STEVENS INSTITUTE of TECHNOLOGY
School of Business

Amir H Gandomi; PhD
Assistant Professor
Stevens Institute of Technology
a.h.gandomi@stevens.edu

Multivariate Data Analysis – BIA 652

Class 7 – Cluster Analysis

Outline – Class 8

- Introduction of Cluster Analysis – Chapter 16
  - Including K-Means and Density Clustering
- Preliminary results of the project (Due April 1)
- This week's assignment is HW: 16.6,16.8 (Due March 25)
- Additional assignment – review what we did earlier on: Matrix Algebra, Eigenvectors
- Next Class (March 25):
  - Lecture about your project writing and presentation

Corporate networking event

- Poster for Corporate networking event will held on Tuesday April 30 at Bissinger room

- Important dates:
  - poster drafts: April 10
  - final versions: April 20
  - Poster event: April 30

- Last event was in November, when we had over 70 industry folk and about 8 companies interviewing.

Project Presentation

- **For BIA-652-A:**
  - Oral presentation in class is mandatory and poster presentation at Corporate networking event is optional (Extra Credit)

- **For BIA-652-WS:**
  - Poster presentation at Corporate networking event is mandatory and oral presentation in class is optional (Extra Credit)

Where we are:

- If there is an outcome variable:
  - Perform a classification or regression analysis
- (Now) To group observations:
  - Perform Cluster Analysis
- (Later) To restructure a group of variables:
  - Perform PCA or Factor Analysis

Clustering

- A process of Grouping observations
  - Observations in the same group call cluster
  - Clustering is based on the similarities!

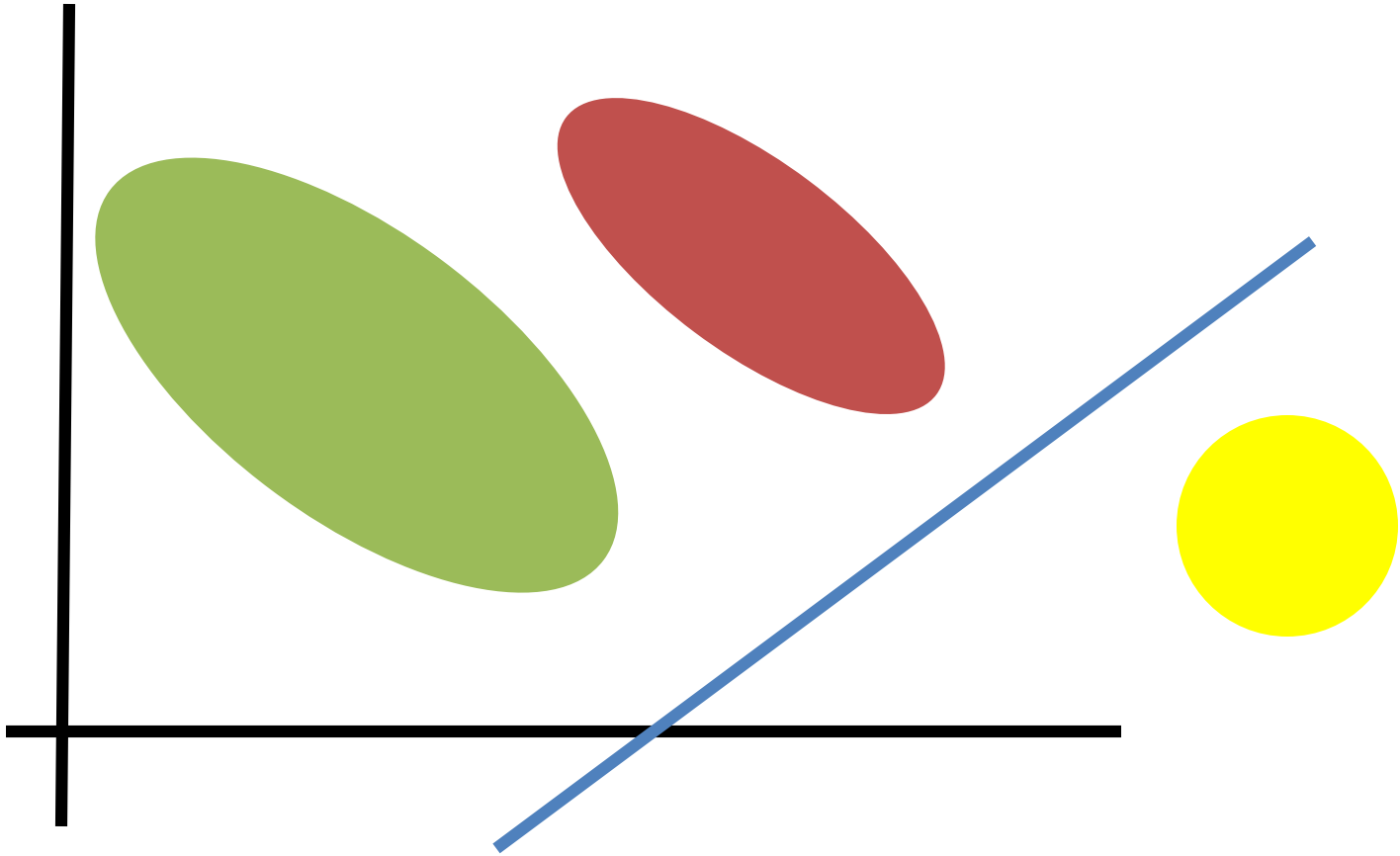
- Cluster analysis itself is not a specific algorithm, and it is could be an algorithm which tries to find undetected relationships within data to perform clustering

- Application in many different fields:
  - Business
  - Bioinformatics

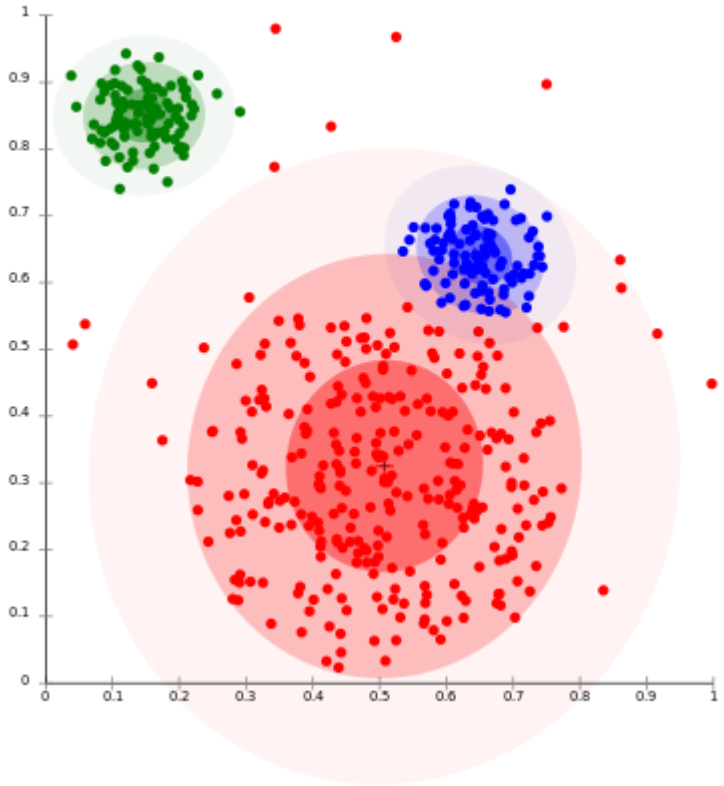
Cluster Analysis

- Graphical Cluster Analysis
- Analytical clustering techniques
  - Hierarchical Cluster Analysis
  - K-means Cluster Analysis
  - Density Cluster Analysis

