THE UNIVERSITY OF
SYDNEY

Lecture 2: Clustering

STAT5003

What is clustering?

- Methods of grouping samples (***x***) that are ***similar*** – according to some pre-defined criteria.

- A form of ***unsupervised learning*** – no label information (y) is used to tell the algorithm which observations should be grouped together.

- It is often used for ***exploratory data analysis*** – a way of looking for patterns or structure in the data that are of interest.

Basic principles of clustering

Aim: to group observations that are "similar" based on predefined criteria.

Issues:

- Data types - counts, ratio, ordinal, categorical and continuous.
- Missing data
- Scaling
- (Dis)similarity metric (a critical step in clustering):
  - Euclidean, Manhattan, Pearson correlation, Spearman correlation etc.

Algorithm:

- Hierarchical clustering
- k-means clustering
- Advanced:
  - Fuzzy c-means clustering, Semi-supervised clustering, bi-clustering

Commonly used (dis)similarity measures

- A metric is a measure of the similarity or dissimilarity between two data objects and it's used to form data points into clusters

- Two main classes of distance:
  - Correlation coefficients (compares shape of expression curves)

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i}^{n} y_i$$

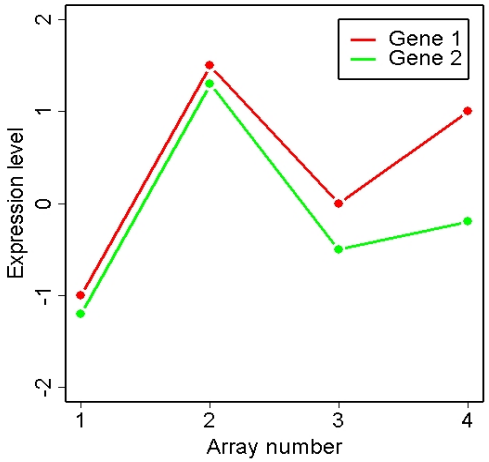
$$d_p = \frac{1-\rho(\mathbf{x},\mathbf{y})}{2}$$

  - Distance metrics
    - City Block (Manhattan) distance: $d(X,Y) = \sum_i |x_i - y_i|$
    - Euclidean distance: $d_{euc}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$

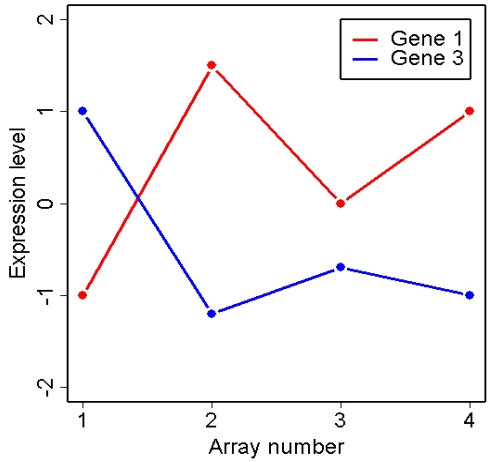
```
> X <- 1:3
> Y <- 1:3 + 1

> dist(cbind(X,Y))
         1        2
2 1.414214
3 2.828427 1.414214

> dist(cbind(X,Y), method="manhattan")
  1 2
2 2
3 4 2

> cor(cbind(X,Y))
  X Y
X 1 1
Y 1 1
```

Correlation (a measure between -1 and 1)

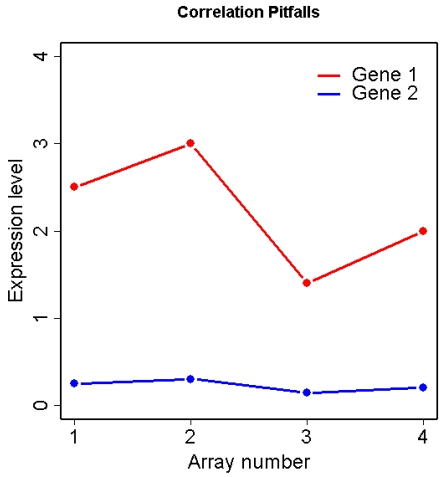
- Others include Spearman's and Kendall's correlation
- You can use absolute correlation to capture both positive and negative correlation



Positive correlation



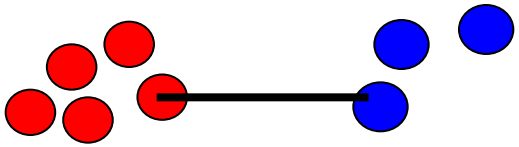
Negative correlation



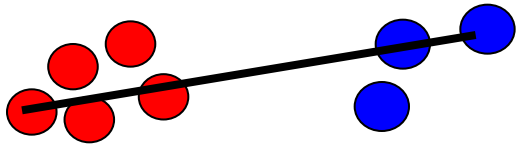
Potential pitfalls

Correlation = 1

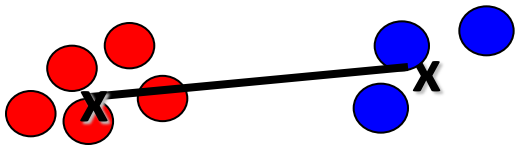
Distance between clusters
(between-cluster dissimilarity measures)



