HIERARCHICAL CLUSTERING

As its name suggests, hierarchical clustering produces not just one partition of a dataset into clusters, but a hierarchy of clusterings.

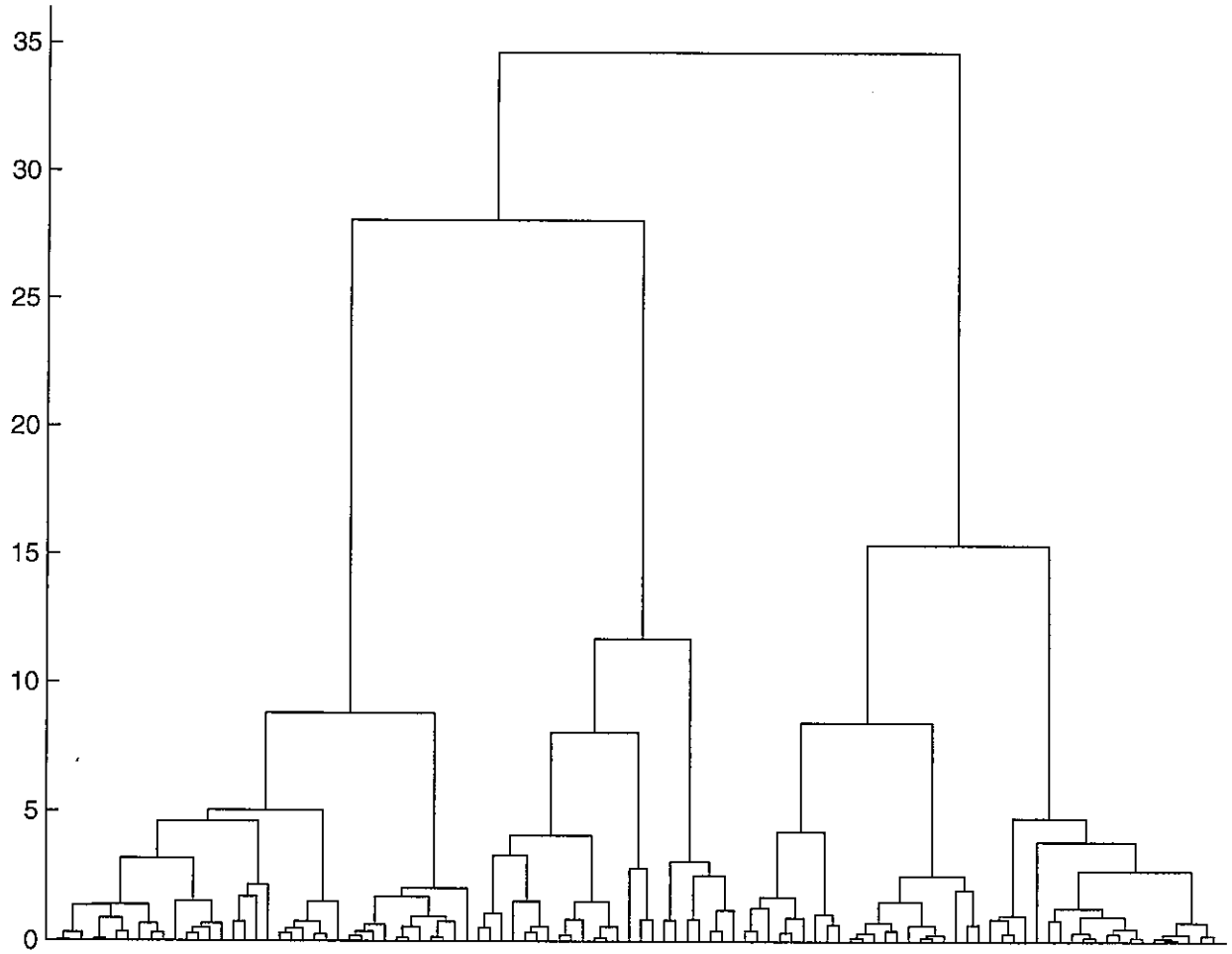
Let the data be $x_1, \ldots, x_n$.