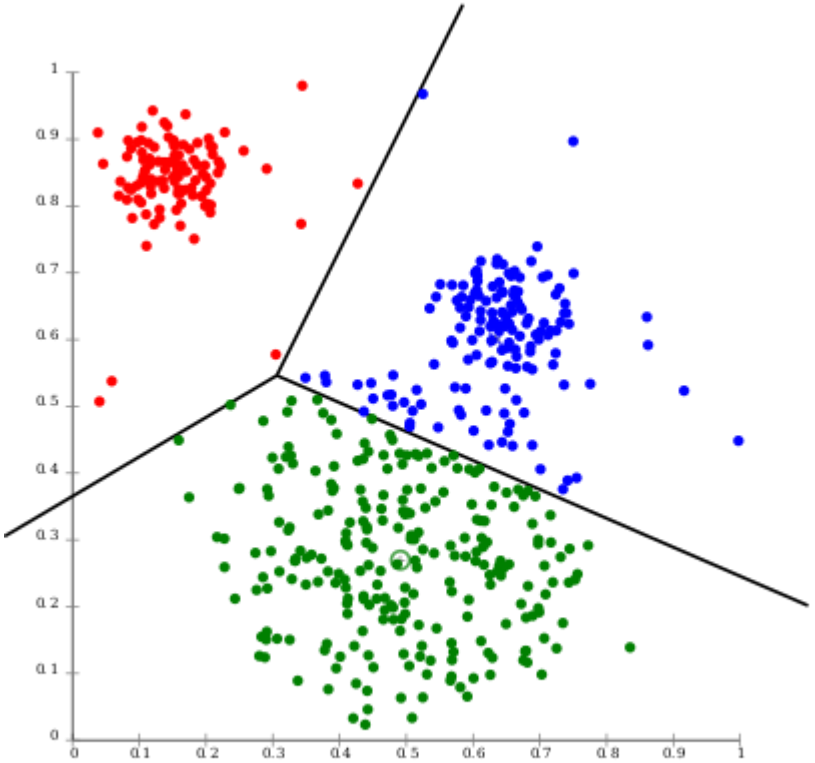
Recall Discriminant Function Picture
Related, but not identical problem



Visualization



Gaussians

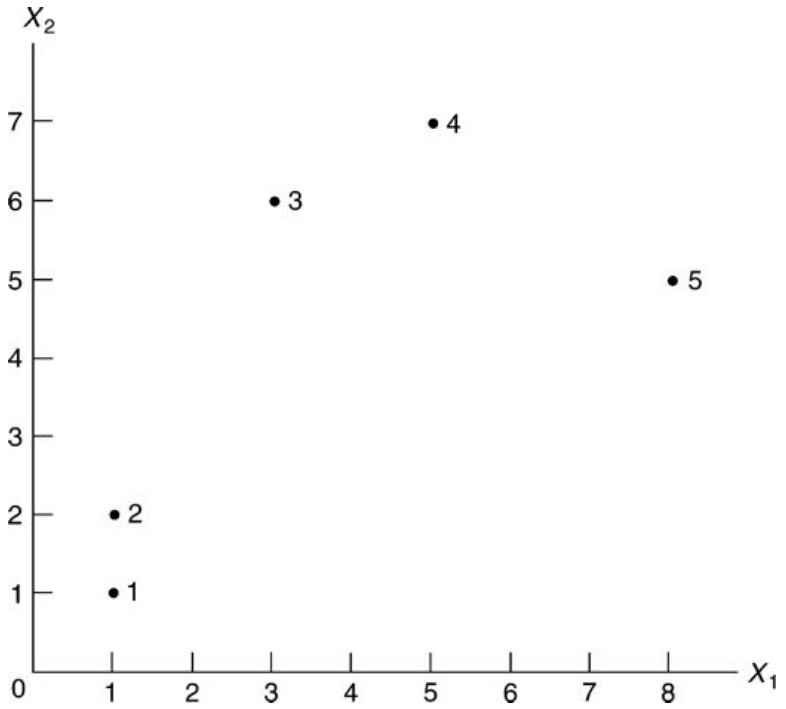


Voronoi-cells

Graphical Methods

Scatter Plots

- initial analysis
  - E.g. scatter diagram of two variables (p 406)



Distance Measures

- Power (Minkowski) Distance:

  $\{|x_{11} - x_{21}|^q + |x_{12} - x_{22}|^q\}^{1/q}$

- Special Cases:
  - Manhattan (rectilinear) Distance: q = 1
  - Euclidean Distance: q = 2

- Distance matrix
  - diagonal are all zero
  - off-diagonal entries are positive
  - It is a symmetric matrix
  - for any i and j, $x_{ij} \leq x_{ik} + x_{kj}$ for all k

Small Scale Example (p 413) Distances in R^2

Euclidian distance between five hypothetical points

	1	2	3	4	5
1	0	1.00	5.39	7.21	8.06
2	1.00	0	4.47	6.40	7.62
3	5.39	4.47	0	2.24	5.10
4	7.21	6.40	2.24	0	3.61
5	8.06	7.62	5.10	3.61	0

Profile diagram/plot

1. Standardize each variable:

$$Z = (X - m(X))/SD(X)$$

2. Compute Z's for each sample
3. Plot all Z's on a single graph

- Example (p 408):
  - Financial performance data for chemical, health, and supermarket companies (Source: *Forbes*, vol. 127, no. 1 (January 5, 1981))
  - Start with P=7 variables (X's)

Example (p 408)

Type	Symbol	Num	ROR5	D/E	SALESGR5	EPS5	NPM1	P/E	PAYOUTR1
Chem	dia	1	13.0	0.7	20.2	15.5	7.2	9	0.426398
Chem	dow	2	13.0	0.7	17.2	12.7	7.3	8	0.380693
Chem	stf	3	13.0	0.4	14.5	15.1	7.9	8	0.406780
Chem	dd	4	12.2	0.2	12.9	11.1	5.4	9	0.568182
Chem	uk	5	10.0	0.4	13.6	8.0	6.7	5	0.324544
Chem	psm	6	9.8	0.5	12.1	14.5	3.8	6	0.508083
Chem	gra	7	9.9	0.5	10.2	7.0	4.8	10	0.378913
Chem	hpc	8	10.3	0.3	11.4	8.7	4.5	9	0.481928
Chem	mtc	9	9.5	0.4	13.5	5.9	3.5	11	0.573248
Chem	acy	10	9.9	0.4	12.1	4.2	4.6	9	0.490798
Chem	cz	11	7.9	0.4	10.8	16.0	3.4	7	0.489130
Chem	ald	12	7.3	0.6	15.4	4.9	5.1	7	0.272277
Chem	rom	13	7.8	0.4	11.0	3.0	5.6	7	0.315646
Chem	rei	14	6.5	0.4	18.7	–3.1	1.3	10	0.384000
Heal	hum	15	9.2	2.7	39.8	34.4	5.8	21	0.390879
Heal	hca	16	8.9	0.9	27.8	23.5	6.7	22	0.161290
Heal	nme	17	8.4	1.2	38.7	24.6	4.9	19	0.303030
Heal	ami	18	9.0	1.1	22.1	21.9	6.0	19	0.303318
Heal	ahs	19	12.9	0.3	16.0	16.2	5.7	14	0.287500
Groc	lks	20	15.2	0.7	15.3	11.6	1.5	8	0.598930
Groc	win	21	18.4	0.2	15.0	11.6	1.6	9	0.578313
Groc	sgl	22	9.9	1.6	9.6	24.3	1.0	6	0.194946
Groc	slc	23	9.9	1.1	17.9	15.3	1.6	8	0.321070
Groc	kr	24	10.2	0.5	12.6	18.0	0.9	6	0.453731
Groc	sa	25	9.2	1.0	11.6	4.5	0.8	7	0.594966
Means			10.4	0.7	16.8	13.2	4.3	10	0.408
SD			2.6	0.5	7.9	8.4	2.2	5	0.124

Example: 25 Companies (p 410)

