Lecture 6:
Hierarchical Clustering; Spectral Clustering

Lester Mackey

April 16, 2014

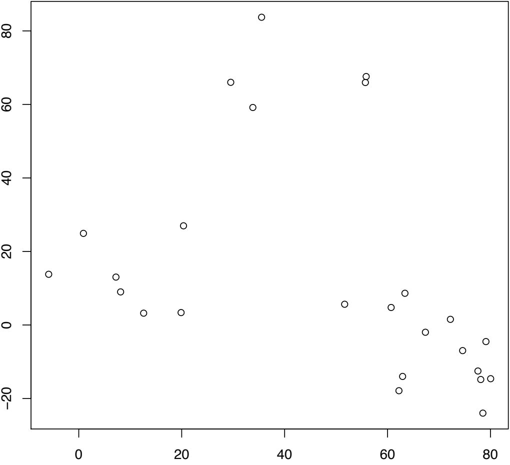
Blackboard discussion

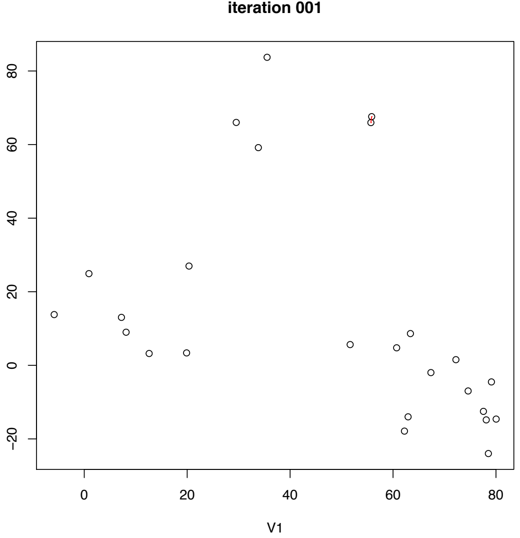
- See lecture notes

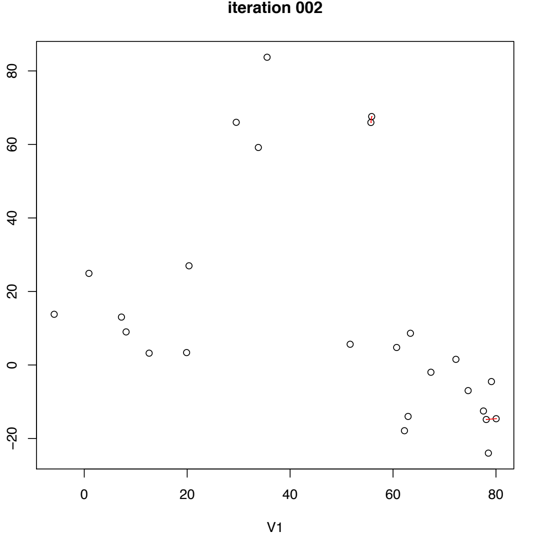
Average linkage agglomerative clustering

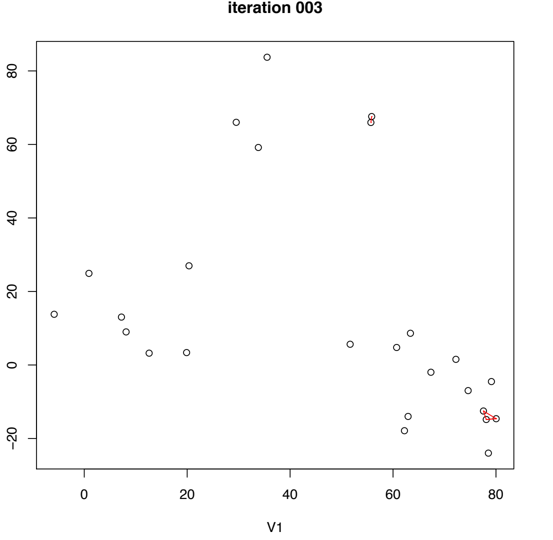
- Example behavior in 2D, Courtesy: Dave Blei

