Cluster Analysis of Microarray Data

4/13/2009

Clustering

- Group *objects* that are *similar* to one another together in a *cluster*.
- Separate *objects* that are *dissimilar* from each other into different *clusters*.
- The *similarity* or *dissimilarity* of two *objects* is determined by comparing the *objects* with respect to one or more *attributes* that can be measured for each *object*.

Data for Clustering

	attribute				
object	1	2	3	...	m
1	4.7	3.8	5.9	...	1.3
2	5.2	6.9	3.8	...	2.9
3	5.8	4.2	3.9	...	4.4
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n	6.3	1.6	4.7	...	2.0

Microarray Data for Clustering

		attribute (time points)				
	object	1	2	3	...	m
genes	1	4.7	3.8	5.9	...	1.3
	2	5.2	6.9	3.8	...	2.9
	3	5.8	4.2	3.9	...	4.4
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	n	6.3	1.6	4.7	...	2.0

estimated expression levels

Microarray Data for Clustering

		attribute (tissue types)				
	object	1	2	3	...	m
genes	1	4.7	3.8	5.9	...	1.3
	2	5.2	6.9	3.8	...	2.9
	3	5.8	4.2	3.9	...	4.4
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	n	6.3	1.6	4.7	...	2.0

estimated expression levels

Microarray Data for Clustering

		attribute (treatment conditions)				
	object	1	2	3	...	m
genes	1	4.7	3.8	5.9	...	1.3
	2	5.2	6.9	3.8	...	2.9
	3	5.8	4.2	3.9	...	4.4
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	n	6.3	1.6	4.7	...	2.0

estimated expression levels

Microarray Data for Clustering

object	attribute 1	2	3	...	m (genes)
1	4.7	3.8	5.9	...	1.3
2	5.2	6.9	3.8	...	2.9
3	5.8	4.2	3.9	...	4.4
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n (samples)	6.3	1.6	4.7	...	2.0

estimated expression levels

7

Clustering: An Example Experiment

- Researchers were interested in studying gene expression patterns in developing soybean seeds.
- Seeds were harvested from soybean plants at 25, 30, 40, 45, and 50 days after flowering (daf).
- One RNA sample was obtained for each level of daf.

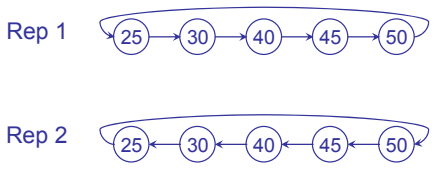
8

An Example Experiment (continued)

- Each of the 5 samples was measured on two two-color cDNA microarray slides using a loop design.
- The entire process we repeated on a second occasion to obtain a total of two independent biological replications.

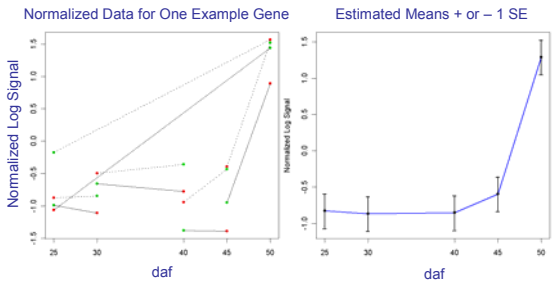
9

Diagram Illustrating the Experimental Design



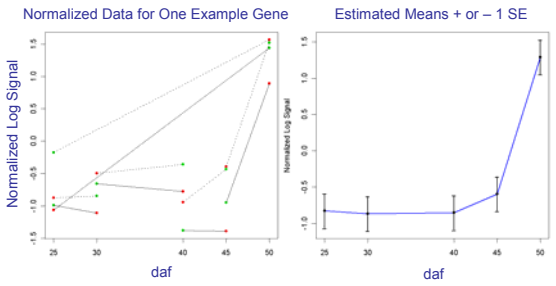
10

The daf means estimated for each gene from a mixed linear model analysis provide a useful summary of the data for cluster analysis.



11

400 genes exhibited significant evidence of differential expression across time (*p*-value<0.01, FDR=3.2%). We will focus on clustering their estimated mean profiles.



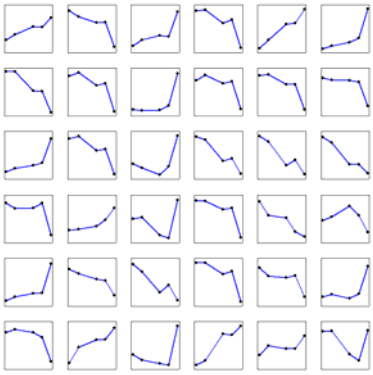
12

We build clusters based on the most significant genes rather than on all genes because...

- Much of the variation in expression is noise rather than biological signal, and we would rather not build clusters on the basis of noise.
- Some clustering algorithms will become computationally expensive if there are a large number of objects (gene expression profiles in this case) to cluster.

13

Estimated Mean Profiles for Top 36 Genes



14

Dissimilarity Measures

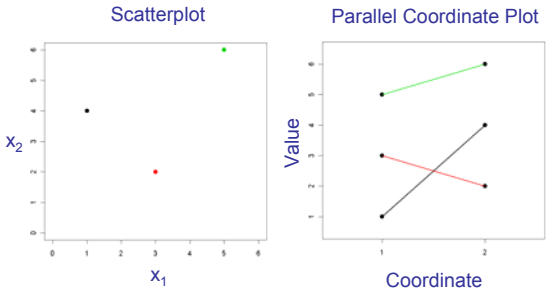
- When clustering objects, we try to put *similar* objects in the same cluster and *dissimilar* objects in different clusters.
- We must define what we mean by *dissimilar*.
- There are many choices.
- Let x and y denote m dimensional objects:

$x=(x_1, x_2, ..., x_m)$ $y=(y_1,y_2, ..., y_m)$

e.g., estimated means at m=5 five time points for a given gene.

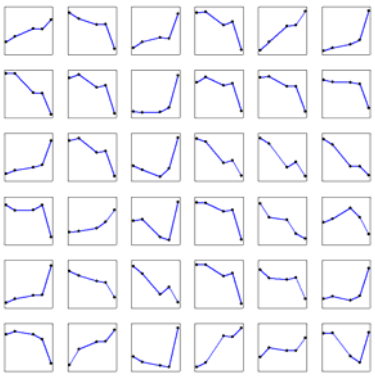
15

Parallel Coordinate Plots



16

These are parallel coordinate plots that each show one point in 5-dimensional space.



17

Euclidean Distance

$$d_E(x,y) = || x - y || = \sqrt{\sum_{j=1}^{m} (x_j - y_j)^2}$$

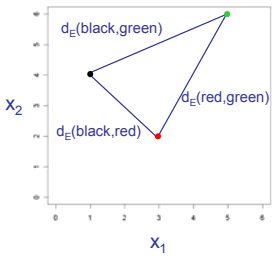
1-Correlation

$$d_{cor}(x,y) = 1 - r_{xy} = 1 - \frac{\sum_{j=1}^{m} (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{j=1}^{m} (x_j - \bar{x})^2} \sqrt{\sum_{j=1}^{m} (y_j - \bar{y})^2}}$$

18

Euclidean Distance

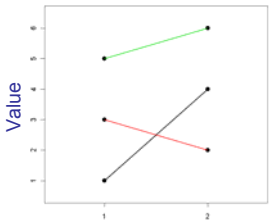
Scatterplot



19

1-Correlation Dissimilarity

Parallel Coordinate Plot



The black and green objects are close together and far from the red object.

