Clustering, Distance Methods, and Ordination

Shyh-Kang Jeng
Department of Electrical Engineering/
Graduate Institute of Communication/
Graduate Institute of Networking and Multimedia

Outlinet

- Introduction
- Similarity Measures
- Hierarchical Clustering Methods
- Nonhierarchical Clustering Methods
- Clustering Based on Statistical Models
- Multidimensional Scaling

Outlinet

- <u>Introduction</u>
- Similarity Measures
- Hierarchical Clustering Methods
- Nonhierarchical Clustering Methods
- Clustering Based on Statistical Models
- Multidimensional Scaling

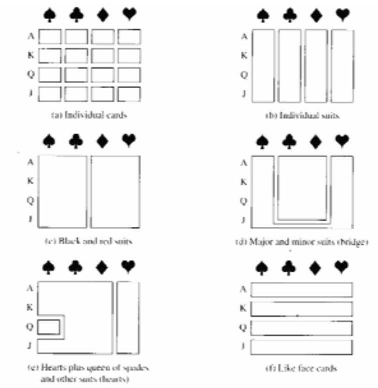
Clustering

- Searching data for a structure of "natural" groupings
- An exploratory technique
- Provides means for
  - Assessing dimensionality
  - Identifying outliers
  - Suggesting interesting hypotheses concerning relationships

Classification vs. Clustering

- Classification
  - Known number of groups
  - Assign new observations to one of these groups
- Cluster analysis
  - No assumptions on the number of groups or the group structure
  - Based on similarities or distances (dissimilarities)

Difficulty in Natural Grouping



6

Outlinet

- Introduction
- <u>Similarity Measures</u>
- Hierarchical Clustering Methods
- Nonhierarchical Clustering Methods
- Clustering Based on Statistical Models
- Multidimensional Scaling

Choice of Similarity Measure

- Nature of variables
  - Discrete, continuous, binary
- Scale of measurement
  - Nominal, ordinal, interval, ratio
- Subject matter knowledge
- Items: proximity indicated by some sort of distance
- Variables: grouped by correlation coefficient or measures of association

Some Well-known Distances

- Euclidean distance

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$$

- Statistical distance

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\mathbf{A}(\mathbf{x}-\mathbf{y})}$$

- Minkowski metric

$$d(\mathbf{x},\mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{1/m}$$

Two Popular Measures of Distance for Nonnegative Variables

- Canberra metric

$$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{x_i + y_i}$$

- Czekanowski coefficient

$$d(\mathbf{x},\mathbf{y}) = 1 - \frac{2\sum_{i=1}^{p}\min(x_i, y_i)}{\sum_{i=1}^{p}(x_i + y_i)}$$

A Caveat

- Use "true" distances when possible
  - i.e., distances satisfying distance properties
- Most clustering algorithms will accept subjectively assigned distance numbers that may not satisfy, for example, the triangle inequality

Example of Binary Variable

	Variable				
	1	2	3	4	5
Item *i*	1	0	0	1	1
Item *j*	1	1	0	1	0

Squared Euclidean Distance for Binary Variables

- Squared Euclidean distance

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sum_{j=1}^{p} \left(x_{ij} - x_{kj}\right)^2$$

$$\left(x_{ij} - x_{kj}\right)^2 = \begin{cases} 0, \text{ if } x_{ij} = x_{kj} = 1 \text{ or } x_{ij} = x_{kj} = 0 \\ 1, \text{ if } x_{ij} \neq x_{kj} \end{cases}$$

- Suffers from weighting the 1-1 and 0-0 matches equally
  - e.g., two people both read ancient Greek is stronger evidence of similarity than the absence of this capability

Contingency Table

		Item k		Totals
		1	0	
Item i	1	a	b	$a + b$
	0	c	d	$c + d$
Totals		$a+c$	$b+d$	$p = a + b + c + d$

Some Binary Similarity Coefficients

Similarity Coefficients for Clustering Items*

Coefficient	Rationale
1. $\frac{a+d}{p}$	Equal weights for 1–1 matches and 0–0 matches.
2. $\frac{2(a+d)}{2(a+d)+b+c}$	Double weight for 1–1 matches and 0–0 matches.
3. $\frac{a+d}{a+d+2(b+c)}$	Double weight for unmatched pairs.
4. $\frac{a}{p}$	No 0–0 matches in numerator.
5. $\frac{a}{a+b+c}$	No 0–0 matches in numerator or denominator. (The 0–0 matches are treated as irrelevant.)
6. $\frac{2a}{2a+b+c}$	No 0–0 matches in numerator or denominator. Double weight for 1–1 matches.
7. $\frac{a}{a+2(b+c)}$	No 0–0 matches in numerator or denominator. Double weight for unmatched pairs.
8. $\frac{a}{b+c}$	Ratio of matches to mismatches with 0–0 matches excluded.

*[p binary variables; see (12-7).]