Data

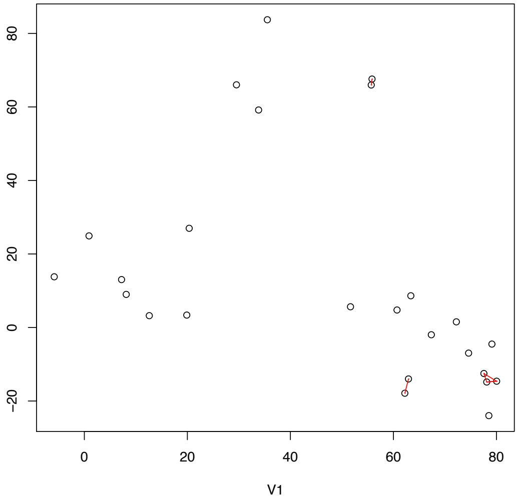



iteration 001
V1
0
20
40
60
80
-20
0
20
40
60
80

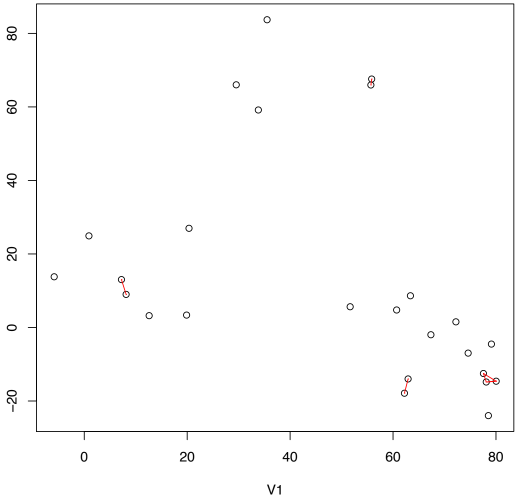

iteration 002
V1
0
20
40
60
80
-20
0
20
40
60
80

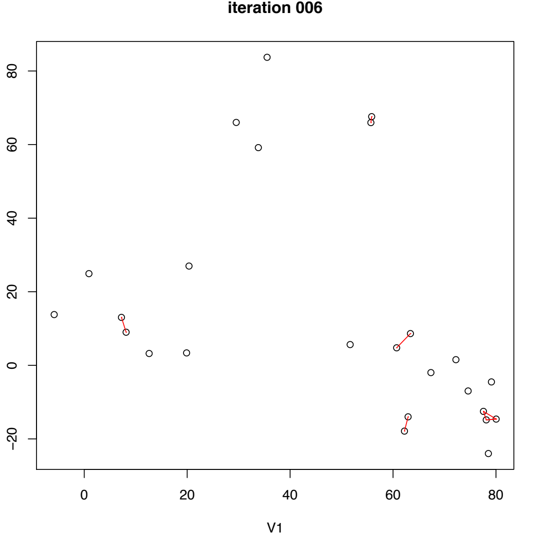

iteration 003
80
60
40
20
0
−20
0
20
40
60
80
V1

iteration 004

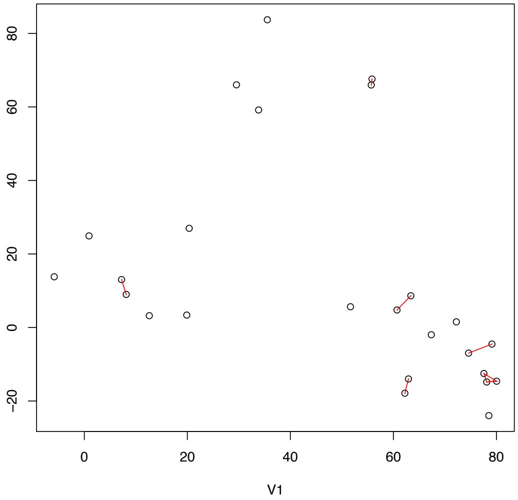


iteration 005

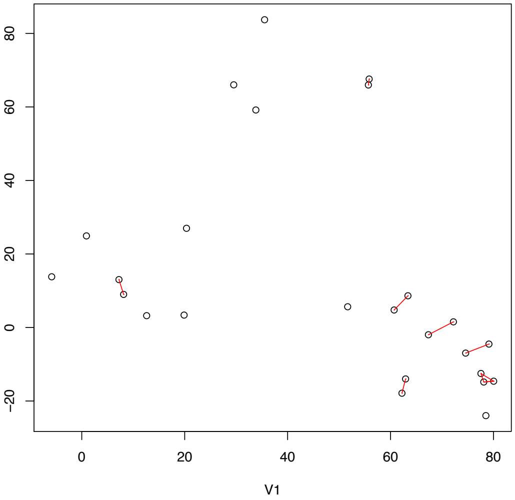


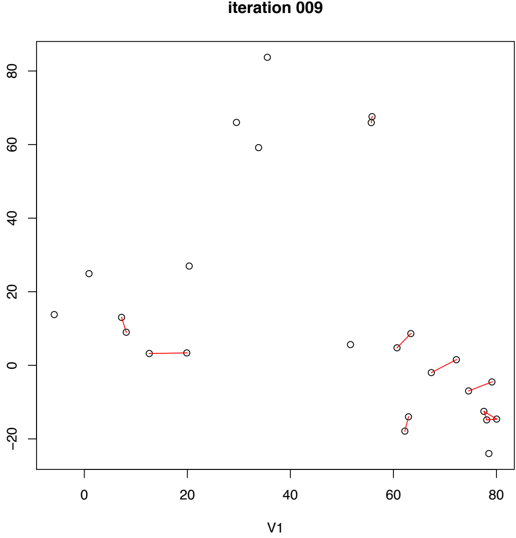

iteration 006
80
60
40
20
0
−20
0
20
40
60
80
V1

