CLUSTERING IN MICROARRAY EXPERIMENTS

- Introduction
- Dissimilarity (measures of)
- Missing Data and Imputation
- Clustering Methods
  - Partitioning methods
  - Hierarchical methods
- Examples

GOAL

- Identification of samples with similar expression level patterns, and genes that are similar across samples.
- Supervised learning task (association between clusters and an outcome; outcome is not used in construction of the clusters)

EXAMPLE

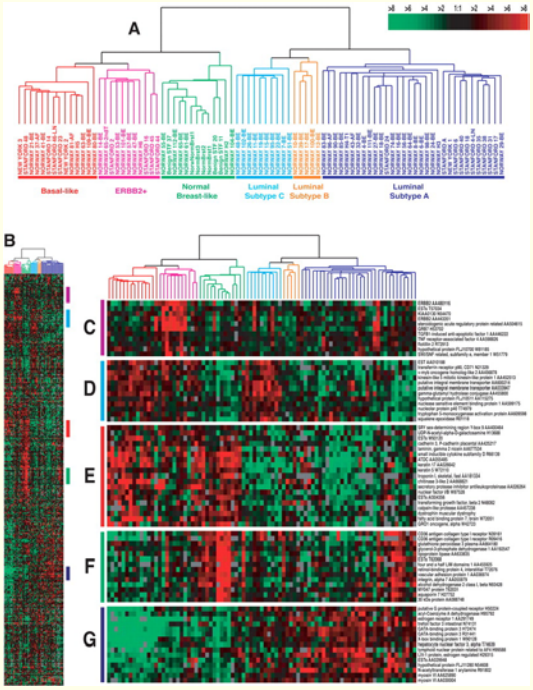
(Sørlie et al., 2001)

- Goal: Classification of breast carcer based on gene expression, and studying associations between tumor characteristics to clinical outcome.
- Data set:
  - cDNA microarrays; 456 clones selected out of 8,102 genes
  - 85 tissue samples (4 normal, 78 carcinomas, 3 fibroadenomas)
  - Heat map (next slide)

Heat map of the microarray data, with rows and columns arranged according to a hierarchical clustering method. Grey pixels represent missing data. Letters and corresponding colored bars represent groups of genes displayed in greater details on right panel.

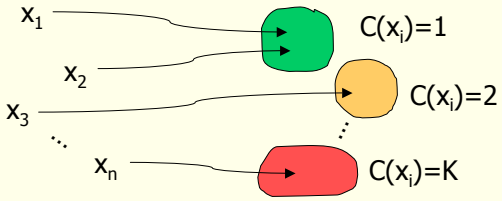
Legend:

- Basal-like
- ERBB2+
- Normal Brest-like
- Luminal Subtype C
- Luminal Subtype B
- Luminal Subtype A



DISSIMILARITY

➡ Cluster:
Group of objects (samples or genes) which are most similar



C(.): encoder function

➡ Common measures of dissimilarity:

① Euclidean distance: $\| x - y \| = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$

or

$$\| x - y \|^2 = \sum_{i=1}^{p} (x_i - y_i)^2$$

② Manhattan distance: $d_{Mn} = \sum_{i=1}^{p} | x_i - y_i |$

③ "1 - correlation" distance:

$$d_\rho = 1 - r_{xy} = 1 - \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{p}(x_i - \bar{x})^2\right]^{1/2}\left[\sum_{i=1}^{p}(y_i - \bar{y})^2\right]^{1/2}}$$

Note:

- $0 \le 1 - r_{xy} \le 2$
- Variations: { using $\bar{x} = \bar{y} = 0$; using $1 - |r_{xy}|$
- If $\tilde{x} = \frac{x - \bar{x}}{\sqrt{\sum (x_i - \bar{x})^2 / p}}$ and $\tilde{y} = \frac{y - \bar{y}}{\sqrt{\sum (y_i - \bar{y})^2 / p}}$

$$\Rightarrow \| \tilde{x} - \tilde{y} \|^2 = 2p(1 - r_{xy}) = 2pd_\rho$$

The location/scale invariance of d_ρ makes it a popular choice for microarray data.