15

Example 12.1

	Height	Weight	Eye color	Hair color	Handedness	Gender
Individual 1	68 in	140 lb	green	blond	right	female
Individual 2	73 in	185 lb	brown	brown	right	male
Individual 3	67 in	165 lb	blue	blond	right	male
Individual 4	64 in	120 lb	brown	brown	right	female
Individual 5	76 in	210 lb	brown	brown	left	male

Example 12.1

$$X_1 = \begin{cases} 1 & \text{height} \ge 72 \text{ in.} \\ 0 & \text{height} < 72 \text{ in.} \end{cases} \quad X_4 = \begin{cases} 1 & \text{blond hair} \\ 0 & \text{not blond hair} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{weight} \ge 150 \text{ lb} \\ 0 & \text{weight} < 150 \text{ lb} \end{cases} \quad X_5 = \begin{cases} 1 & \text{right handed} \\ 0 & \text{left handed} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{brown eyes} \\ 0 & \text{otherwise} \end{cases} \quad X_6 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

Example 12.1

		X_1	X_2	X_3	X_4	X_5	X_6
Individual	1	0	0	0	1	1	1
	2	1	1	1	0	1	0

		Individual 2		
		1	0	Total
Individual 1	1	1	2	3
	0	3	0	3
	Totals	4	2	6

Example 12.1: Similarity Matrix with Coefficient 1

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & & & & \\ 1/6 & 1 & & & \\ 4/6 & 3/6 & 1 & & \\ 4/6 & 3/6 & 2/6 & 1 & \\ 0 & 5/6 & 2/6 & 2/6 & 1 \end{bmatrix} \end{array}$$

Conversion of Similarities and Distances

- Similarities from distances
  - e.g., $\tilde{s}_{ik} = 1/(1+d_{ik})$
- "True" distances from similarities
  - Matrix of similarities must be nonnegative definite
  - e.g., $d_{ik} = \sqrt{2(1-\tilde{s}_{ik})}, \quad \tilde{s}_{ii} = 1$

Contingency Table

		Variable k		Totals
		1	0	
Variable i	1	a	b	$a + b$
	0	c	d	$c + d$
Totals		$a+c$	$b+d$	$n = a + b + c + d$

Product Moment Correlation as a Measure of Similarity

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

- Related to the chi-square statistic ($r^2 = \chi^2/n$) for testing independence
  - For n fixed, large similarity is consistent with presence of dependence

22

Example 12.2 Similarities of 11 Languages

Numerals in 11 Languages

English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neljä
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

23

Example 12.2 Similarities of 11 Languages

Concordant First Letters for Numbers in 11 Languages

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	10										
N	8	10									
Da	8	9	10								
Du	3	5	4	10							
G	4	6	5	5	10						
Fr	4	4	4	1	3	10					
Sp	4	4	5	1	3	8	10				
I	4	4	5	1	3	9	9	10			
P	3	3	4	0	2	5	7	6	10		
H	1	2	2	2	1	0	0	0	0	10	
Fi	1	1	1	1	1	1	1	1	1	2	10

24

Agglomerative Methods

- Initially a many clusters as objects
- The most similar objects are first grouped
- Initial groups are merged according to their similarities
- Eventually, all subgroups are fused into a single cluster

25

Outlinet

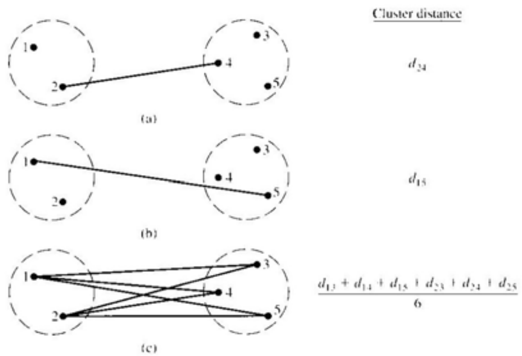
- Introduction
- Similarity Measures
- <u>Hierarchical Clustering Methods</u>
- Nonhierarchical Clustering Methods
- Clustering Based on Statistical Models
- Multidimensional Scaling

Divisive Methods

- Initial single group is divided into two subgroups such that objects in one subgroup are “far from” objects in the other
- These subgroups are then further divided into dissimilar subgroups
- Continues until there are as many subgroups as objects

27

Inter-cluster Distance for Linkage Methods



:8

Example 12.3: Single Linkage

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{bmatrix} \end{array}$$

29

Example 12.3: Single Linkage

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ \textcircled{3} & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

30

Example 12.3: Single Linkage

$$\begin{array}{c} \\ (135) \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (135) & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & \textcircled{5} & 0 \end{bmatrix} \end{array}$$

