Chapter 12 - Cluster Analysis

- Goal: Group individual observations into a set of "natural" sets. Unlike Discriminant Analysis, groups are not "known" in advance. Grouping made on similarities/Distances between individual cases.

12.2 - Similarity Measures

Distances for Pairs of Items

- Euclidean: $d(\underset{\sim}{X},\underset{\sim}{Y}) = \sqrt{\sum_{i=1}^{p}(X_i - Y_i)^2} = \sqrt{(\underset{\sim}{X}-\underset{\sim}{Y})'(\underset{\sim}{X}-\underset{\sim}{Y})}$

- Statistical: $d(\underset{\sim}{X},\underset{\sim}{Y}) = \sqrt{(\underset{\sim}{X}-\underset{\sim}{Y})'A(\underset{\sim}{X}-\underset{\sim}{Z})}$ $A \equiv S^{-1}$ typically

  Problematic when groups are not defined.

- Minkowski: $d(\underset{\sim}{X},\underset{\sim}{Y}) = \left[\sum_{i=1}^{p}|X_i - Y_i|^m\right]^{1/m}$

Binary outcomes

	Variable 1	2	3	4	5
item i	1	0	0	1	1
item k	1	1	0	1	0

$X_{ij} \equiv$ score for j^{th} var on item i

$X_{kj} \equiv$ " " " " " " k

$$\sum_{j=1}^{p}(X_{ij} - X_{kj}^2) = (1-1)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 = 2$$

$\equiv$ # mismatches. (weights 1-1 and 0-0 matches equally)

Contingency table for items i, k

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

Table 12.1 Similarity Coefficients for Clustering Items (binary vars)

1) $\frac{a+d}{p}$ (Equal weights for 1-1, 0-0 matches)

2) $\frac{2(a+d)}{2(a+d)+b+c}$ (Double weight for 1-1, 0-0 matches)

3) $\frac{a+d}{a+d+2(b+c)}$ (Double weight for mis-matches)

4) $\frac{a}{p}$ (No 0-0 matches in numerator)

5) $\frac{a}{a+b+c}$ (0-0 matches removed)

6) $\frac{2a}{2a+b+c}$ (0-0 removed, double weight for 1-1)

7) $\frac{a}{a+2(b+c)}$ (0-0 removed, double weight for mis-matches)

8) $\frac{a}{b+c}$ (0-0 removed, ratio of matches / mismatches).

Example - Diversity of Artifacts of 8 forts in Western Canada

~~Total Artifact Diversity Lux Goods Div. # of artifacts occupation years~~

Vars: (1) Total Artifact Diversity (# of Types) $X_1 = \begin{cases} 1 & \text{if } \geq 48 \\ 0 & \text{if } < 48 \end{cases}$

(2) Lux Goods Diversity $X_2 = \begin{cases} 1 & \text{if } \geq 10 \\ 0 & \text{if } < 10 \end{cases}$, (3) # artifacts $X_3 = \begin{cases} 1 & \text{if } \geq 1500 \\ 0 & \text{if } < 1500 \end{cases}$

(4) Occupation years $X_4 = \begin{cases} 1 & \text{if } \geq 10 \\ 0 & \text{if } < 10 \end{cases}$ (5) Ethnicity $X_5 = \begin{cases} 1 & \text{if Orkney} \\ 0 & \text{if French Canadian} \end{cases}$ (6) Region $\begin{cases} X_6 = 1 & \text{if Saskatchewan} \\ 0 & \text{if Athabaska} \end{cases}$

Data

Fact	X_1	X_2	X_3	X_4	X_5	X_6
1	0	0	1	0	0	0
2	0	0	0	0	0	0
3	1	1	1	0	1	0
4	1	1	0	0	1	0
5	1	0	0	1	1	1
6	0	0	1	0	0	1
7	1	1	1	1	1	0
8	0	0	0	1	0	0

Fact 4	1	0	Total
1	3	1	4
0	0	2	2
Total	3	3	6

Similarity index 1: $\frac{a+d}{p} = \frac{3+2}{6} = \frac{5}{6}$

Similarity matrix (measure 1)