(A) A hierarchical clustering has ___ levels. Each level corresponds to a different partition or cluster map. These levels are hierarchical in the sense that

- Level n $\rightarrow$ $\{x_1\}, \{x_2\}, \ldots, \{x_n\}$
- Level 1 $\rightarrow$ $\{x_1, \ldots, x_n\}$
- Level k, $1 \leq k < n$ $\longrightarrow$ Formed by merging two clusters at level $k+1$

The primary reason why people like hierarchical clustering is because of a graphical representation

(B) called a _____. Any horizontal line across this graph corresponds to a particular partition in the hierarchy.



- Horizontal axis: no physical meaning, just shows organization of clusters (not unique)
- Vertical axis: dissimilarity of _____ (C) clusters.

Compared to non hierarchical clustering algorithms such as K-means, hierarchical clustering has the following advantages:

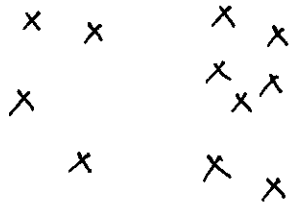
- Clusters may exist at multiple scales, i.e., clusters may have _____.

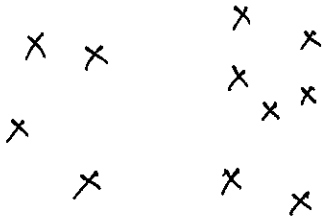
(D)

- Does not require specifying the number of clusters in advance.

There are two classes of algorithms for HC:

- bottom-up or _____.
- top-down or _____.





Agglomerative Hierarchical Clustering

Both agglomerative and divisive HC require

(E) input in the form of a _____

matrix

$$D = \left[d_{ij} \right]_{i,j=1}^{n}, \quad d_{ij} = d(x_i, x_j)$$

The dissimilarity matrix is then used to define the dissimilarity between two _____.

There are several ways to do this.

Example

Suppose A, B are clusters. We may define the dissimilarity between A and B to be

$$d_{avg}(A, B) :=$$

Note In these notes we speak of clusters as subsets of $\{x_1, \ldots, x_n\}$ as opposed to subsets of $\mathbb{R}^d$.

Agglom. HC implements the following algorithm, where

$\mathcal{H}_k$ = set of clusters at level k

Initialize $\mathcal{H}_n = \{\{x_1\}, \{x_2\}, \ldots, \{x_n\}\}$

For $k = n-1$ down to 1

- Select cluster $A, B \in \mathcal{H}_{k+1}$ for which $d(A,B)$ is _____.
- Set $\mathcal{H}_k$ to be $\mathcal{H}_{k+1}$ with A and B deleted and $A \cup B$ added ....

End

(F)

In other words, we iteratively _____ the two least dissimilar clusters until we have one cluster.

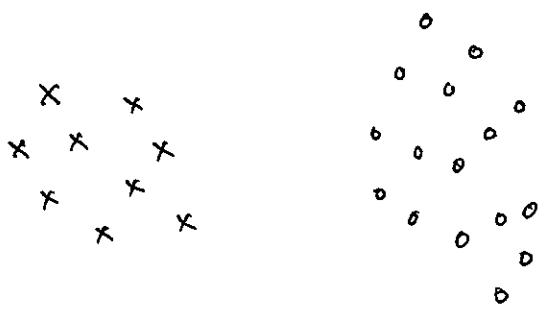
Linkage

The formula that relates point dissimilarities to cluster dissimilarities is called the

(6) _____ function.