20

Relationship between Euclidean Distance and 1-Correlation Dissimilarity

$$\text{Let } \tilde{x}_j = \frac{x_j - \bar{x}}{s_x} \quad \text{and let } \tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m).$$

$$\text{Let } \tilde{y}_j = \frac{y_j - \bar{y}}{s_y} \quad \text{and let } \tilde{y} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_m).$$

$$\| \tilde{x} - \tilde{y} \| = \sqrt{\sum_{j=1}^{m} (\tilde{x}_j - \tilde{y}_j)^2} = \sqrt{\sum_{j=1}^{m} (\tilde{x}_j^2 + \tilde{y}_j^2 - 2\tilde{x}_j\tilde{y}_j)}$$

$$= \sqrt{\sum_{j=1}^{m} \tilde{x}_j^2 + \sum_{j=1}^{m} \tilde{y}_j^2 - 2\sum_{j=1}^{m} \tilde{x}_j\tilde{y}_j}$$

$$= \sqrt{2(m-1)}\sqrt{1 - r_{xy}}$$

21

Thus Euclidean distance for standardized objects is proportional to the square root of the 1-correlation dissimilarity.

- We will standardize our mean profiles so that each profile has mean 0 and standard deviation 1 (i.e., we will convert each x to $\tilde{x}$).
- We will cluster based on the Euclidean distance between standardized profiles.
- Original mean profiles with similar patterns are "close" to one another using this approach.

22

Clustering methods are often divided into two main groups.

1. Partitioning methods that attempt to optimally separate n objects into K clusters.
2. Hierarchical methods that produce a nested sequence of clusters.

23

Some Partitioning Methods

1. K-Means
2. K-Medoids
3. Self-Organizing Maps (SOM)

   (Kohonen, 1990; Tomayo, P. et al., 1998)

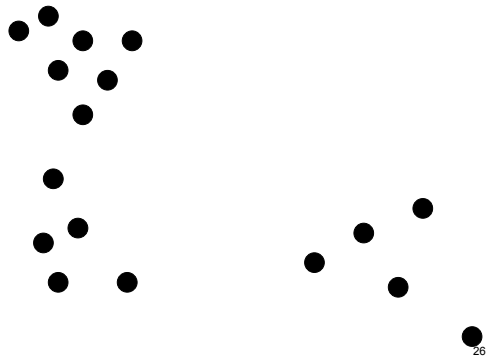
24

K Medoids Clustering

0. Choose K of the n objects to represent K cluster centers (a.k.a., *medoids*).
1. Given a current set of K medoids, assign each object to the nearest medoid to produce an assignment of objects to K clusters.
2. For a given assignment of objects to K clusters, find the new medoid for each cluster by finding the object in the cluster that is the closest on average to all other objects in its cluster.
3. Repeat steps 1 and 2 until the cluster assignments do not change.

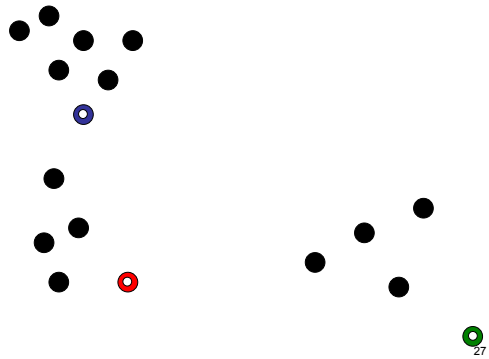
25

Example of K Medoids Clustering



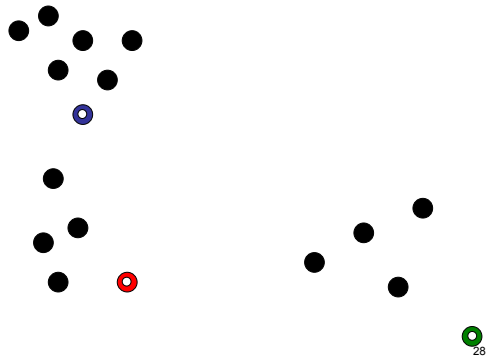
26

Start with K Medoids



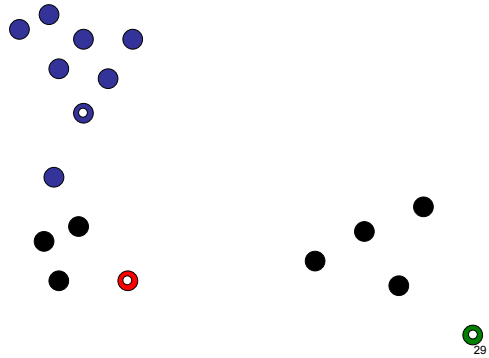
27

Assign Each Point to Closest Medoid



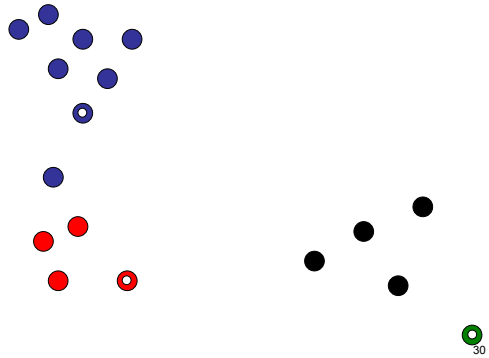
28

Assign Each Point to Closest Medoid



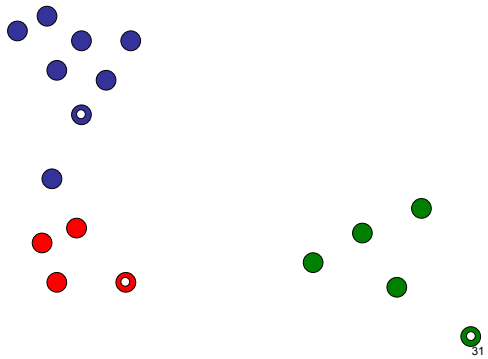
29

Assign Each Point to Closest Medoid



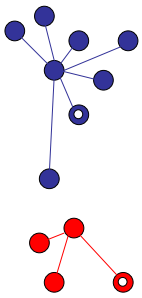
30

Assign Each Point to Closest Medoid

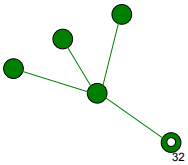


31

Find New Medoid for Each Cluster

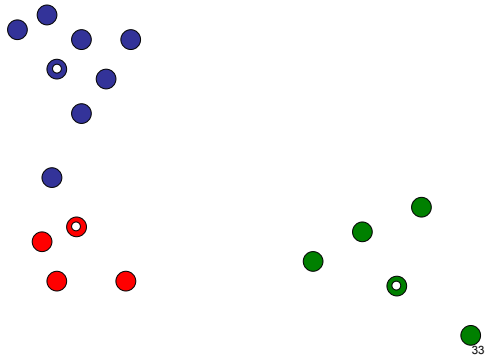


New medoids have smallest average distance to other points in their cluster.



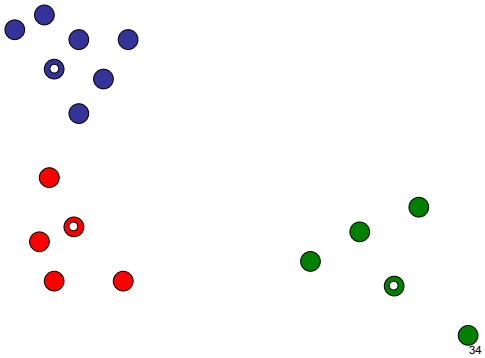
32

Reassign Each Point to Closest Medoid



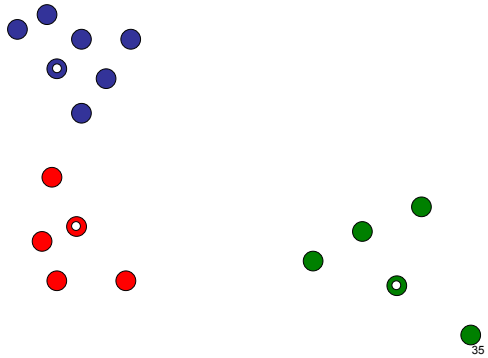
33

Reassign Each Point to Closest Medoid



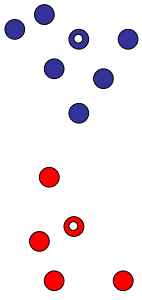
34

Find New Medoid for Each Cluster

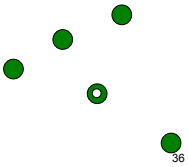


35

Reassign Each Point to Closest Medoid

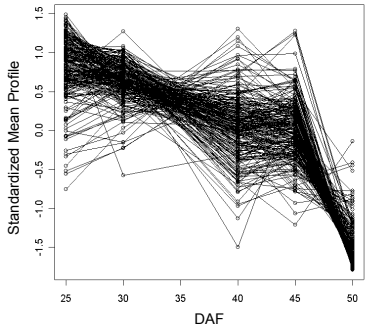


No reassignment is needed, so the procedure stops.