➡ Missing Data and Imputation

- **Two alternatives:**
  - Clustering methods (or modification of) that can deal with missing values
  - Imputation as a pre step

Modification of methods:

① Calculate dissimilarities between two objects with pairwise deletion (if there is missing value)

② Adjust (if necessary) dissimilarity measures for the number of terms used (e.g. using a multiplier p/c, where c is the number of complete pairs)

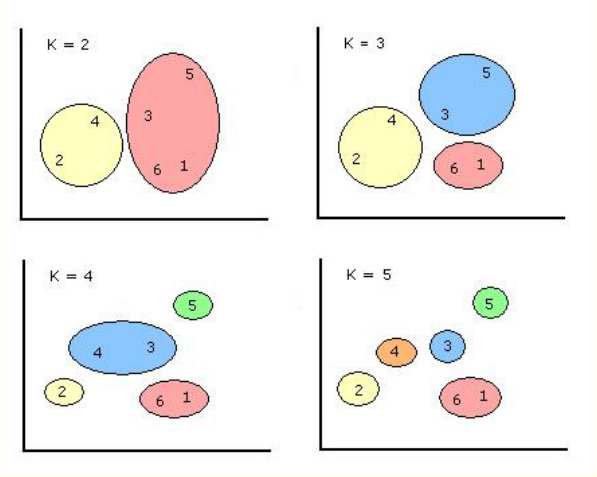
Imputation:

- Replacement of missing values with the mean level for that gene, or
- Predict missing values for each variable using a regression approach with (all) other variables as predictors, or
- Iterative approach:

missing data imputation → Singular Value Decomposition (SVD)

➡ Clustering Methods

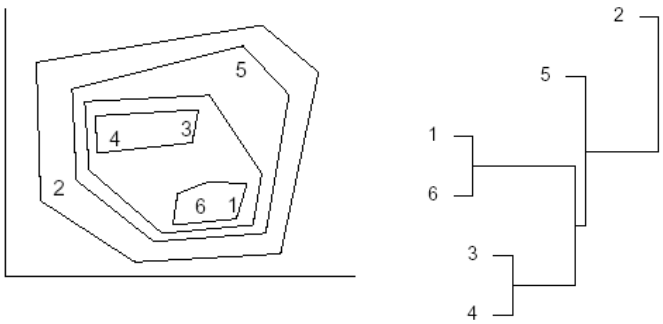
Partitioning methods: Division of objects into a fixed number of clusters



K-means clustering example, with K = 2 to 5 clusters.
Adapted from Chipman et al. (2003)

➡ Clustering Methods

Hierarchical methods: Nested sequence of clusters; Dendogram



Hierarchical clustering example. Points in two-dimensional space are illustrated on the left, and the corresponding hierarchical clustering is depicted on the right. Source: Chipman et al. (2003)

➡ Partitioning Methods

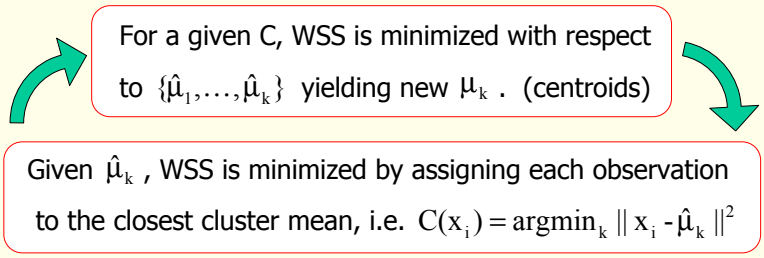
- Seek to minimize some measure of within-group dissimilarity for a fixed number (K) of groups.