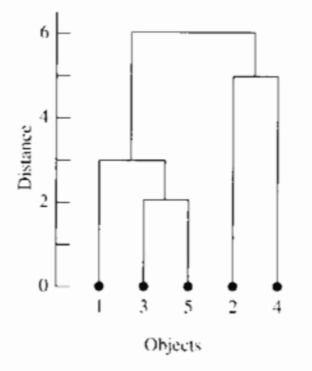
31

Example 12.3: Single Linkage

$$\begin{array}{c} \\ (135) \\ (24) \end{array} \begin{array}{c} \begin{array}{cc} (135) & (24) \end{array} \\ \begin{bmatrix} 0 & \\ \textcircled{6} & 0 \end{bmatrix} \end{array}$$

32

Example 12.3: Single Linkage Resultant Dendrogram



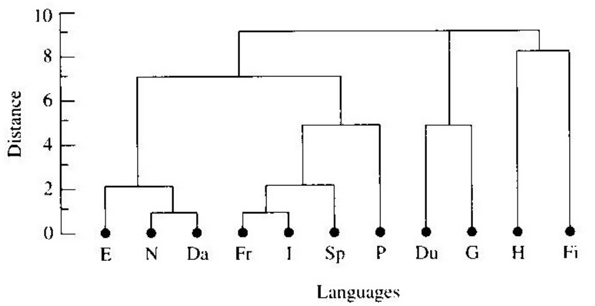
33

Example 12.4 Single Linkage of 11 Languages

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	0										
N	2	0									
Da	2	(1)	0								
Du	7	5	6	0							
G	6	4	5	5	0						
Fr	6	6	6	9	7	0					
Sp	6	6	5	9	7	2	0				
I	6	6	5	9	7	(1)	(1)	0			
P	7	7	6	10	8	5	3	4	0		
H	9	8	8	8	9	10	10	10	10	0	
Fi	9	9	9	9	9	9	9	9	9	8	0

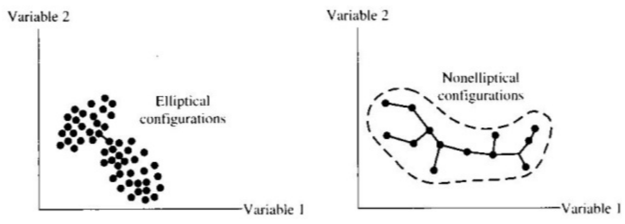
34

Example 12.4 Single Linkage of 11 Languages



35

Pros and Cons of Single Linkage



(a) Single linkage confused by near overlap

(b) Chaining effect

36

Example 12.5: Complete Linkage

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{bmatrix} \end{array}$$

37

Example 12.5: Complete Linkage

$$\begin{array}{c|cccc} & (35) & 1 & 2 & 4 \\ \hline (35) & 0 & & & \\ 1 & 11 & 0 & & \\ 2 & 10 & 9 & 0 & \\ 4 & 9 & 6 & \textcircled{5} & 0 \end{array}$$

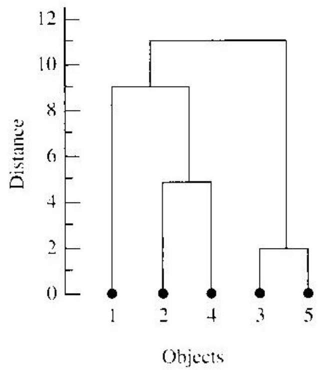
38

Example 12.5: Complete Linkage

$$\begin{array}{c|ccc} & (35) & (24) & 1 \\ \hline (35) & 0 & & \\ (24) & 10 & 0 & \\ 1 & 11 & \textcircled{9} & 0 \end{array}$$

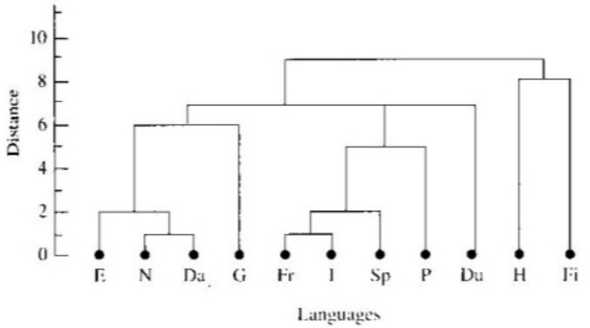
39

Example 12.5: Complete Linkage



40

Example 12.6
Complete Linkage of 11 Languages



Example 12.7
Clustering Variables

Public Utility Data (1975)