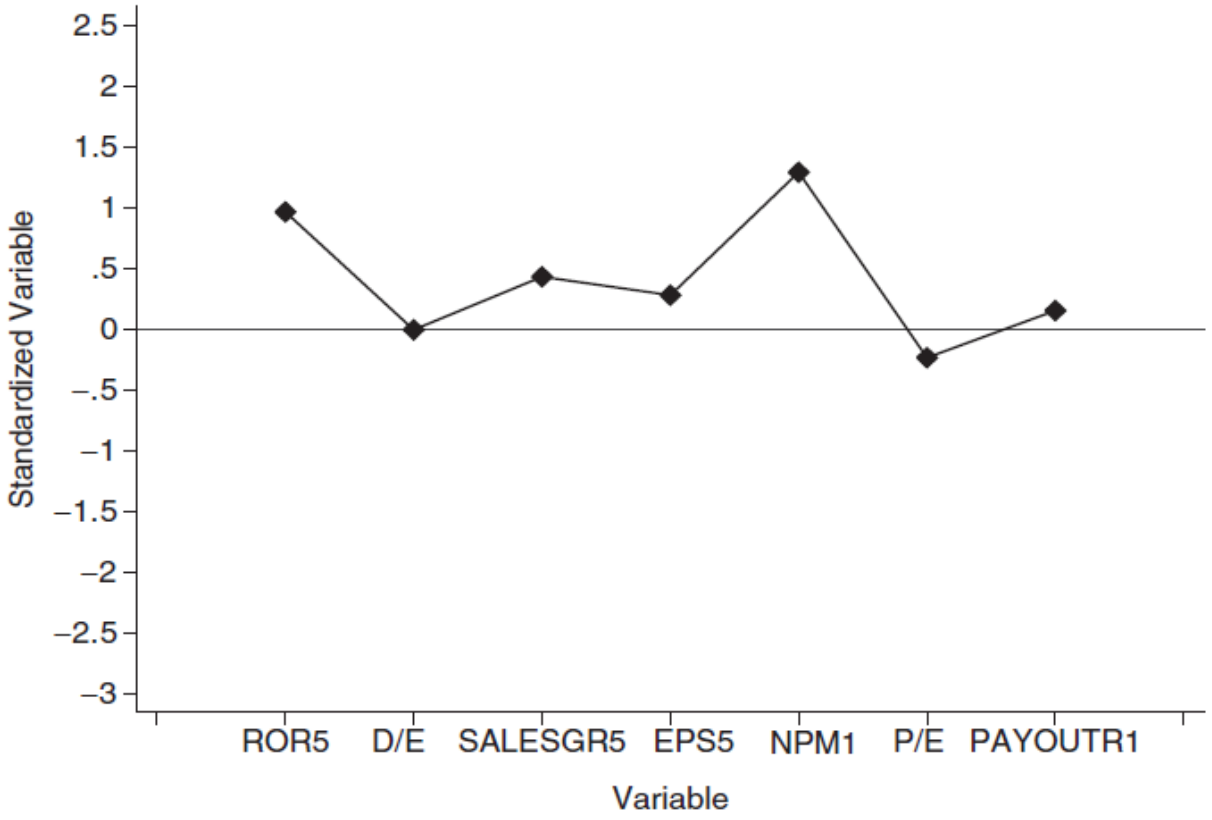
Type	Symbol	Num	ROR5	D/E	SALESGR5	EPS5	NPM1	P/E	PAYOUTR1
Chem	dia	1	0.963	−0.007	0.431	0.277	1.289	−0.237	0.151
Chem	dow	2	0.963	−0.007	0.052	−0.057	1.334	−0.442	−0.193
Chem	stf	3	0.963	−0.559	−0.290	0.230	1.601	−0.442	−0.007
Chem	dd	4	0.661	−0.927	−0.492	−0.248	0.488	−0.237	1.291
Chem	uk	5	−0.171	−0.559	−0.403	−0.618	1.067	−1.056	−0.668
Chem	psm	6	−0.246	−0.375	−0.593	0.158	−0.224	−0.851	0.807
Chem	gra	7	−0.209	−0.375	−0.833	−0.737	−0.221	−0.033	−0.231
Chem	hpc	8	−0.057	−0.743	−0.681	−0.534	0.087	−0.237	0.597
Chem	mtc	9	−0.360	−0.559	−0.416	−0.869	−0.358	0.172	1.331
Chem	acy	10	−0.209	−0.559	−0.593	−1.072	0.132	−0.237	0.668
Chem	cz	11	−0.964	−0.589	−0.757	0.337	−0.402	−0.647	0.655
Chem	ald	12	−1.191	−0.191	−0.176	−0.988	0.354	−0.647	−1.089
Chem	rom	13	−1.002	−0.559	−0.732	−1.215	0.577	−0.647	−0.740
Chem	rci	14	−1.494	−0.559	0.241	−1.943	−1.337	−0.033	−0.190
Heal	hum	15	−0.473	3.672	2.908	2.534	0.666	2.218	−0.135
Heal	hca	16	−0.587	0.361	1.366	1.233	1.067	2.422	−1.981
Heal	nme	17	−0.775	0.913	2.769	1.364	0.265	1.809	−0.841
Heal	ami	18	−0.549	0.729	0.671	1.042	0.755	1.809	−0.839
Heal	ahs	19	0.925	−0.743	−0.100	0.361	0.621	0.786	0.966
Groc	lks	20	1.794	−0.007	−0.189	−0.188	−1.248	−0.442	1.538
Groc	win	21	3.004	−0.927	−0.226	−0.188	−1.204	−0.237	1.372
Groc	sgl	22	−0.209	1.649	−0.909	1.328	−1.471	−0.851	−1.710
Groc	slc	23	−0.209	0.729	0.140	0.254	−1.204	−0.442	−0.696
Groc	kr	24	−0.095	−0.375	−0.530	0.576	−1.515	−0.851	0.370
Groc	sa	25	−0.473	0.545	−0.656	−1.036	−1.560	−0.647	1.506

Example: Profile Plot for Co. #1 Standardized Variables (Z's)

FIGURE 16.3

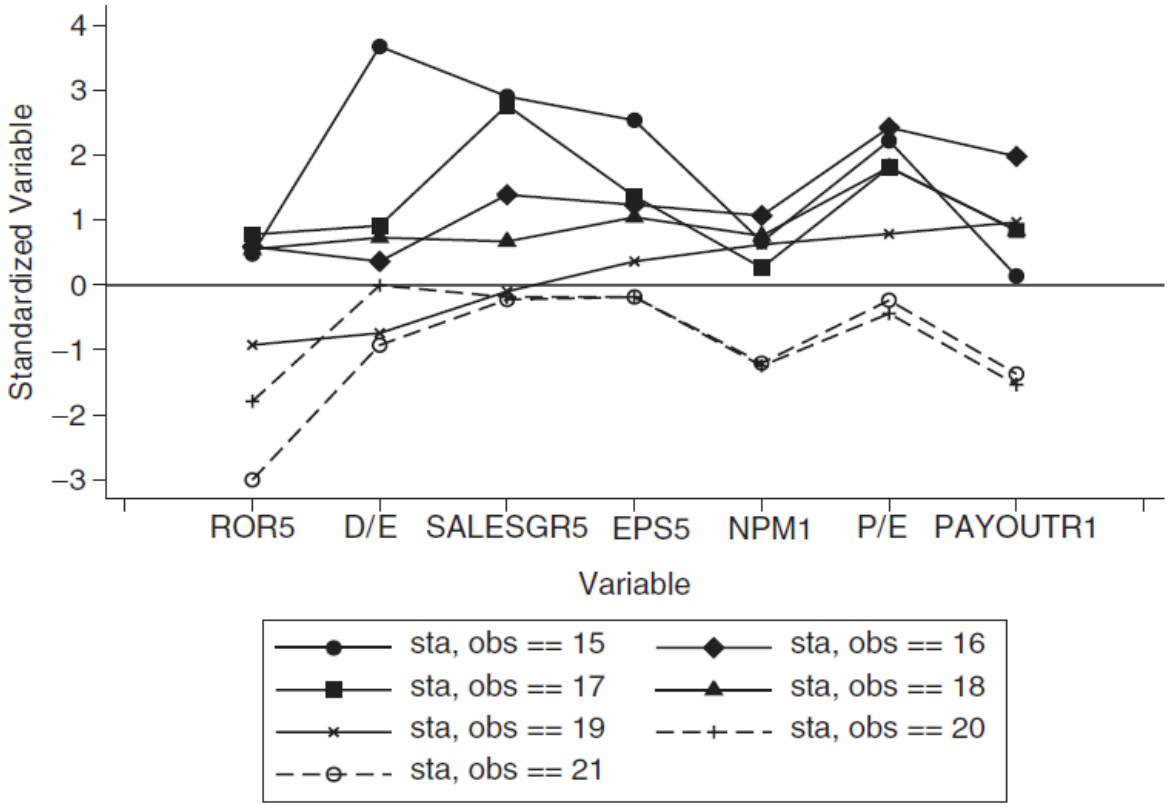
Profile Diagram of a Chemical Company (dia) Using Standardized Financial Performance Data