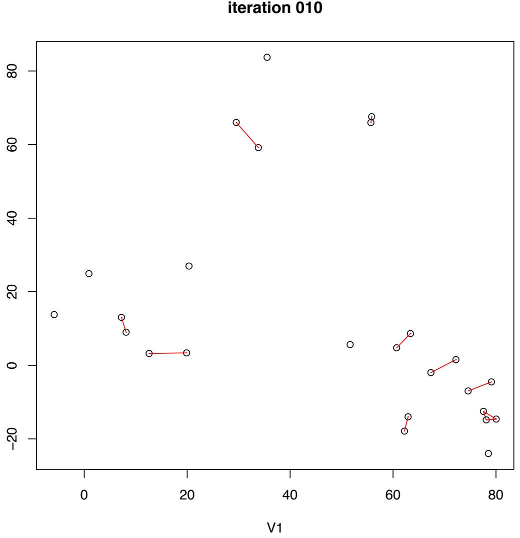
iteration 007



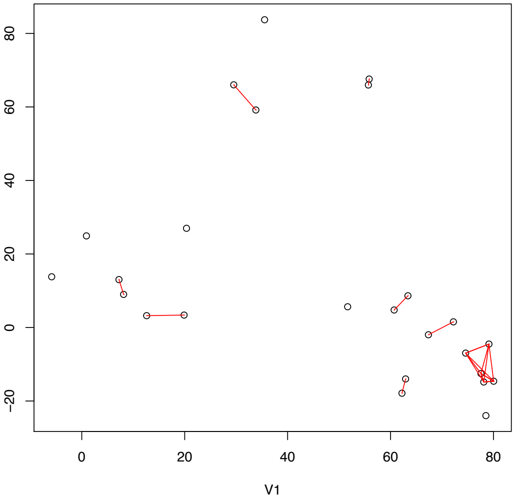
iteration 008



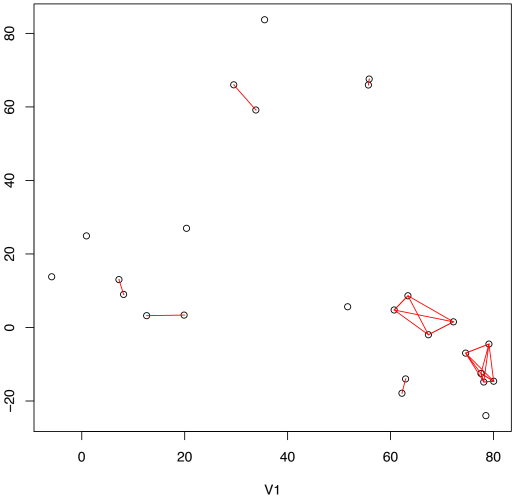

iteration 009
80
60
40
20
0
−20
0
20
40
60
80
V1

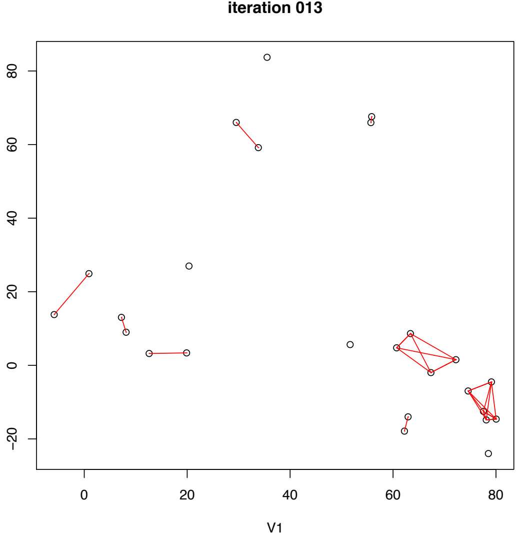

iteration 010
V1
0
20
40
60
80
-20
0
20
40
60
80

iteration 011



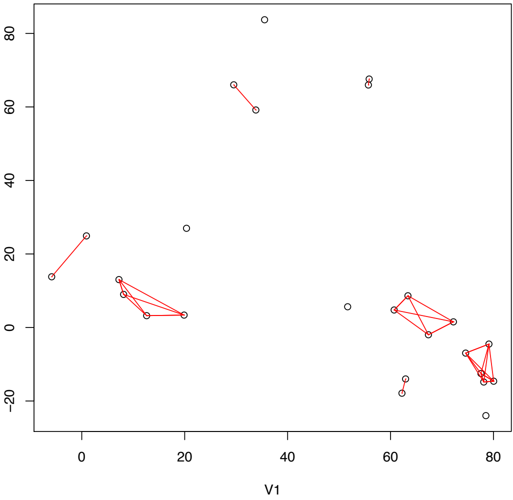
iteration 012



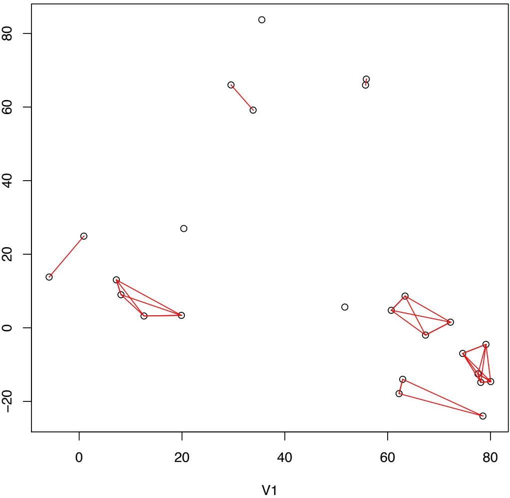


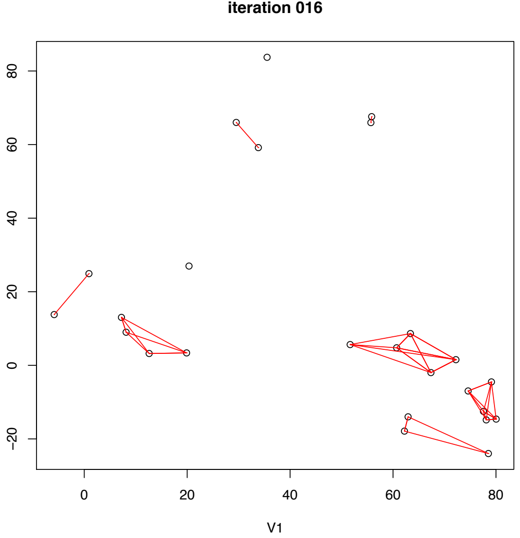
iteration 013
80
60
40
20
0
−20
0
20
40
60
80
V1

