇒ Use Correlations

$$R = \begin{pmatrix} 1, r_{12}, \dots, r_{1G} \\ r_{21}, 1, \dots, r_{2G} \\ \dots \dots \dots \\ r_{G1}, r_{G2}, \dots, 1 \end{pmatrix}$$

2. $\text{Dim}(R) = G \times G$ and G is between 1000 and 25000, this is too big
⇒ Dimension reduction.

3. $\text{Rank}(R) = p$

Gene expression matrix X :

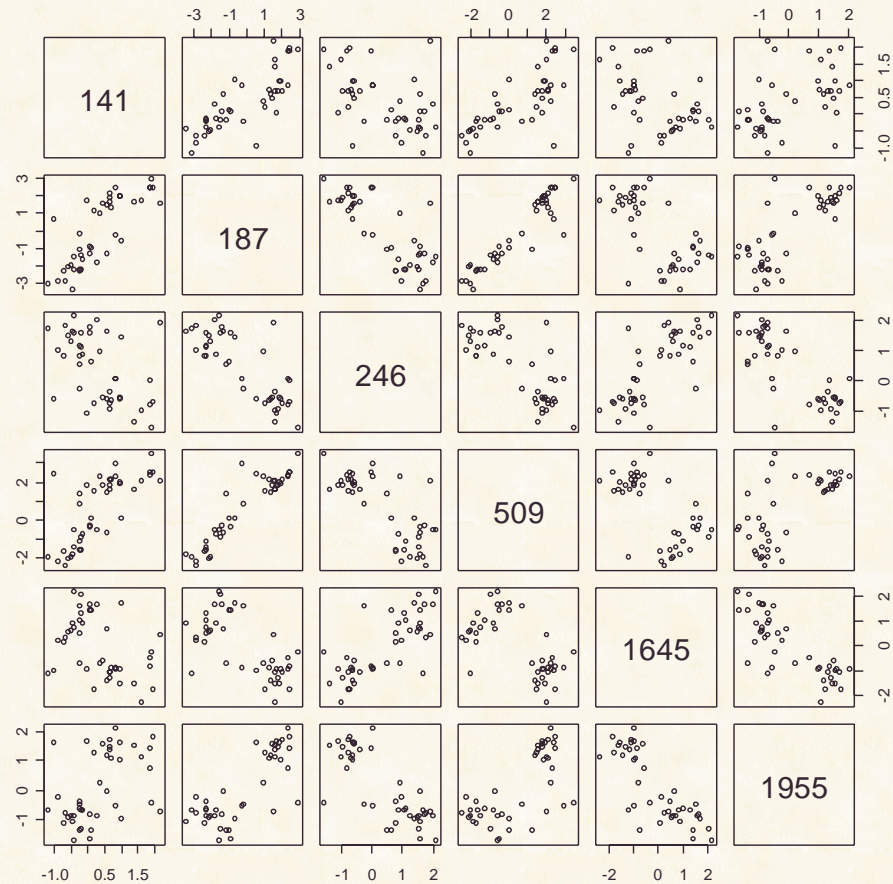
Rows = Genes = Variables

Columns = Microarrays = Subjects = Observations

Sample Correlation Matrix

	Gene 141	Gene 187	Gene 246	Gene 509	Gene 1645	Gene 1955
Gene 141	1.0000	0.7983 (0.000)	-0.5058 (0.001)	0.7463 (0.000)	-0.4049 (0.007)	0.4676 (0.002)
Gene 187	0.7983 (0.000)	1.0000	-0.8111 (0.000)	0.9357 (0.000)	-0.6621 (0.000)	0.7891 (0.000)
Gene 246	-0.5058 (0.001)	-0.8111 (0.000)	1.0000	-0.7717 (0.000)	0.7624 (0.000)	-0.7977 (0.000)
Gene 509	0.7463 (0.000)	0.9357 (0.000)	-0.7717 (0.000)	1.0000	-0.6388 (0.000)	0.6827 (0.000)
Gene 1645	-0.4049 (0.007)	-0.6621 (0.000)	0.7624 (0.000)	-0.6388 (0.000)	1.0000	-0.8143 (0.000)
Gene 1955	0.4676 (0.002)	0.7891 (0.000)	-0.7977 (0.000)	0.6827 (0.000)	-0.8143 (0.000)	1.0000

Scatterplot Matrix



Linear Algebra

Linear algebra is useful to write computations in a convenient way.

Singular Value Decomposition: $X = U D V'$
 $G_{xp} \quad G_{xp} \quad p \times p \quad p \times p$

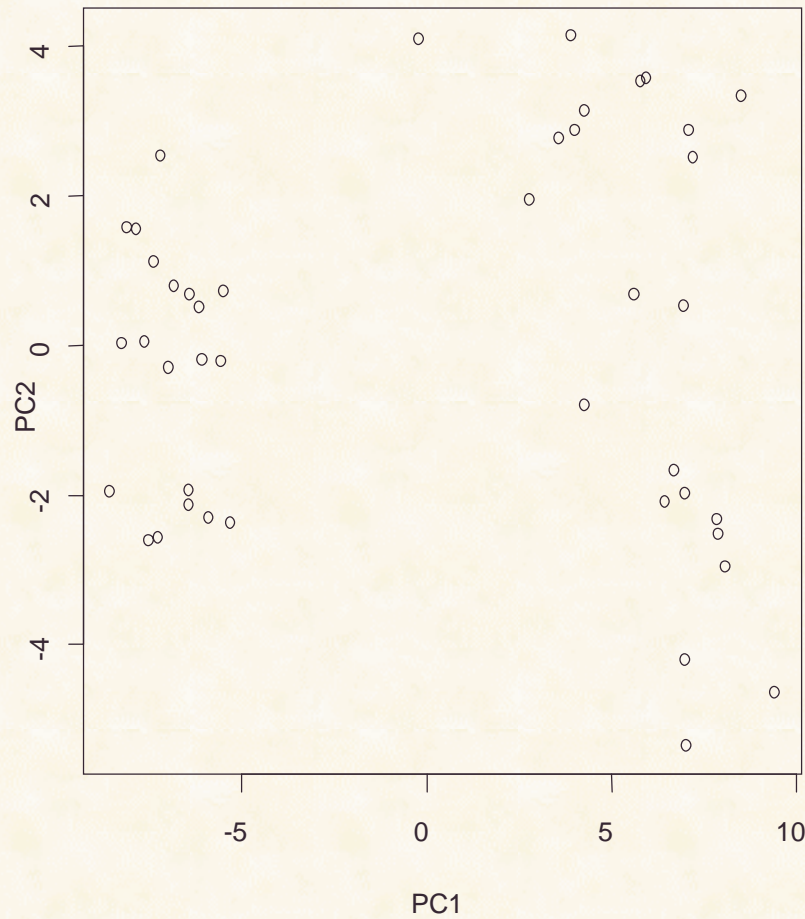
Covariance Matrix: $S = U D^2 U'$
 $G_{xG} \quad G_{xp} \quad p \times p \quad p \times G$

Correlation Matrix: Subtract mean of rows of X and divide by standard deviation and calculate the covariance

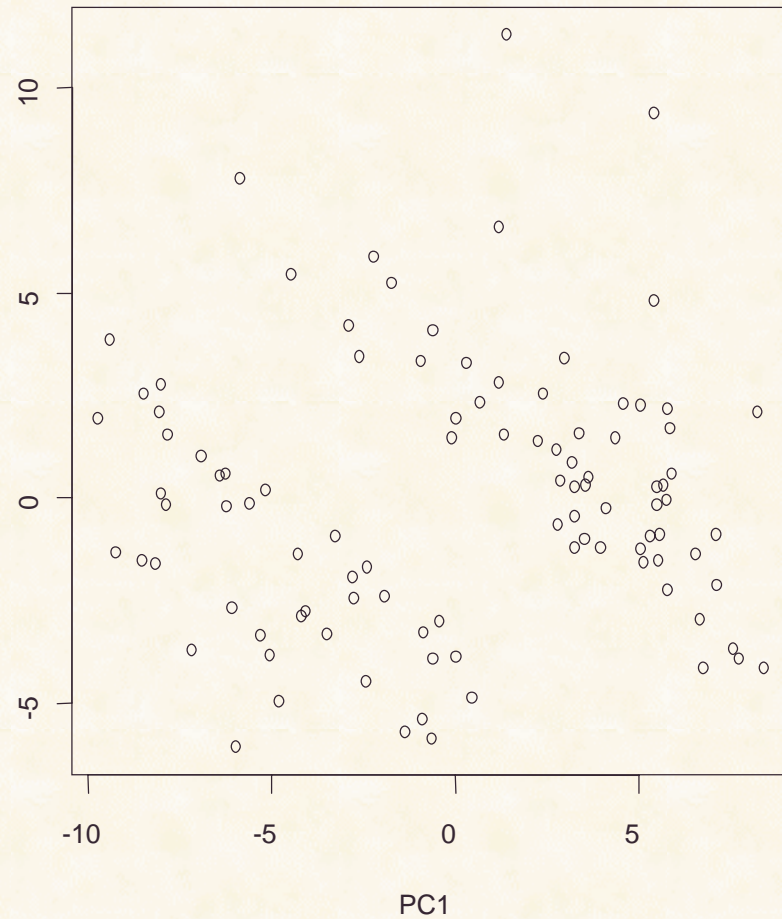
Principal Components(PC): Columns of U.

Eigenvalues (Variance of PC's): Diagonal elements of D

Principal components of 100 genes. PC2 Vs PC1.



(a) Cells are the observations
Genes are the variables



(b) Genes are the observations
Cells are the variables

Dimension reduction:

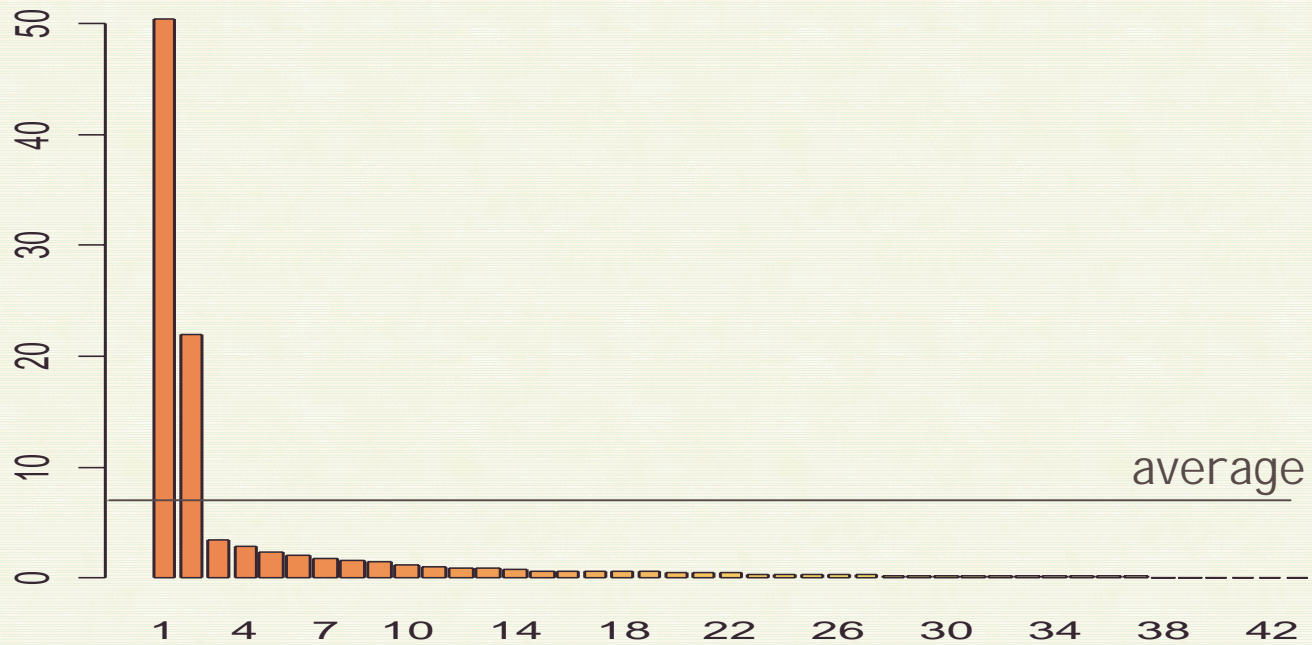
Choosing the number of PC's

1. k components explain some percentage of the variance: 70%,80%.
2. k eigenvalues are greater than the average (1)
3. *Scree plot*: Graph the eigenvalues and look for the last sharp decline and choose k as the number of points above the cut off.
4. Test the null hypothesis that the last m eigenvalues are equal (0)

$$u = (G - (2m + 1) / 6)(m \times \log \bar{\lambda} - \sum_{i=p-m+1}^p \log \lambda_i)$$

The same idea can be applied to factor analysis.