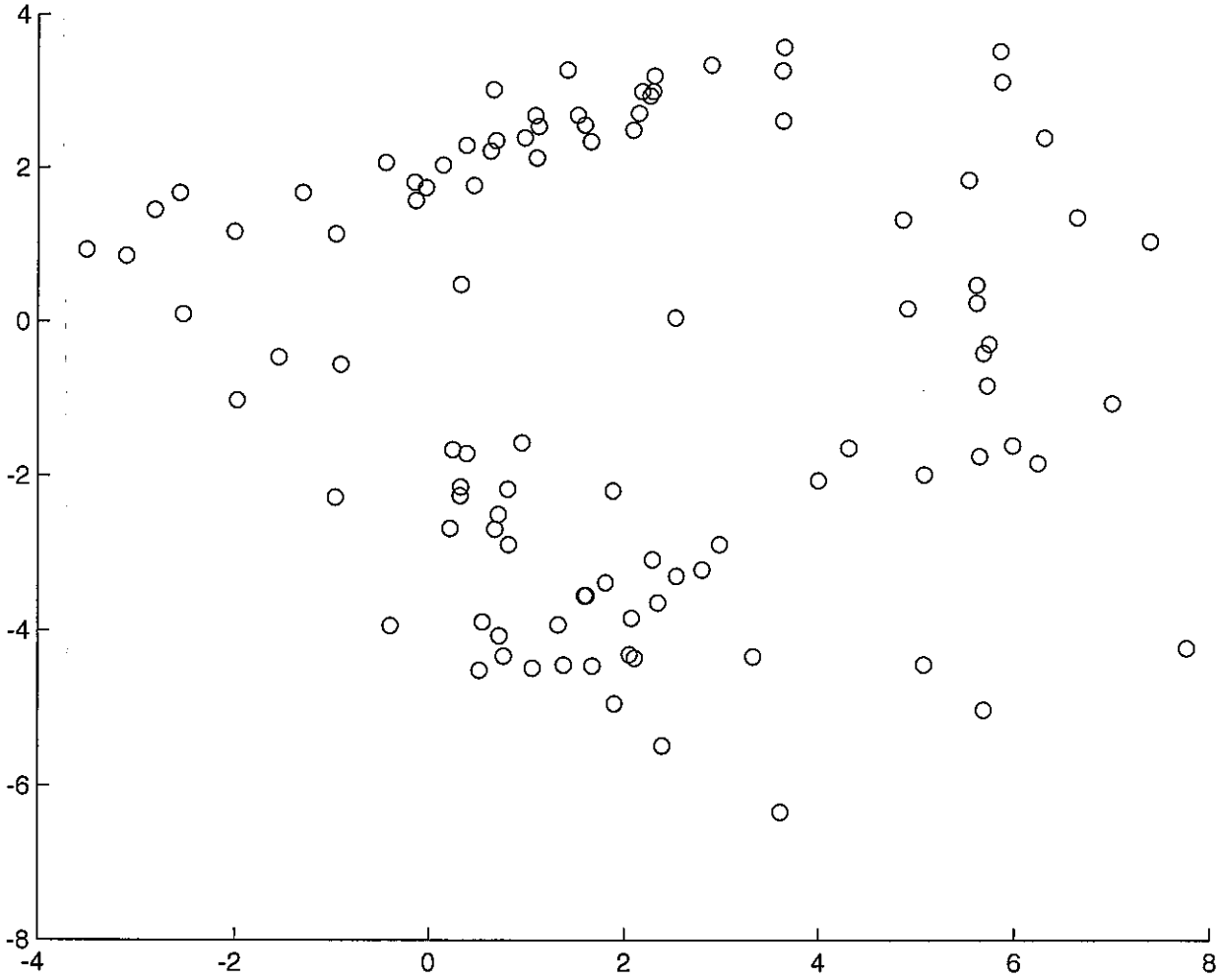
Examples

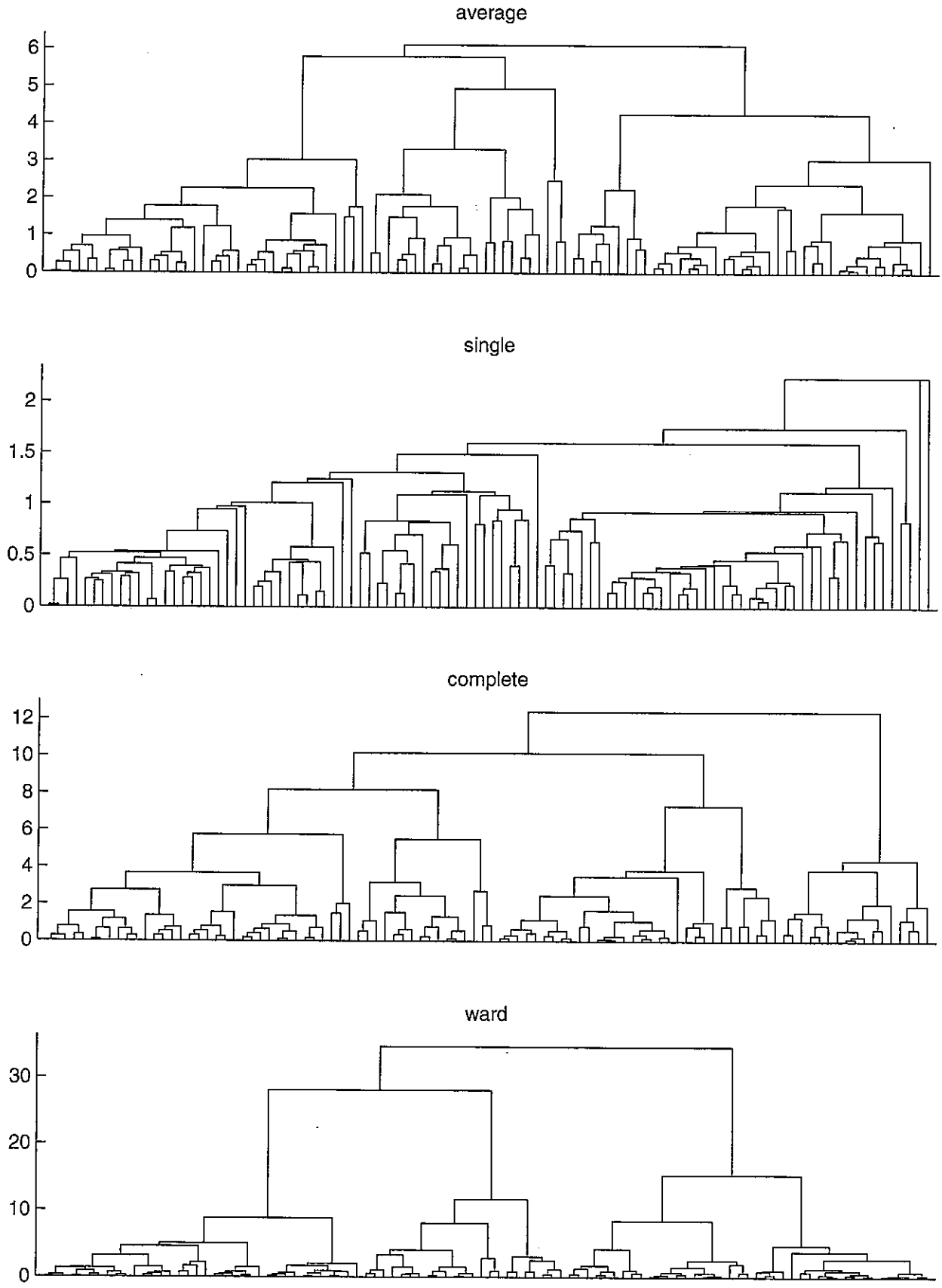
- average $d_{avg}(A,B) =$
- single $d_{min}(A,B) =$
- complete $d_{max}(A,B) =$
- centroid $d_{cent}(A,B) =$
- Ward's $d_{ward}(A,B) = \sqrt{\frac{n_A n_B}{n_A + n_B}} \| \bar{x}_A - \bar{x}_B \|$



Remarks

- The centroid and Ward's linkage are not built out of an underlying point dissimilarity
- The average, single, and complete linkages can be applied to cluster _____ data as long as point dissimilarities can be defined.
- The choice of linkage function has a major effect on the HC. Furthermore, there is often no clear choice which one to use.