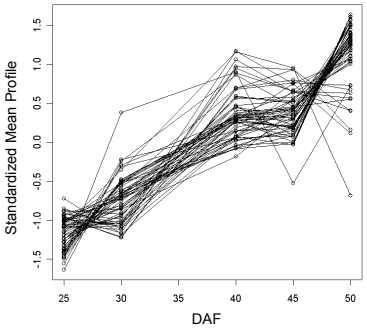
36

Cluster 1 of 3 from K-Medoids Algorithm Applied to the Top 400 Genes from the Two-Color Array Data



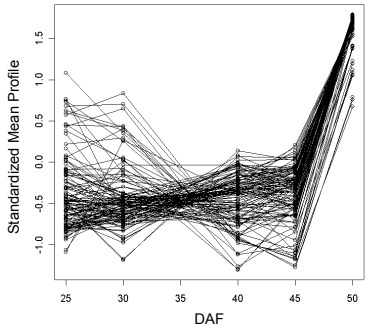
37

Cluster 2 of 3 from K-Medoids Algorithm Applied to the Top 400 Genes from the Two-Color Array Data



38

Cluster 3 of 3 from K-Medoids Algorithm Applied to the Top 400 Genes from the Two-Color Array Data



39

Choosing the Number of Clusters K

- Choose K that maximizes the average *silhouette width*.

Rousseeuw, P.J. (1987). *Journal of Computational and Applied Mathematics*, **20**, 53-65.

Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

- Choose K according to the *gap statistic*.

Tibshirani, R., Walther, G., Hastie, T. (2001). *Journal of the Royal Statistics Society, Series B-Statistical Methodology*, **63**, 411-423.

40

Silhouette Width

- The silhouette width of an object is

(B-W)/max(B,W)

where W=average distance of the object to all other objects within its cluster and B=average distance of the object to all objects in its nearest neighboring cluster.

- The silhouette width will be between -1 and 1.

41

Silhouette Width = (B-W)/max(B,W)

- Values near 1 indicate that an object is near the center of a tight cluster.
- Values near 0 indicate that an object is between clusters.
- Negative values indicate that an object may be in the wrong cluster.

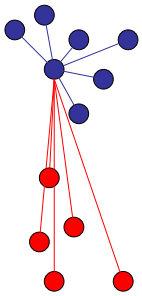
42

Silhouette Width = (B-W)/max(B,W)

- The silhouette widths of clustered objects can be averaged.
- A clustering with a high average silhouette width is preferred.
- For a given method of clustering, we may wish to choose the value of K that maximizes the average silhouette width.

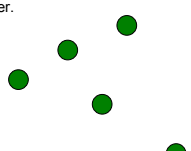
43

For a Given K Compute Silhouette Width for Each Point



Find W=average distance from point to all others within its cluster.

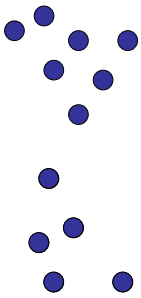
Find B=average distance from point to all others in its nearest neighboring cluster.



Silhouette width is $\frac{B-W}{max(B,W)}$

44

Choice of K

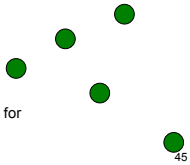


Silhouette width is computed for all points and averaged.

K with largest average silhouette width is preferred.

K=3: Average Silhouette Width=0.640

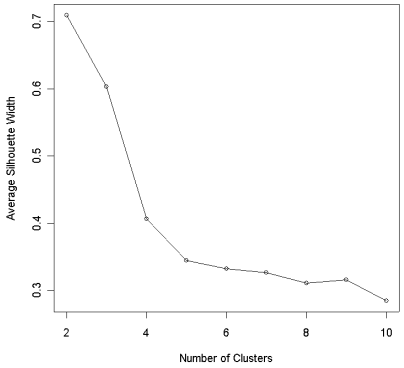
K=2: Average Silhouette Width=0.646



Slight preference for K=2 in this case.

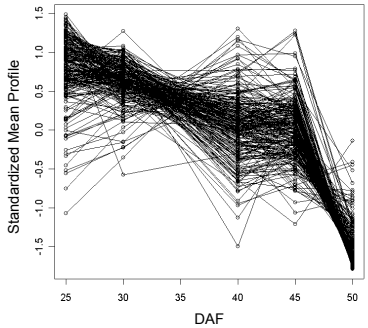
45

Average Silhouette Width vs. K for the K-Medoids Algorithm Applied to the Top 400 Genes from the Two-Color Array Data



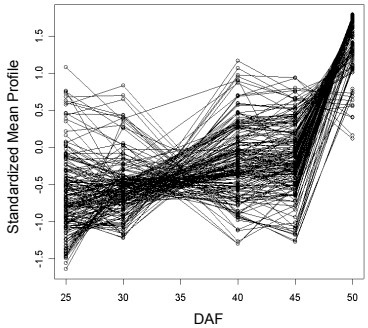
46

Cluster 1 of 2 from K-Medoids Algorithm Applied to the Top 400 Genes from the Two-Color Array Data



47

Cluster 2 of 2 from K-Medoids Algorithm Applied to the Top 400 Genes from the Two-Color Array Data



48

Gap Statistic

- For a given clustering of n objects $x_1,...,x_n$, the distance $d(x_i,x_j)$ between objects x_i and x_j is called a *within-cluster distance* if x_i and x_j are within the same cluster.
- Let D_r = the sum of all within-cluster distances in the r^{th} cluster, and let n_r denote the number of objects in the r^{th} cluster.
- For a given clustering of n objects into k clusters, let $W_k = \Sigma_{r=1}^{k} D_r / n_r$.

49

Gap Statistic (continued)

- For a given clustering method, compute $\log W_1$, $\log W_2$,..., $\log W_K$.
- Let $\min_j$ denote the minimum of the j^{th} component of all n objects clustered.
- Let $\max_j$ denote the maximum of the j^{th} component of all n objects to be clustered.
- Generate n random objects uniformly distributed on the m dimensional rectangle

$$[\min_1,\max_1] \times \cdots \times [\min_m,\max_m].$$

50

Gap Statistic (continued)

- Using the random uniform data, compute $\log W_1^*$, $\log W_2^*$,..., $\log W_K^*$.
- Randomly generate new uniform data multiple times (20 or more) and use the results to obtain $\overline{\log W_1^*}$, $\overline{\log W_2^*}$,..., $\overline{\log W_K^*}$ and $S_1, S_2,...,S_K$; the averages and standard deviations of the simulated log W values.
- Let $G(k) = \overline{\log W_k^*} - \log W_k$.