1. *The top 5 eigenvalues explain 81% of variability.*
2. *Five eigenvalues greater than the average 2.5%*
3. *Scree Plot*



4. *Test statistic is 4 significant for 6 and highly significant for 2.*

$p-m$	24	20	15	9	8	7	6	5	4	3	2	1
u	0.1	5	32	146	182	222	279	340	425	554	1632	3260
χ^2	9.2	37	94	195	215	237	259	282	307	332	358	386

Biplots

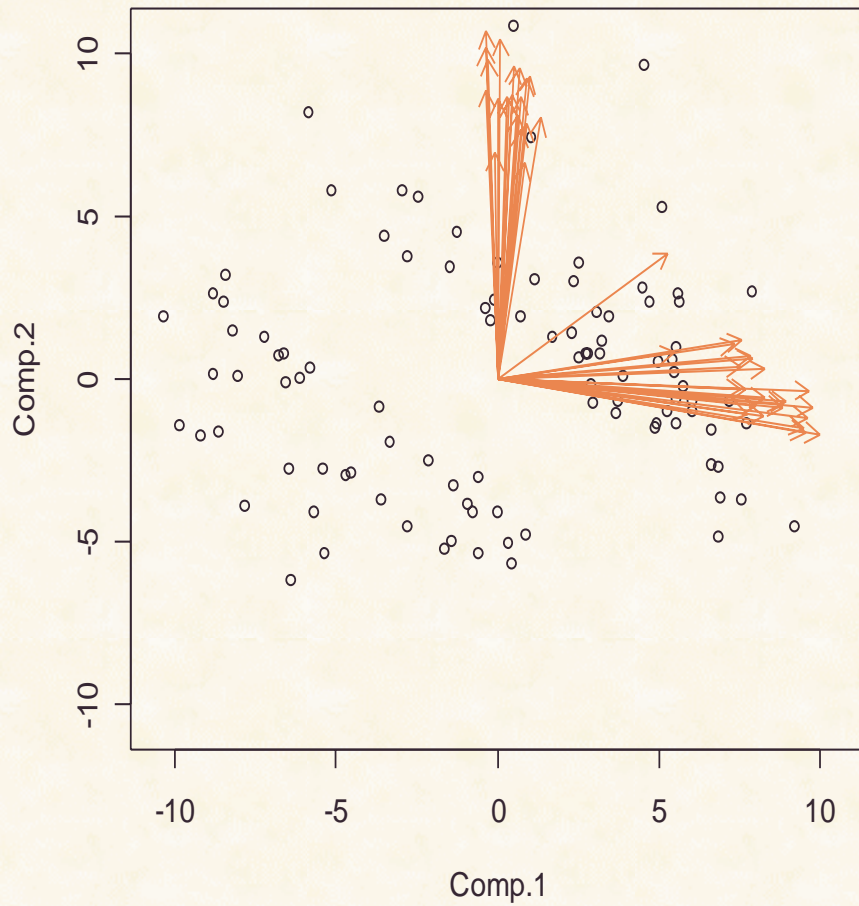
Graphical display of X in which two sets of markers are plotted.
One set of markers a_1, \dots, a_G represents the rows of X
The other set of markers, b_1, \dots, b_p , represents the columns of X .

$$\text{For example: } X = UDV' \Rightarrow X_2 = U_2 D_2 V_2'$$

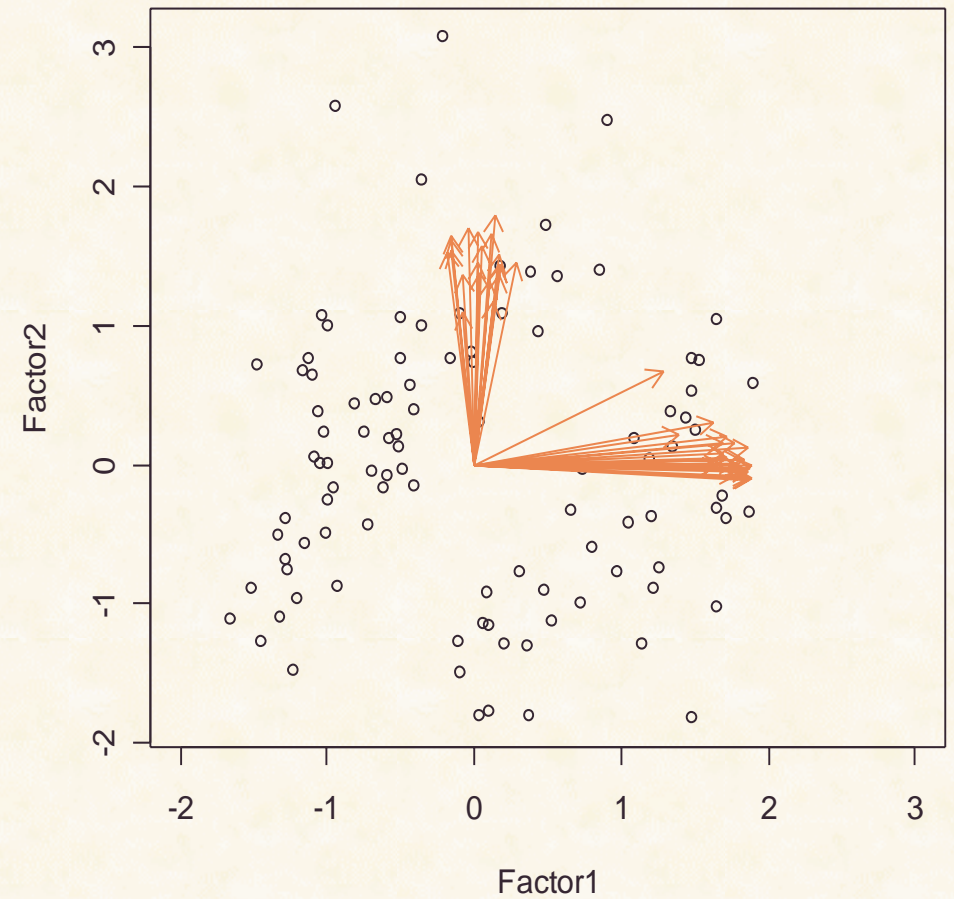
$$A = U_2 D_2^a \text{ and } B = V_2 D_2^b, a+b=1 \text{ so } X=AB'$$

The biplot is the graph of A and B together in the same graph.

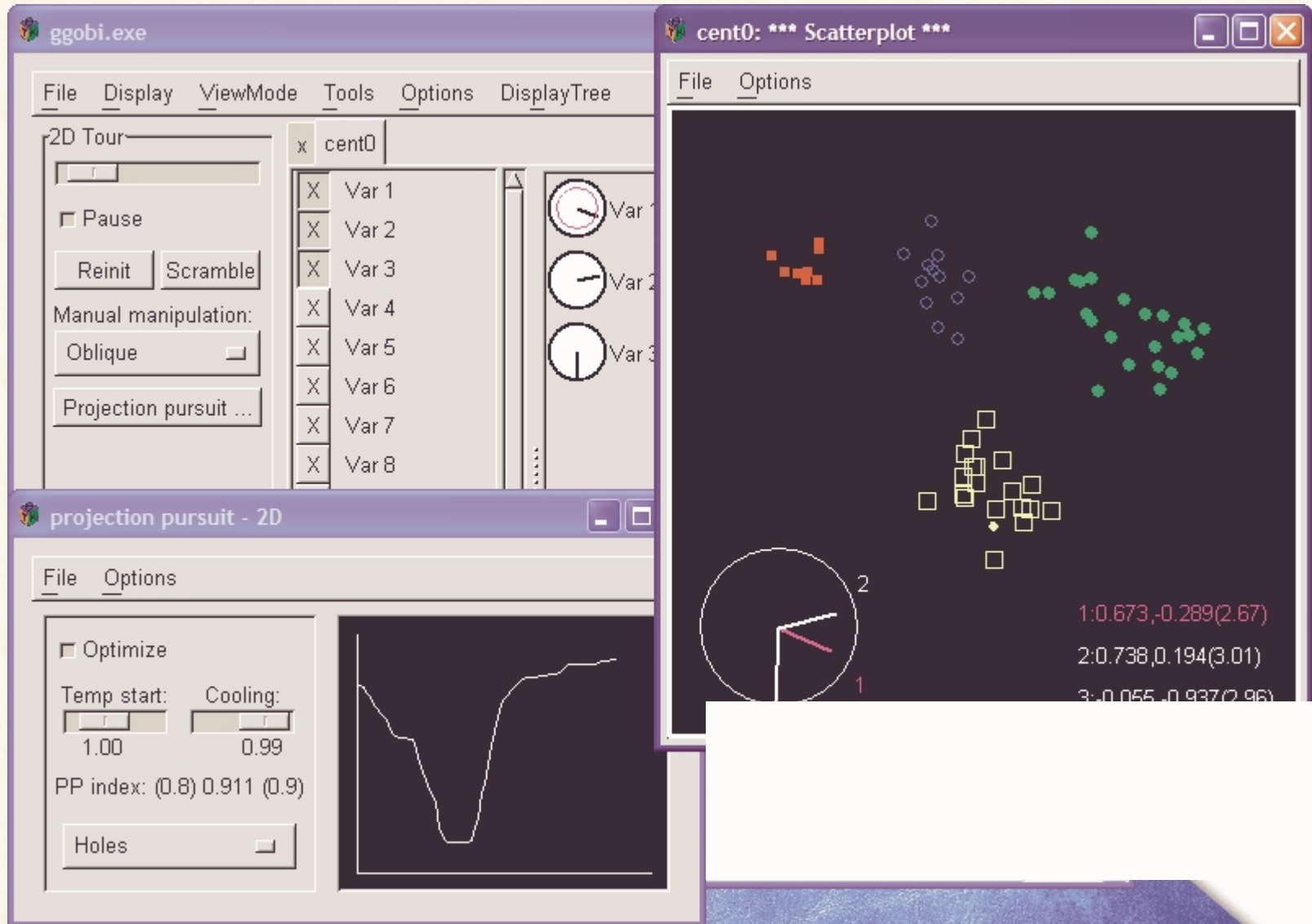
Biplot of the first two principal components.