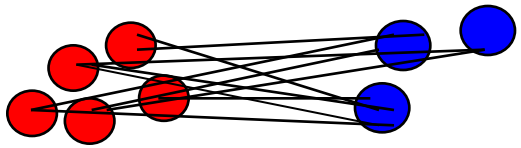
Single (minimum)



Complete (maximum)



Distance between centroids

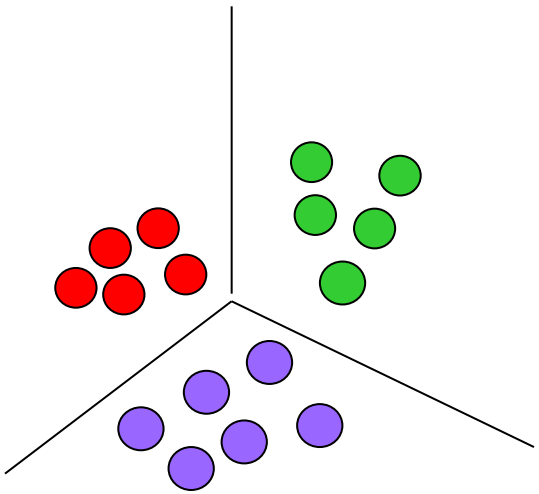


Average (Mean) linkage

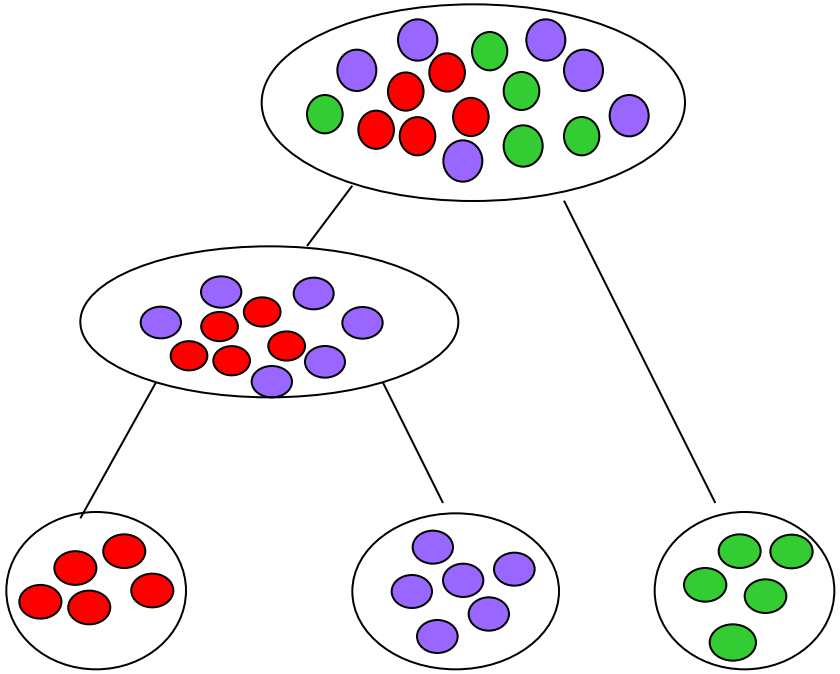
Clustering algorithms

- Clustering algorithm comes in 2 basic flavors

Partitioning



Hierarchical

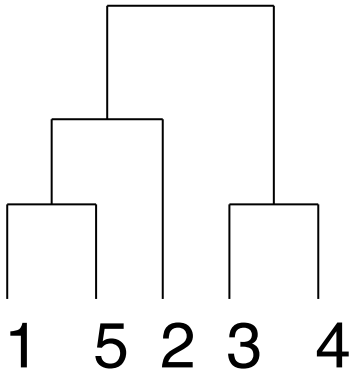


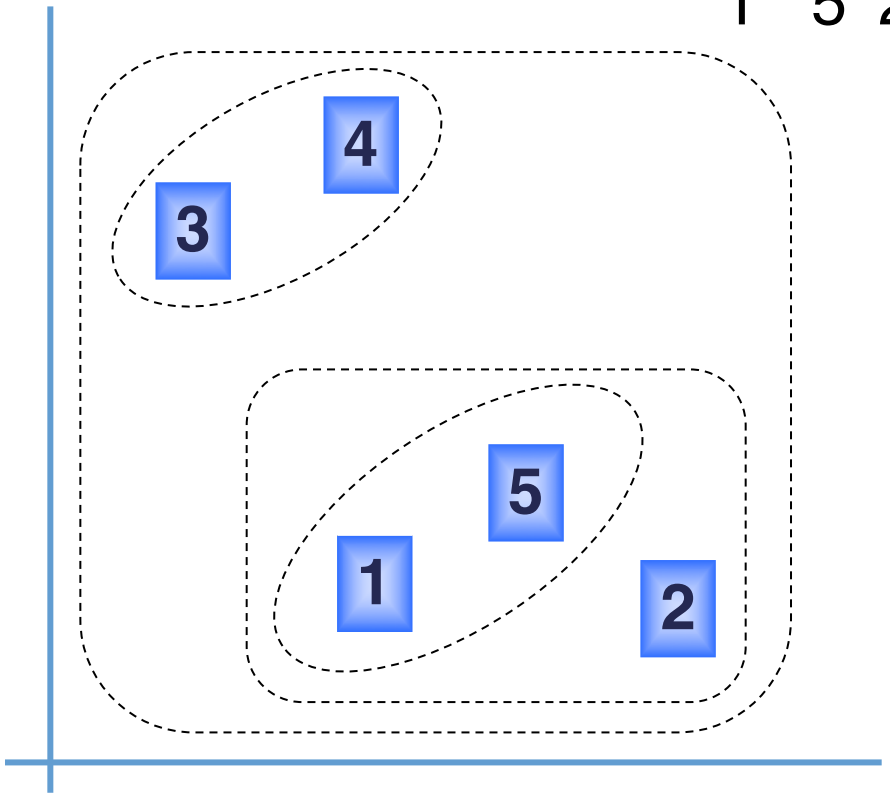
Hierarchical methods

- Hierarchical clustering methods produce a tree or dendrogram.

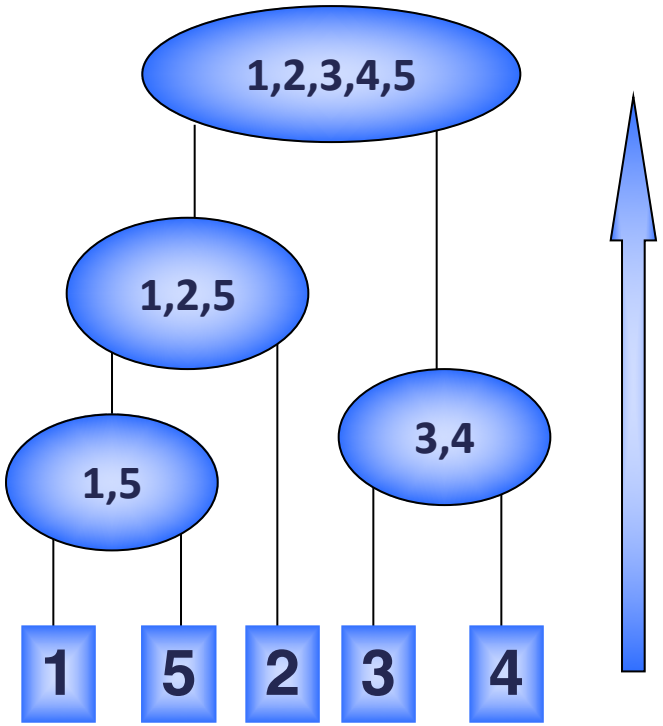
- They avoid specifying how many clusters are appropriate by providing a partition for each k obtained from cutting the tree at some level.

An illustration of hierarchical clustering





Bottom-up