51

Estimate of Best K Using the Gap Statistic

- An approximate standard error for G(k) is

$$S_k\sqrt{1+1/N}$$

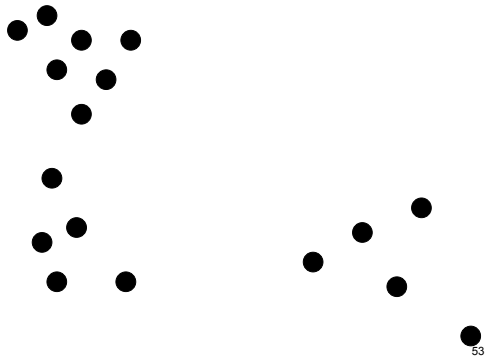
where N denotes the number of randomly generated data sets.

- An estimate of the best K is given by

$$\hat{K} = \min \{ k : G(k) \geq G(k+1) - S_{k+1}\sqrt{1+1/N} \}.$$

52

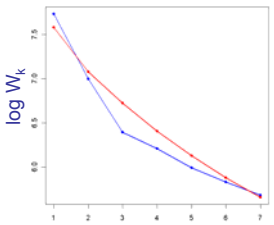
Simple Example Data Revisited



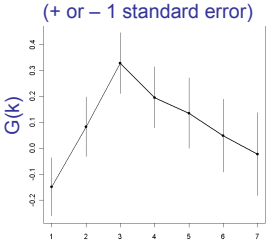
53

Gap Analysis for the Simple Example (N=1000)

$\overline{\log W_k^*}$ and $\log W_k$ vs. k

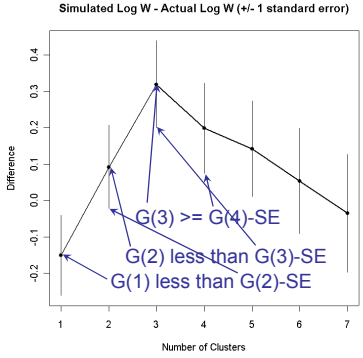


$G(k)=\overline{\log W_k^*} - \log W_k$ vs. k
(+ or – 1 standard error)



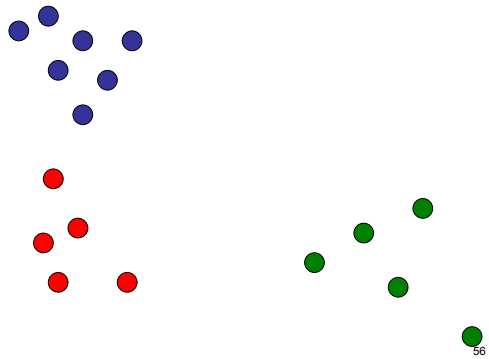
54

The Gap Statistic Suggests K=3 Clusters



55

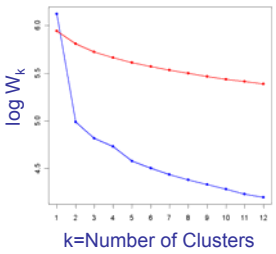
The Gap Statistic Suggests K=3 Clusters



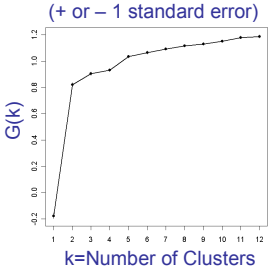
56

Gap Analysis for Two-Color Array Data (N=100)

$\overline{\log W_k^*}$ and $\log W_k$ vs. k

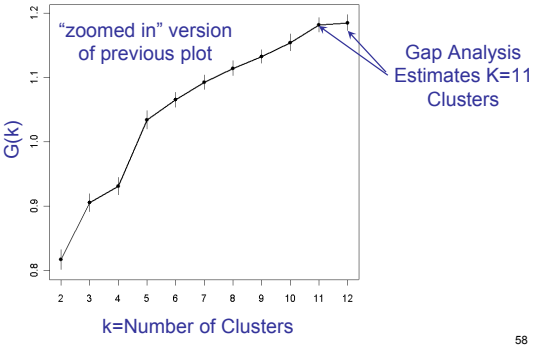


$G(k)=\overline{\log W_k^*} - \log W_k$ vs. k
(+ or – 1 standard error)

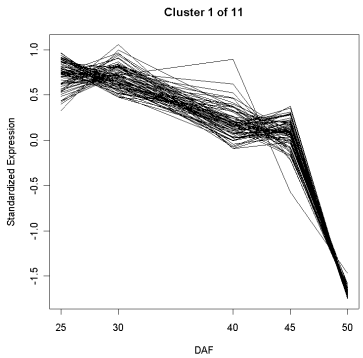


57

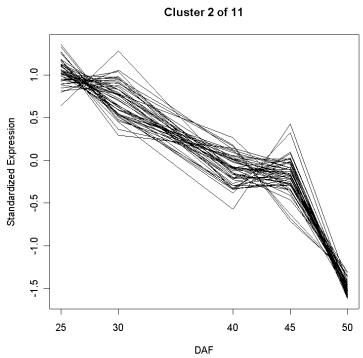
Gap Analysis for Two-Color Array Data (N=100)