4
2
0
-2
-4
-6
-8
-4
-2
0
2
4
6
8


average
6
5
4
3
2
1
0
single
2
1.5
1
0.5
0
complete
12
10
8
6
4
2
0
ward
30
20
10
0

More Remarks

- single linkage
  - (I) → generates a _____ _____ _____
  - → sensitive to outliers: tends to merge them at the very end
  - → chaining: tends to produce elongated clusters
- complete linkage
  - → discourages elongated clusters, favors clusters with small _____
- average
  - → compromise between single and complete
  - → affected by monotone scaling of d_{ij}
- centroid
  - → easy to compute
  - → dendrogram can be non-monotone
- Ward's
  - → corrects centroid's monotonicity problem

Monotonicity

Certain linkages have a monotonicity property that allows us to assign a quantitative value to the height of nodes in the

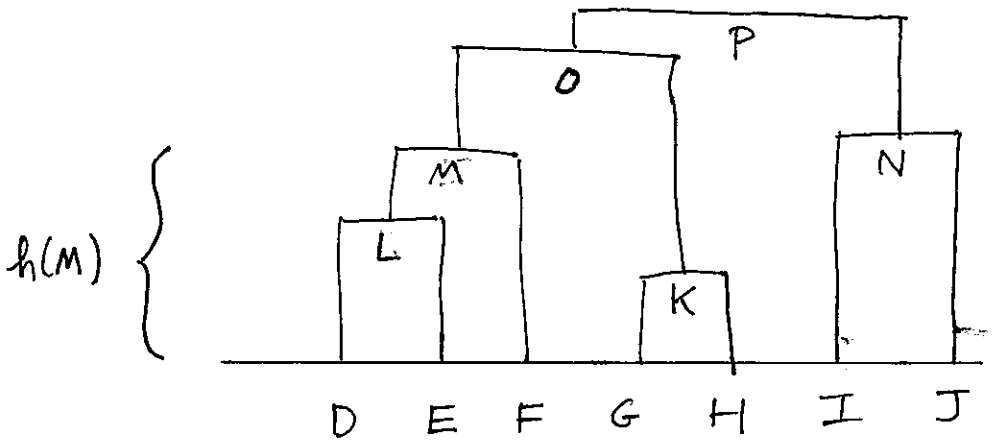
(5) _____.

In particular, suppose a node was formed by merging two clusters A and B. Then the height of $A \cup B$ is defined to be

Definition A linkage d is monotone if, for any cluster $\{A \cup B\} \cup C$ produced by HC, we have

Denotes order of merging

Consider a simple example:



(K)

Denote h = height

$h(M) =$

$h(O) =$

$h(P) =$

Example | The single linkage has the monotone property.

To see this, suppose HC produces the cluster $\{A \cup B\} \cup C$.

Then

$$d(A \cup B, C) = \min_{\substack{x \in A \cup B \\ z \in C}} d(x, z)$$

$$= \min \left\{ \min_{\substack{x \in A \\ z \in C}} d(x, z), \min_{\substack{y \in B \\ z \in C}} d(y, z) \right\}$$

(L)

$$\geq$$

- complete is monotone: same proof as for single with min → max

- average is monotone:

$$d_{avg}(A \cup B, C) = \frac{1}{n_C} \cdot \frac{1}{n_A + n_B} \sum_{z \in C} \sum_{x \in A \cup B} d(z, x)$$

$$= \frac{1}{n_C} \sum_{z \in C} \cdot \left(\frac{1}{n_A + n_B} \sum_{x \in A} d(z, x) + \frac{1}{n_A + n_B} \sum_{y \in B} d(z, y) \right)$$

$$= \frac{n_A}{n_A + n_B} d(A, C) + \frac{n_B}{n_A + n_B} d(B, C)$$

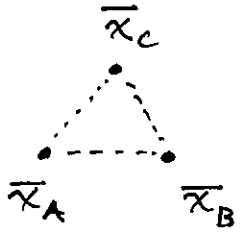
$$\geq \frac{n_A}{n_A + n_B} d(A, B) + \frac{n_B}{n_A + n_B} d(A, B)$$

[otherwise C would have merged with A or B]

$$= d(A, B)$$

- centroid is not monotone

counterexample: equilateral triangle →



- Ward's is monotone: proof based on connection to within-class scatter (see K-means lecture)

Global criterion?

HC defines a cluster to be the output of a certain algorithm.

Can we view HC as an algorithm for (approximately) optimizing a global objective function?

Let $\mathcal{J}_k$ be an objective function that assess the quality of a clustering into k clusters.