Biplot of the first two Factors (rotated).



Ggobi display finding four clusters of tumors using the PP index on the set of 63 cases. The main panel shows the two dimensional projection selected by the PP index with the four clusters in different colors and glyphs. The top left panel shows the main controls and the left bottom panel displays the controls and the graph of the PP index that is been optimized. The graph shows the index value for a sequence of projection ending at the current one.



Class prediction or supervised classification

Example:

Golub *et al* (1999): Separate two different types of leukemia, (ALL, AML) based on gene expression information.

Khan *et al* (2001): 4 types of small round blue cell tumors (SRBCT)
Neuroblastoma (NB)
Rhabdomyosarcoma (RMS)
Ewing family of tumors (EWS)
Burkitt lymphomas (BL)

Arrays: Training set= 63 arrays(23 EWS, 20 RMS, 12 NB, 8 BL)
Testing set= 25 arrays(6 EWS, 5 RMS, 6 NB, 3 BL, 5 other)

Genes: Of 6567 initial genes, 2308 genes were selected because they showed minimal expression levels.

Class prediction or supervised classification

Objectives:

Classification of new patients into one of the groups.
Achieve a low misclassification rate.

Problem:

Too many genes.

Data reduction:

Using fewer genes (30-50)
Reducing the dimension with PA and FA (1-10).

Few genes makes the classification rule more practical.

Variable Reduction: four ways of doing the variable reduction

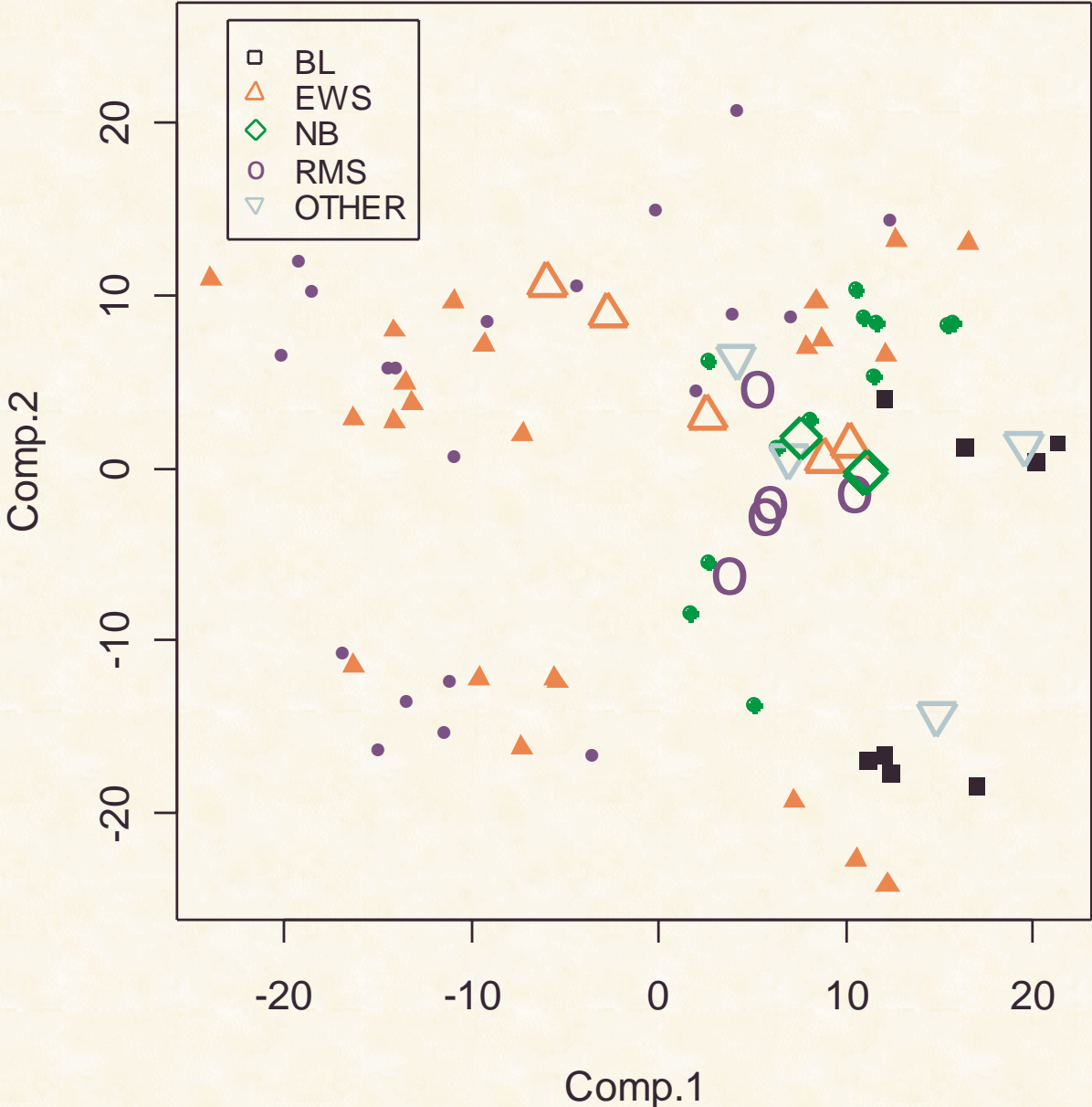
(i) First 10 PCA basis using all genes

(ii) 10 Principal Components for 450 significant genes.

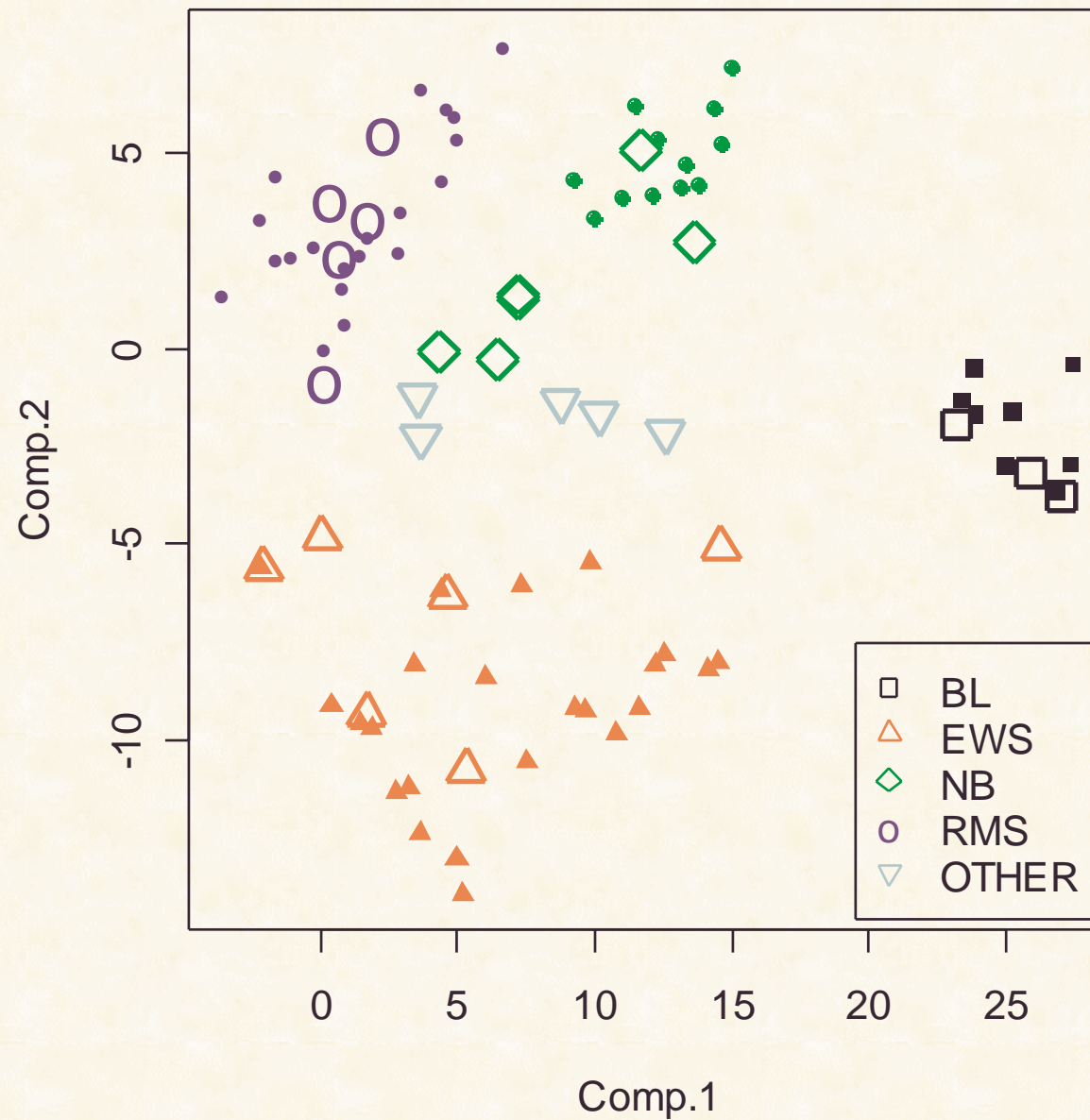
(iii) 10 Cluster means of 50 significant genes.

(iv) Principal Components for the top 30 significant genes.