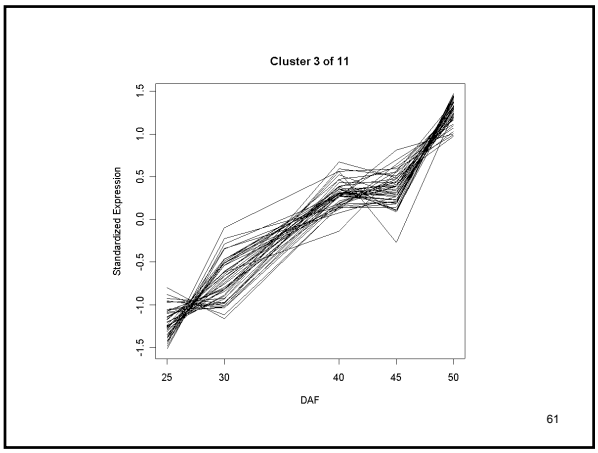
58

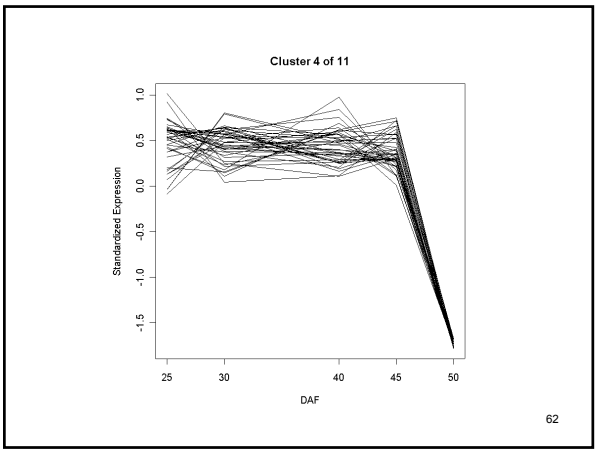


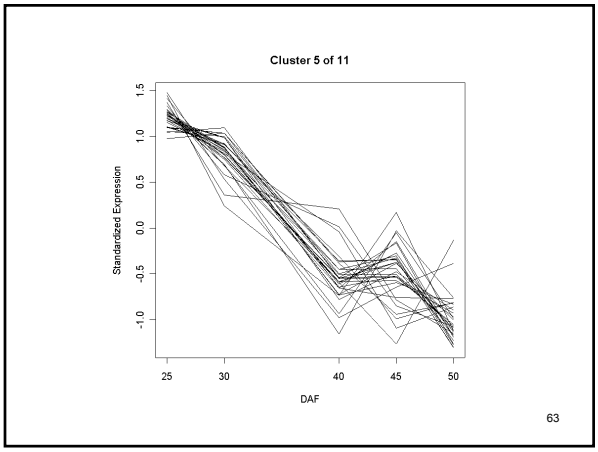
59

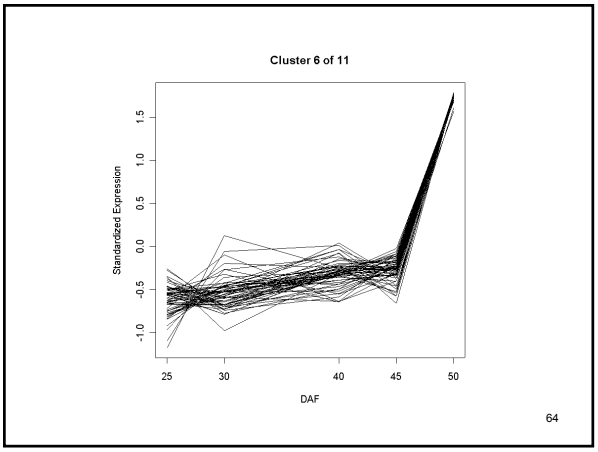


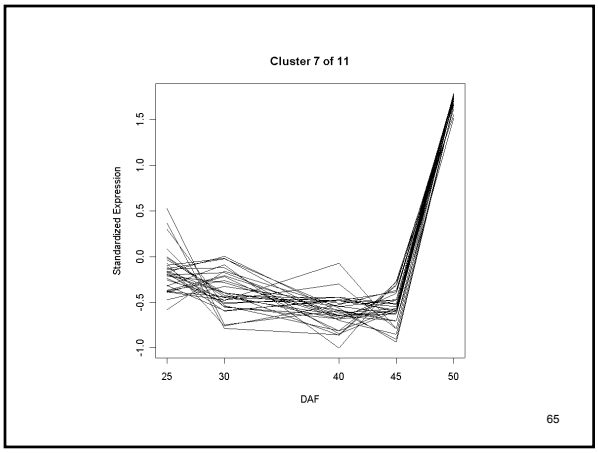
60

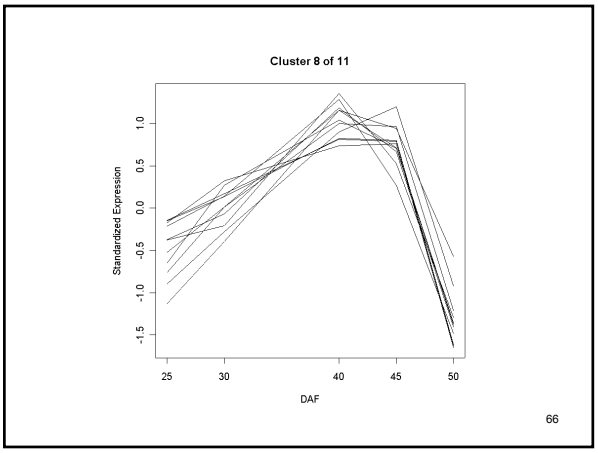

Cluster 3 of 11
Standardized Expression
DAF
61

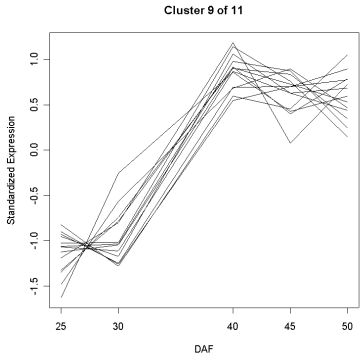

Cluster 4 of 11
Standardized Expression
DAF
62


Cluster 5 of 11
Standardized Expression
DAF
63

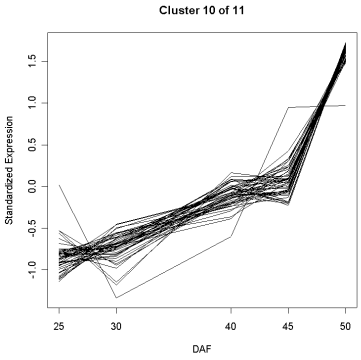

Cluster 6 of 11
Standardized Expression
DAF
64


Cluster 7 of 11
Standardized Expression
DAF
65

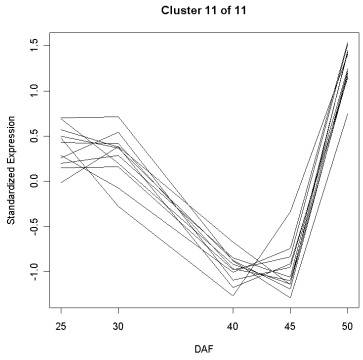

Cluster 8 of 11
Standardized Expression
DAF
66



67

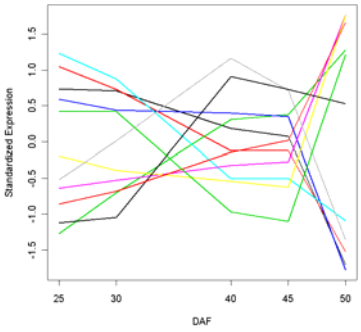


68



69

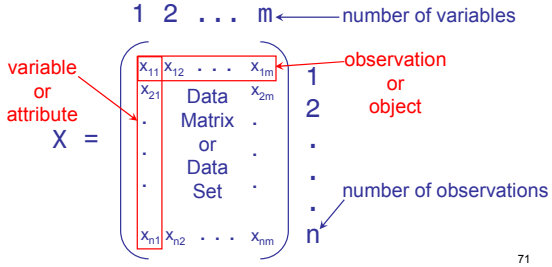
Plot of Cluster Medoids



70

Principal Components

- Principal components can be useful for providing low-dimensional views of high-dimensional data.



71

Principal Components (continued)

- Each principal component of a data set is a variable obtained by taking a linear combination of the original variables in the data set.
- A linear combination of m variables x_1, x_2, ..., x_m is given by $c_1x_1 + c_2x_2 + \cdots + c_mx_m$.
- For the purpose of constructing principal components, the vector of coefficients is restricted to have unit length, i.e., $c_1^2 + c_2^2 + \cdots + c_m^2 = 1$.

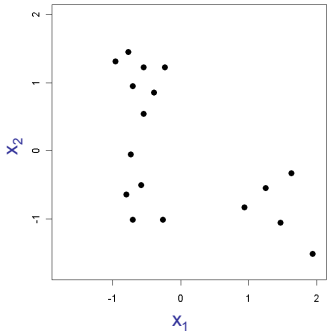
72

Principal Components (continued)

- The first principal component is the linear combination of the variables that has maximum variation across the observations in the data set.
- The j^{th} principal component is the linear combination of the variables that has maximum variation across the observations in the data set subject to the constraint that the vector of coefficients be orthogonal to coefficient vectors for principal components 1, ..., j-1.

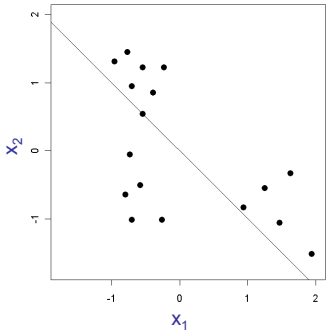
73

The Simple Data Example



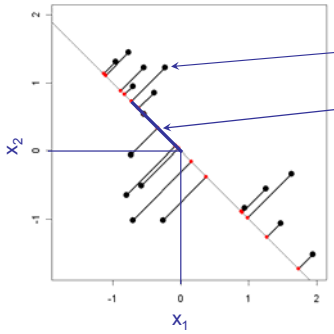
74

The First Principal Component Axis



75

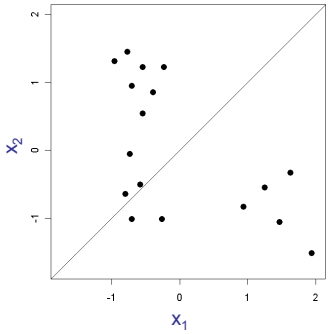
The First Principal Components



1st PC for this point is signed distance between its projection onto the 1st PC axis and the origin.