Bottom-up tree building procedure

- Start with n sample (or m feature) clusters
- At each step, merge the two closest clusters using a measure of between-cluster dissimilarity which reflects the shape of the clusters
- The distance between clusters is defined by the method used (e.g., if complete linkage, the distance is defined as the distance between furthest pair of points in the two clusters)

?hclust

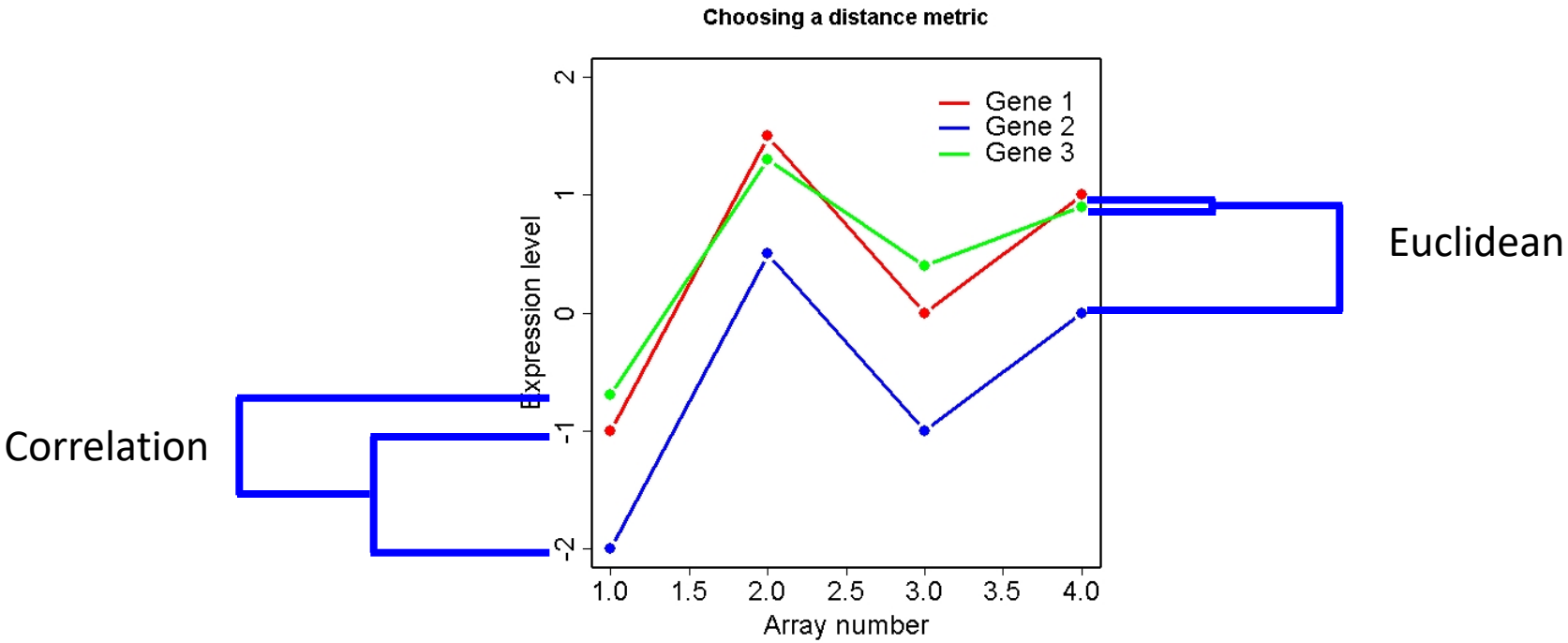
Violent Crime Rates by US State

```
?USArrests

par(mfrow=c(2,2))
hc <- hclust(dist(USArrests, method="euclidean"), method="ave")
plot(hc)
hc <- hclust(dist(USArrests, method="manhattan"), method="single")
plot(hc)
hc <- hclust(dist(USArrests), method="complete")
plot(hc)
```