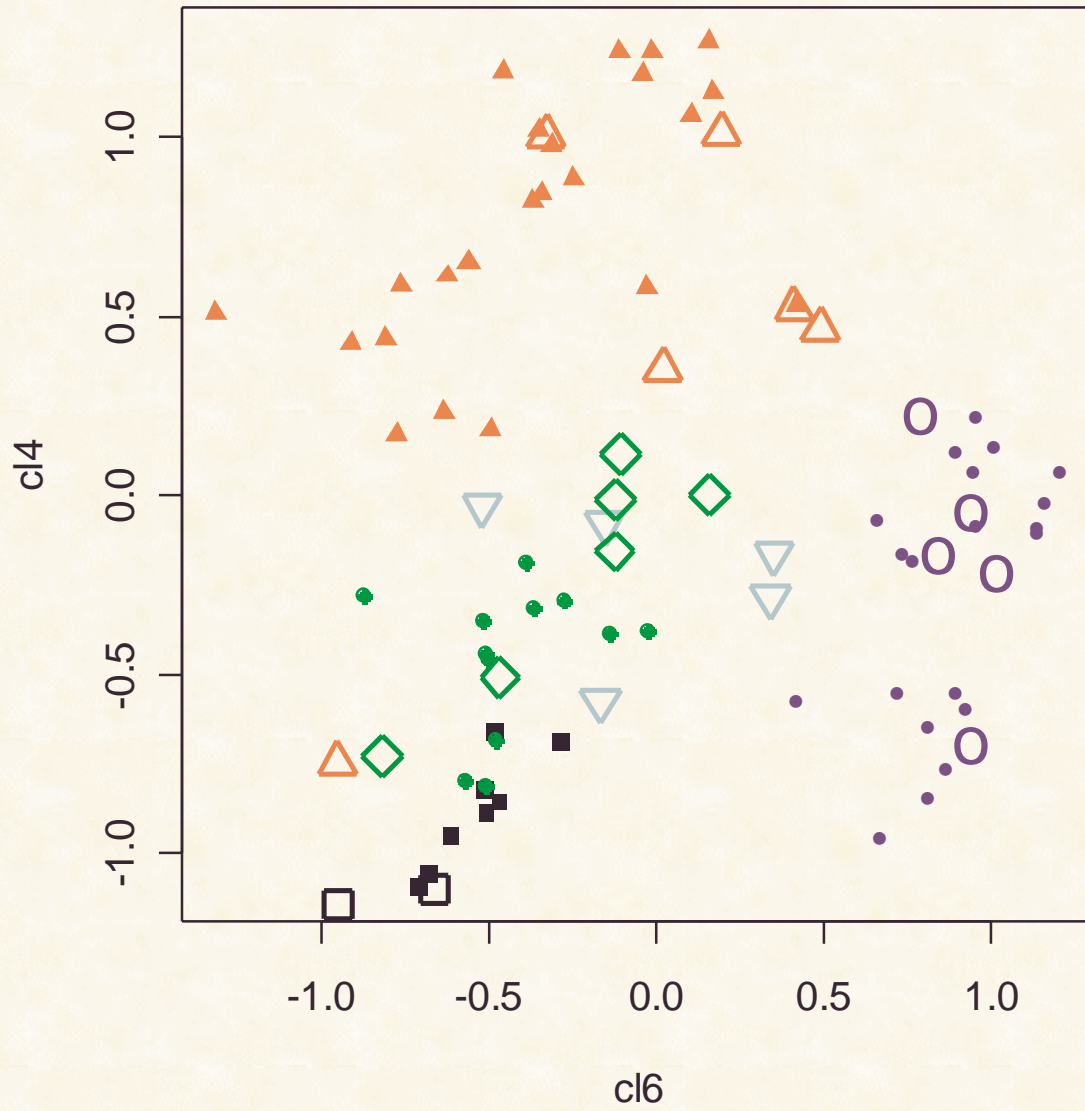
(i) 63 training samples and the 25 test samples in the coordinates of the first two PCA basis using all genes



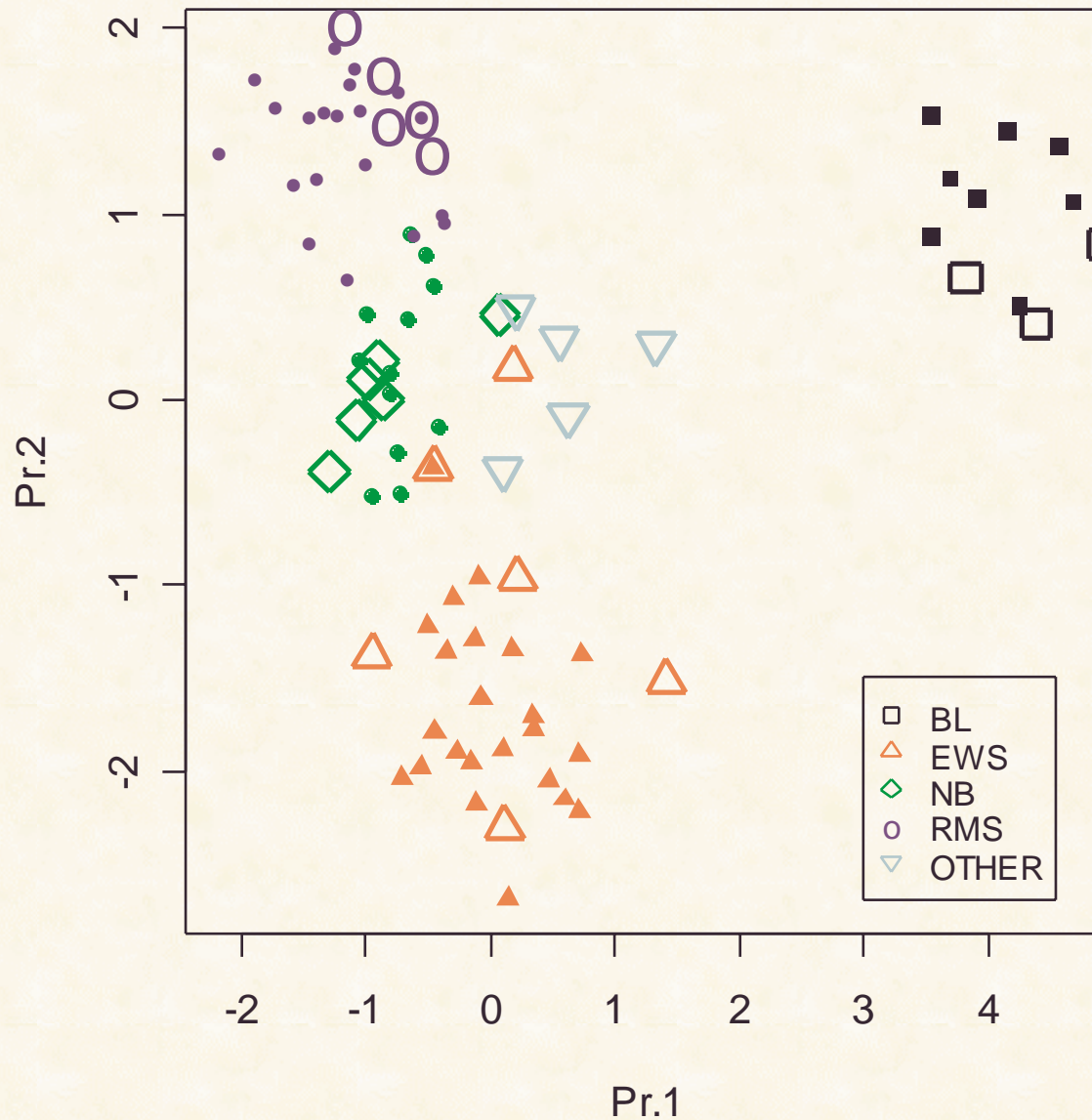
(ii) Principal Components for 450 significant genes. 63 training samples and the 25 test samples in the two first PCA basis.



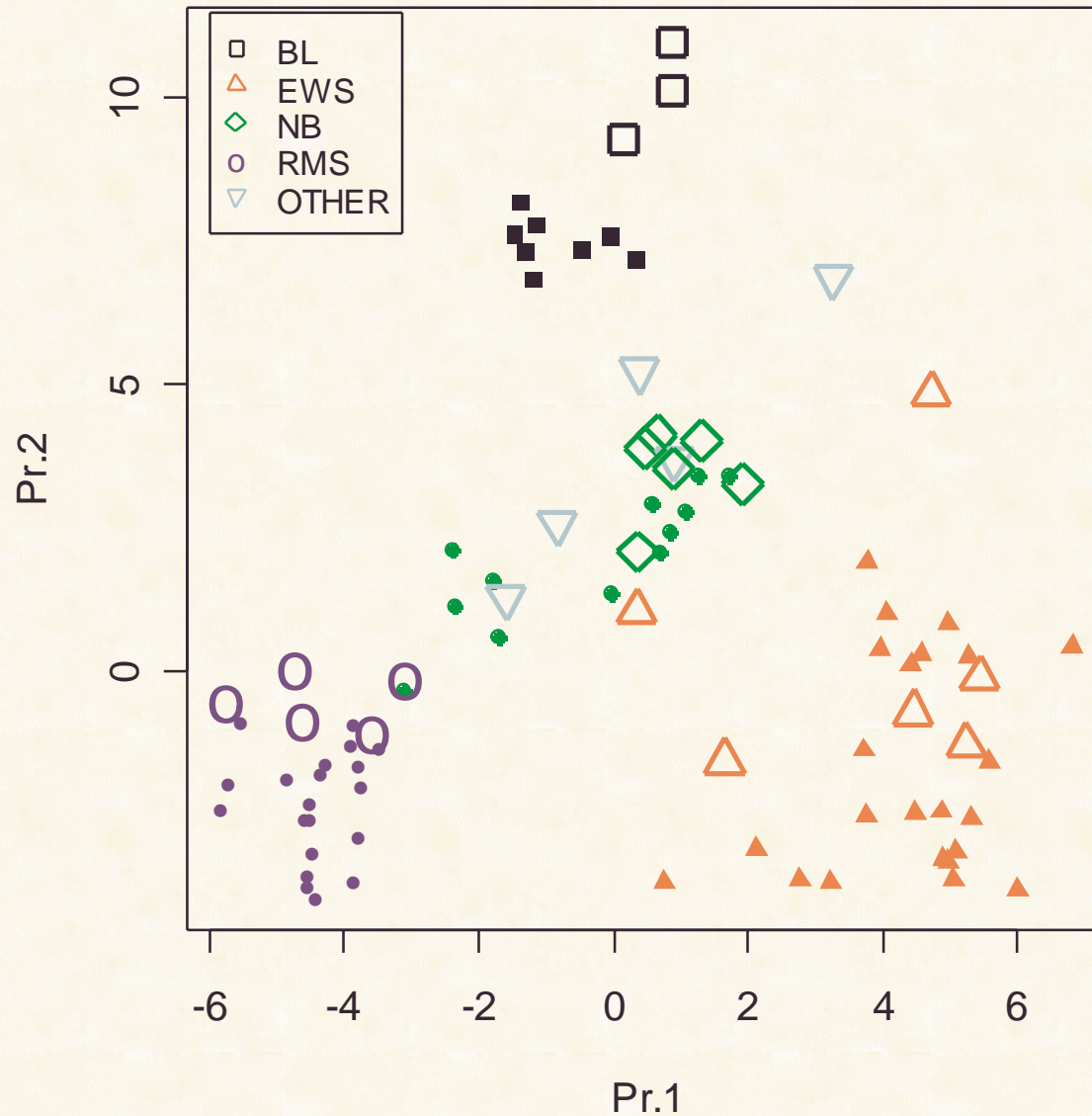
(iii) 2 Cluster means of 50 significant genes. 63 training samples and the 25 test samples in the two first PCA basis.



(iii)b First two Principal Components for the means of 10 clusters obtained from a subset of the top 50 significant genes. 63 training samples and the 25 test samples in the two first PCA basis.



(iv) Principal Components for the top 30 significant genes. 63 training samples and the 25 test samples in the two first PCA basis.