Initialize $\mathcal{H}_n = \{\{x_1\}, \{x_2\}, \ldots, \{x_n\}\}$

For $k = n-1$ down to 1

- Find $A, B \in \mathcal{H}_{k+1}$ such that merging A and B to form $\mathcal{H}_k$ yields the smallest $\mathcal{J}_k$

End

Does this _____ algorithm ever coincide with HC? Sometimes. (M)

Examples

- $d = d_{max}$ = complete linkage

$\mathcal{T}_k(\mathcal{H}) =$

- $d = d_{ward}$ = Ward's linkage

$\mathcal{T}_k(\mathcal{H}) =$

↗ requires a little algebra to verify this.

Divisive Hierarchical Clustering

Initialize $\mathcal{H}_1 = \{x_1, \ldots, x_n\}$

For $k = 2:n$

- Select a cluster $C \in \mathcal{H}_{k-1}$
- Split C into clusters A and B
- Set $\mathcal{H}_k$ to be $\mathcal{H}_{k-1}$ with C replaced by A and B

End

Comments

- Less common than agglomerative methods
- Splits must be chosen carefully to ensure a monotone dendrogram
- Can be faster than agglomeration if only a small number of clusters is desirable.

Additional Issues

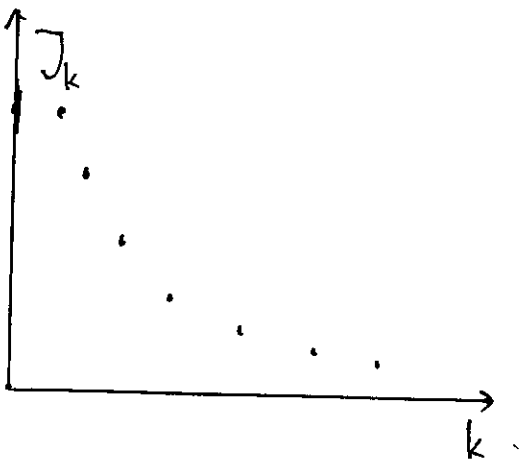
Other uses

- Initialization for other clustering methods such as K-means

Choosing K

Although HC produces a nice tree, we may want to choose a specific level. We can

- use the same method used for K-means



- look for a large jump in the dendrogram

Instability

(N) Like _____ _____, HC is sensitive to perturbations of the data

Interpretation

Dendrogram = summary of _____

≠ summary of _____

To what extent does dendrogram represent the actual structure of the data?

Model-Based Interpretation

HC may be viewed as a greedy method for maximum likelihood estimation of cluster parameters, where different generative models correspond to different linkages. See

Kamvar, Klein, and Manning, "Interpreting and Extending Classical Agglomerative Clustering Algorithms Using a Model-Based Approach."

Key A. n B. dendrogram C. child

D. subclusters ; agglomerative / divisive

E. dissimilarity , clusters , $d_{avg}(A,B) = \frac{1}{|A|\cdot|B|}\sum_{x\in A}\sum_{y\in B} d(x,y)$

F. minimal , merge G. linkage

$$d_{min}(A,B) = \min_{\substack{x\in A \\ y\in B}} d(x,y)$$

$$d_{max}(A,B) = \max_{\substack{x\in A \\ y\in B}} d(x,y)$$

$$d_{cent}(A,B) = \|\bar{x}_A - \bar{x}_B\|$$

H. non-Euclidean I. minimal spanning tree , diameter

J. dendrogram ; $d(A,B)$; $d(A\cup B, C) \geq d(A,B)$

K. $h(M) = d(L,F) = d(D\cup E, F)$

$\geq d(D,E) = h(L)$

$h(O) = d(M,K) = d(L\cup F, K)$

$\geq d(L,F) = h(M)$

$h(P) = d(O,N) \ldots \geq h(O)$ and $h(N)$

L. $\geq \min_{\substack{x\in A \\ y\in B}} d(x,y) = d(A,B)$, otherwise A, B would have merged with C

M. greedy, $\tilde{J}_k(\mathcal{H}) = \max_{A \in \mathcal{H}} \left(\max_{x,y \in A} d(x,y) \right)$

max cluster "diameter"

$\tilde{J}_k(\mathcal{H})$ = within-cluster scatter (as in K-means)

N. Decision trees ; algorithm, data