Fact	1	2	3	4	5	6	7	8
1	1							
2	$\frac{5}{6}$	1						
3	$\frac{3}{6}$	$\frac{2}{6}$	1					
4	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{5}{6}$	1				
5	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	1			
6	$\frac{5}{6}$	$\frac{4}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	1		
7	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{4}{6}$	$\frac{3}{6}$	$\frac{1}{6}$	1	
8	$\frac{4}{6}$	$\frac{5}{6}$	$\frac{0}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{3}{6}$	$\frac{2}{6}$	1

Highest Similarities $\{1,2\}$, $\{1,6\}$, $\{2,8\}$, $\{3,4\}$, $\{3,7\}$

Possible Subgroups $\{1,2,6,8\}$, $\{3,4,7\}$, $\{5\}$

Similarity Measures: $\tilde{s}_{ik} = \frac{1}{1+d_{ik}}$ d_{ik} = distance measure

$$0 < \tilde{s}_{ik} \leq 1$$

If Similarity matrix is and, max Similarity $\tilde{s}_{ii} = 1$,

then $d_{ik} = \sqrt{2(1-\tilde{s}_{ik})}$ has properties of distance

Similarity and Association Measures for Pairs of Variables

Binary Case

		Variable k: 1	Variable k: 0	
variable i	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	a+b+c+d = n

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

~~[illegible]~~ For Example on variables

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	8					
X_2	7	8				
X_3	4	5	8			
X_4	5	4	3	8		
X_5	6	7	4	5	8	
X_6	4	3	4	5	4	8

a+d

Variable

12.3 Hierarchical Clustering Methods

- **Agglomerative** - Start w/ individual objects (units or variables) and combine until a single cluster formed.

Linkage Methods for Combining clusters

- **Single Linkage** - Minimum distance between objects in clusters.
- **Complete Linkage** - Maximum distance between objects in Clusters
- **Average Linkage** - Mean distance between objects in clusters.

- **Divisive** - start w/ single cluster and split off until each object is a cluster

Dendogram - 2-Dimensional Diagram of the Clustering process.

EXAMPLE - AVERAGE (x_3) POINTS, (x_1) REBOUNDS, (x_2) ASSISTS of $n = 5$ WNBA Players (2014): Maya Moore, Skylar Diggins, Candace Parker, Angel McCoughtry, Tina Charles

$$\text{Distance} = \sqrt{\sum_{i=1}^{3} (x_{ji} - y_i)^2}$$

	MM	SD	CP	AM	TC
MM	0				
SD	6.92	0			
CP	4.63	4.71	0		
AM	6.09	3.40	(2.21)	0	
TC	6.78	7.93	3.73	4.59	0

Single Linkage

- 2 Closest Items: C, A $d_{CA} = 2.21$

$\Rightarrow$ Merge C, A into cluster (CA)

$d_{(CA)M} = \min\{d_{CM}, d_{AM}\} = \min\{4.63, 6.09\} = 4.63$

$d_{(CA)S} = \min\{d_{CS}, d_{AS}\} = \min\{4.71, 3.40\} = 3.40$

$d_{(CA)T} = \min\{d_{CT}, d_{AT}\} = \min\{3.73, 4.59\} = 3.73$

D^* =

	(CA)	M	S	T
(CA)	0			
M	4.63	0		
S	(3.40)	6.92	0	
T	3.73	6.78	7.93	0

Smallest distance in $D^* = d_{(CA)S} = 3.40$

$\Rightarrow$ merge (CA, S) into (CAS)