Linear discriminant analysis

The classification rule is:

$$\text{If } w'x > w'(\bar{x}_1 + \bar{x}_2)/2$$

then x is classified as belonging to Class 1,

where

$$w = S^{-1}(\bar{x}_1 - \bar{x}_2)$$

	<i>(i) 10 PC of 2308 genes.</i>		<i>(ii) 10 PC of 450 genes.</i>		<i>(iii) 10 cluster means of 50 genes.</i>		<i>(iv) 10 PC of 30 genes.</i>	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
2 Classifiers	35	13	0	2	4	2	3	2
3 Classifiers	5	5	0	0	0	1	1	1
4 Classifiers	0	3	0	0	0	0	0	0
10 Classifiers	0	3	0	0	0	0	0	0

Quadratic discriminant analysis

The classification rule is:

x is classified as belonging to class with largest value of:

$$Q_h(x) = (x - \bar{x}_h) S_h^{-1} (x - \bar{x}_h)' + \log(S_h)$$

	<i>(i) 10 PC of 2308 genes.</i>		<i>(ii) 10 PC of 450 genes.</i>		<i>(iii) 10 cluster means of 50 genes.</i>		<i>(iv) 10 PC of 30 genes.</i>	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
2 Classifiers	26	14	0	2	3	2	4	6
3 Classifiers	0	6	0	1	1	1	0	5
4 Classifiers	0	6	0	1	0	0	0	0
7 Classifiers	0	3	0	0	0	0	0	0

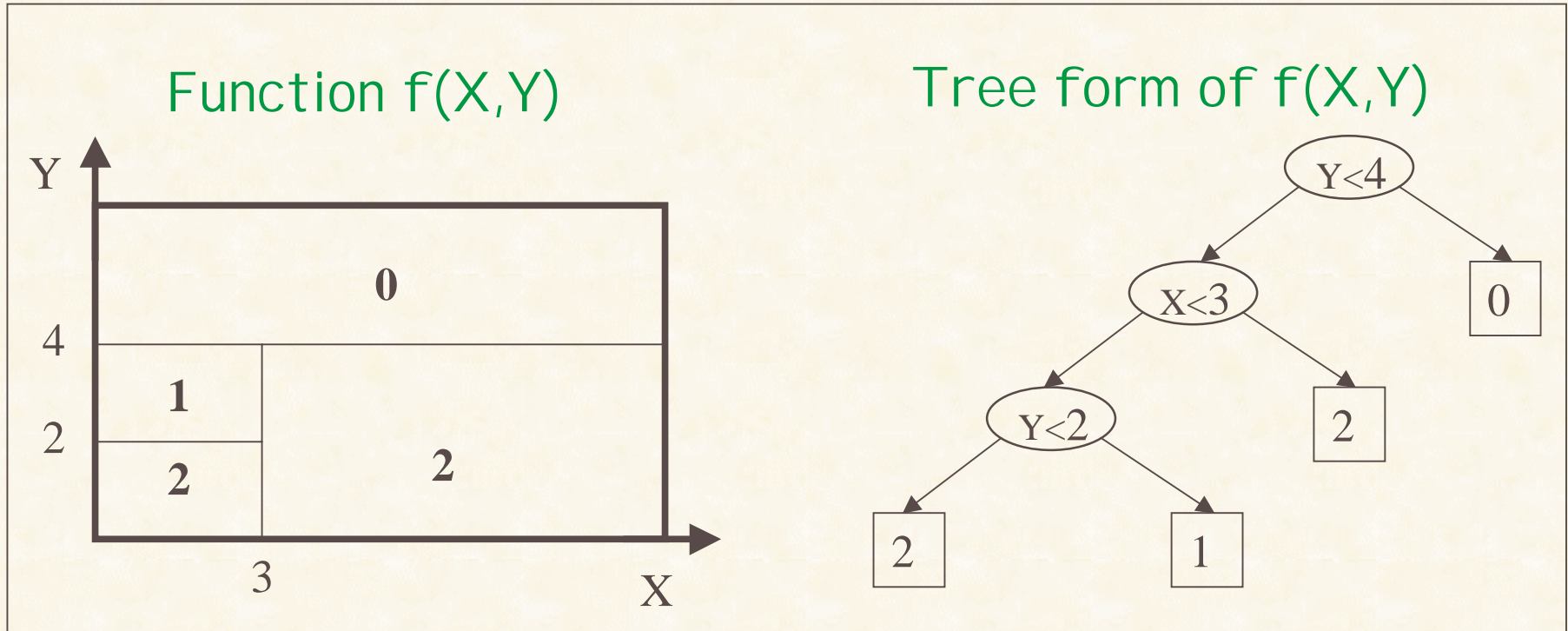
K -nearest neighbors

The classification rule is:

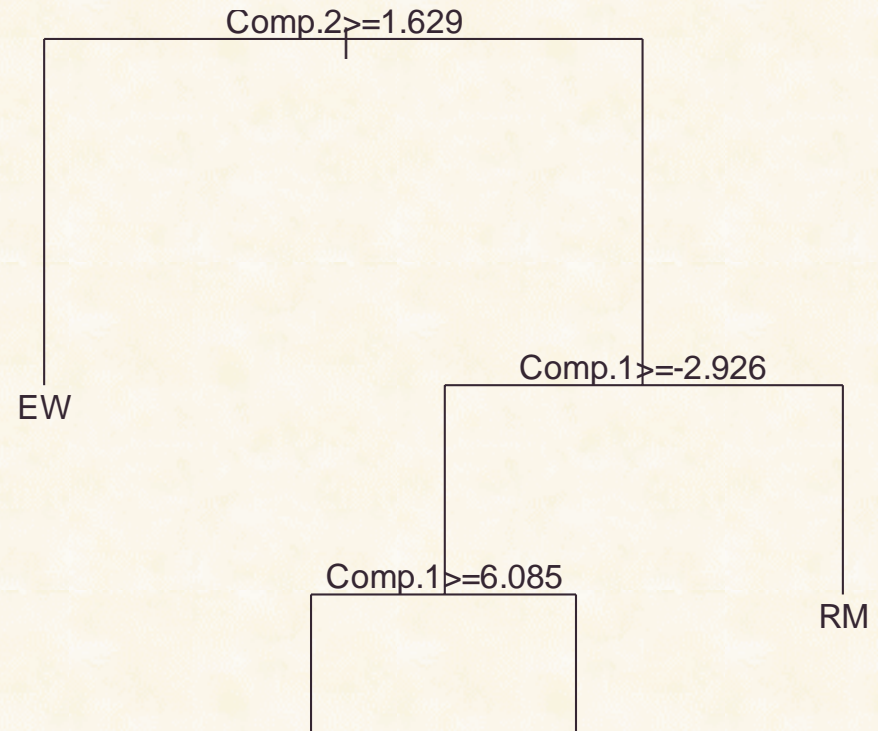
x is classified as belonging to class that is more frequent among the k nearest neighbors of x.

	<i>(i) 10 PC of 2308 genes.</i>		<i>(ii) 10 PC of 450 genes.</i>		<i>(iii) 10 cluster means of 50 genes.</i>		<i>(iv) 10 PC of 30 genes.</i>	
	K=1,2	K=10	K=1,4,...,10	K=2,3	K=1,2	K=10	K=1,2,3	K=10
2 Classifiers	14	10	2	1	2	2	2	2
3 Classifiers	10	8	1	0	0	0	1	1
4 Classifiers	3	2	1	1	0	0	1	1
10 Classifiers	6	3	1	1	0	0	1	1

Classification trees



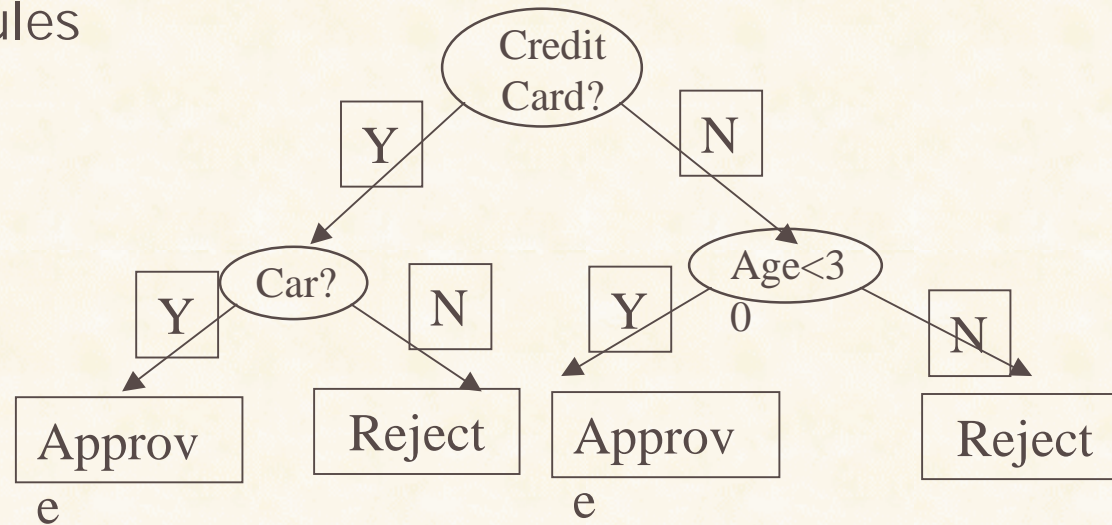
Classification tree for the cancer groups using 10 principal components of the top 100 cancer genes. The classification rule produces zero mistakes in the training set and five mistakes in the testing set.



	<i>(i) 10 PC of 2308 genes.</i>		<i>(ii) 10 PC of 450 genes.</i>		<i>(iii) 10 cluster means of 50 genes.</i>		<i>(iv) 10 PC of 30 genes.</i>	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
2 Classifiers	18	14	0	2	0.5	1.5	0	3.5
3 Classifiers	8	14	0	1	0	0.5	0	1.5
4 Classifiers	0	3	0	0.5	0	0.5	0	1.5
10 Classifiers	0	8	0	0.5	0	0.5	0	1.5

Tree methods: Dependent variable is categorical

- Classification trees (e.g., CART, C5, Firm, Tree)
- Decision Trees
- Decision Rules