Company	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1. Arizona Public Service	1.06	9.2	151	54.4	1.6	9077	0.	.628
2. Boston Edison Co.	.89	10.3	202	57.9	2.2	5088	25.3	1.555
3. Central Louisiana Electric Co.	1.43	15.4	113	53.0	3.4	9212	0.	1.058
4. Commonwealth Edison Co.	1.02	11.2	168	56.0	.3	6423	34.3	.700
5. Consolidated Edison Co. (N.Y.)	1.49	8.8	192	51.2	1.0	3300	15.6	2.044
6. Florida Power & Light Co.	1.32	13.5	111	60.0	−2.2	11127	22.5	1.241
7. Hawaiian Electric Co.	1.22	12.2	175	67.6	2.2	7642	0.	1.652
8. Idaho Power Co.	1.10	9.2	245	57.0	3.3	13082	0.	.309
9. Kentucky Utilities Co.	1.34	13.0	168	60.4	7.2	8406	0.	.862
10. Madison Gas & Electric Co.	1.12	12.4	197	53.0	2.7	6455	39.2	.623
11. Nevada Power Co.	.75	7.5	173	51.5	6.5	17441	0.	.768
12. New England Electric Co.	1.13	10.9	178	62.0	3.7	6154	0.	1.897
13. Northern States Power Co.	1.15	12.7	199	53.7	6.4	7179	50.2	.527
14. Oklahoma Gas & Electric Co.	1.09	12.0	96	49.8	1.4	9673	0.	.588
15. Pacific Gas & Electric Co.	.96	7.6	164	62.2	−0.1	6468	.9	1.400
16. Puget Sound Power & Light Co.	1.16	9.9	252	56.0	9.2	15991	0.	.620
17. San Diego Gas & Electric Co.	.76	6.4	136	61.9	9.0	5714	8.3	1.920
18. The Southern Co.	1.05	12.6	150	56.7	2.7	10140	0.	1.108
19. Texas Utilities Co.	1.16	11.7	104	54.0	2.1	13507	0.	.636
70. Wisconsin Electric Power Co.	1.20	11.8	148	59.9	3.5	7287	41.1	.702
21. United Illuminating Co.	1.04	8.6	204	61.0	3.5	6650	0.	2.116
22. Virginia Electric & Power Co.	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

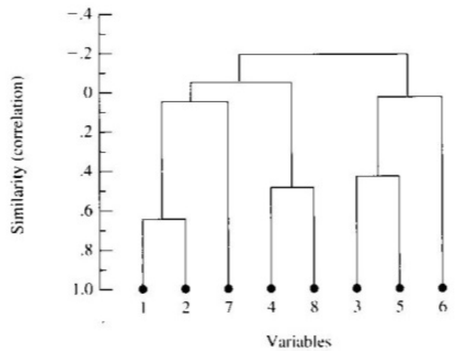
KEY: X_1: Fixed-charge coverage ratio (income/debt).
X_2: Rate of return on capital.
X_3: Cost per KW capacity in place.
X_4: Annual load factor.
X_5: Peak kWh demand growth from 1974 to 1975.
X_6: Sales (kWh use per year).
X_7: Percent nuclear.
X_8: Total fuel costs (cents per kWh).

Source: Data courtesy of H. E. Thompson.

Example 12.7
Correlations of Variables

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1.000							
.643	1.000						
−.103	−.348	1.000					
−.082	−.086	.100	1.000				
−.259	−.260	.435	.034	1.000			
−.152	−.010	.028	−.288	.176	1.000		
.045	.211	.115	−.164	−.019	−.374	1.000	
−.013	−.328	.005	.486	−.007	−.561	−.185	1.000