0.963
-0.007
0.431
0.277
1.289
-0.237
0.151



Example:
Profile Plot for several companies Standardized Variables (Z's)(p 412)

Profile Plot of Health and Supermarket Companies with Standardized Financial Performance Data



Analysis

On can distinguish 3 clusters:

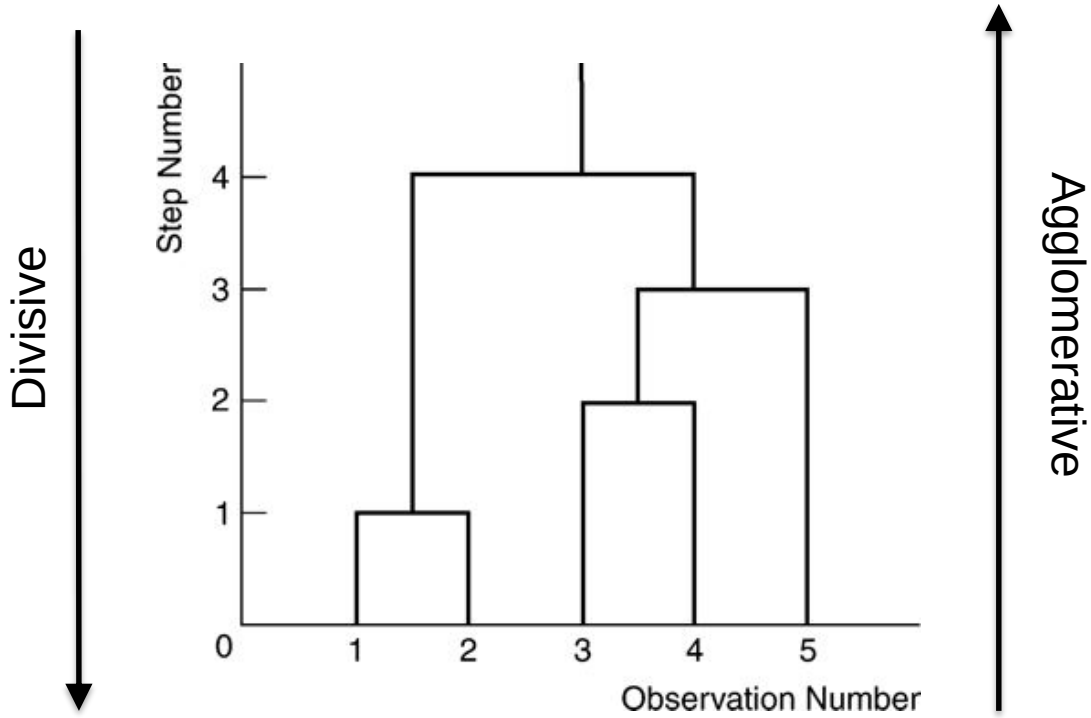
- Health companies: #15-18
- Grocery companies: #20, 21
- Hospital supply companies: #19

Analytical clustering techniques

Hierarchical Clustering

Hierarchical Methods

- **Agglomerative methods** start with *N* clusters and combine the two closest clusters, thus reducing the number of clusters by one in each step.
- **Divisive methods** start with one cluster and split off the cases that are most dissimilar to the remaining ones.

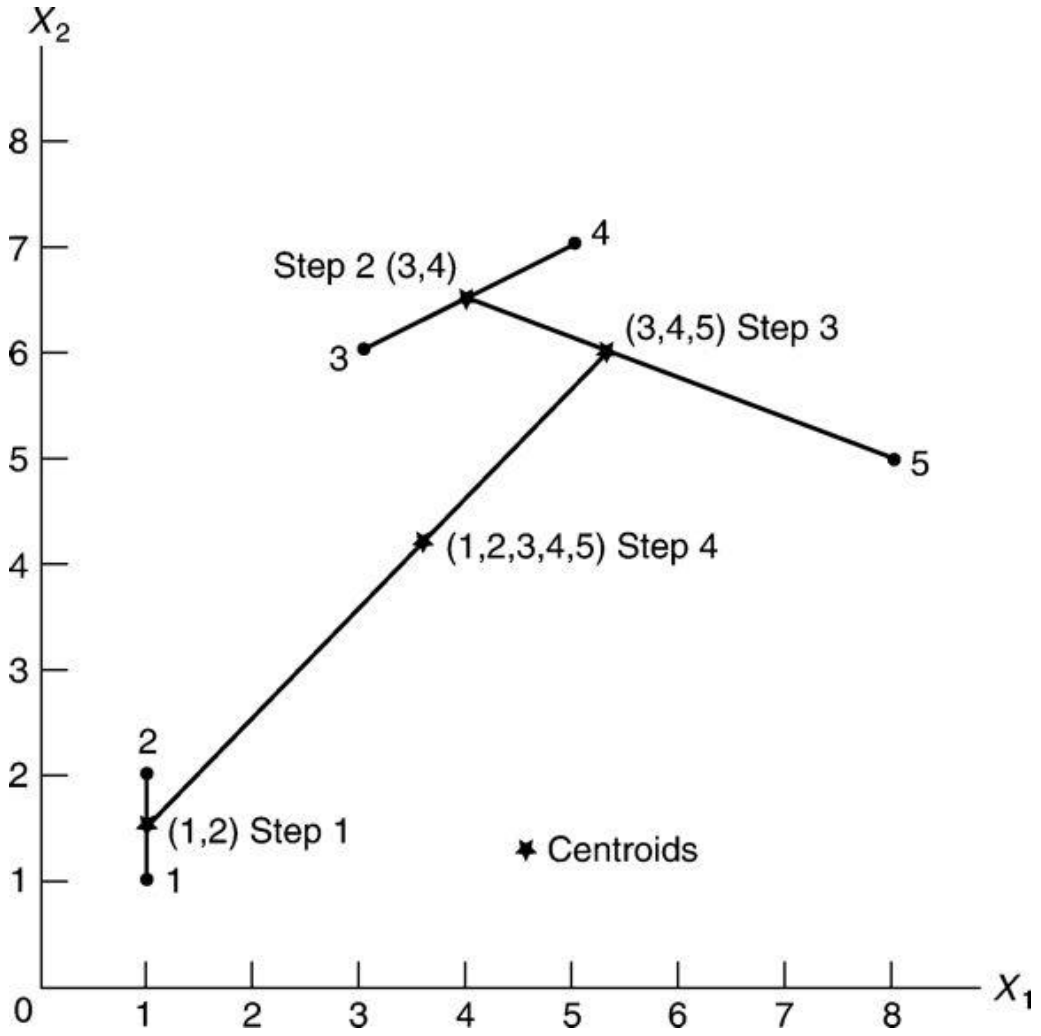


Popular Procedure

- Most of the commonly used programs are of the **agglomerative** type.
- The centroid procedure is a widely used example of agglomerative methods.
- In the centroid method the distance between two clusters is defined as the distance between the group centroids.
- The process proceeds by combining groups according to the distance between their centroids.
- The groups with the shortest distance being combined first.