Compare the trees using different agglomeration method.

Euclidean vs Correlation

- Euclidean distance
- Correlation



Demonstrate hierarchical clustering

Gene expression data

Gene expression data on p genes for n samples

samples (e.g. patients, cell types)

Genes	sample1	sample2	sample3	sample4	sample5	...
1	0.46	0.30	0.80	1.51	0.90	...
2	-0.10	0.49	0.24	0.06	0.46	...
3	0.15	0.74	0.04	0.10	0.20	...
4	-0.45	-1.03	-0.79	-0.56	-0.32	...
5	-0.06	1.06	1.35	1.09	-1.09	...

Gene expression level of gene i in sample j

Clustering in gene expression data analysis

- Clustering leads to readily interpretable figures and can be helpful for identifying patterns in time or space.

Examples:

- We can cluster cell samples (cols),

  e.g. 1) for identification (profiles). Here, we might want to estimate the number of different neuron cell types in a set of samples, based on gene expression.

  2) the identification of new / unknown tumor classes using gene expression profiles.

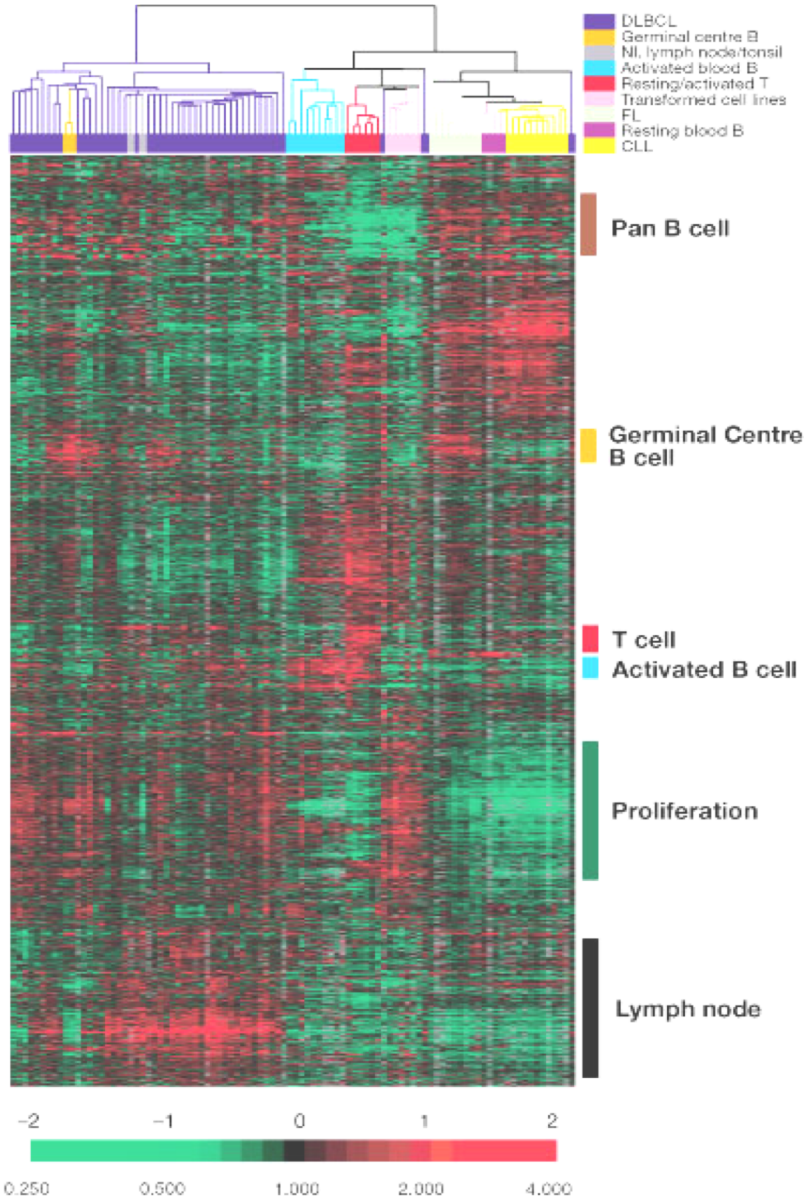
- We can cluster genes (rows) ,

  e.g. 1) using large numbers of yeast experiments, to identify groups of co-regulated genes.