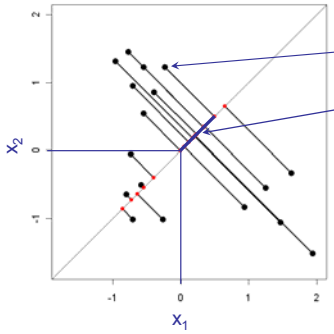
76

The Second Principal Component Axis



77

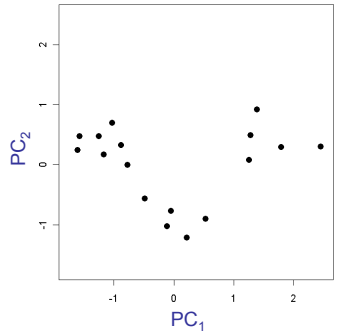
The Second Principal Component



2nd PC for this point is signed distance between its projection onto the 2nd PC axis and the origin.

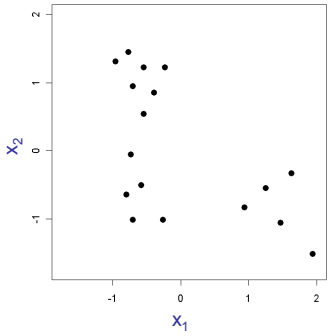
78

Plot of PC1 vs. PC2



79

Compare the PC plot to the plot of the original data below.



Because there are only two variables here, the plot of PC2 vs. PC1 is just a rotation of the original plot.

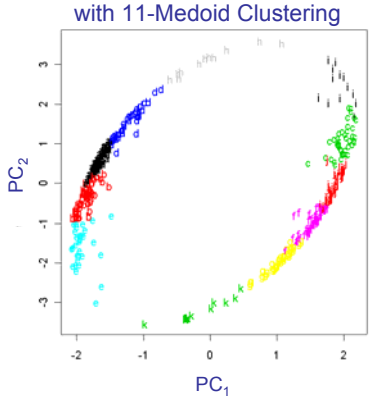
80

There is more to be gained when the number of variables is greater than 2.

- Consider the principal components for the 400 significant genes from our two-color microarray experiment.
- Our data matrix has n=400 rows and m=5 columns.
- We have looked at this data using parallel coordinate plots.
- What would it look like if we projected the data points to 2-dimensions?

81

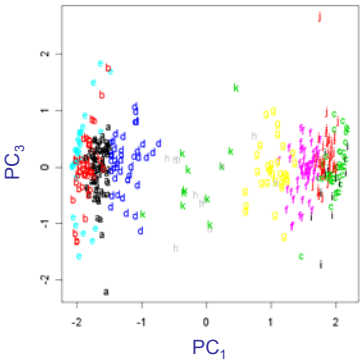
Projection of Two-Color Array Data with 11-Medoid Clustering



a=1
b=2
c=3
d=4
e=5
f=6
g=7
h=8
i=9
j=10
k=11

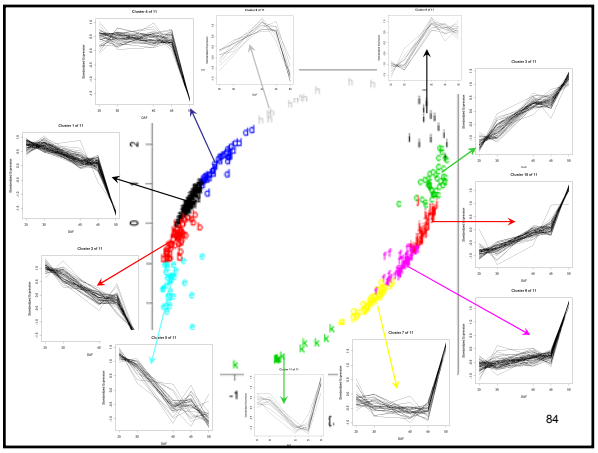
82

Projection of Two-Color Array Data with 11-Medoid Clustering



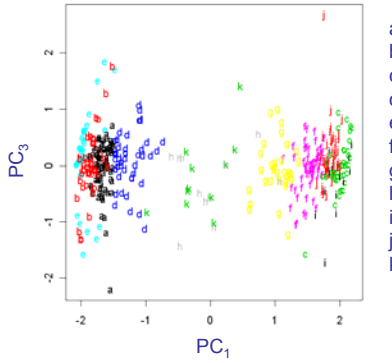
a=1
b=2
c=3
d=4
e=5
f=6
g=7
h=8
i=9
j=10
k=11

83



84

Projection of Two-Color Array Data with 11-Medoid Clustering



a=1
b=2
c=3
d=4
e=5
f=6
g=7
h=8
i=9
j=10
k=11

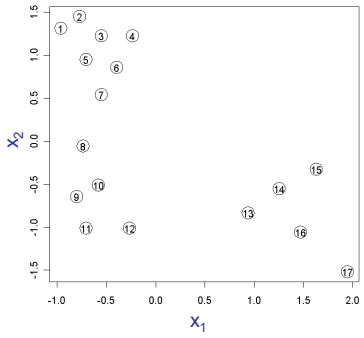
85

Hierarchical Clustering Methods

- Hierarchical clustering methods build a nested sequence of clusters that can be displayed using a *dendrogram*.
- We will begin with some simple illustrations and then move on to a more general discussion.

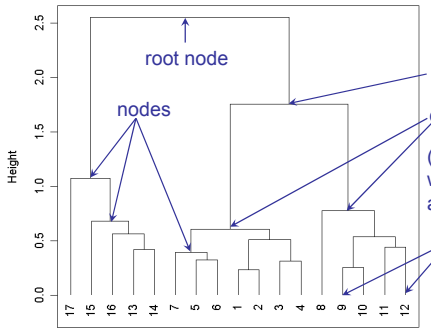
86

The Simple Example Data with Observation Numbers



87

Dendrogram for the Simple Example Data



Tree Structure

a parent node

daughter nodes

(daughter nodes with same parent are sister nodes)

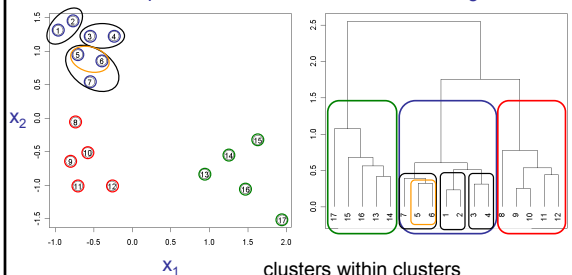
terminal nodes or leaves corresponding to objects

88

A Hierarchical Clustering of the Simple Example Data

Scatterplot of Data

Dendrogram

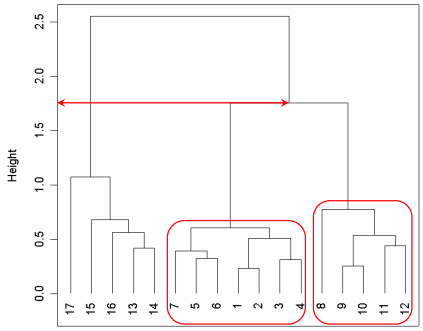


clusters within clusters

within clusters...

89

Dendrogram for the Simple Example Data

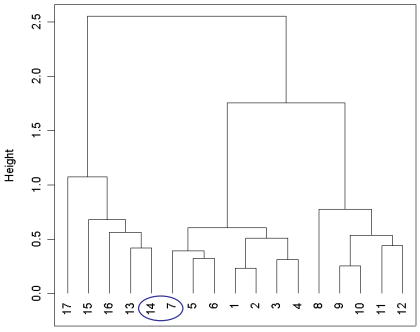


The height of a node represents the dissimilarity between the two clusters merged together at the node.

These two clusters have a dissimilarity of about 1.75.