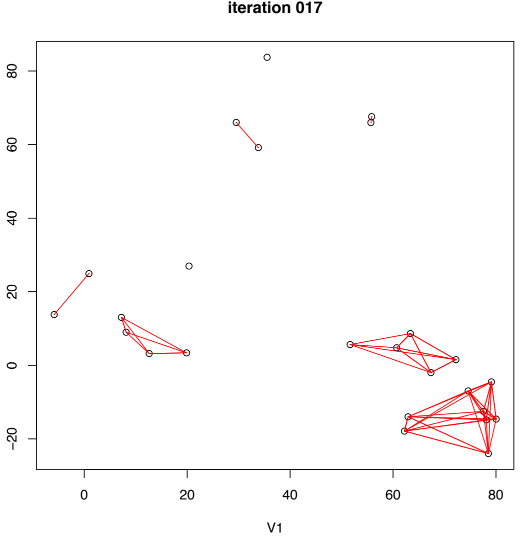
iteration 014

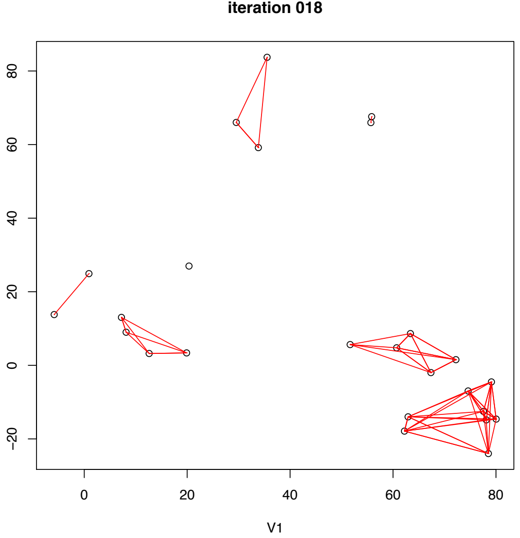


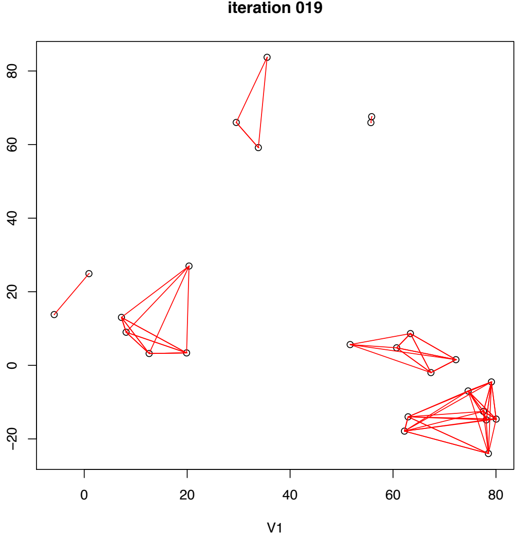
iteration 015

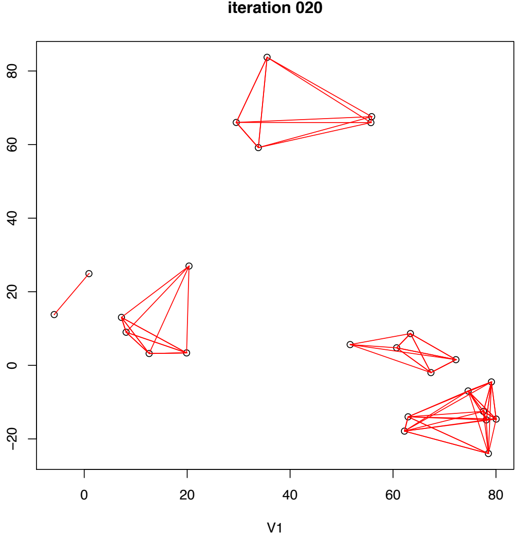



iteration 016
V1
0
20
40
60
80
-20
0
20
40
60
80

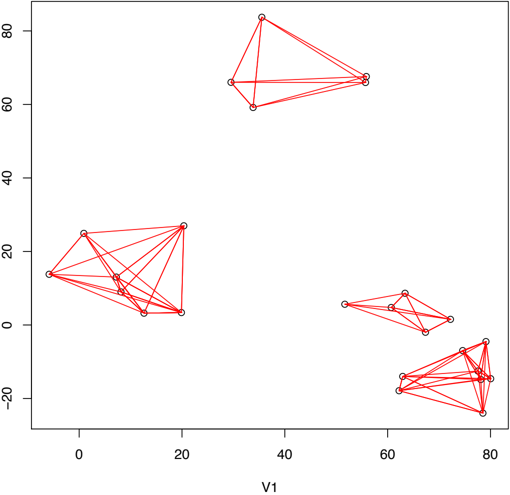

iteration 017
80
60
40
20
0
−20
0
20
40
60
80
V1


iteration 018
80
60
40
20
0
−20
0
20
40
60
80
V1

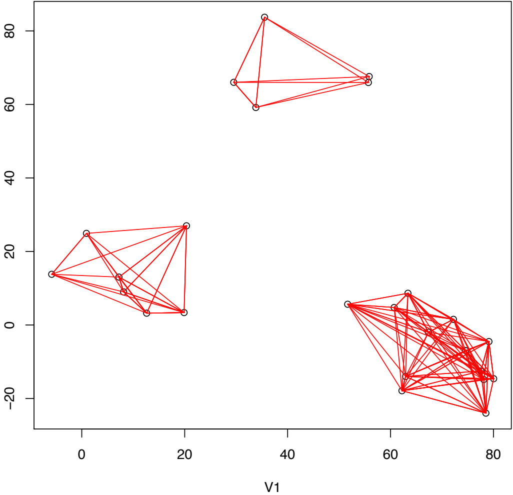

iteration 019