Hierarchical Clustering (p 415) $O(n^3)$

Hierarchical Cluster Analysis



Steps (p 414)

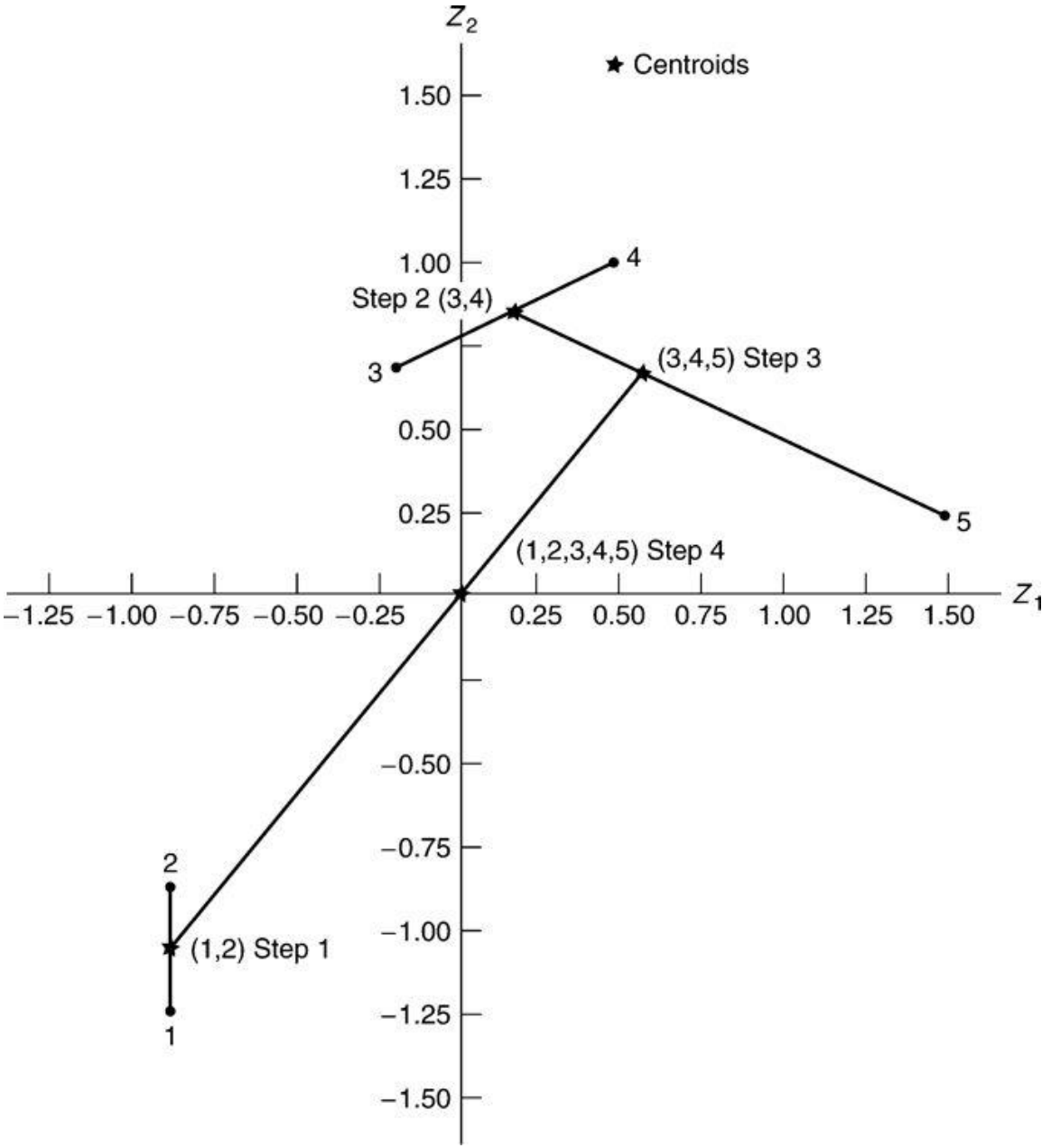
I. Initially, the closest two centroids (points) of the five hypothetical observations are points 1 and 2, so they are combined first and their centroid is obtained in step 1.

II. In step 2, centroids (points) 3 and 4 are combined (and their centroid is obtained), since they are the closest now that points 1 and 2 have been replaced by their centroid.

III. At step 3 the centroid of points 3 and 4 and centroid (point) 5 are combined, and the centroid is obtained.

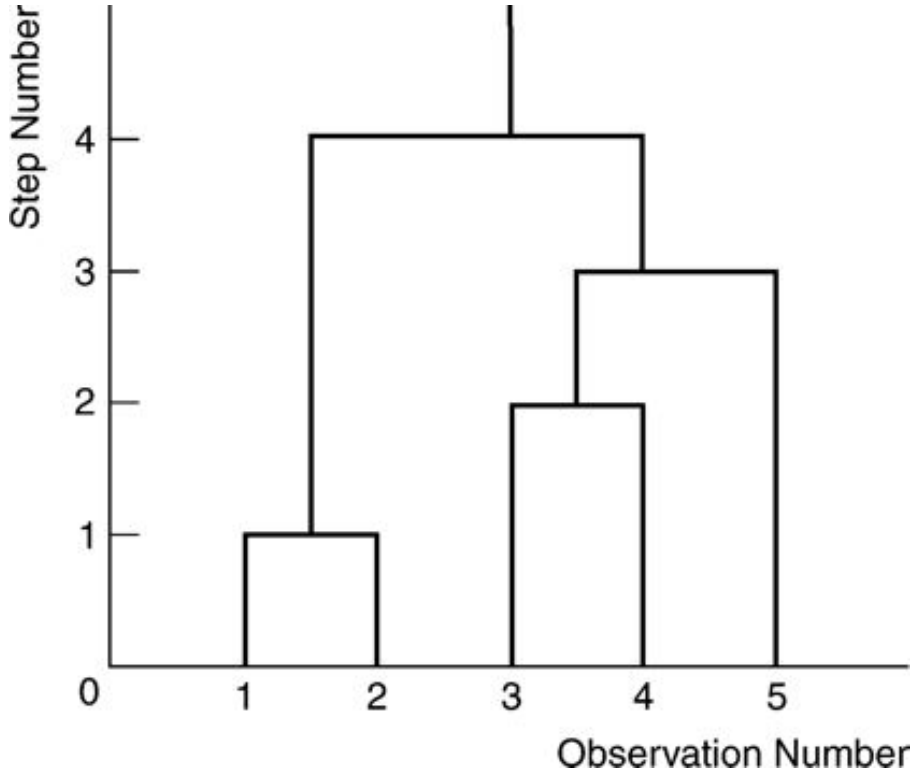
IV. Finally, at the last step the centroid of points 1 and 2 and the centroid of points 3, 4, and 5 are combined to form a single group.

Hierarchical Cluster Analysis Using Standardized Hypothetical Data Set (p 416)



Example Continued:
Dendrogram (or tree graph) for Hierarchical Cluster Analysis (p 417)

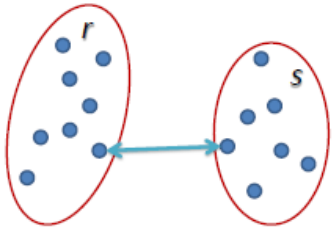
Dendrogram for Hierarchical Cluster Analysis of Hypothetical Data Set