  2) we can cluster genes to reduce redundancy (cf. variable selection) in predictive models.

Clustering both cells and genes

Taken from Nature February, 2000 Paper by Alizadeh et al. Distinct types of diffuse large B-cell lymphoma clustered by gene expression profiling



Partitioning methods

- Partition the data into a pre-specified number k of mutually exclusive and exhaustive groups.
- Iteratively reallocate the observations to clusters until some criterion is met, e.g. minimize within cluster sums of squares.
- Examples:
  - k-means, k-medoids, fuzzy c-means clustering, self-organizing maps (SOM), etc.;

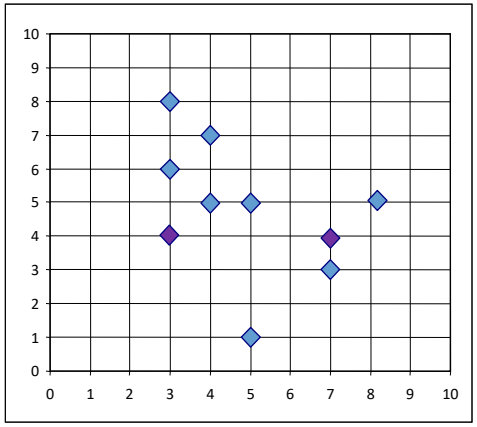
A typical k-means clustering algorithm

- Arbitrarily choose k objects as the initial cluster centers

- Until no change, do
  - (Re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster

  - Update the cluster means, i.e., calculate the mean value of the objects for each cluster