Example 12.7:
Complete Linkage Dendrogram

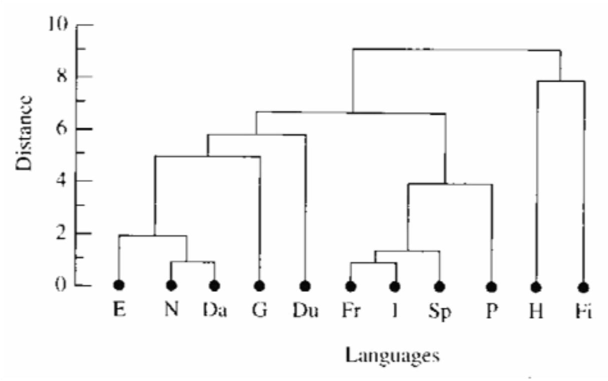


Average Linkage

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}$$

Example 12.8

Average Linkage of 11 Languages



Example 12.9

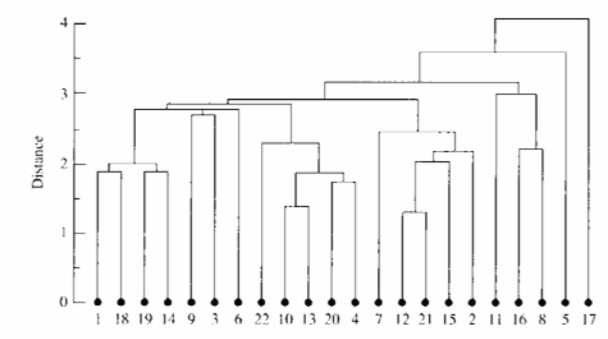
Average Linkage of Public Utilities

Table 12.6 Distances Between 22 Utilities

Firm no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	.00																					
2	3.10	.00																				
3	3.68	4.92	.00																			
4	2.46	2.16	4.11	.00																		
5	4.12	3.85	4.47	4.13	.00																	
6	3.61	4.22	2.99	3.20	4.60	.00																
7	3.90	3.45	4.22	3.97	4.60	3.35	.00															
8	2.74	3.89	4.99	3.69	5.16	4.91	4.36	.00														
9	3.25	3.96	2.75	3.75	4.49	3.73	2.80	3.59	.00													
10	3.10	2.71	3.93	1.49	4.05	3.83	4.51	3.67	3.57	.00												
11	3.49	4.79	5.90	4.86	6.46	6.00	6.00	3.46	5.18	5.08	.00											
12	3.22	2.43	4.03	3.50	3.60	3.74	1.66	4.06	2.74	3.94	5.21	.00										
13	3.96	3.43	4.39	2.58	4.76	4.55	5.01	4.14	3.66	1.41	5.31	4.50	.00									
14	2.11	4.32	2.74	3.23	4.82	3.47	4.91	4.34	3.82	3.61	4.32	4.34	4.39	.00								
15	2.59	2.50	5.16	3.19	4.26	4.07	2.93	3.85	4.11	4.26	4.74	2.33	5.10	4.24	.00							
16	4.03	4.84	5.26	4.97	5.82	5.84	5.04	2.20	3.63	4.53	3.43	4.62	4.41	5.17	5.18	.00						
17	4.40	3.62	6.36	4.89	5.63	6.10	4.58	5.43	4.90	5.48	4.75	3.50	5.61	5.56	3.40	5.56	.00					
18	1.88	2.90	2.72	2.65	4.34	2.85	2.95	3.24	2.43	3.07	3.95	2.45	3.78	2.30	3.00	3.97	4.43	.00				
19	2.41	4.63	3.18	3.46	5.13	2.58	4.52	4.11	4.11	4.13	4.52	4.41	5.01	1.88	4.03	5.23	6.09	2.47	.00			
20	3.17	3.00	3.73	1.82	4.39	2.91	3.54	4.09	2.95	2.05	5.35	3.43	2.23	3.74	3.78	4.82	4.87	2.92	3.90	.00		
21	3.45	2.32	5.09	3.88	3.64	4.63	2.68	3.98	3.74	4.36	4.88	1.38	4.94	4.93	2.10	4.57	3.10	3.19	4.97	4.15	.00	
22	2.51	2.42	4.11	2.58	3.77	4.03	4.00	3.24	3.21	2.56	3.44	3.00	2.74	3.51	3.35	3.46	3.63	2.55	3.97	2.62	3.01	.00

Example 12.9

Average Linkage of Public Utilities



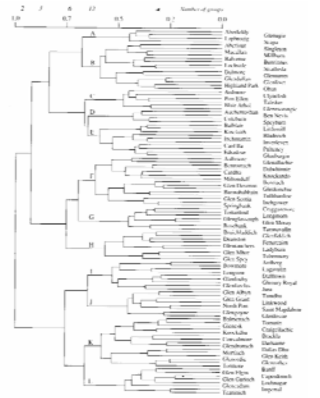
Ward's Hierarchical Clustering Method

- For a given cluster k, let ESS_k be the sum of the squared deviation of every item in the cluster from the cluster mean
- At each step, the union of every possible pair of clusters is considered
- The two clusters whose combination results in the smallest increase in the sum of Ess_k are joined

49

Example 12.10

Ward's Clustering Pure Malt ScotchWhiskies



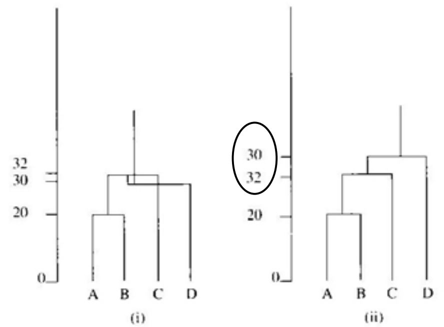
50

Final Comments

- Sensitive to outliers, or "noise points"
- No reallocation of objects that may have been "incorrectly" grouped at an early stage
- Good idea to try several methods and check if the results are roughly consistent
- Check stability by perturbation

51

Inversion



52

Outlinet

- Introduction
- Similarity Measures
- Hierarchical Clustering Methods
- Nonhierarchical Clustering Methods
- Clustering Based on Statistical Models
- Multidimensional Scaling

K-means Method

- Partition the items into *K* initial clusters
- Proceed through the list of items, assigning an item to the cluster whose centroid is nearest
- Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item
- Repeat until no more reassignment

54

Example 12.11 *K*-means Method

	Observations	
Item	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

55

Example 12.11 *K*-means Method

	Coordinates of Centroid	
Cluster	x1	x2
(AB)	(5+(-1))/2 = 2	(3+1)/2 = 2
(CD)	(1+(-3))/2=-1	(-2+(-2))/2=-2