Distance between clusters - 1

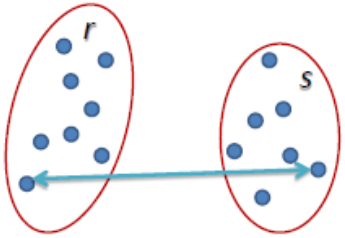
- **Single Linkage**

$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$



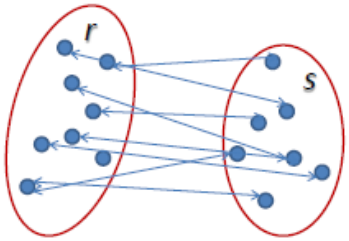
- **Complete Linkage**

$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$



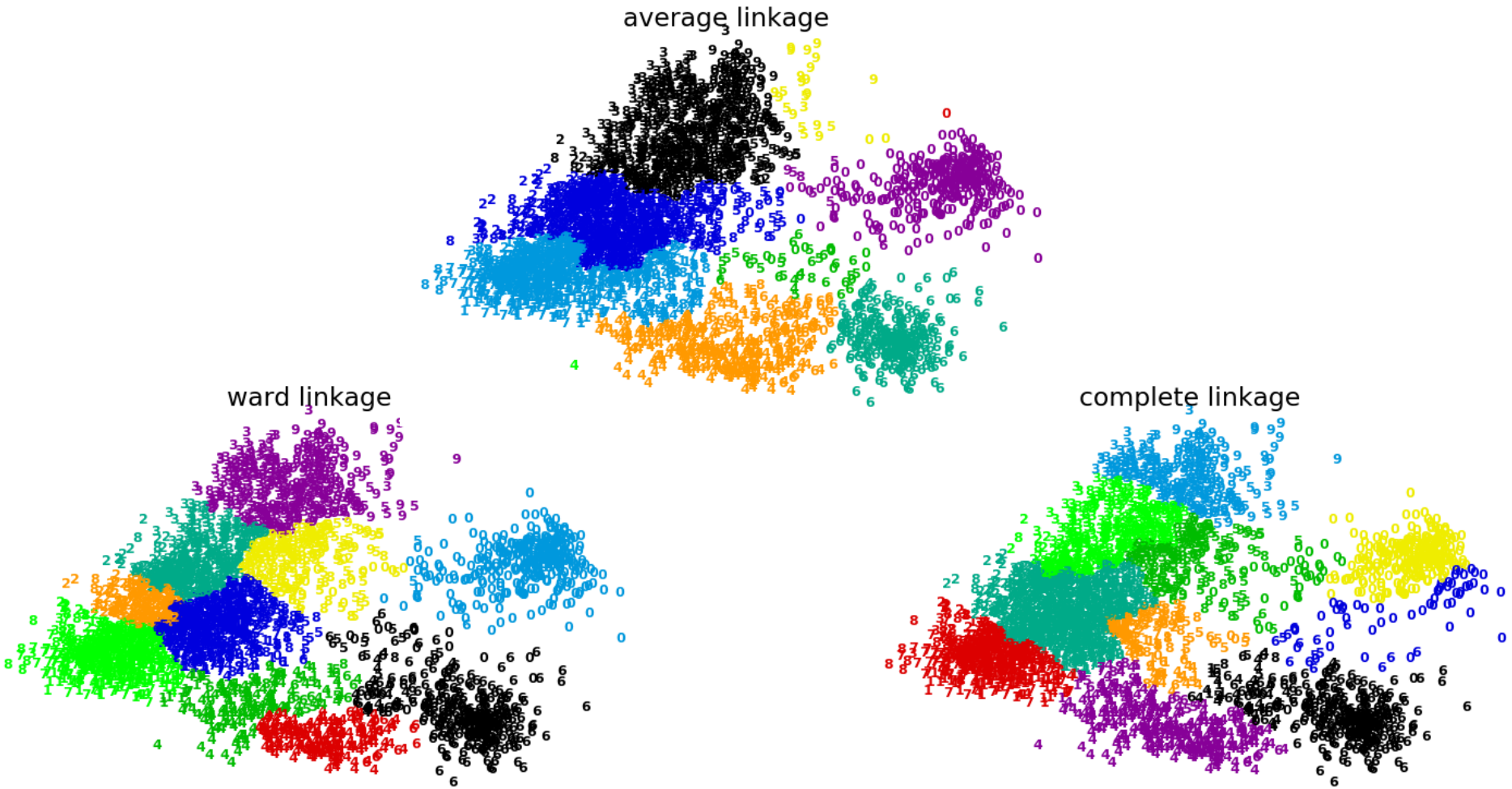
- **Average Linkage:**

$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$



- **Centroid linkage**
- **Ward's linkage: SSE between two clusters**

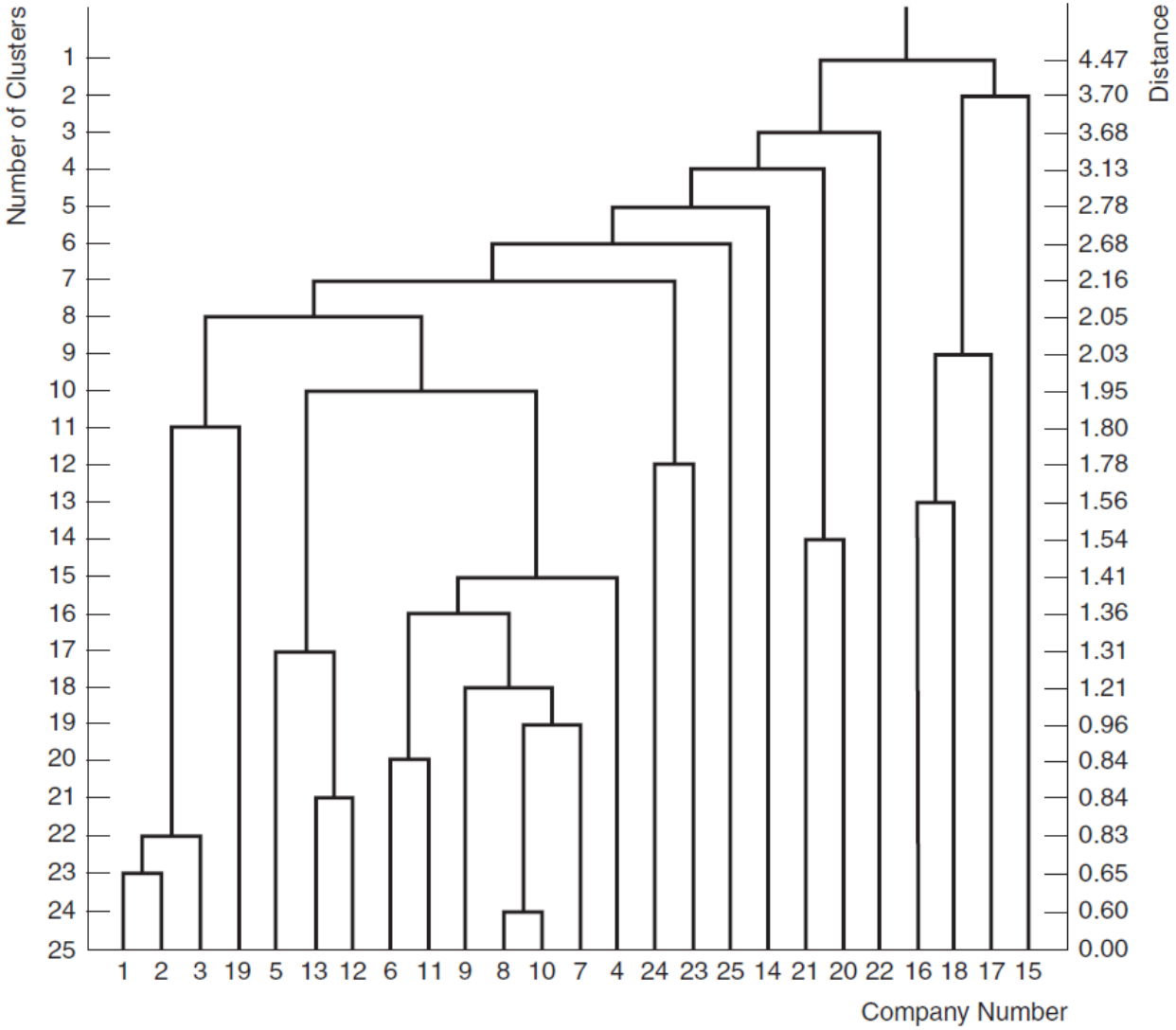
Distance between clusters - 2



Example:

FIGURE 16.9

Dendrogram of Standardized Financial Performance Data Set



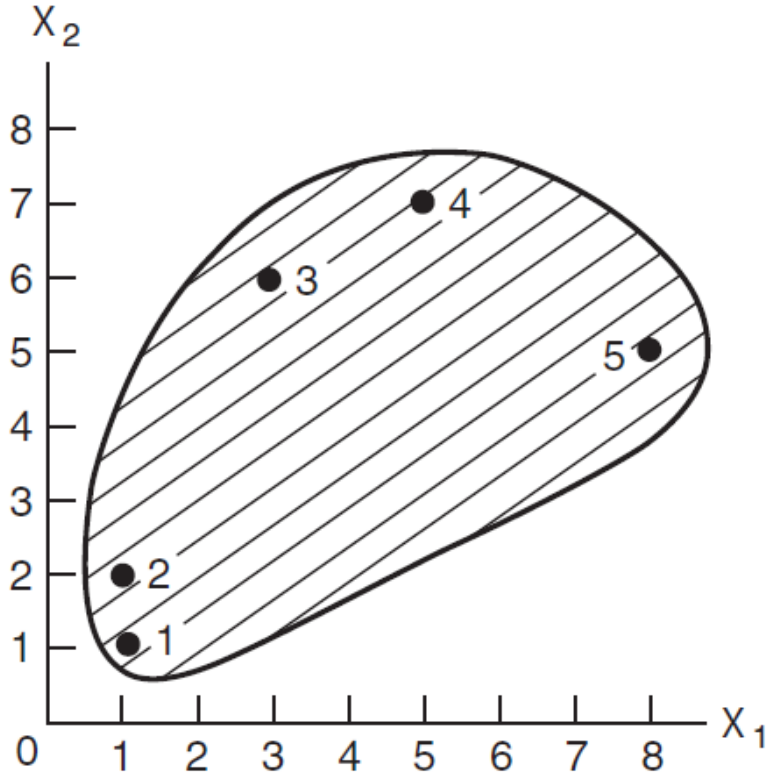
K-means Clustering

Steps of *K*-means clustering

1. Divide the data into *K* initial clusters.
2. Calculate the means or centroids of the *K* clusters.
3. For a given case, calculate its distance to each centroid. If the case is closest to the centroid of its own cluster, leave it in that cluster; otherwise, reassign it to the cluster whose centroid is closest to it.
4. Repeat step 3 for each case.
5. Repeat steps 2, 3, and 4 until no cases are reassigned.