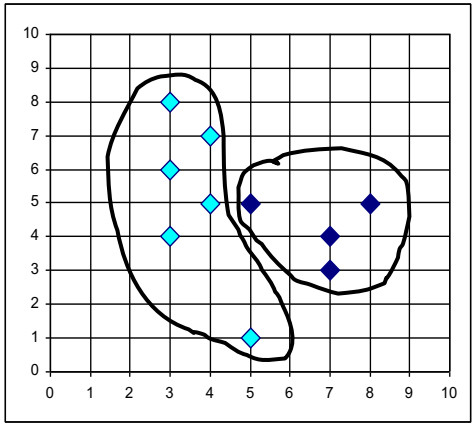
k-means: an example



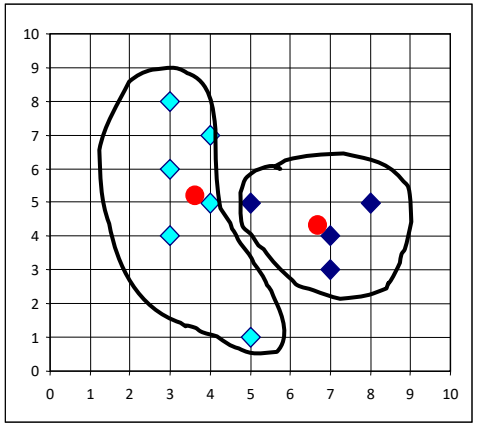
K=2

Arbitrarily choose K object as initial cluster center

Assign each objects to most similar center

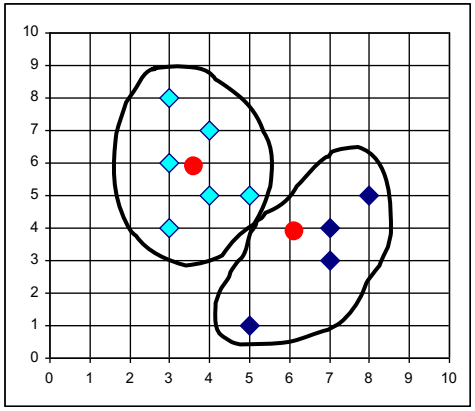


Update the cluster means

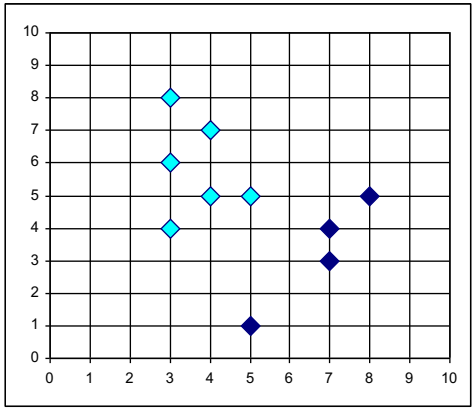


reassign

reassign



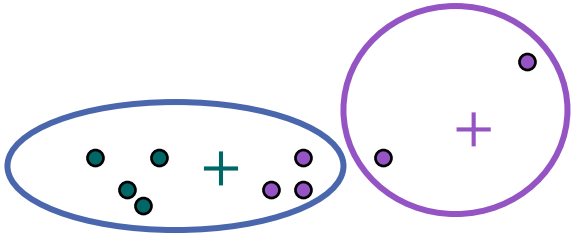
Update the cluster means



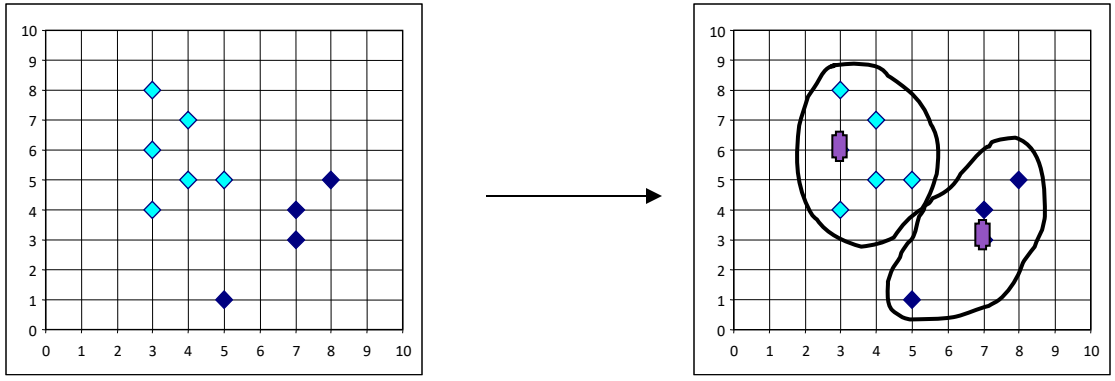
A problem of k-means clustering

- Sensitive to outliers

objects with extremely large values May substantially distort the distribution of the data



- K-medoids: the most centrally located object in a cluster



Demonstrating k-means clustering

Fuzzy c-means clustering

The fuzzy *c*-means algorithm is very similar to the *k*-means algorithm in that the the objective functions are virtually identical

However they differ in that *k*-means algorithm assigns each sample (black and white) to a single cluster whereas fuzzy c-means algorithm assigns each sample with a vector confidence to each of all clusters.

Fuzzy c-means clustering: the algorithm

1. Randomly initialise the membership matrix:

$$\mathrm{M}^{(0)} = \sum_{j=1}^{C} \mu_{ij}\,, \quad i = 1, 2 \ldots k$$

2. Calculate the Centroid as follows:

$$c_j = \frac{\sum_i [\mu_{ij}]^m \, x_i}{\sum_i [\mu_{ij}]^m}$$

3. Update $\mathrm{M}^{(t)}, \mathrm{M}^{(t+1)}$

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left|\left| x_i - c_j \right|\right|}{\left|\left| x_i - c_k \right|\right|} \right)^{\frac{2}{m-1}}}$$

4. If $\left|\left| \mathrm{M}^{(t+1)} - \mathrm{M}^{(t)} \right|\right| < \varepsilon$ then stop, otherwise return to step 2.

Partitioning vs. hierarchical clustering

Partitioning:

Advantages

- Optimal for certain criteria.
- Samples automatically assigned to clusters

Disadvantages

- Need initial *k;*
- Often require long computation times.
- All samples are forced into a cluster.

Hierarchical

Advantages

- Faster computation.
- Visual.

Disadvantages

- Unrelated objects are eventually joined
- Rigid, cannot correct later for erroneous decisions made earlier.
- Hard to define clusters.