56

Example 12.11
K-means Method

$$\bar{x}_{i,new} = \frac{n\bar{x}_i + x_{ji}}{n+1}$$

if the *j*th item is added to a group

$$\bar{x}_{i,new} = \frac{n\bar{x}_i - x_{ji}}{n-1}$$

if the *j*th item is removed from a group

Example 12.11
Final Clusters

	Squared distances to group centroids			
	Item			
Cluster	A	B	C	D
A	0	40	41	89
(BCD)	52	4	5	5

58

F Score

$$F_{nuc} = \frac{\text{mean square percent nuclear between clusters}}{\text{mean square percent nuclear within clusters}}$$

59

Outlinet

- Introduction
- Similarity Measures
- Hierarchical Clustering Methods
- Nonhierarchical Clustering Methods
- Clustering Based on Statistical Models
- Multidimensional Scaling

Normal Mixture Model

$$f_{Mix}(\mathbf{x}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x})$$

$$p_k \geq 0, \quad \sum_{k=1}^{K} p_k = 1, \quad f_k(\mathbf{x}) : N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$f_{Mix}(\mathbf{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)$$

$$= \sum_{k=1}^{K} \frac{p_k}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right)$$

61

Likelihood

$$L(p_1, \cdots, p_k, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)$$

$$= \prod_{j=1}^{N} f_{Mix}(\mathbf{x}_j \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)$$

$$= \prod_{j=1}^{N} \sum_{k=1}^{K} \frac{p_k}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x}_j - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_k) \right)$$

62

Statistical Approach

Obtain the maximum likelihood estimates and

$$L_{\max} = L(\hat{p}_1, \cdots, \hat{p}_k, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1, \cdots, \hat{\boldsymbol{\mu}}_K, \hat{\boldsymbol{\Sigma}}_K)$$

Determine K via maximizing

$$\text{AIC} = 2 \ln L_{\max} - 2N\left(K \frac{1}{2}(p+1)(p+2) - 1 \right)$$

or

$$\text{BIC} = 2 \ln L_{\max} - 2\ln(N)\left(K \frac{1}{2}(p+1)(p+2) - 1 \right)$$

63

BIC for Special Structures

Assumed form for $\boldsymbol{\Sigma}_k$	Total number of parameters	BIC
$\boldsymbol{\Sigma}_k = \eta \mathbf{I}$	$K(p+1)$	$\ln L_{\max} - 2\ln(N)K(p+1)$
$\boldsymbol{\Sigma}_k = \eta_k \mathbf{I}$	$K(p+2) - 1$	$\ln L_{\max} - 2\ln(N)(K(p+2) - 1)$
$\boldsymbol{\Sigma}_k = \eta_k \, Diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$	$K(p+2) + p - 1$	$\ln L_{\max} - 2\ln(N)(K(p+2) + p - 1)$

64

Software Package *MCLUST*

- Combines hierarchical clustering, EM algorithm, and BIC
- In the E step of EM, a matrix is created whose *j*th row contains the estimates of the conditional probabilities that observation $\mathbf{x}_j$ belongs to cluster 1, 2, . . ., *K*
- At convergence $\mathbf{x}_j$ is assigned to cluster *k* for which the conditional probability of membership is largest

65

Example 12.13 Clustering of Iris Data

$K = 3$

$p = 4, \quad \mathbf{\Sigma}_k = \eta_k \mathbf{I}, \quad k = 1,2,3$

$\hat{\eta}_1 = 0.076, \quad \hat{\eta}_2 = 0.163, \quad \hat{\eta}_3 = 0.163$

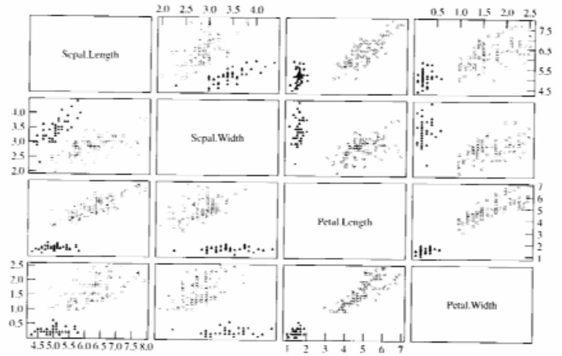
$\hat{p}_1 = 0.3333, \quad \hat{p}_2 = 0.4133, \quad \hat{p}_3 = 0.2534$

$\text{BIC} = -853.8$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 5.01 \\ 3.43 \\ 1.46 \\ 0.25 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 5.90 \\ 2.75 \\ 4.40 \\ 1.43 \end{bmatrix}, \quad \boldsymbol{\mu}_3 = \begin{bmatrix} 6.85 \\ 3.07 \\ 5.73 \\ 2.07 \end{bmatrix},$$