90

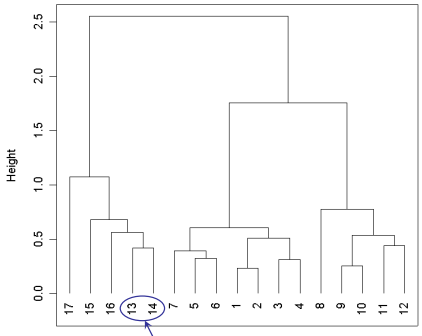
The appearance of a dendrogram is not unique.



Any two sister nodes could trade places without changing the meaning of the dendrogram.

Thus 14 next to 7 does not imply that these objects are similar.

The appearance of a dendrogram is not unique.

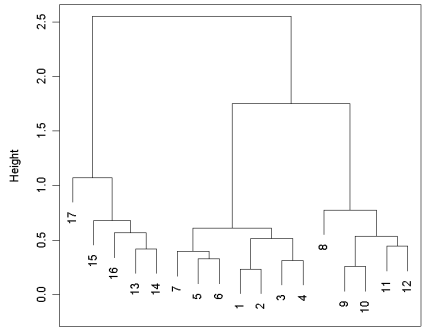


By convention, R dendrograms show the lower sister node on the left.

Ties are broken by observation number.

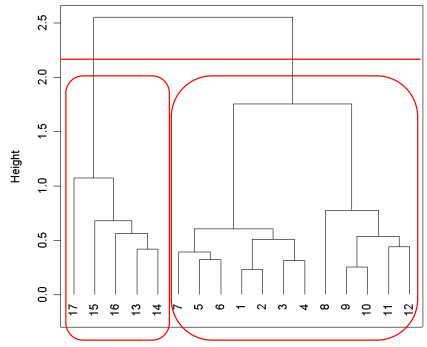
e.g., 13 is to the left of 14

The appearance of a dendrogram is not unique.



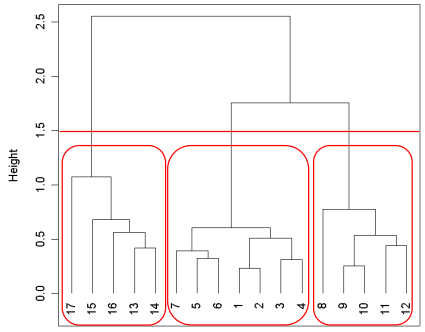
The lengths of the branches leading to terminal nodes have no particular meaning in R dendrograms.

Cutting the tree at a given height will correspond to a partitioning of the data into k clusters.



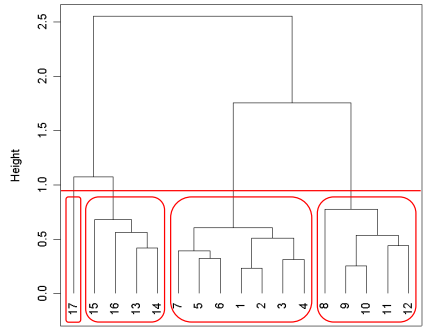
k=2 Clusters

Cutting the tree at a given height will correspond to a partitioning of the data into k clusters.



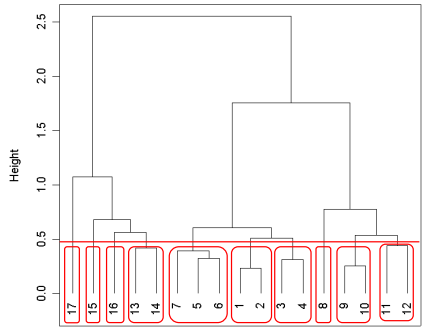
k=3 Clusters

Cutting the tree at a given height will correspond to a partitioning of the data into k clusters.



k=4 Clusters

Cutting the tree at a given height will correspond to a partitioning of the data into k clusters.



Agglomerative (Bottom-Up) Hierarchical Clustering

- Define a measure of distance between any two clusters. (An individual object is considered a cluster of size one.)
- Find the two nearest clusters and merge them together to form a new cluster.
- Repeat until all objects have been merged into a single cluster.

98

Common Measures of Between-Cluster Distance

- Single Linkage a.k.a. Nearest Neighbor: the distance between any two clusters A and B is the minimum of all distances from an object in cluster A to an object in cluster B.
- Complete Linkage a.k.a Farthest Neighbor: the distance between any two clusters A and B is the maximum of all distances from an object in cluster A to an object in cluster B.

99

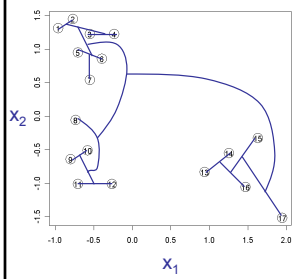
Common Measures of Between-Cluster Distance

- Average Linkage: the distance between any two clusters A and B is the average of all distances from an object in cluster A to an object in cluster B.
- Centroid Linkage: the distance between any two clusters A and B is the distance between the centroids of cluster A and B. (The centroid of a cluster is the componentwise average of the objects in a cluster.)

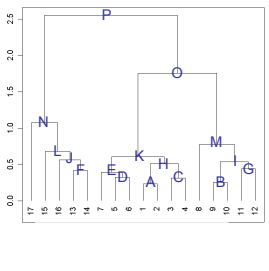
100

Agglomerative Clustering Using Average Linkage for the Simple Example Data Set

Scatterplot of Data

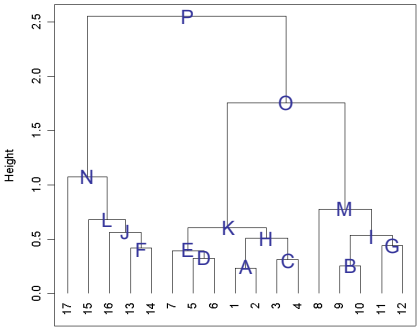


Dendrogram



101

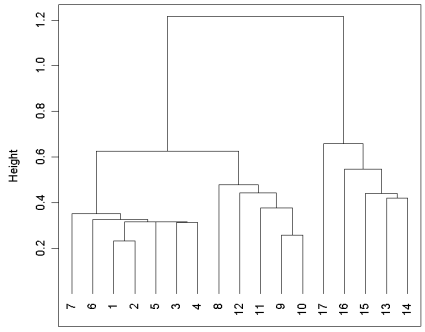
Agglomerative Clustering Using Average Linkage for the Simple Example Data Set



A. 1-2
B. 9-10
C. 3-4
D. 5-6
E. 7-(5,6)
F. 13-14
G. 11-12
H. (1,2)-(3,4)
I. (9,10)-(11,12)
etc....

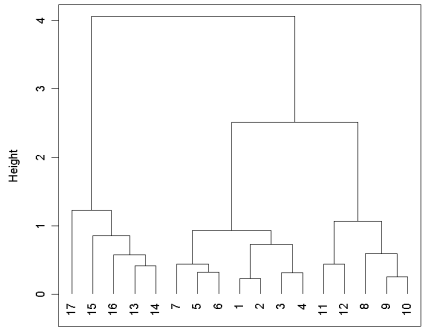
102

Agglomerative Clustering Using Single Linkage for the Simple Example Data Set



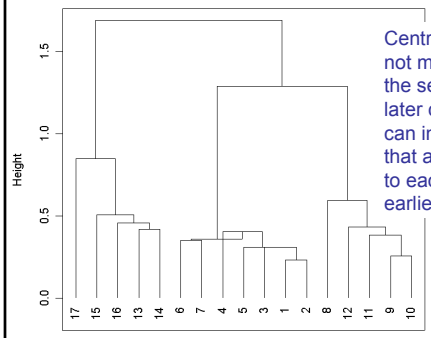
103

Agglomerative Clustering Using Complete Linkage for the Simple Example Data Set



104

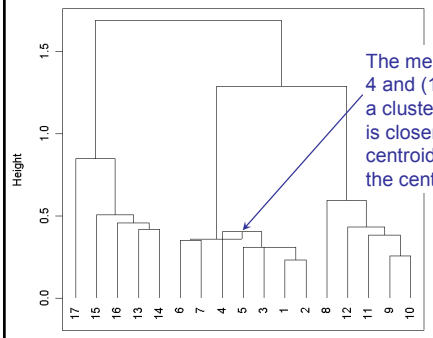
Agglomerative Clustering Using Centroid Linkage for the Simple Example Data Set



Centroid linkage is not monotone in the sense that later cluster merges can involve clusters that are more similar to each other than earlier merges.