A Small Example: K-means Clustering (NP), (p 418)

a. Start with All Points in One Cluster



K-means Clustering – Initial Step

Divide the data into K initial clusters:

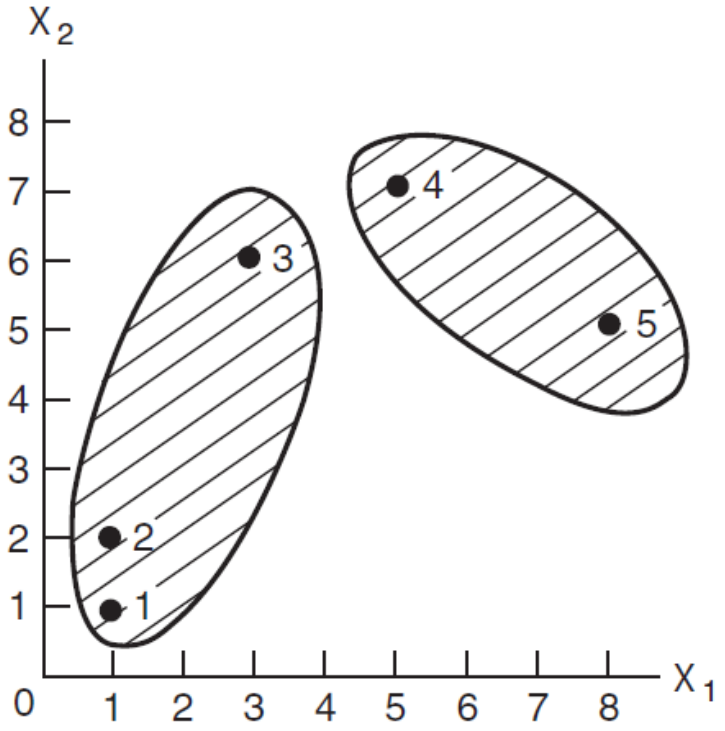
- K specified by user, or
- K specified by program

- **Disadvantage**: There is no global theorem for the optimum number of clusters

- For this small example K = 2

A Small Example: K-means Clustering Initial Step (p 418)

b. Cluster Is Split into Two Clusters at Midrange of X_1 (Variable with Largest Variance)

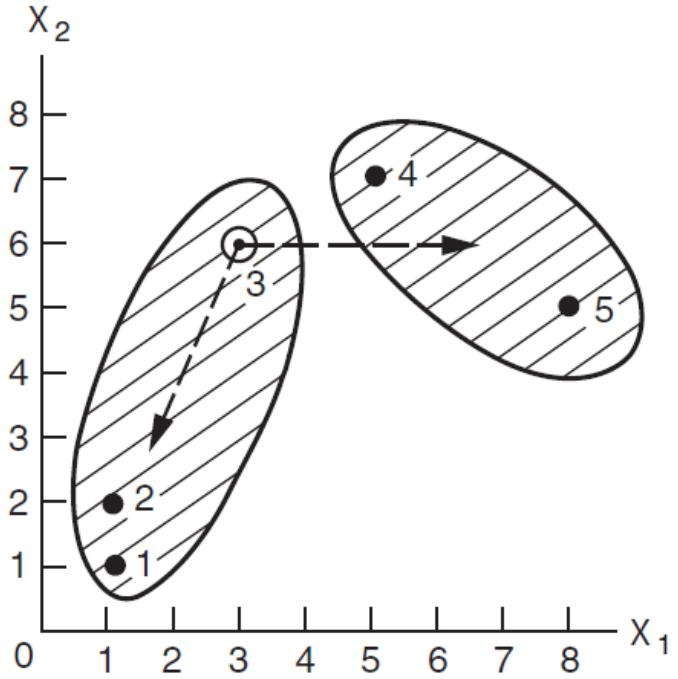


K-means Clustering – Succeeding Steps

- Calculate the means of each of the K clusters
  - Note: Means are vectors
- For each case: calculate distance to the mean of each cluster:
  - Assign case to the cluster for which it is closest to the mean.
- Repeat these two steps until no cases are reassigned.

A Small Example: K-means Clustering Succeeding Step (p 418)

c. Point 3 Is Closer to Centroid of Cluster (1,2,3) and Stays assigned to Cluster (1,2,3)

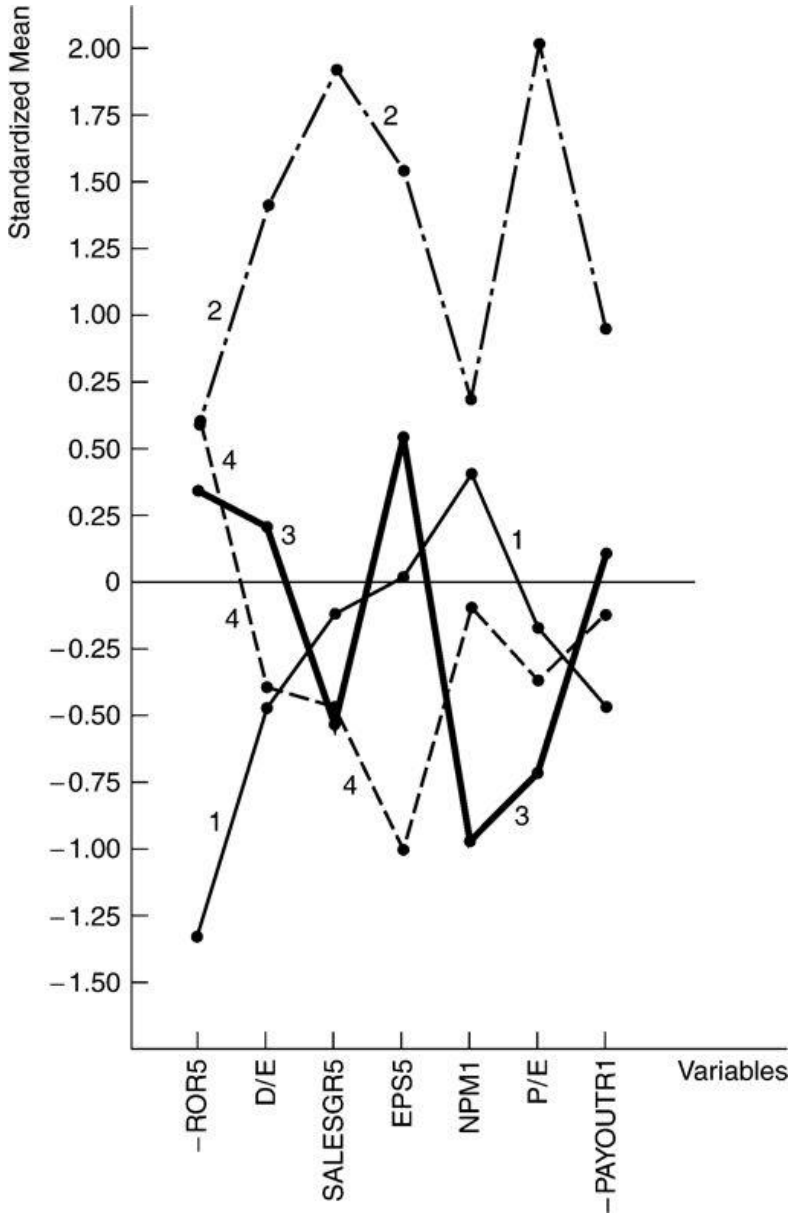


A Slightly Larger Example: 25 Companies K-means Clustering (p 422)