① K-means

- Most popular partitioning method
- Minimization of the sum of squared (Euclidean) distances:

$$WSS = \sum_{k=1}^{K} \sum_{C(x_i)=k} \| x_i - \hat{\mu}_k \|^2 \qquad \text{(cluster variance)}$$

- Algorithm:

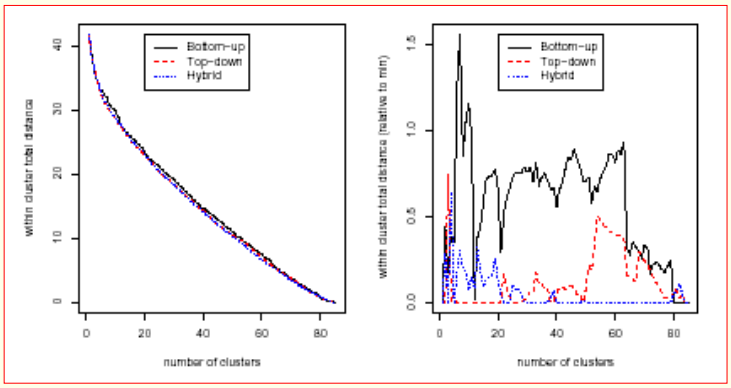


① K-means (cont'ed)

- K-means algorithm is fast: don't need to evaluate all n(n-1)/2 pairwise dissimilarities. At each iteration, K × n dissimilarities are evaluated and K centroids updated.
- Different solutions will be achieved for different starting values; it is goo practice to use multiple runs of the algorithm.
- Using Euclidean distances instead of squared Euclidean distances may give more robustness to outliers.
- There is a c connection between K-means (squared distances) and multivariate Gaussian mixtures (ML with spherical covariance structure).
  Advantage of mixture models: probability models can be used in criteria to choose K.

☞ Choosing K (number of clusters)

- Typically, use different K and compare results (loss measure).



Comparison of within-group dissimilarities (left) for various clustering methods, as functions of cluster size. The right plot gives difference from the minimum. (Chipman et al., 2003)

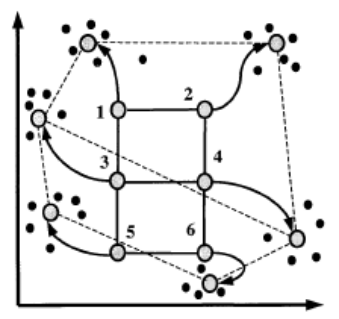
② K-medoids

- Instead of using average of points within each cluster (K-means), centroids are taken to be representative objects in each group.
- Indicated mostly for categorical or discrete variables.
- Robust to outliers.

③ Self-Organizing Maps (SOMs)

- Partitioning algorithms that are constrained so that clusters may be represented in a regular, low-dimensional structure, such as a grid.

Principle of SOMs. Initial geometry of nodes in 3 x 2 rectangular grid is indicated by solid lines connecting the nodes. Hypothetical trajectories of nodes as they migrate to fit the data during successive iterations of SOM algorithm is shown. Data points are represented by black dots, six nodes of SOM by large circles, and trajectories by arrows. (Tamayo et al., 1999)



⇨ Algorithm: SOM clustering (two-dimensional rectangular grid)

- Select the number of rows (q_1) and columns (q_2) in the grid. Then there will be $K = q_1 q_2$ clusters.
- Initialize step size $\alpha = 1$ and grid radius $r = 2$ (for example).
- Initialize prototype vectors M_j, $j \in (1, \ldots, q_1) \times (1, \ldots, q_2)$ by assigning points to prototypes. This assignment might be based on a partitioning of the data after projection onto principal components.
- Loop over the entire dataset
  - Loop over each data point x_i

    (a) Identify the index vector j^* of the prototype M_j nearest to x_i.

    (b) Identify a set S of neighboring prototypes of M_{j^*}, i.e.,

    $$S = \{j : \text{distance}(j, j^*) < r\}.$$

    The distance might be Euclidean or some other metric.