$d_{(CAS)M} = \min\{d_{(CA)M}, d_{SM}\} = \min\{4.63, 6.92\} = 4.63$

$d_{(CAS)T} = \min\{d_{(CA)T}, d_{ST}\} = \min\{3.73, 7.93\} = 3.73$

D^{**}

	(CAS)	M	T
(CAS)	0		
M	4.63	0	
T	(3.73)	6.78	0

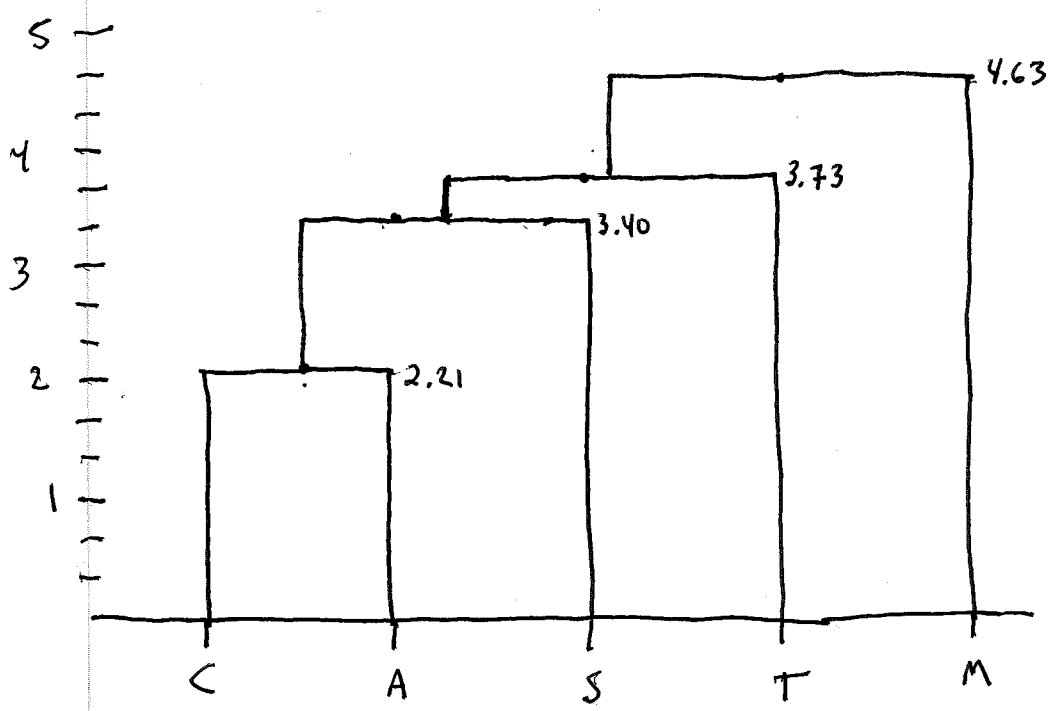
Smallest Distance in $D^{**} = d_{(CAS)T} \Rightarrow$ merge (CAS, T) into (CAST)

$d_{(CAST)M} = \min\{d_{(CAS)M}, d_{TM}\}$
$= \min\{4.63, 6.78\} = 4.63$

D^{***}

	(CAST)	M
(CAST)	0	
M	(4.63)	0

Dendogram



Complete Linkage

D

	M	S	C	A	T
M	0				
S	6.92	0			
C	4.63	4.71	0		
A	6.09	3.40	(2.21)	0	
T	6.78	7.93	3.73	4.59	0

Step 1 => CA (2.21)

(Same as single linkage)

$d_{CA(M)} = \max\{d_{CM}, d_{AM}\} = \max\{4.63, 6.09\} = 6.09$

$d_{CA(S)} = \max\{d_{CS}, d_{AS}\} = \max\{4.71, 3.40\} = 4.71$

$d_{CA(T)} = \max\{d_{CT}, d_{AT}\} = \max\{3.73, 4.59\} = 4.59$

=> $D^* =$

	(CA)	M	S	T
(CA)	0			
M	6.09	0		
S	4.71	6.92	0	
T	(4.59)	6.78	7.93	0

=> Combine CA, T into (CAT)