Type of Company	FASTCLUS *K* = 3	S-PLUS *K* = 3	Stata *K* = 3	Summary of three runs
1 Chem	1	1	1	1
2 Chem	1	1	1	1
3 Chem	1	1	1	1
4 Chem	1	1	1	1
5 Chem	3	1	1	1,3
6 Chem	1	1	1	1
7 Chem	3	1	1	1,3
8 Chem	1	1	1	1
9 Chem	1	1	1	1
10 Chem	1	1	1	1
11 Chem	3	1	3	1,3
12 Chem	3	1	1	1,3
13 Chem	3	1	1	1,3
14 Chem	3	1	3	1,3
15 Heal	2	2	2	2
16 Heal	2	2	2	2
17 Heal	2	2	2	2
18 Heal	2	2	2	2
19 Heal	1	1	1	1
20 Groc	1	3	1	1,3
21 Groc	1	3	1	1,3
22 Groc	3	1	3	1,3
23 Groc	3	1	3	1,3
24 Groc	3	1	3	1,3
25 Groc	3	1	3	1,3

Profile plots of means

Profile of Cluster Means for Four Clusters (Financial Performance Data Set) (p 424)



Other forms of Clustering:

Density Based Clustering

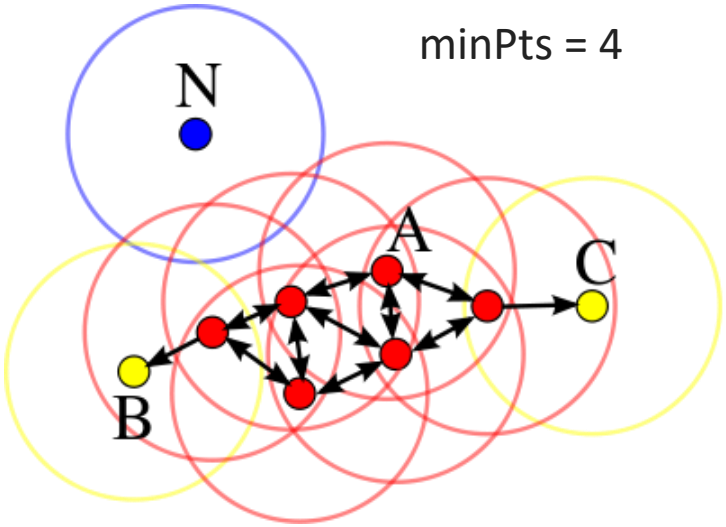
http://truecluster.com/TrueclusterExamples.pdf

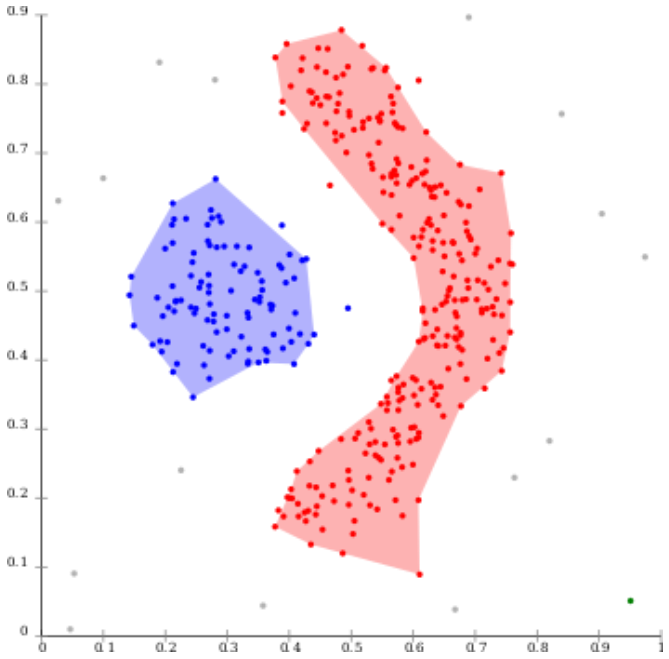
Density Based Clustering Basic Idea - $O(n^2)$

- Most common algorithms: DBSCAN, OPTICS
- Concepts:
  - Direct Density Reachability
  - Density Reachability
  - Neighborhood

- Two parameters:
  - ε : Neighborhood
  - minPts: minimum number of points

Density Based Clustering





A (red points): are core points (density-reachable)
B and C (yellow points): are border points (not density-reachable) as they have < minPts
N (blue point): it is noise as it has no point in its neighborhood

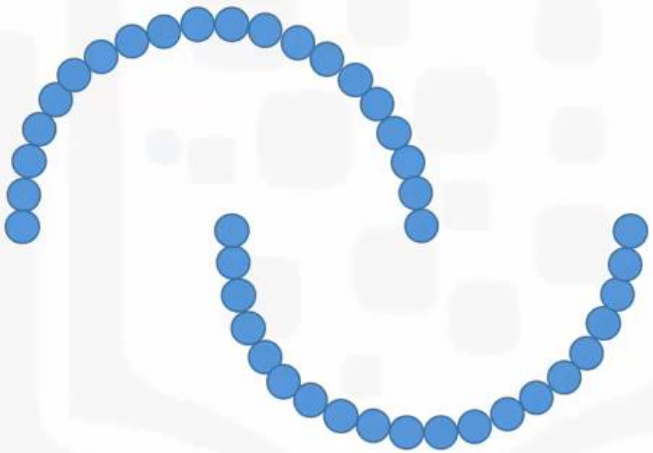
https://en.wikipedia.org

Density-based spatial clustering of applications with noise (DBSCAN)

- Algorithm:
  - Starting with a point p,
    - If there no points in P's ε-Neighborhood, then mark p as noise
    - Else if there are < minPts in P's ε-Neighborhood, then mark p as border
    - Else mark P as a core
  - Repeat for all points q in P's ε-Neighborhood, until there are no neighborhoods dense enough to add. (i.e. transitive closure)

Visualization Density Based Clustering and Comparison with K-mean

Density-Based Clustering vs K-means



BIG DATA UNIVERSITY

IBM Analytics Education 13

DBSCAN
Advantages & Disadvantages

- Advantages:
  - No need to specify number of clusters a priori
  - Can detect arbitrarily shaped clusters
  - Has a notion of **noise**
  - Requires just two parameters, and mostly insensitive to order in database
- Disadvantages:
  - “Curse of Dimensionality” for Euclidean distance.
  - Problems occur where clusters have large differences in density
  - Problems occur when clusters are very close together

Analysis in your project

- Methods:
  - Regression
  - Classification
  - Clustering

- Steps:
  - Pre-processing
  - Model development
  - Post-processing: explain and investigate the results

Per-processing

- Type of variables
- Statistical tests
- Scatter Plots
- Correlations
- Finding outliers
- imputing missing values
- Investigating the multicollinearity (for multiple regression)
- variable dependencies
- Standardizing variables
- etc!

"Perform each if required"

Modelling: Regression

- Variable selection
  - Forward selection
  - Backward elimination
  - Stepwise selection
- Using correlation coefficient and error functions (e.g. RMSE) to evaluate and compare models
- Extra Criteria to compare to models with different number of IVs
  - Mallows C_p
  - Akaike Information Criterion (AIC)
- Model Validation
- Plot the prediction results and residual errors

Modelling: Classification

- Classifiers
  - Linear Discriminant Analysis
  - K Nearest Neighbor
  - Logistic Regression
  - Naïve Bayes
  - Ensemble

- Performance Criteria
  - Cross-validation (if it is not Big Data)
  - Confusion Matrix
  - ROC Curves

Dimension Reduction

- Additional step!
- Techniques
  - Principal Component Analysis (PCA)
  - Factor Analysis
- You can find the optimum number of reduction!
- Two classes (April 1 and 8)

Post-processing

- Visualize as much as possible!
- Interpret the results
  - For example: LG model shows the first variable is more important than the second variable which is expected based on nature of the problem/variable.
- Use the concept/assumptions to explain the results
  - For example: method A outperforms method B because assumption of the normality of variables in method B is not satisfied.

STEVENS
INSTITUTE of TECHNOLOGY
School of Business

stevens.edu

Amir H Gandomi; PhD
Assistant Professor of Information Systems
a.h.gandomi@stevens.edu