V1
0
20
40
60
80
-20
0
20
40
60
80

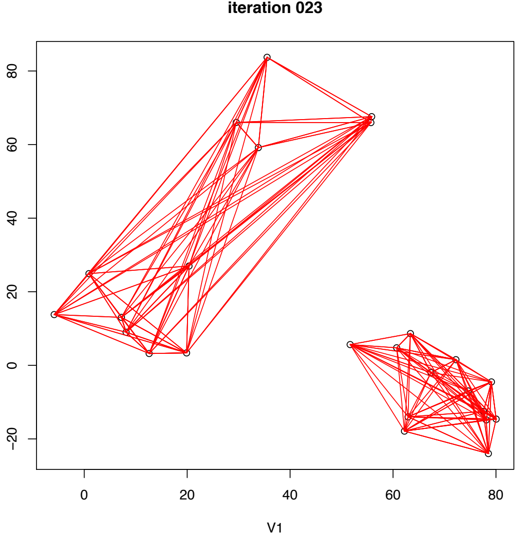

iteration 020
80
60
40
20
0
−20
0
20
40
60
80
V1

iteration 021

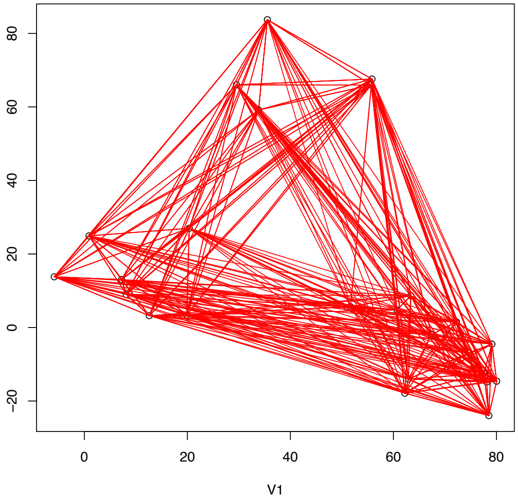


iteration 022




iteration 023
80
60
40
20
0
−20
0
20
40
60
80
V1

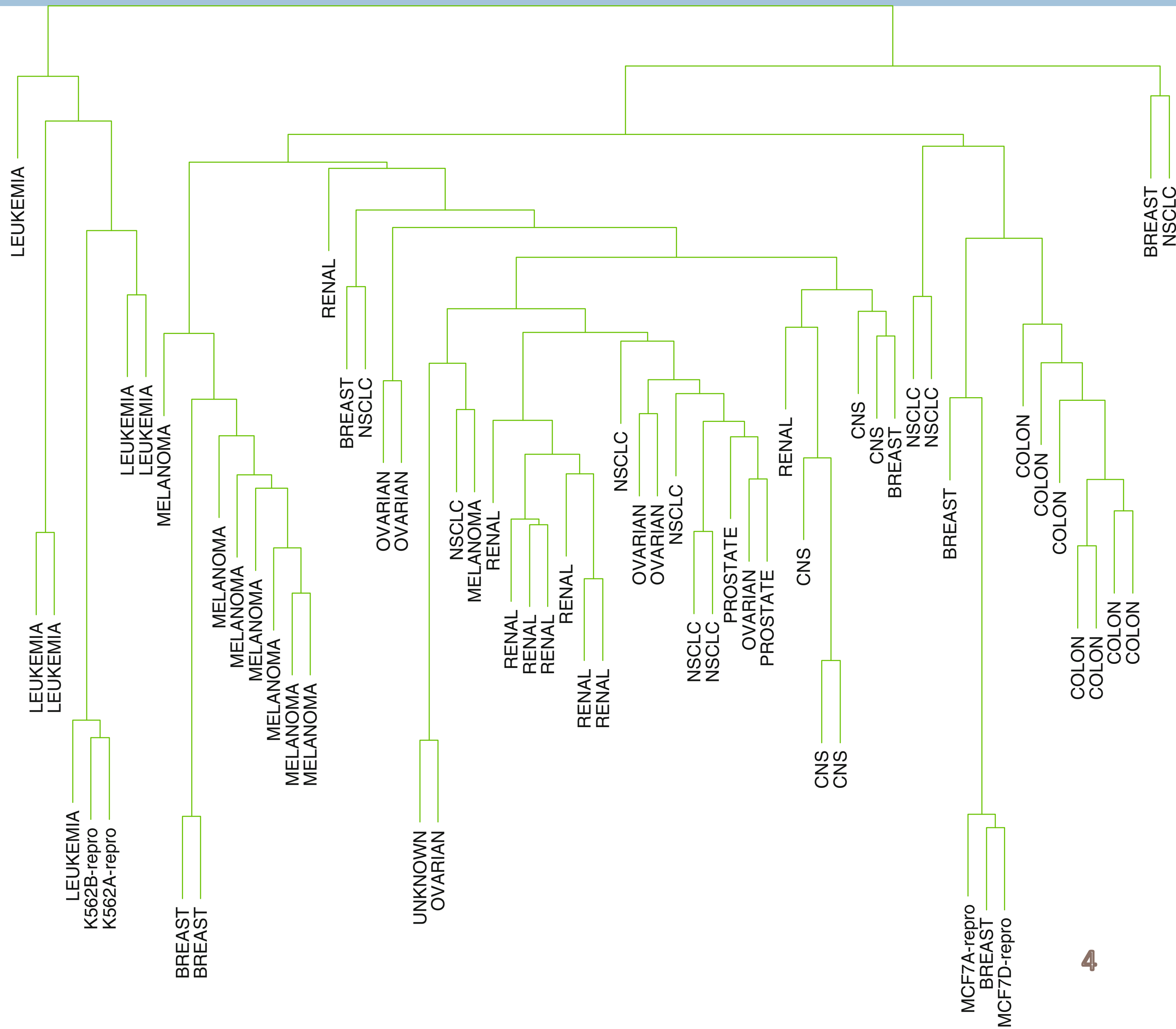
iteration 024



Clustering human tumor microarray data

Dendrogram from agglomerative hierarchical clustering with average linkage (Source: ESL)