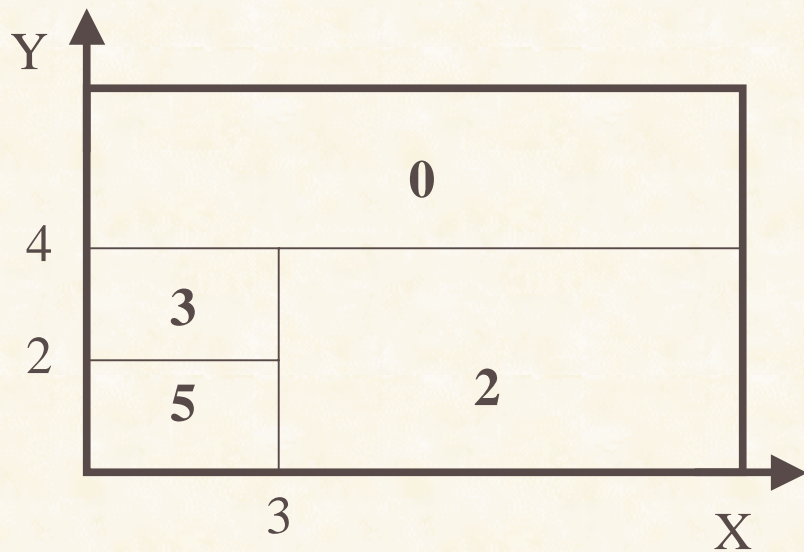


Tree methods: Dependent variable is numeric

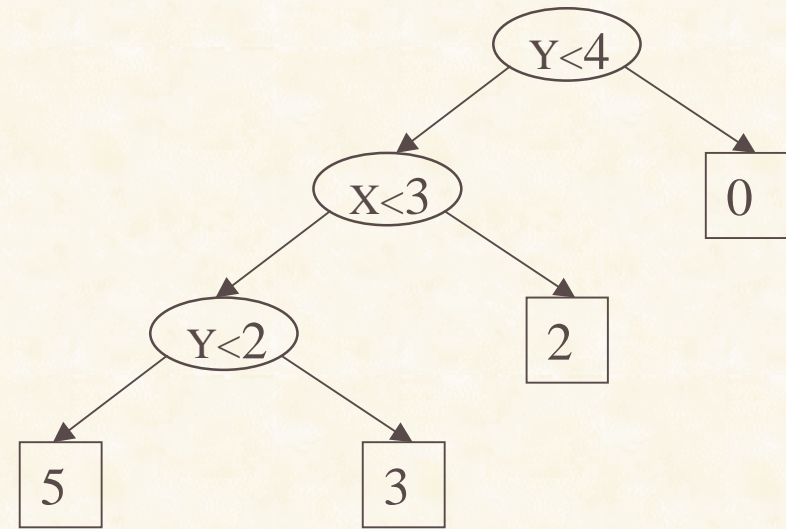
- Regression Trees

Trees

Function $f(X,Y)$



Tree form of $f(X,Y)$



Classification & Regression Trees

- Fit a tree model to data.
- Recursive Partitioning Algorithm.
- At each node we perform a split: we chose a variable X and a value t that minimizes a criteria.
- The split: $L = \{X < t\}$; $R = \{X \geq t\}$

Regression Tree for log(Sales)

```
HIP95 < 40.5 [Ave: 1.074, Effect: -0.76 ]
  HIP96 < 16.5 [Ave: 0.775, Effect: -0.298 ]
    RBEDS < 59 [Ave: 0.659, Effect: -0.117 ]
      HIP95 < 0.5 [Ave: 1.09, Effect: +0.431 ] -> 1.09
      HIP95 >= 0.5 [Ave: 0.551, Effect: -0.108 ]
        KNEE96 < 3.5 [Ave: 0.375, Effect: -0.175 ] -> 0.375
        KNEE96 >= 3.5 [Ave: 0.99, Effect: +0.439 ] -> 0.99
      RBEDS >= 59 [Ave: 1.948, Effect: +1.173 ] -> 1.948
    HIP96 >= 16.5 [Ave: 1.569, Effect: +0.495 ]
      FEMUR96 < 27.5 [Ave: 1.201, Effect: -0.368 ] -> 1.201
      FEMUR96 >= 27.5 [Ave: 1.784, Effect: +0.215 ] -> 1.784
  HIP95 >= 40.5 [Ave: 2.969, Effect: +1.136 ]
    KNEE95 < 77.5 [Ave: 2.493, Effect: -0.475 ]
      BEDS < 217.5 [Ave: 2.128, Effect: -0.365 ] -> 2.128
      BEDS >= 217.5 [Ave: 2.841, Effect: +0.348 ]
        OUTV < 53937.5 [Ave: 3.108, Effect: +0.267 ] -> 3.108
        OUTV >= 53937.5 [Ave: 2.438, Effect: -0.404 ] -> 2.438
    KNEE95 >= 77.5 [Ave: 3.625, Effect: +0.656 ]
      SIR < 9451 [Ave: 3.213, Effect: -0.412 ] -> 3.213
      SIR >= 9451 [Ave: 3.979, Effect: +0.354 ] -> 3.979
```

- For regression trees two criteria functions are:

$$\mathbf{h} = \frac{N_L \hat{\sigma}_L^2 + N_R \hat{\sigma}_R^2}{N_L + N_R}$$

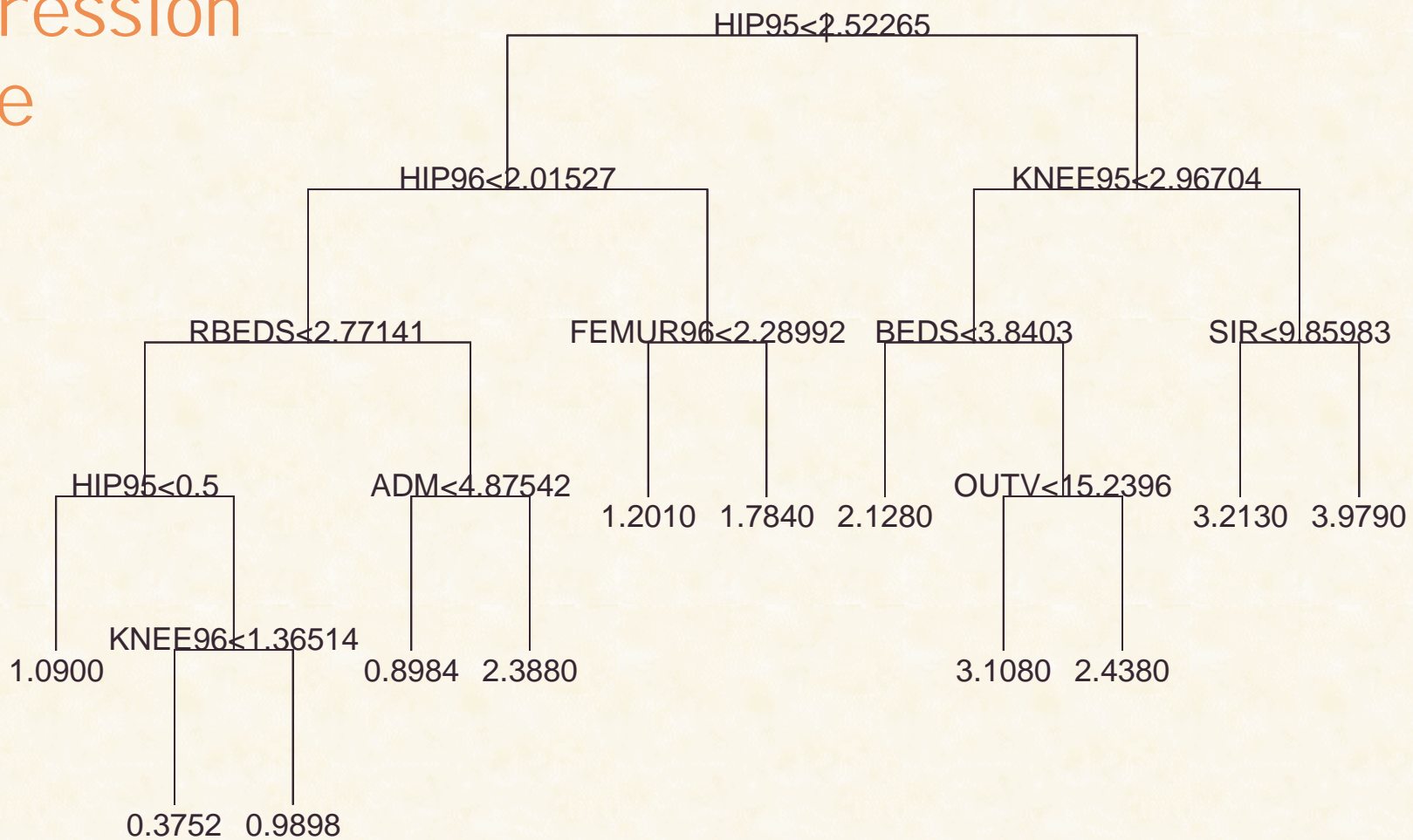
- For classification trees: criteria functions

$$\mathbf{h} = p_L \min(p_L^0, p_L^1) + p_R \min(p_R^0, p_R^1)$$

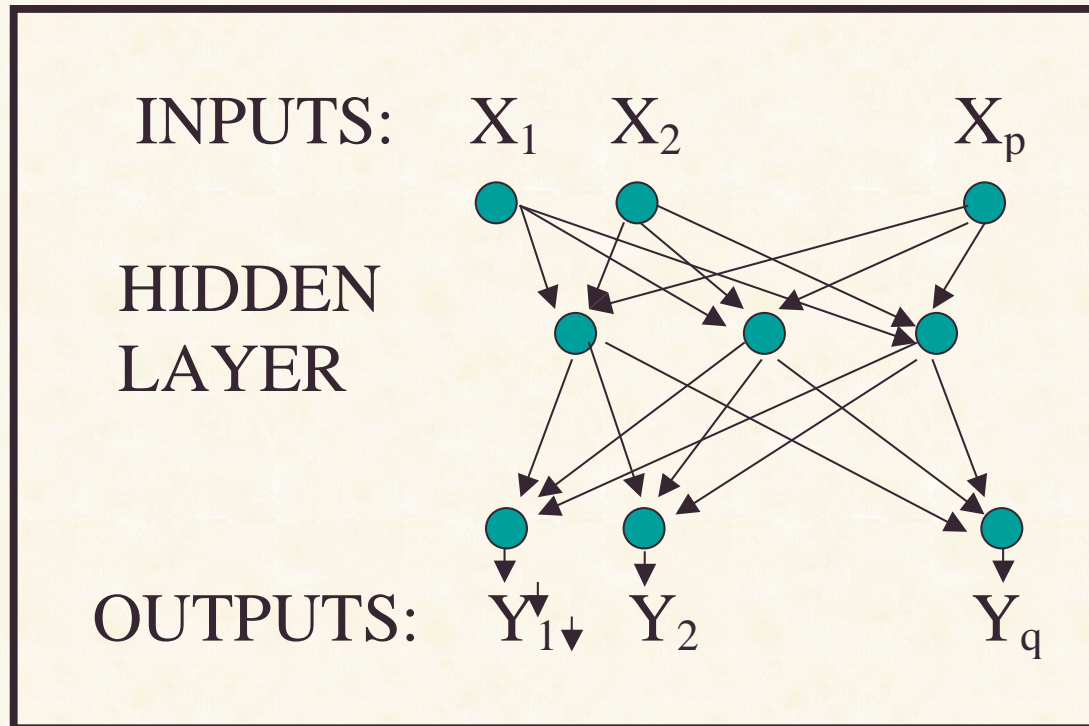
$$\mathbf{h} = p_L (-p_L^0 \log p_L^0 - p_L^1 \log p_L^1) + p_R (-p_R^0 \log p_R^0 - p_R^1 \log p_R^1) \text{ (C5)}$$

$$\mathbf{h} = p_L p_L^0 p_L^1 + p_R p_R^0 p_R^1 \text{ (CART)}$$

Regression Tree



Graph of an ANN with one hidden layer.



	<i>(i) 10 PC of 2308 genes.</i>		<i>(ii) 10 PC of 450 genes.</i>		<i>(iii) 10 cluster means of 50 genes.</i>		<i>(iv) 10 PC of 30 genes.</i>	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
2 Classifiers	18	14	0	2	0.5	1.5	0	3.5
3 Classifiers	8	14	0	1	0	0.5	0	1.5
4 Classifiers	0	3	0	0.5	0	0.5	0	1.5
10 Classifiers	0	8	0	0.5	0	0.5	0	1.5

Support Vector Machines(SVM)

Generalizations of the linear classifier methods (such as LDA)
Machine learning literature.