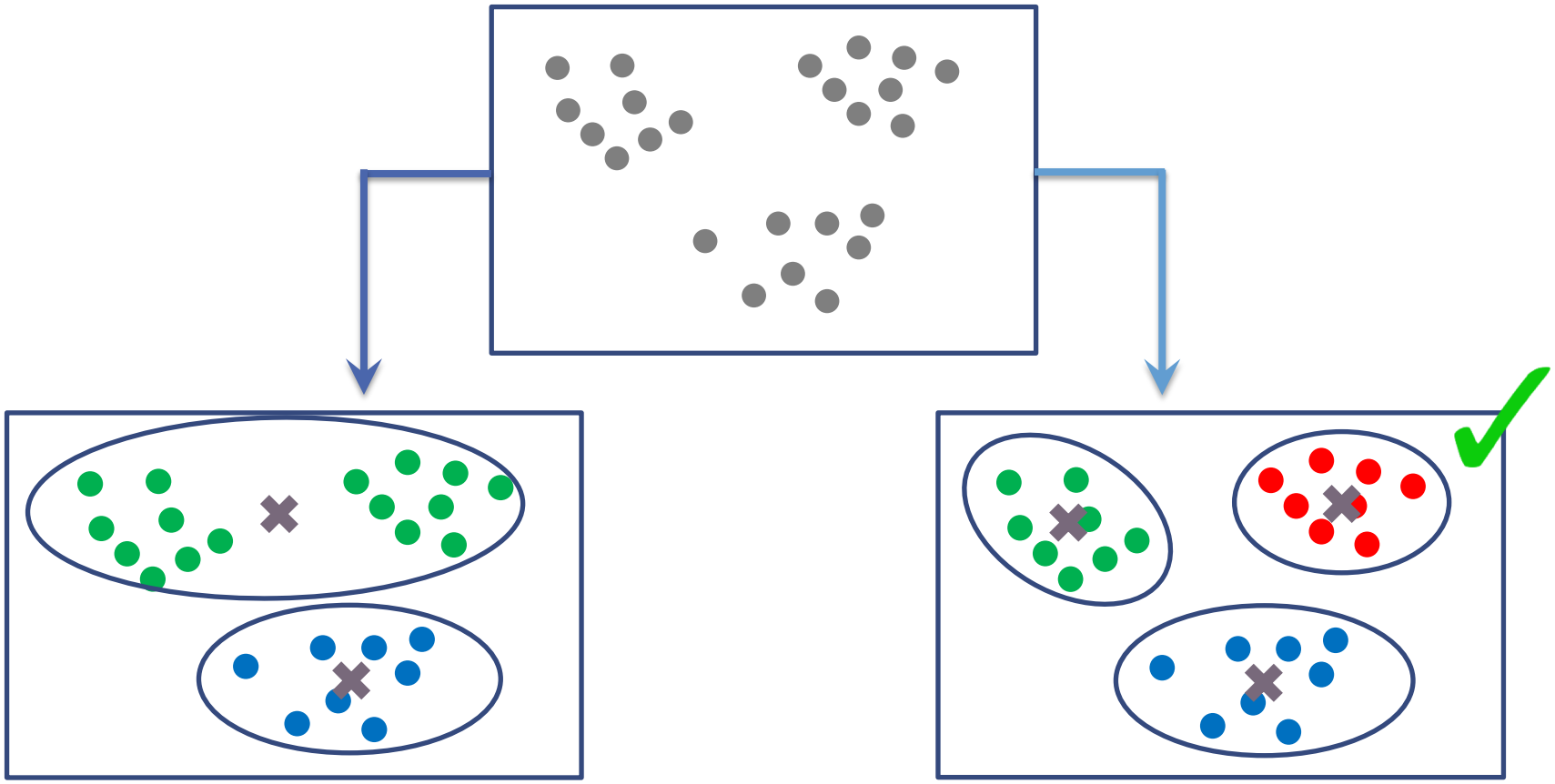
List of R cluster analysis packages

- **cclust**: convex clustering methods.
- **class**: self-organizing maps (**SOM**).
- **cluster**:
  - AGglomerative NESting (**agnes**),
  - Clustering LARe Applications (**clara**),
  - DIvisive ANAlysis (**diana**),
  - Fuzzy Analysis (**fanny**),
  - MONothetic Analysis (**mona**),
  - Partitioning Around Medoids (**pam**).
- **e1071**:
  - fuzzy C-means clustering (**cmeans**),
  - bagged clustering (**bclust**).
- **flexmix**: flexible mixture modeling.
- **fpc**: fixed point clusters, clusterwise regression and discriminant plots.
- **GeneSOM**: self-organizing maps.
- **mclust, mclust98**: model-based cluster analysis.
- **mva**:
  - hierarchical clustering (**hclust**),
  - k-means (**kmeans**).
- **gplots**: heatmap.2

