

Clustering Methods

Applications of Multivariate Statistical Analysis

Jiangsheng Yu

©School of Electronics Engineering and Computer Science

Peking University, Beijing, 100871

yujs@pku.edu.cn, <http://icl.pku.edu.cn/yujs>



Topics

1. What's clustering?
2. Similarity Measures
3. Hierarchical Clustering Methods
4. Nonhierarchical Clustering Methods
5. Multidimensional Scaling
6. Correspondence Analysis
7. Biplots for Viewing Sampling Units and Variables
8. Procrustes Analysis: A Method
9. Conclusion
10. References

Clustering Problem

Exploratory procedures are helpful in understanding the complex nature of multivariate relationships.

Problem 1 Given a set of data $\{\mathbf{x}_i \in \mathbb{R}^p\}$, satisfying

1. the number of classes is unknown;
2. the class of any individual is unknown.

We intend to

1. define some suitable statistics;
2. clarify the number of classes K ;
3. find a reasonable clustering method; and
4. classify the data into K categories.^a

^aSo, clustering is also called **unsupervised classification**.

Example of Clustering

Partition a given set by some similarity:

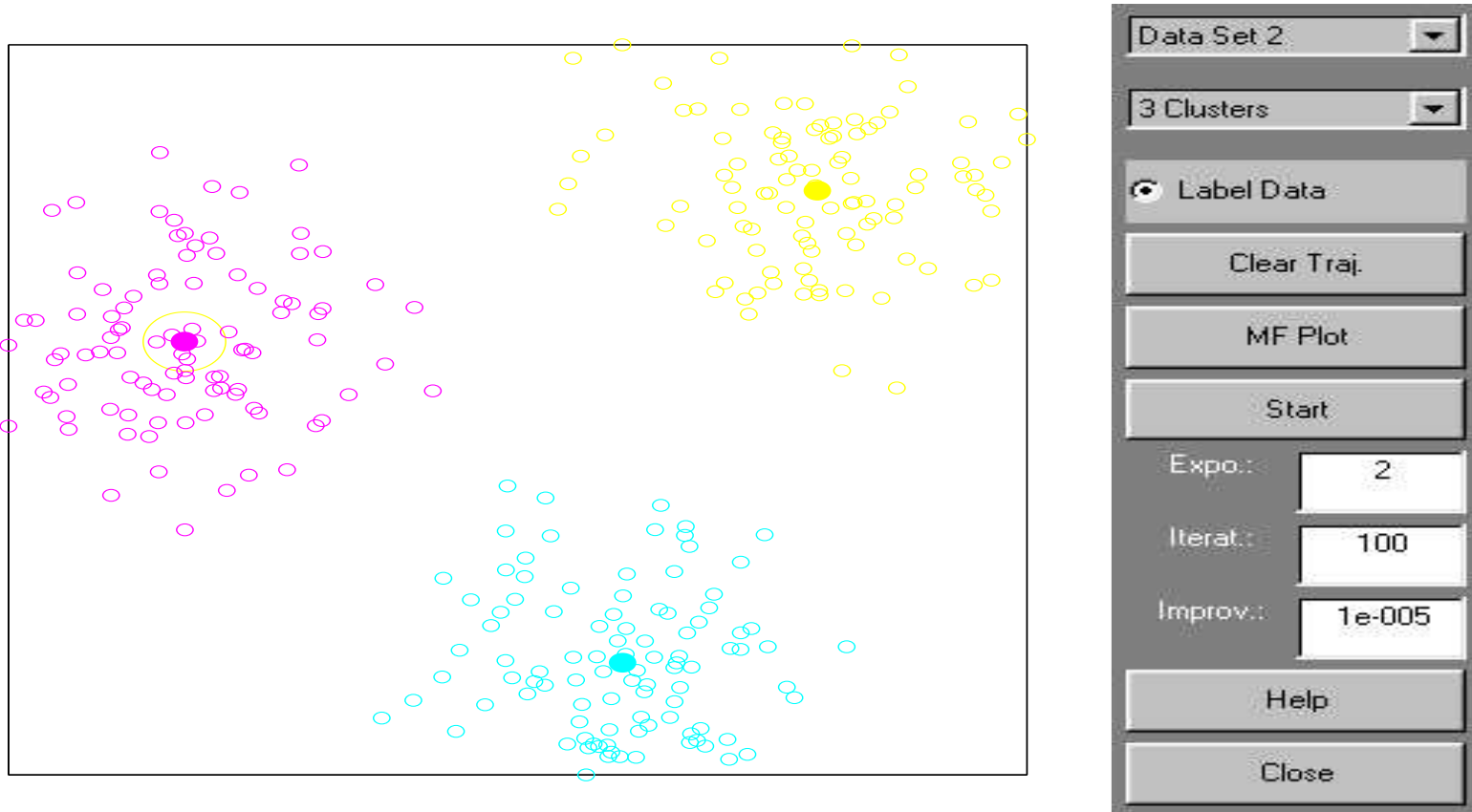


Figure 1: fuzzy c-means clustering

Observation Matrix

Given n sample points, each has m variables:

	X_1	\dots	X_j	\dots	X_m
\mathbf{x}_1	x_{11}	\dots	x_{1j}	\dots	x_{1m}
\vdots	\vdots				
\mathbf{x}_i	x_{i1}	\dots	x_{ij}	\dots	x_{im}
\vdots	\vdots				
\mathbf{x}_n	x_{n1}	\dots	x_{nj}	\dots	x_{nm}
mean	\bar{x}_1	\dots	\bar{x}_i	\dots	\bar{x}_m
std	s_1	\dots	s_i	\dots	s_m

Table 1: Observation data

No Best Clustering Method



Cluster Methods

1. System Method: merge the most similar classes, update the data and repeat the procedure till all data are classified.
2. Dynamic Method: give an initial classification of data firstly, then adjust the classes by least value of loss function till no improvement can made.
3. Fuzzy Method: for instance fuzzy c-means clustering, usually works well for data with fuzzy characteristics.
4. Method of Minimum Spanning Tree
5. ...

Transformation of Data

Centralization: make the mean 0, and the variance-covariance matrix unchanged.

$$x_{ij}^* = x_{ij} - \bar{x}_j \quad (1)$$

$$S^* = S = (s_{ij})_{m \times m} \quad (2)$$

$$s_{ij} = \frac{1}{n-1} \sum_{t=1}^n (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j) = \frac{1}{n-1} \sum_{t=1}^n x_{ti}^* x_{tj}^* \quad (3)$$

Standardization: make the mean 0, and sdt 1.

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (4)$$

Measuring Similarity

1. Distance
 - (a) Minkowski Distance
 - (b) Statistical Distance
 - (c) Lance Distance
 - (d) Mahalanobis Distance, . . .
2. Measure (not necessary distance)
 - (a) Canberra Measure
 - (b) Czekanowski Coefficient, . . .
3. Similarity Coefficient
 - (a) Cosine Similarity
 - (b) Correlation Coefficient, . . .

Minkowski Distance

1. Minkowski Distance:

$$d_k(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{t=1}^m |x_{it} - x_{jt}|^k \right]^{1/k} \quad (5)$$

2. Euclidean Distance:

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{t=1}^m (x_{it} - x_{jt})^2} \quad (6)$$

3. Chebyshev Distance:

$$d_\infty(\mathbf{x}_i, \mathbf{x}_j) = \max_{1 \leq t \leq m} |x_{it} - x_{jt}| \quad (7)$$

Distances without Measure Unit

1. Statistical Distance:

$$d_k^*(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{t=1}^m \left| \frac{x_{it} - x_{jt}}{s_t} \right|^k \right]^{1/k} \quad (8)$$

2. Lance Distance:

$$d_L(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{m} \sum_{t=1}^m \frac{|x_{it} - x_{jt}|}{x_{it} + x_{jt}} \quad (9)$$

3. Mahalanobis Distance:

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top S^{-1} (\mathbf{x}_i - \mathbf{x}_j) \quad (10)$$

Measures

1. Canberra Measure:

$$m(\mathbf{x}_i, \mathbf{x}_j) = \sum_{t=1}^m \frac{|x_{it} - x_{jt}|}{x_{it} + y_{jt}} \quad (11)$$

2. Czekanowski Coefficient:

$$m(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{2 \sum_{t=1}^m \min(x_{it}, x_{jt})}{\sum_{t=1}^m (x_{it} + x_{jt})} \quad (12)$$

