Part I: Introduction and overview

**Concept:** cluster analysis divides data into groups that are meaningful or useful.

**Managerial implications:** market segmentation

**Why CA?**

1) Summarization;

2) Compression;

3) Efficiency in finding nearest neighbors.

**Goal:** within a group, the more similarity, the better; between groups, the greater differences, the better.

**Types of classification:**

1) Hierarchical vs. partitional;

2) Exclusive vs. overlapping vs. fuzzy;

3) Complete vs. partial

**Types of clusters:**

1) Well-separated

2) prototype-based

3) graph-based

4) density-based

Part III: Agglomerative Hierarchical Clustering



Agglomerative: from individual clusters, merge the closest pairs

Divisive: from all-inclusive cluster, split a cluster until singletons

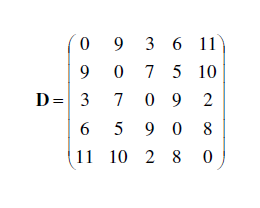
**Key Question: How to define cluster proximity?**

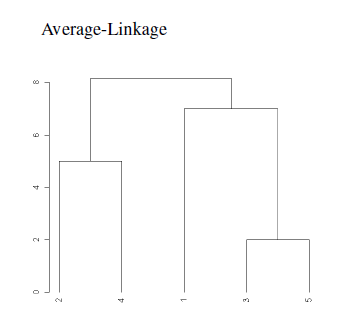
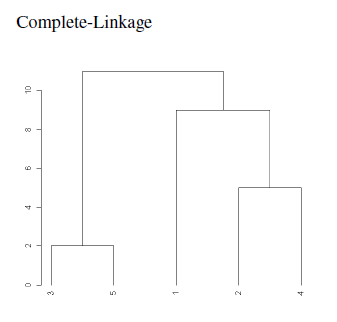
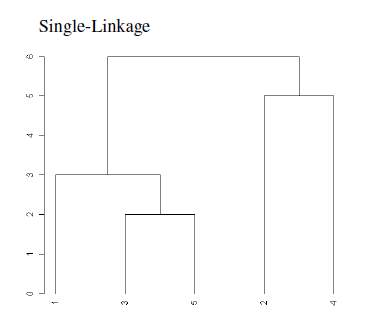
Single Link (MIN): The minimum of the distance between any two points in the two different clusters

Complete Link (MAX): The maximum of the distance between any two points in the two different clusters

Group Average: the average pairwise proximity among all pairs of points in different clusters

Proximity Matrix (Distance Matrix)





**Comparison between MIN, MAX, average**

MIN: good at non-elliptical shapes; sensitive to noise and outlier

Average: intermediate

MAX: robust to noise and outlier; break large clusters and favour globular shapes

**Ward’s Method (*Entropy in thermodynamics)***

The proximity between two clusters is defined as the increase in the squared error from merging two clusters. (similar to average method)

Variation within groups + variation between groups.

Problem: Inversions (two clusters merged may be more similar than the pairs of clusters that were merged in a previous step)

**Critiques for Hierarchical Clustering**

Lack of a global objective function (decisions are made locally)

Cluster sizes (weighted vs. unweights in terms of data points)

Merging decisions are final (Local Optimal 🡪 Global Optimal )

Expensive in computational complexity

Noise and outliers (undoing is prohibited)

Part V: Cluster Evaluation (Cluster Validation)

**Types** Supervised (Internal Indices); Unsupervised (External Indices); Relative (comparison)

**Unsupervised Evaluation**

**Graph-based view of cohesion and separation**

**Prototype-based view of cohesion and separation**

**Measuring Cluster Validity via Correlation of Proximity Matrix**

*The correlation between the similarity matrix and an ideal version of the similarity matrix based on the cluster labels*

**Hierarchical Clustering**

*CoPhenetic Correlation Coefficients*

**Number of Clusters -- SSE**

**Supervised Evaluation**

Classification-oriented: *entropy, purity, precision, recall, F-measure*

Similarity-oriented: correlation between cluster and class matrices

**General Remarks**

Cluster validity measures are hard to interpret.

It is difficult to calculate statistical significance levels for both unsupervised, supervised and relative measures.