Linear Classification rule is of the form $r(x) = \text{Sign}(\beta'x - \beta_0)$.

Nonlinear: $r(x) = \text{Sign}(\sum_{i=1}^p \beta_i h(x_i, x) - \beta_0)$,

where the most popular forms for h are:

Radial basis functions: $h(x,y) = \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)$

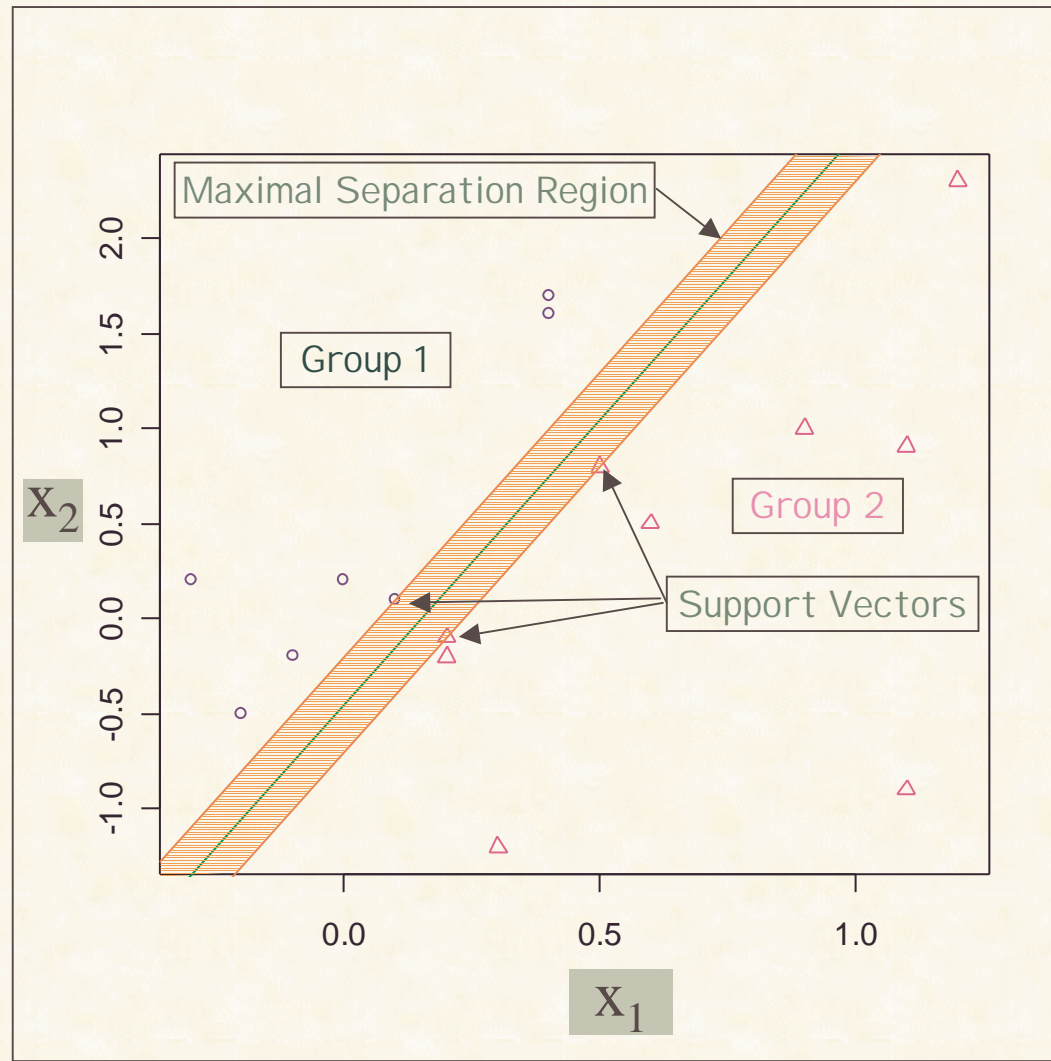
Sigmoidal functions: $h(x,y) = \tanh(\alpha_0 + \alpha_1 (x'y)^2)$

The solution of the estimation problem can be expressed as a function of a few of the samples that are called support vectors

.

Support Vector Machine:

The shaded area represents the separation region.
The arrows indicate the location of the support vectors.



Prediction Analysis for Microarrays PAM

$$d_j = \frac{\bar{x}_j - \bar{x}}{m_j(s + s_0)}$$

$$m_j = \sqrt{1/n_j + 1/n}$$

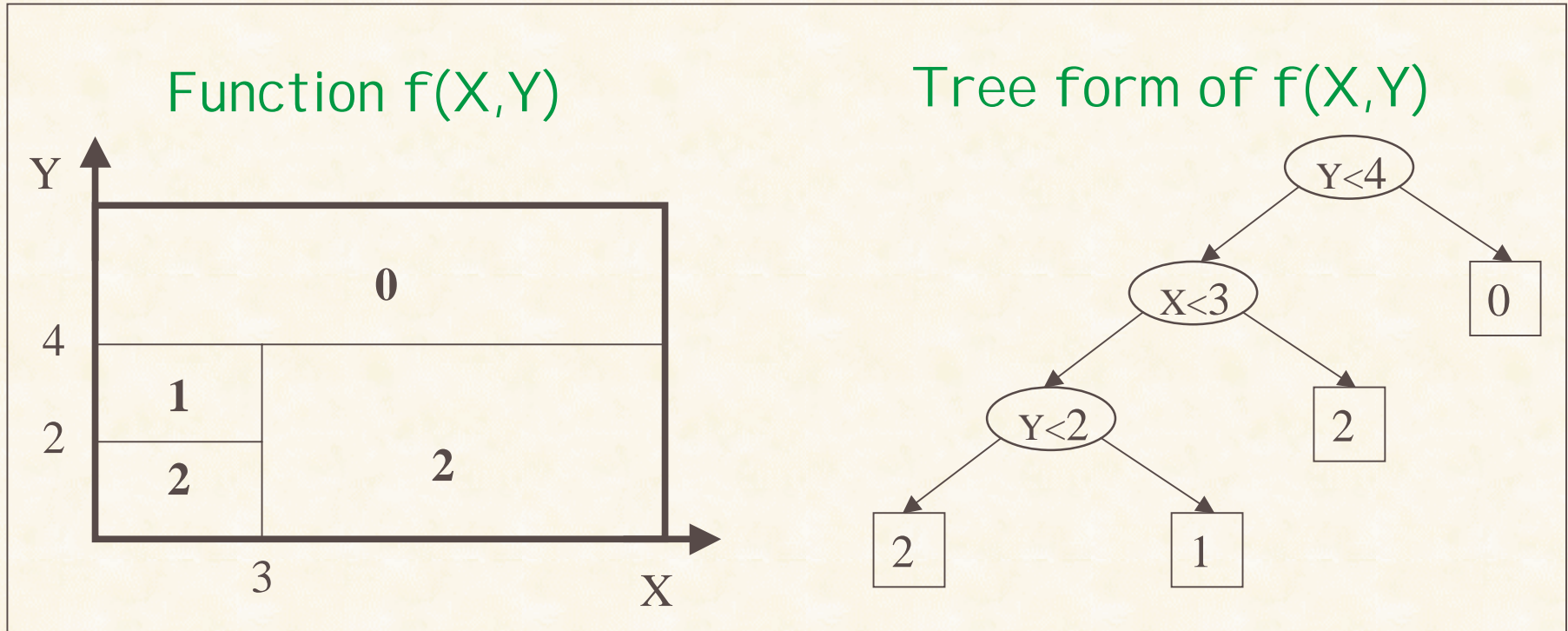
s_0 median of the components of s

$$\bar{x}_j^* = \bar{x} + m_j(s + s_0)d'_j$$

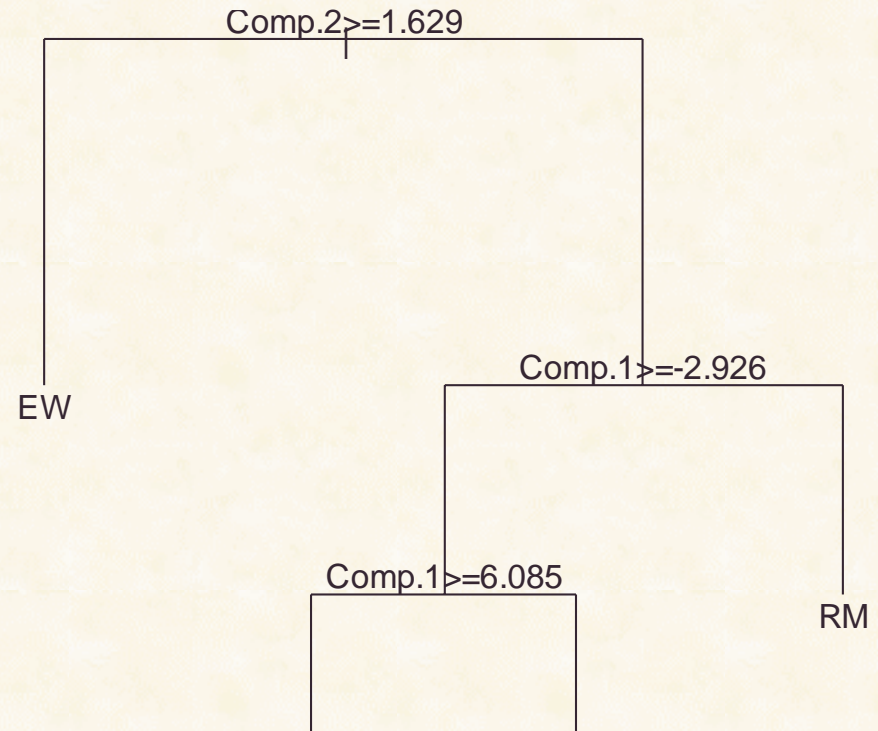
$$d'_j = \text{sign}(d_j)(|d_j| - \Delta)_+$$

The value of Δ is chosen according to the method of cross-validation

Classification trees



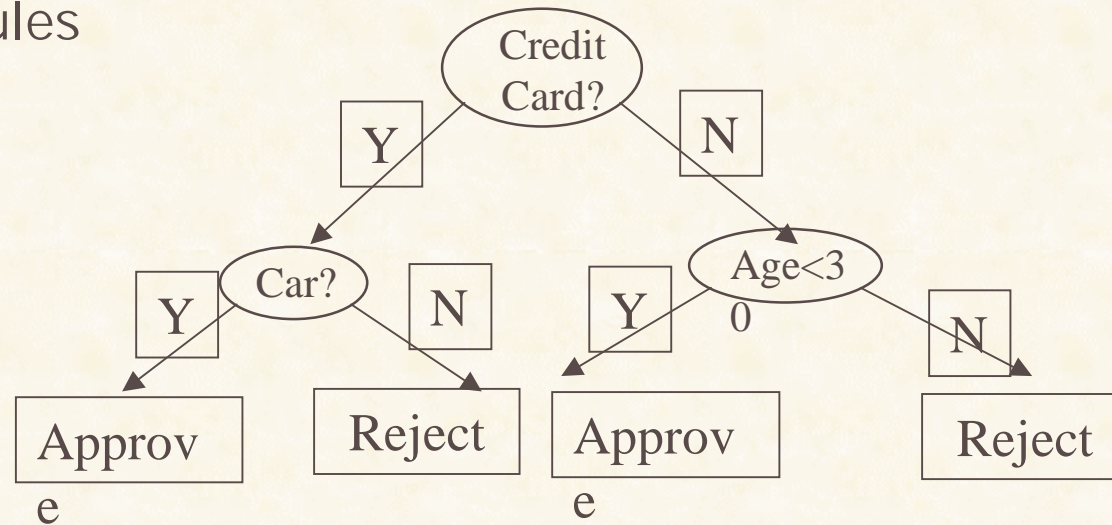
Classification tree for the cancer groups using 10 principal components of the top 100 cancer genes. The classification rule produces zero mistakes in the training set and five mistakes in the testing set.