66

Example 12.13 Clustering of Iris Data



67

Example 12.13 Clustering of Iris Data

$K = 2$ is the best solution

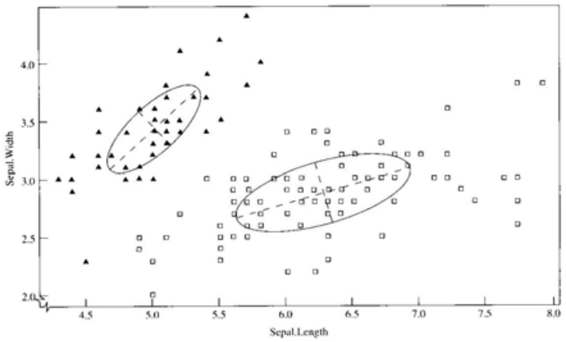
$\hat{p}_1 = 0.3333, \quad \hat{p}_2 = 0.6667$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 5.01 \\ 3.43 \\ 1.46 \\ 0.25 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 6.26 \\ 2.87 \\ 4.91 \\ 1.68 \end{bmatrix}$$

$$\hat{\mathbf{\Sigma}}_1 = \begin{bmatrix} .1218 & .0972 & .0160 & .0101 \\ .0972 & .1408 & .0115 & .0091 \\ .0160 & .0115 & .0296 & .0059 \\ .0101 & .0091 & .0059 & .0109 \end{bmatrix} \quad \hat{\mathbf{\Sigma}}_2 = \begin{bmatrix} .4530 & .1209 & .4489 & .1655 \\ .1209 & .1096 & .1414 & .0792 \\ .4489 & .1414 & .6748 & .2858 \\ .1655 & .0792 & .2858 & .1786 \end{bmatrix}$$

68

Example 12.13
Clustering of Iris Data



69

Outlinet

- Introduction
- Similarity Measures
- Hierarchical Clustering Methods
- Nonhierarchical Clustering Methods
- Clustering Based on Statistical Models
- Multidimensional Scaling

Multidimensional Scaling (MDS)

- Displays (transformed) multivariate data in low-dimensional space
- Different from plots based on PC
  - Primary objective is to "fit" the original data into low-dimensional system
  - Distortion caused by reduction of dimensionality is minimized
- Distortion
  - Similarities or dissimilarities among data

71

Multidimensional Scaling

- Given a set of similarities (or distances) between every pair of *N* items
- Find a representation of the items in few dimensions
- Inter-item proximities "nearly" match the original similarities (or distances)

72

Non-metric and Metric MDS

- Non-metric MDS
  - Uses only the rank orders of the $N(N-1)/2$ original similarities and not their magnitudes
- Metric MDS
  - Actual magnitudes of original similarities are used
  - Also known as principal coordinate analysis

73

Objective

N items, $M = N(N-1)/2$ similarities

Assume no ties, and arrange

$s_{i_1k_1} < s_{i_2k_2} < \cdots < s_{i_Mk_M}$

Find a q-dimensional configuration, such that

$d_{i_1k_1}^{(q)} > d_{i_2k_2}^{(q)} > \cdots > d_{i_Mk_M}^{(q)}$

74

Kruskal's Stress

$$\text{Stress}(q) = \left\{ \frac{\sum_k \sum_{i<k} \left(d_{ik}^{(q)} - \hat{d}_{ik}^{(q)} \right)^2}{\sum_k \sum_{i<k} \left[d_{ik}^{(q)} \right]^2} \right\}^{1/2}$$

$\hat{d}_{ik}^{(q)}$ are numbers known to satisfy the ordering

They are not distances, and merely reference numbers

75

Takane's Stress

$$\text{SStress} = \left\{ \frac{\sum_k \sum_{i<k} \left(d_{ik}^2 - \hat{d}_{ik}^2 \right)^2}{\sum_k \sum_{i<k} d_{ik}^4} \right\}^{1/2}$$

76

Basic Algorithm

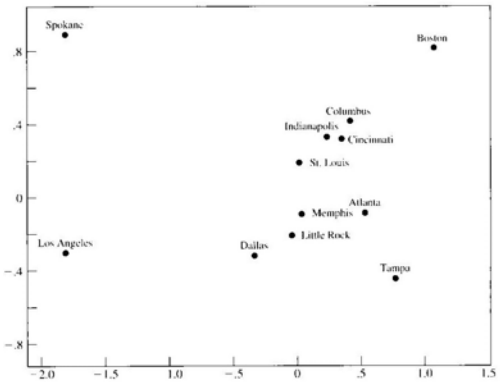
- Obtain and order the M pairs of similarities
- Try a configuration in q dimensions
  - Determine inter-item distances and reference numbers
  - Minimize Kruskal's or Takane's stress
- Move the points around to obtain an improved configuration
- Repeat until minimum stress is obtained