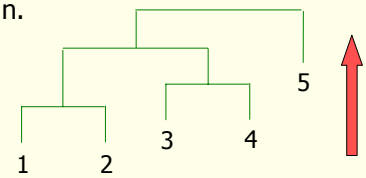
    (c) Update each element of S by moving the corresponding prototype toward x_i: $M_j \leftarrow M_j + \alpha(x_i - M_j) \quad \text{any } j \in S$
  - Decrease the values of α and r by some predetermined amount. Typically $\alpha = 0$, $r = 1$ upon completion of the outer loop.

➡ Hierarchical Methods

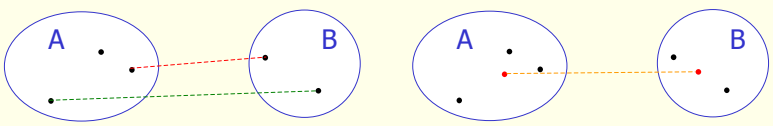
- Tree representation: dendogram; different levels of detail may be explored.

① Bottom-up Methods

- Start with each object representing one cluster of size 1.
- At each step, the closest two clusters are joined until all objects are in a single cluster of size n.

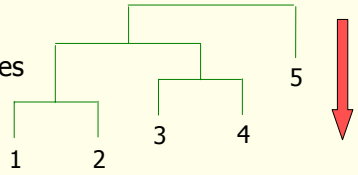


- "Closeness": *single linkage* (minimum distance) *complete linkage* (maximum distance), *mean distance* and *distance between centroids*.



② Top-down Methods

- Especially applicable if interest focuses on identifying a few clusters.

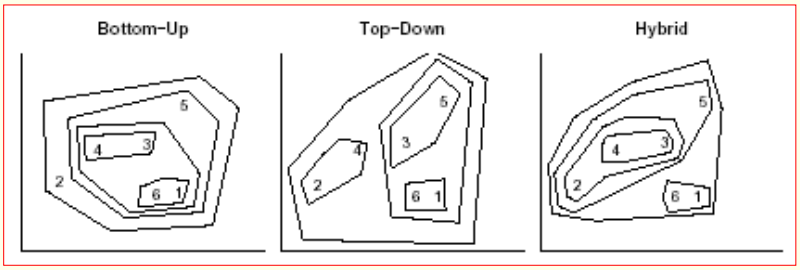


Tree Structural Vector Quantization

- Recursevely partition of the data using 2-means algorithm.

Macnaughton-Smith algorithm

- Select point with greatest mean dissimilarity to all other points to form a "splinter group".
- Swap objects that are closer to the splinter group.



Simple example to illustrate different clustering methods. Six points in two dimensions are hierarchically clustered using three different methods. Polygons are used to indicate nested clusters. (Chipman and Tibishirani ,2003)

➡ Two-way Clustering

- Previous methods cluster samples independently of the genes and vice versa.
- Typically these two operations are both used.
- Two-way clustering methods: use both samples and genes simultaneously to extract joint information about both of them.
- Methods are not yet well-developed, and are not yet in widespread use.
  - ① Coupled two-way clustering
  - ② Block clustering
  - ③ Plaid models

REFERENCES

Sørlie T. et al (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS 98(19), 10869-10874. Supplemental information online at http://genome-www.stanford.edu/.

Chipman, H., E. George, and R. McCulloch (1998) Making sense of a forest of tree, in Proc. 30th Symposium on the Interface, S. Weisberg, Ed., Interface Foundation of North America, Fairfax Station, VA 84-92.

Tamayo, P. et al. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. PNAS 96, 2907-2912.

Chipman, H., T. J. Hastie and R. Tibshirani (2003) Clustering microarray data, in Statistical Analysis of Gene Expression Microarray Data, T. Speed (Ed.), Chapman & Hall/CRC, Boca Raton, FL, p. 159-200.

Dudoit, S. and J. Fridlyand (2003) Classification in microarray experiments, in Statistical Analysis of Gene Expression Microarray Data, T. Speed (Ed.), Chapman & Hall/CRC, Boca Raton, FL, p. 93-158.