6830 gene expression values from 64 tumors of 12 types



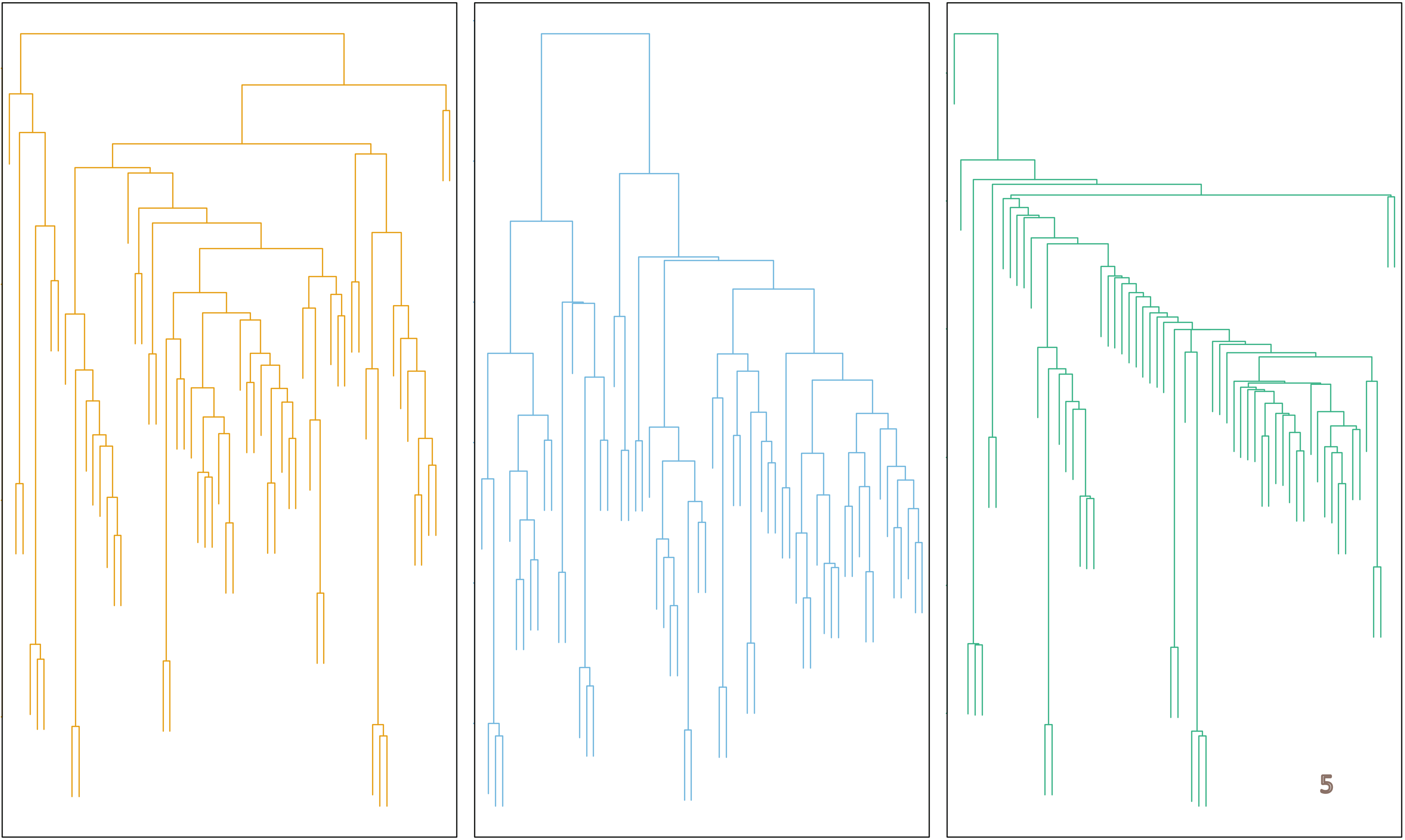
Clustering human tumor microarray data

Source: ESL

- Average Linkage

Complete Linkage

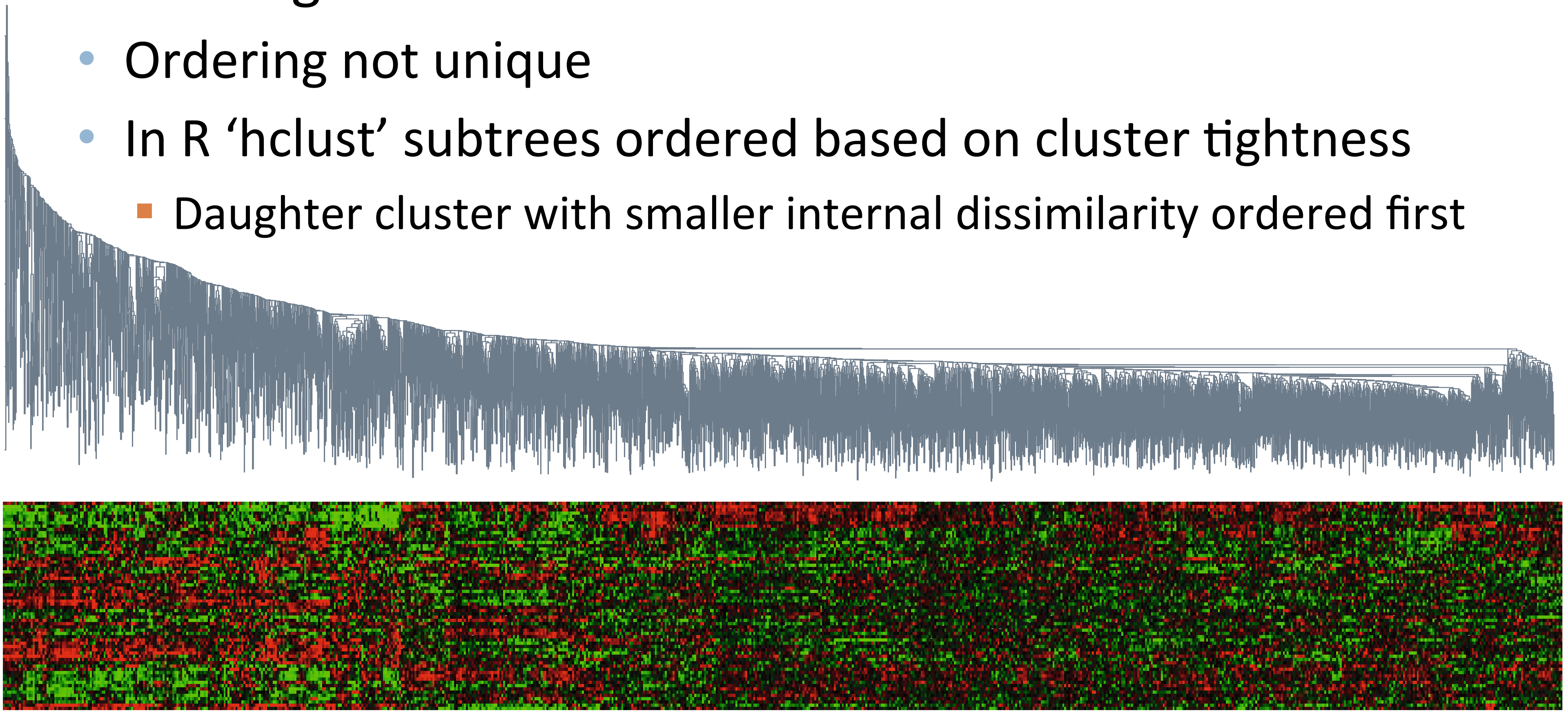
Single Linkage



Clustering human tumor microarray data

Source: ESL

- Can also cluster genes (instead of tumors) based on similar expression patterns across tumors
- Heatmap columns have been reordered based on clustering
  - Ordering not unique
  - In R 'hclust' subtrees ordered based on cluster tightness
    - Daughter cluster with smaller internal dissimilarity ordered first



Choosing *k*

Source: Tibshirani et al. (2001)

- Microarray data
- Avg. linkage
- Gap statistic used to select truncation level / number of clusters
- Cautionary tale?
    - Approximate local maximum at k = 2
    - Gap rises again after k = 6
    - Reflects smaller clusters within large separated clusters

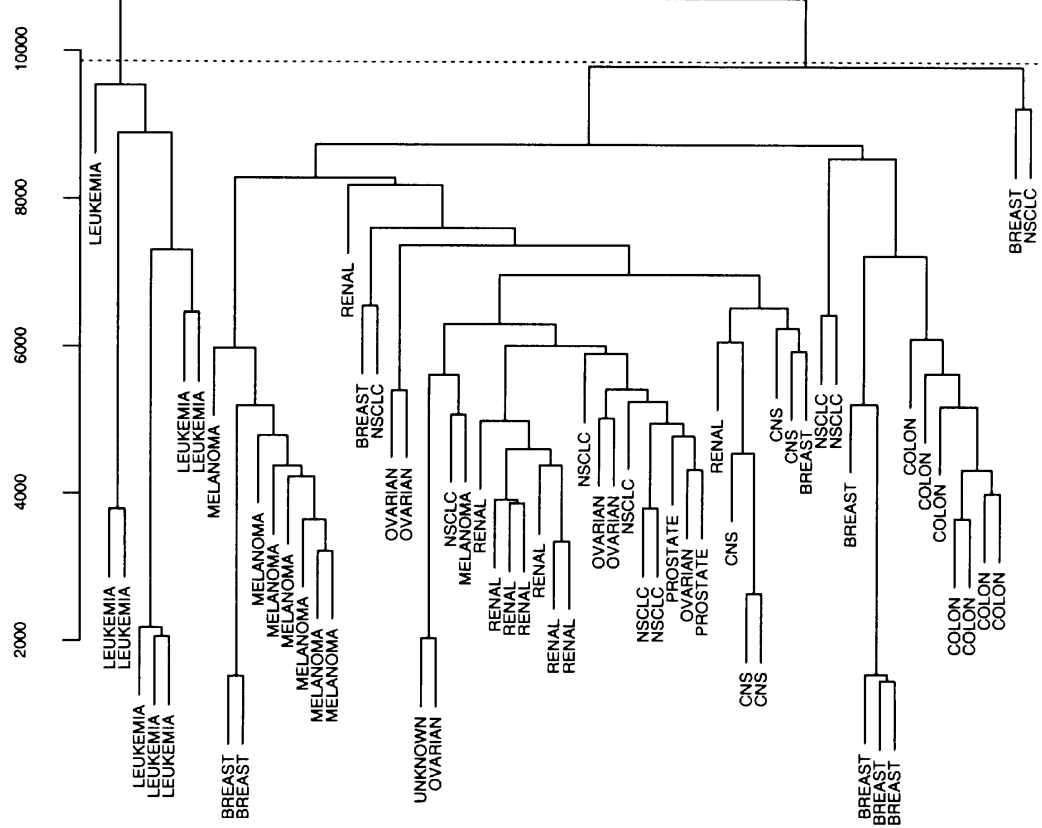


Fig. 3. Dendrogram from the deoxyribonucleic acid (DNA) microarray data: the dotted line cuts the tree, leaving two clusters as suggested by the gap statistic