$d_{(CAT)M} = \max\{d_{(CA)M}, d_{TM}\} = \max\{6.09, 6.78\} = 6.78$

$d_{(CAT)S} = \max\{d_{(CA)S}, d_{TS}\} = \max\{4.71, 7.93\} = 7.93$

D^{**}

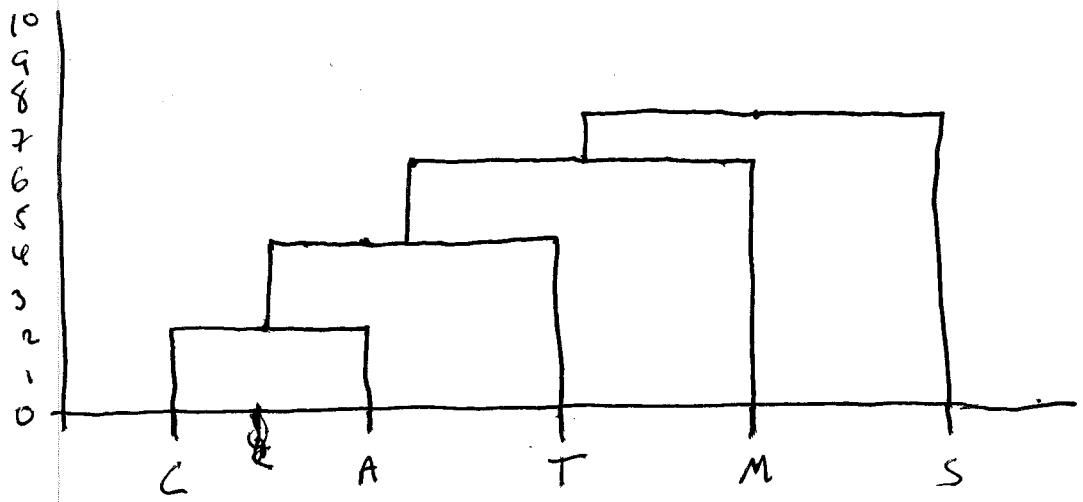
	(CAT)	M	S
(CAT)	0		
M	(6.78)	0	
S	7.93	6.92	0

⇒ Combine CAT, M into CATM

$d_{(CATM)S} = \max\{d_{(CAT)S}, d_{MS}\} = \max\{7.93, 6.92\} = 7.93$

D^{***}

	(CATM)	S
(CATM)	0	
S	(7.93)	0



Average Linkage

D

	M	S	C	A	T
M	0				
S	6.92	0			
C	4.63	4.71	0		
A	6.09	3.40	(2.21)	0	
T	6.78	7.93	3.73	4.59	0

Again choose to combine C, A into (CA)

$$d_{(CA)M} = \frac{d_{CM} + d_{AM}}{2} = \frac{4.63 + 6.09}{2} = 5.36$$

$$d_{(CA)S} = \frac{d_{CS} + d_{AS}}{2} = \frac{4.71 + 3.40}{2} = 4.06$$

$$d_{(CA)T} = \frac{d_{CT} + d_{AT}}{2} = \frac{3.73 + 4.59}{2} = 4.16$$

D*

	(CA)	M	S	T
(CA)	0			
M	5.36	0		
S	(4.06)	6.92	0	
T	4.16	6.78	7.93	0

Combine (CA), S into (CAS)

$$d_{(CAS)M} = \frac{d_{(CA)M} + d_{SM}}{2} = \frac{5.36 + 6.92}{2} = 6.14$$

$$d_{(CAS)T} = \frac{d_{(CA)T} + d_{ST}}{2} = \frac{4.16 + 7.93}{2} = 6.05$$

D**:

	(CAS)	M	T
(CAS)	0		
M	6.14	0	
T	(6.05)	6.78	0

Combine (CAS), T into (CAST)

Average Linkage

D

	M	S	C	A	T
M	0				
S	6.92	0			
C	4.63	4.71	0		
A	6.09	3.40	(2.21)	0	
T	6.78	7.93	3.73	4.59	0

Again choose to combine C, A into (CA)

$$d_{(CA)M} = \frac{d_{CM} + d_{AM}}{2} = \frac{4.63 + 6.09}{2} = 5.36$$

$$d_{(CA)S} = \frac{d_{CS} + d_{AS}}{2} = \frac{4.71 + 3.40}{2} = 4.06$$

$$d_{(CA)T} = \frac{d_{CT} + d_{AT}}{2} = \frac{3.73 + 4.59}{2} = 4.16$$

D*

	(CA)	M	S	T
(CA)	0			
M	5.36	0		
S	(4.06)	6.92	0	
T	4.16	6.78	7.93	0

Combine (CA), S into (CAS)

$$d_{(CAS)M} = \frac{d_{(CA)M} + d_{SM}}{2} = \frac{5.36 + 6.92}{2} = 6.14$$

$$d_{(CAS)T} = \frac{d_{(CA)T} + d_{ST}}{2} = \frac{4.16 + 7.93}{2} = 6.05$$

D**:

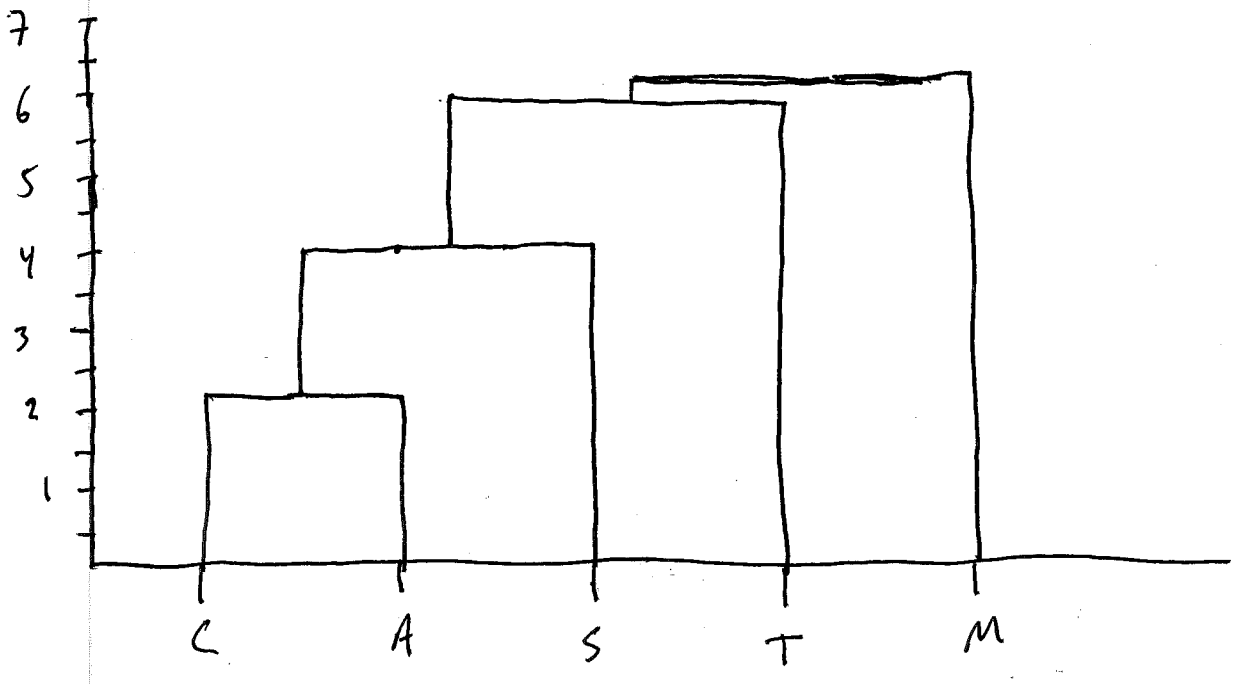
	(CAS)	M	T
(CAS)	0		
M	6.14	0	
T	(6.05)	6.78	0

Combine (CAS), T into (CAST)

$$d_{(CAST)M} = \frac{d_{(CAS)M} + d_{TM}}{2} = \frac{6.14 + 6.78}{2} = 6.46$$

$D^{(3)}$

	(CAST)	M
(CAST)	0	
M	6.46	0



12.4 - Nonhierarchical Clustering Methods

Goal - Group items (not variables) into a group of K clusters. K can be specified in advance or determined from the procedure.

- Distance/similarity matrix does not need to be computed.
- Basic data not stored during run (Computationally more efficient than hierarchical methods).

K-Means Method

1) Partition the Items into K initial clusters (random)
2) Proceed through list of items, assigning each to the nearest cluster centroid. (Euclidean distance w/ standardized or unstandardized observations). Recalculate cluster centroids when observation shifts clusters.
3) Repeat Step 2) until no observations are reassigned.

Several initial partitions should be used, to see whether the clusters are stable.

Optimization Problem: $\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$

where $|C_k|$ ≡ # of items in cluster k.