Example 12.14
MDS of U.S. Cities

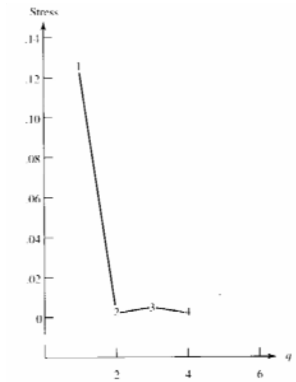
Airline-Distance Data

	Atlanta (1)	Boston (2)	Cincinnati (3)	Columbus (4)	Dallas (5)	Indianapolis (6)	Little Rock (7)	Los Angeles (8)	Memphis (9)	St. Louis (10)	Spokane (11)	Tampa (12)
(1)	0											
(2)	1068	0										
(3)	461	867	0									
(4)	549	769	107	0								
(5)	805	1819	943	1050	0							
(6)	508	941	108	172	882	0						
(7)	505	1494	618	725	325	562	0					
(8)	2197	3052	2186	2245	1403	2080	1701	0				
(9)	366	1355	502	586	464	436	137	1831	0			
(10)	558	1178	338	409	645	234	353	1848	294	0		
(11)	2467	2747	2067	2131	1891	1959	1988	1227	2042	1820	0	
(12)	467	1379	928	985	1077	975	912	2480	779	1016	2821	0

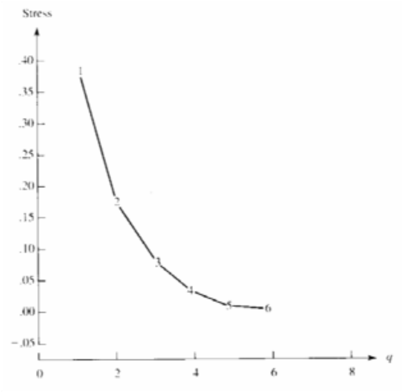
Example 12.14
MDS of U.S. Cities



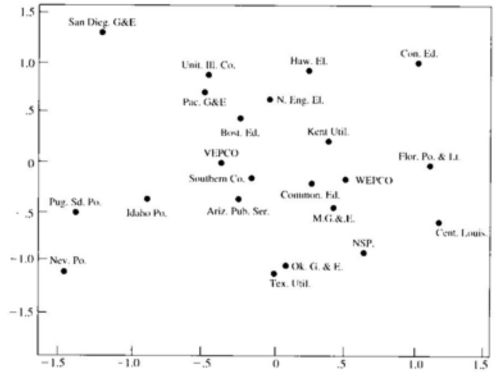
Example 12.14
MDS of U.S. Cities



Example 12.15
MDS of Public Utilities



Example 12.15
MDS of Public Utilities



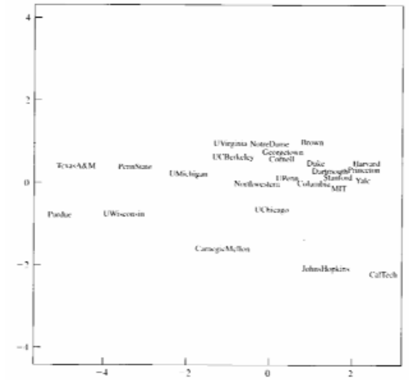
Example 12.16
MDS of Universities

Data on Universities

University	SAT	Top10	Accept	SFRatio	Expenses	Grad
Harvard	14.00	91	14	11	39.525	97
Princeton	13.75	91	14	8	30.220	95
Yale	13.75	95	19	11	43.514	96
Stanford	13.60	90	20	12	36.450	93
MIT	13.80	94	30	10	34.870	91
Duke	13.15	90	30	12	31.585	95
CalTech	14.15	100	25	6	63.575	81
Dartmouth	13.40	89	23	10	32.162	95
Brown	13.10	89	22	13	22.704	94
JohnsHopkins	13.05	75	44	7	58.691	87
UChicago	12.90	75	50	13	38.380	87
UPenn	12.85	80	36	11	27.553	90
Cornell	12.80	83	33	13	21.864	90
Northwestern	12.60	85	39	11	28.052	89
Columbia	13.10	76	24	12	31.510	88
NotreDame	12.55	81	42	13	15.122	94
UVirginia	12.25	77	44	14	13.349	92
Georgetown	12.55	74	24	12	20.126	92
CarnegieMellon	12.60	62	59	9	25.026	72
UMichigan	11.80	65	68	16	15.470	85
UCBerkeley	12.40	95	40	17	15.140	78
UWisconsin	10.85	40	69	15	11.857	71
PennState	10.81	38	54	18	10.185	80
Purdue	10.05	28	90	19	9.066	69
TexasA&M	10.75	49	67	25	8.704	67

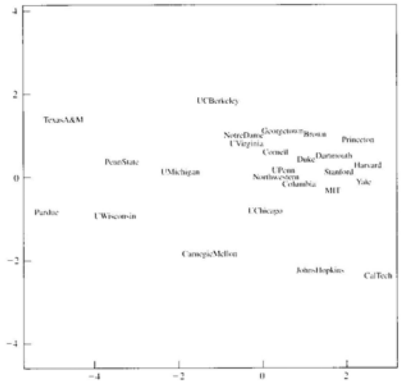
Source: *U.S. News & World Report*, September 18, 1995, p. 126.

Example 12.16
Metric MDS of Universities



Example 12.16

Non-metric MDS of Universities



85