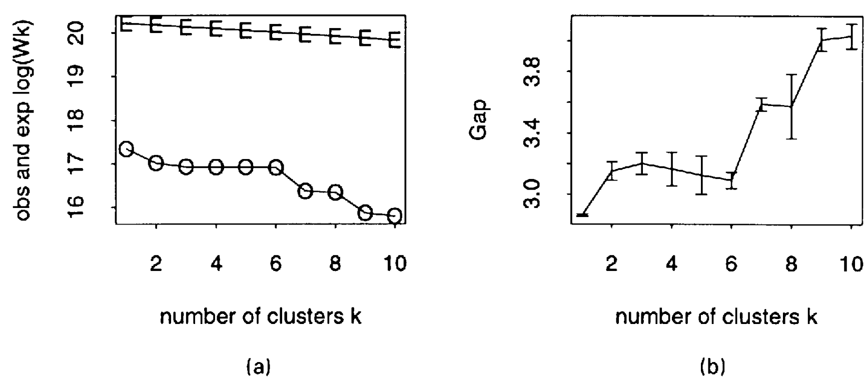
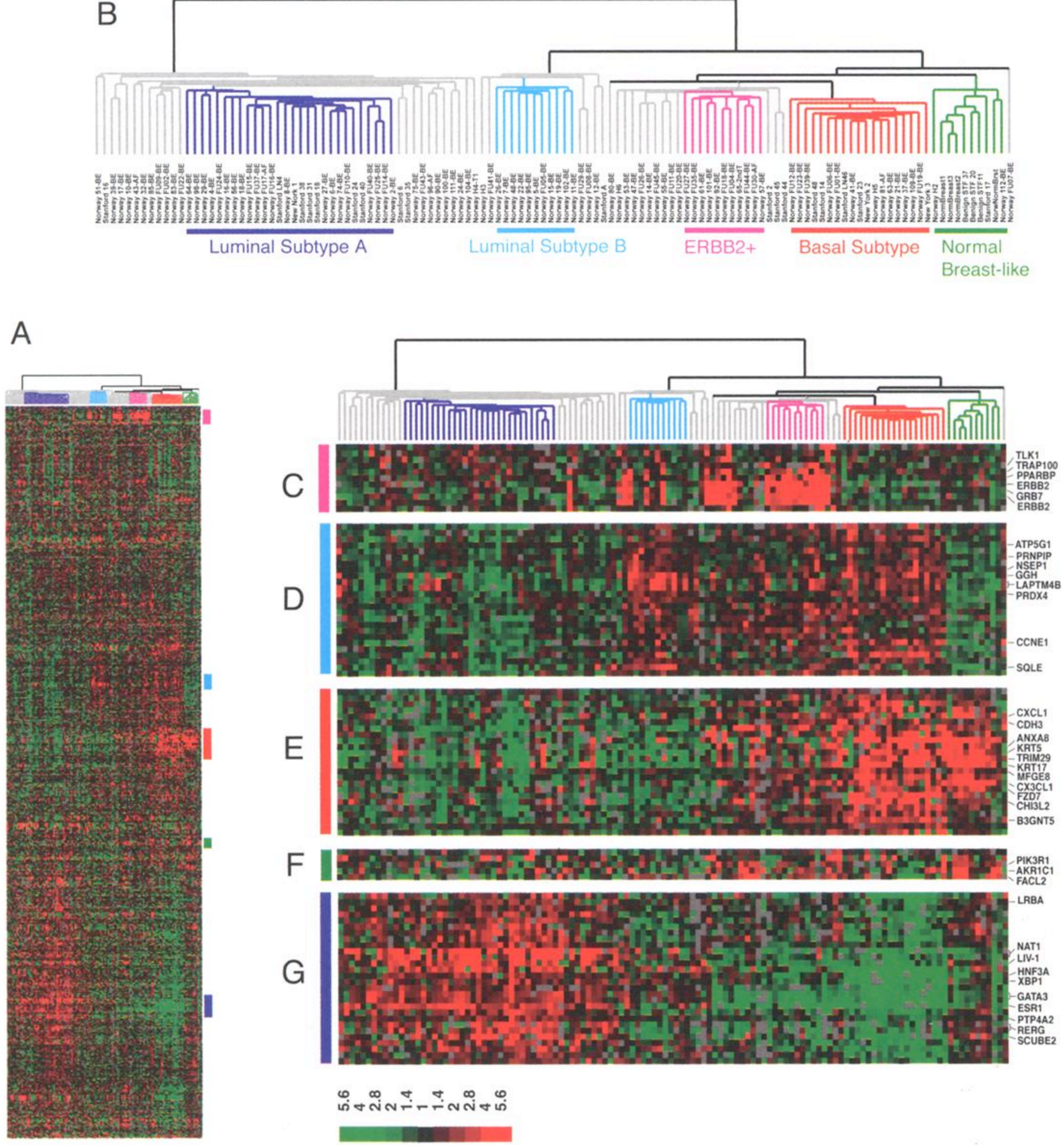


Fig. 4. (a) Logarithmic observed (O) and expected (E) within sum of squares curves and (b) the gap statistic for the DNA microarray data

In the wild

"Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Data Sets" (Sorlie et al., 2003)

- Evidence of multiple disease subtypes based on separate clustering results on several datasets
- Identified highly expressed genes per subtype
- Generated testable hypotheses



Hierarchical clustering in the wild

"The Statistical Analysis of Aesthetic Judgment: An Exploration" (Davenport and Studdert-Kennedy, 1972)

- Clustered 57 paintings rated for composition, drawing, color, & expression
- Results "at odds with conventional expectation"
- "Exploration suggests that there could be productive applications in the comparative analysis of subjective judgment"
- "The value of this analysis...will depend on any interesting speculation it may provoke."