	<i>(i) 10 PC of 2308 genes.</i>		<i>(ii) 10 PC of 450 genes.</i>		<i>(iii) 10 cluster means of 50 genes.</i>		<i>(iv) 10 PC of 30 genes.</i>	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
2 Classifiers	18	14	0	2	0.5	1.5	0	3.5
3 Classifiers	8	14	0	1	0	0.5	0	1.5
4 Classifiers	0	3	0	0.5	0	0.5	0	1.5
10 Classifiers	0	8	0	0.5	0	0.5	0	1.5

Tree methods: Dependent variable is categorical

- Classification trees (e.g., CART, C5, Firm, Tree)
- Decision Trees
- Decision Rules

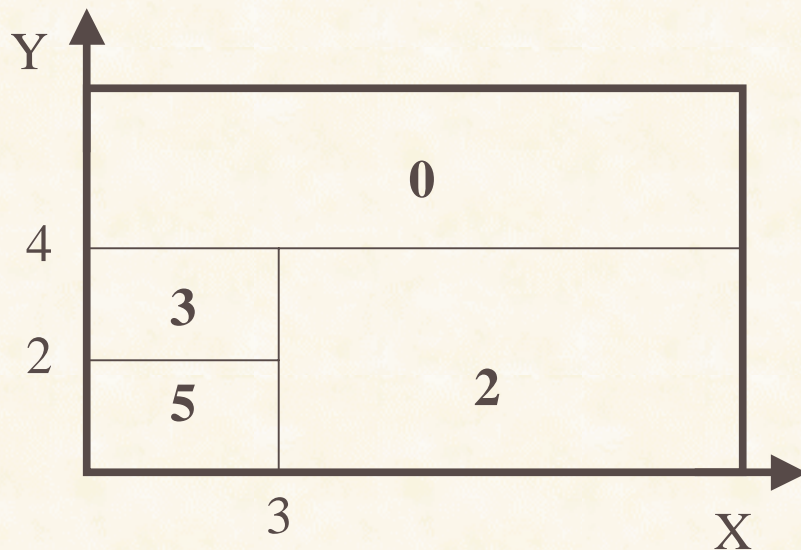


Tree methods: Dependent variable is numeric

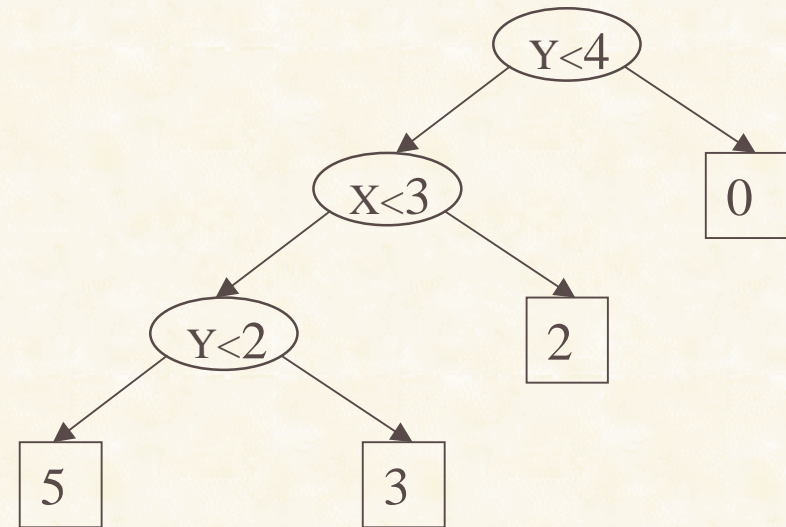
- Regression Trees

Trees

Function $f(X,Y)$



Tree form of $f(X,Y)$



Classification & Regression Trees

- Fit a tree model to data.
- Recursive Partitioning Algorithm.
- At each node we perform a split: we chose a variable X and a value t that minimizes a criteria.
- The split: $L = \{X < t\}$; $R = \{X \geq t\}$

Regression Tree for log(Sales)

```
HIP95 < 40.5 [Ave: 1.074, Effect: -0.76 ]
  HIP96 < 16.5 [Ave: 0.775, Effect: -0.298 ]
    RBEDS < 59 [Ave: 0.659, Effect: -0.117 ]
      HIP95 < 0.5 [Ave: 1.09, Effect: +0.431 ] -> 1.09
      HIP95 >= 0.5 [Ave: 0.551, Effect: -0.108 ]
        KNEE96 < 3.5 [Ave: 0.375, Effect: -0.175 ] -> 0.375
        KNEE96 >= 3.5 [Ave: 0.99, Effect: +0.439 ] -> 0.99
      RBEDS >= 59 [Ave: 1.948, Effect: +1.173 ] -> 1.948
    HIP96 >= 16.5 [Ave: 1.569, Effect: +0.495 ]
      FEMUR96 < 27.5 [Ave: 1.201, Effect: -0.368 ] -> 1.201
      FEMUR96 >= 27.5 [Ave: 1.784, Effect: +0.215 ] -> 1.784
  HIP95 >= 40.5 [Ave: 2.969, Effect: +1.136 ]
    KNEE95 < 77.5 [Ave: 2.493, Effect: -0.475 ]
      BEDS < 217.5 [Ave: 2.128, Effect: -0.365 ] -> 2.128
      BEDS >= 217.5 [Ave: 2.841, Effect: +0.348 ]
        OUTV < 53937.5 [Ave: 3.108, Effect: +0.267 ] -> 3.108
        OUTV >= 53937.5 [Ave: 2.438, Effect: -0.404 ] -> 2.438
    KNEE95 >= 77.5 [Ave: 3.625, Effect: +0.656 ]
      SIR < 9451 [Ave: 3.213, Effect: -0.412 ] -> 3.213
      SIR >= 9451 [Ave: 3.979, Effect: +0.354 ] -> 3.979
```

- For regression trees two criteria functions are:

$$\mathbf{h} = \frac{N_L \hat{\sigma}_L^2 + N_R \hat{\sigma}_R^2}{N_L + N_R}$$

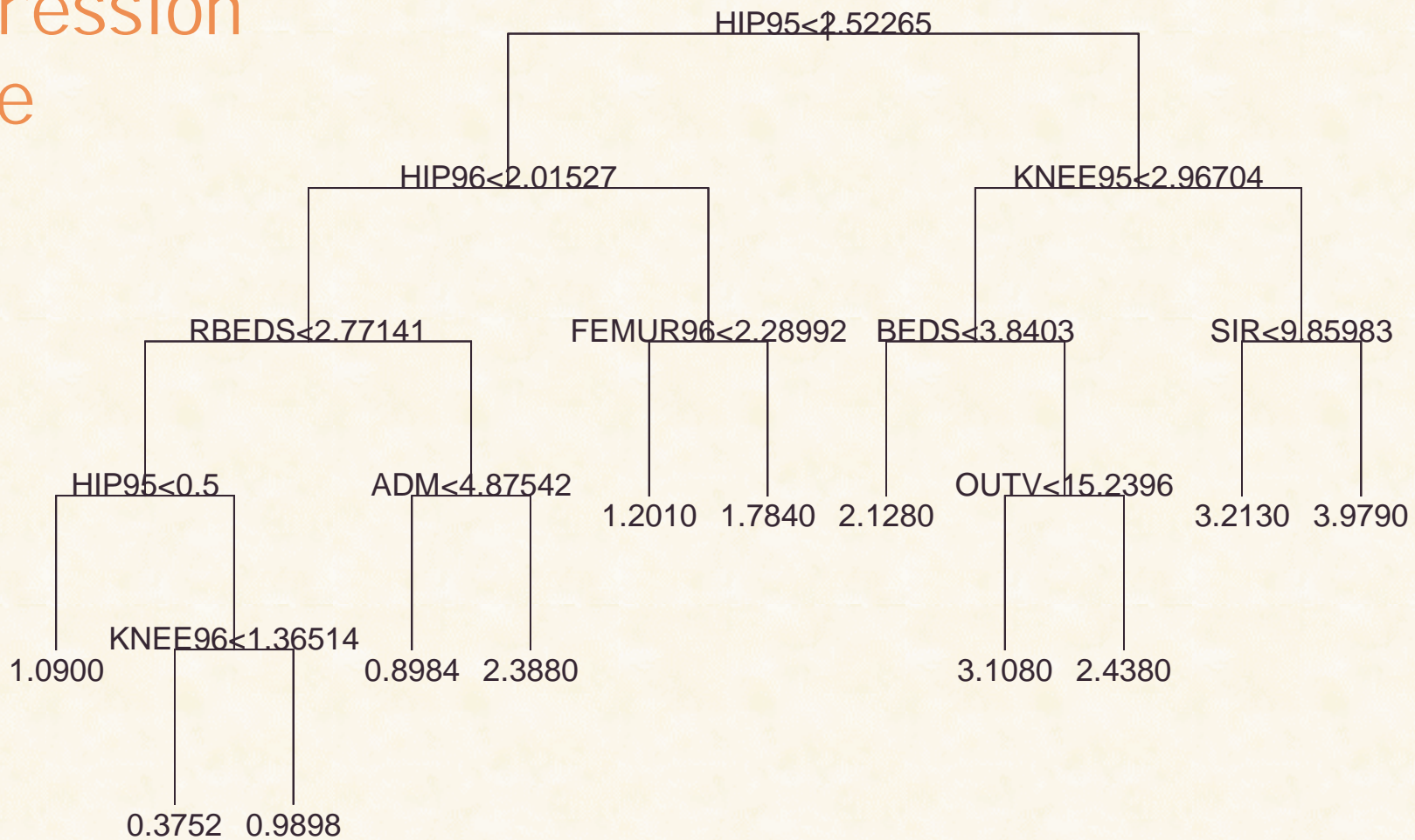
- For classification trees: criteria functions

$$\mathbf{h} = p_L \min(p_L^0, p_L^1) + p_R \min(p_R^0, p_R^1)$$

$$\mathbf{h} = p_L (-p_L^0 \log p_L^0 - p_L^1 \log p_L^1) + p_R (-p_R^0 \log p_R^0 - p_R^1 \log p_R^1) \text{ (C5)}$$

$$\mathbf{h} = p_L p_L^0 p_L^1 + p_R p_R^0 p_R^1 \text{ (CART)}$$

Regression Tree



Other methods:

Regularized Discriminant Analysis

Bayesian Discriminant Analysis

Flexible Discriminant Analysis.