105

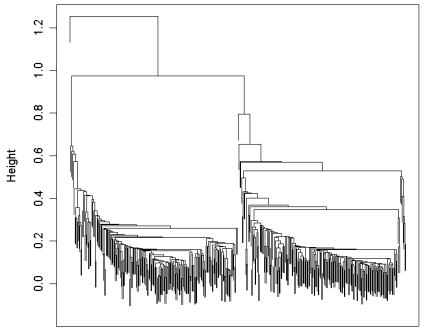
Agglomerative Clustering Using Centroid Linkage for the Simple Example Data Set



The merge between 4 and (1,2,3,5) creates a cluster whose centroid is closer to the (6,7) centroid than 4 was to the centroid of (1,2,3,5).

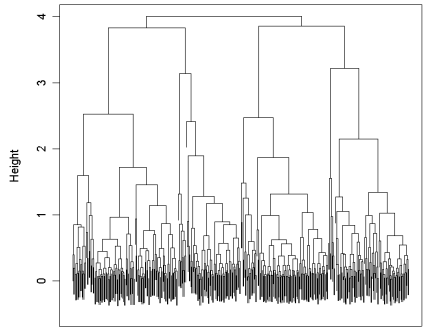
106

Agglomerative Clustering Using Single Linkage for the Two-Color Microarray Data Set



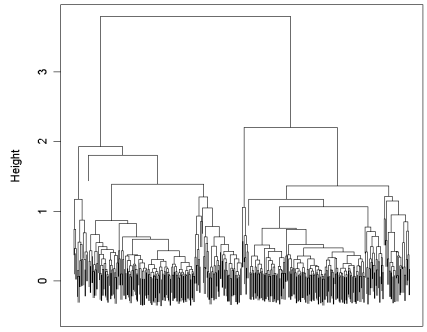
107

Agglomerative Clustering Using Complete Linkage for the Two-Color Microarray Data Set



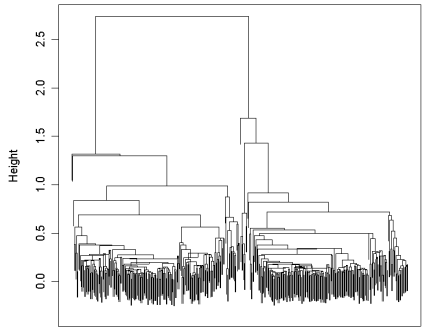
108

Agglomerative Clustering Using Average Linkage for the Two-Color Microarray Data Set



109

Agglomerative Clustering Using Centroid Linkage for the Two-Color Microarray Data Set



110

Which Between-Cluster Distance is Best?

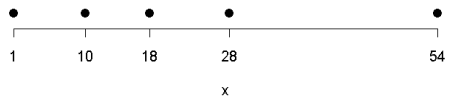
- Depends, of course, on what is meant by "best".
- Single linkage tends to produce "long stringy" clusters.
- Complete linkage produces compact spherical clusters but might result in some objects that are closer to objects in clusters other than their own. (See next example.)
- Average linkage is a compromise between single and complete linkage.
- Centroid linkage is not monotone.

111

1. Conduct agglomerative hierarchical clustering for this data using Euclidean distance and complete linkage.

2. Display your results using a dendrogram.

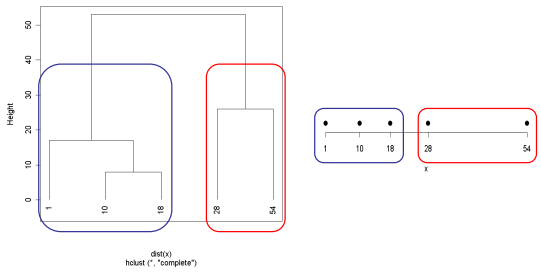
3. Identify the k=2 clustering using your results.



112

Results of Complete-Linkage Clustering

Results for k=2 Clusters



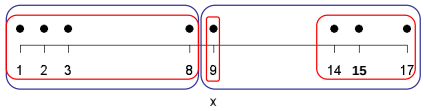
113

Divisive (Top-Down) Hierarchical Clustering

- Start with all data in one cluster and divide it into two clusters (using, e.g., 2-means or 2-medoids clustering).
- At each subsequent step, choose one of the existing clusters and divide it into two clusters.
- Repeat until there are n clusters each containing a single object.

114

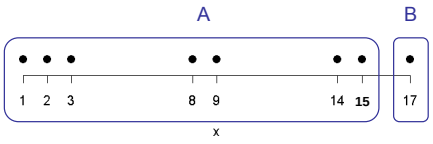
Potential Problem with Divisive Clustering



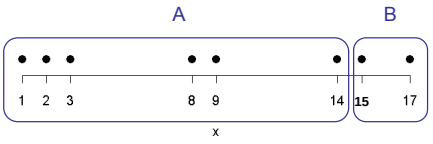
Macnaughton-Smith et al. (1965)

1. Start with objects in one cluster A.
2. Find the object with the largest average dissimilarity to all other objects in A and move that object to a new cluster B.
3. Find the object in cluster A whose average dissimilarity to other objects in cluster A minus its average dissimilarity to objects in cluster B is maximum. If this difference is positive, move the object to cluster B.
4. Repeat step 3 until no objects satisfying 3 are found.
5. Repeat steps 1 through 4 to one of the existing clusters (e.g., the one with the largest average within-cluster dissimilarity) until n clusters of 1 object each are obtained.

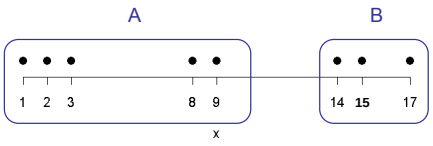
Macnaughton-Smith Divisive Clustering



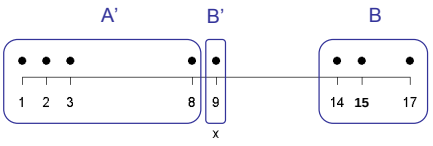
Macnaughton-Smith Divisive Clustering



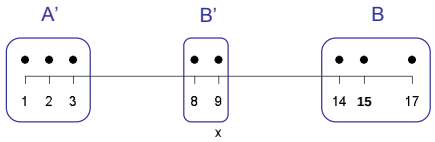
Macnaughton-Smith Divisive Clustering



Macnaughton-Smith Divisive Clustering



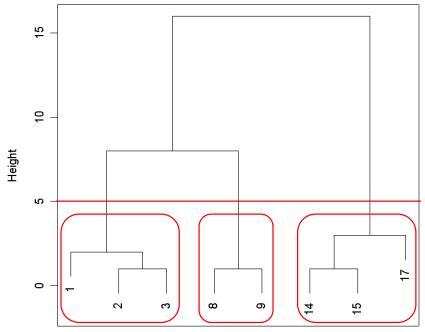
Macnaughton-Smith Divisive Clustering



Next continue to split each of these clusters until each object is in a cluster by itself.

121

Dendrogram for the Macnaughton-Smith Approach



122

Agglomerative vs. Divisive Clustering

- Divisive clustering has not been studied as extensively as agglomerative clustering.
- Divisive clustering may be preferred if only a small number of large clusters is desired.
- Agglomerative clustering may be preferred if a large number of small clusters is desired.

123