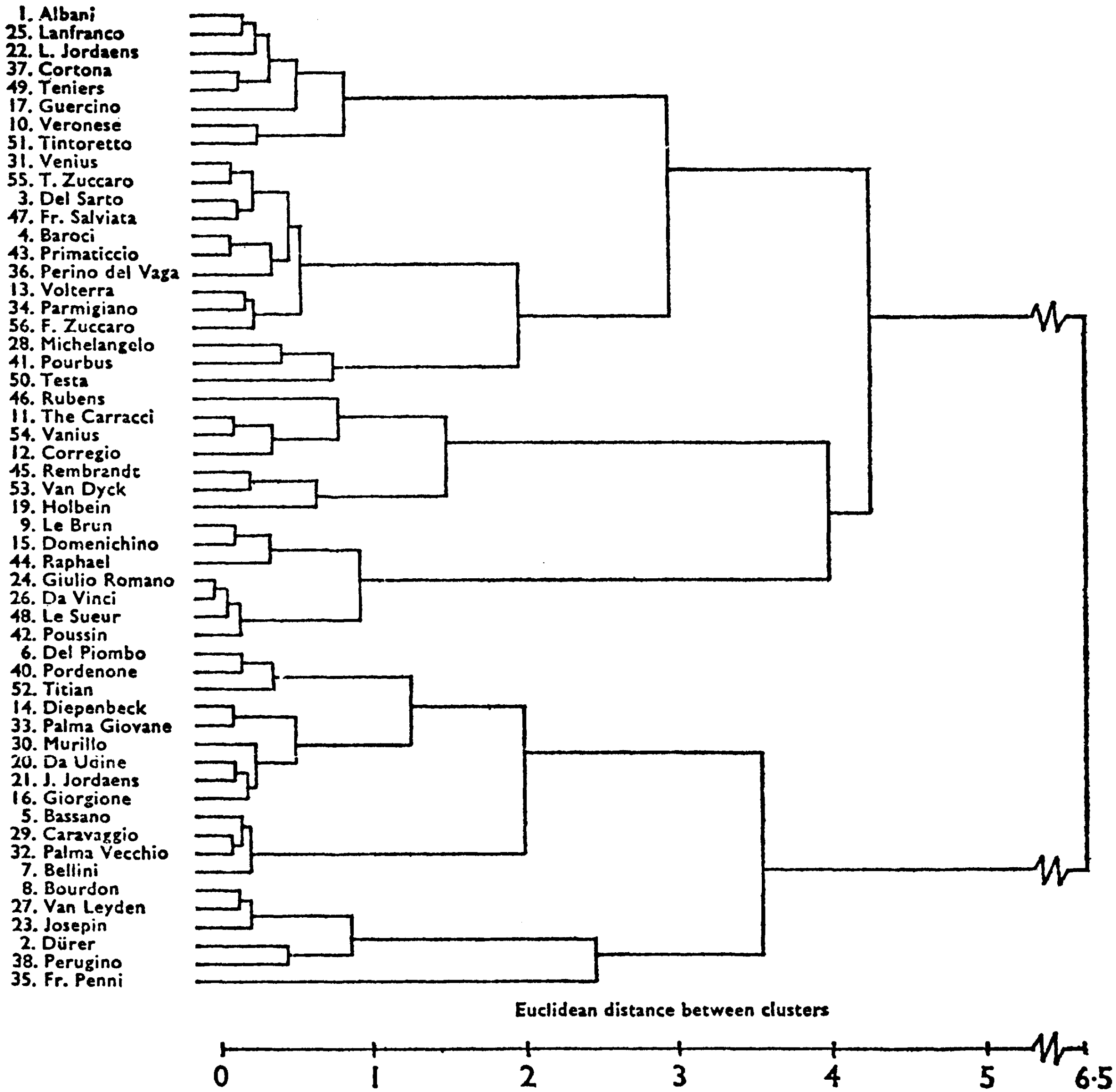


FIG. 1.

Practicalities

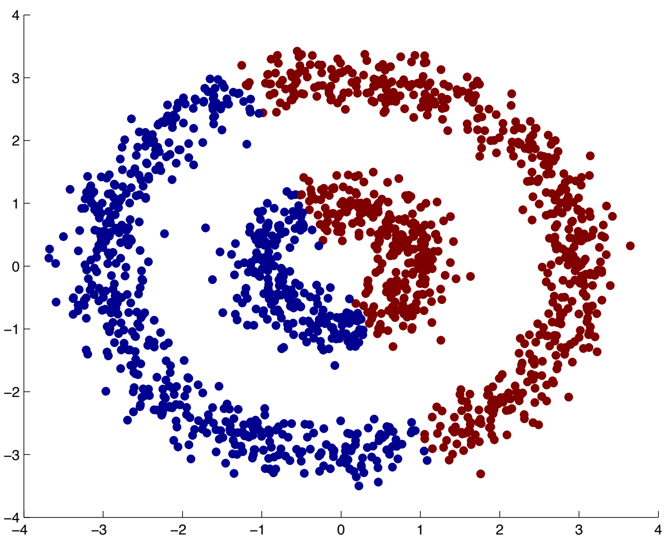
- Model selection (truncation level) is still necessary to achieve a single clustering
  - No single satisfying solution, but many of the methods discussed in *k*-means setting also apply here
- Interpretation of dendrograms difficult for large datasets
  - One solution: label each interior node with a prototype datapoint
    - Choose point with minimal maximum dissimilarity to any other point in cluster (Bien & Tibshirani, 2011: Hierarchical Clustering with Prototypes via Minimax Linkage)
    - Use minimal maximum dissimilarity as cluster dissim. measure: **minimax linkage**
    - Yields interpretable cluster summary at every level

Extensions

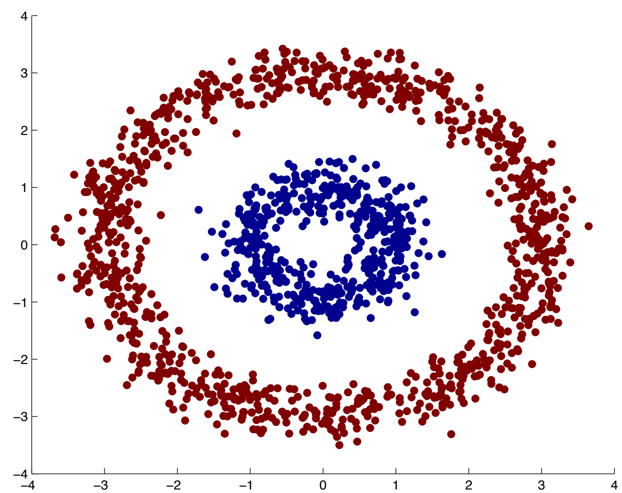
- Could use alternative measures of cluster dissimilarity, even those that do not arise from pairwise observation dissimilarity
- We have discussed **model-free** approaches to hierarchical clustering (akin to *k*-means), but **probabilistic, model-based** approaches (closer in spirit to mixture modeling) also exist

Spectral clustering

- Motivation
  - Methods like *k*-means well-suited for spherical or elliptical clusters but often fail to capture non-convex clusters
    - Example: points in concentric circles
  - **Spectral clustering** is designed for such situations, where clusters are connected but perhaps not compact



***k*-means, 2 clusters**



Spectral clustering, 2 clusters

Blackboard discussion

- See lecture notes