Download from CRAN

Cluster validation

Data structure based metrics to determine optimal clustering



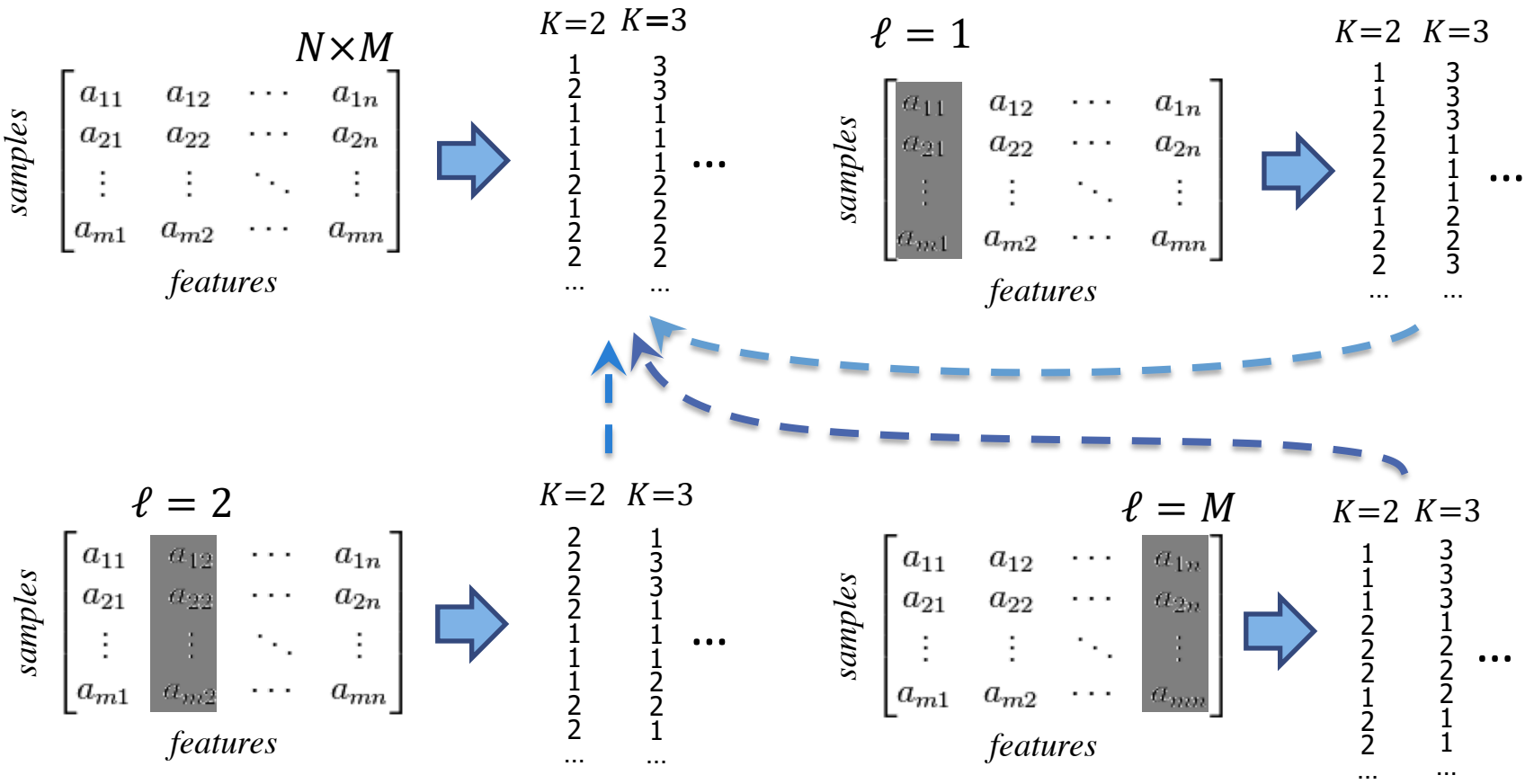
Compactness

1. **Small intra cluster distance**
2. **Large inter cluster distance**

$$intra = \frac{1}{N}\sum_{i=1}^{K}\sum_{x \in C_i} ||x - z_i||^2$$

$$inter = \min\left(||z_i - z_j||^2\right), i = 1,2,\dots,K-1; \quad j = i+1,\dots,K$$

Stability based metrics to determine optimal clustering



$$stability = \frac{1}{M}\sum_{\ell=1}^{M} d(\boldsymbol{C}_K^o, \boldsymbol{C}_K^\ell)$$

Incorporate external information for cluster validation

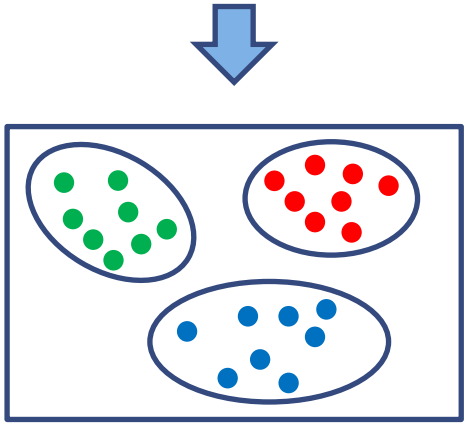
- External information could be utilised for cluster validation. For example, meaning if there are certain samples that you know should be clustered into the same cluster (or separate clusters), then such information could be used to guide the choice of “k”.

Example of ClueR

phosphorylation

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

time points



Kinase-substrate annotation

Contain kinases

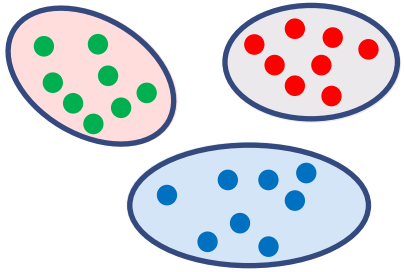
$j = 1, \ldots, m$

$$p_{ij} = \frac{\binom{a_{ij}+b_{ij}}{a_{ij}}\binom{c_{ij}+d_{ij}}{c_{ij}}}{\binom{a_{ij}+b_{ij}+c_{ij}+d_{ij}}{a_{ij}+c_{ij}}}$$

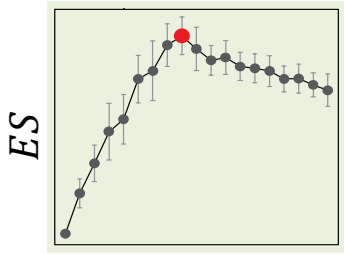
$$p(cl_i) = \min_{j=1\ldots m}(p_{ij})$$

for each cluster

$p(cl_1) = 0.03$ $p(cl_3) = 0.01$



$p(cl_2) = 0.06$



$K = 2 \ldots$

$$P_K = \mathrm{P}\left(\chi^2_d > -2\sum_{i=1}^{K}\log\left(p(cluster_i)\right)\right)$$

$$ES(K) = -log_{10}\,(P_K) - \alpha \times K$$

overall enrichment