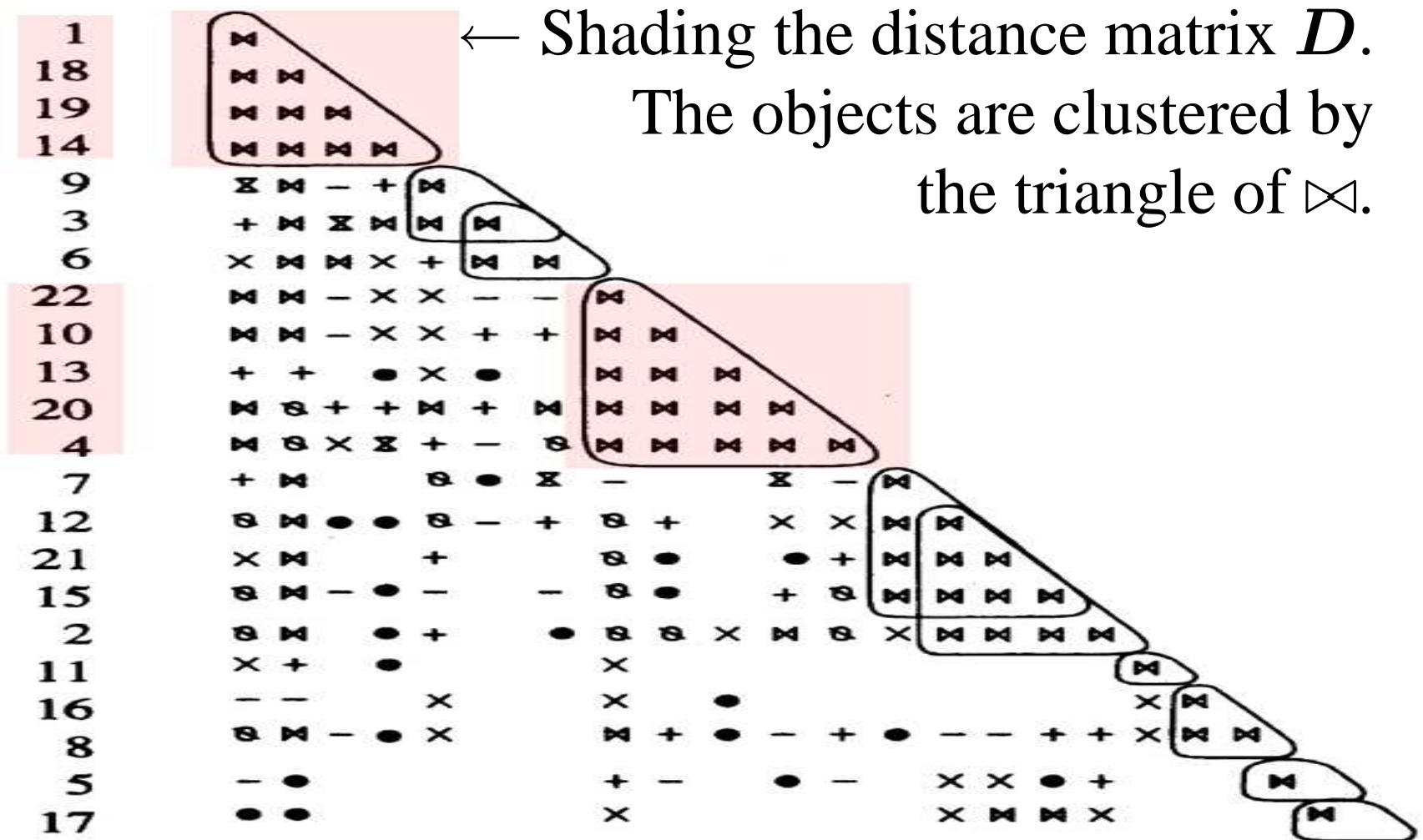
Neither of them is distance.

Shortcomings of Distances

1. Both Minkowski Distance and Lance Distance assume the independency between random variables (r.v.), which talk about the distance in an orthogonal space. But in practice, the r.v.'s are relative. Mahalanobis Distance overcomes this shortcoming, but
2. Mahalanobis Distance works badly if the covariance matrix S is calculated by all data. Concentrated to a particular class, its performance is good. While, we know nothing about classes before clustering.
3. The mathematics of non-distance measures is not beautiful.

Clustering by Shading D-Matrix

Assume $D = (d_{ij})$ denote the distance between the i -th and j -th objects, replaced by a prescribed class.



Example of Shading Method

	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

How to Describe the Difference?

Consider the 0-1 feature vector based on m variables.

Item	0-1 valued variables				
	1	...	l	...	m
i	x_{i1}	...	x_{il}	...	x_{im}
j	x_{j1}	...	x_{jl}	...	x_{jm}

Table 2: Comparison between item i and item j

The difference between item i and item j can be measured by $\sum_{l=1}^m (x_{il} - x_{jl})^2$. But it suffers from weighting the 1-1 and 0-0 matches equally.


Example 1 In grouping people, the characteristic of doing Algebraic Geometry is significant.

Contingency Table

To illustrate the matches and mismatches, we arrange the amounts into a contingency table.

Item $i \backslash$ Item j	1	0	Totals
1	n_{11}	n_{12}	$n_{1.} = n_{11} + n_{12}$
0	n_{21}	n_{22}	$n_{2.} = n_{21} + n_{22}$
Totals	$n_{.1}$	$n_{.2}$	$N = n_{1.} + n_{2.}$

Table 3: Contingency table

In Example 1, it might be reasonable to discount the 0-0 matches or even disregard them completely. There are several suggested schemes of defining similarity. See the next slide. 

Similarity Coefficients for Items

Coefficient	Rationale
1. $\frac{n_{11}+n_{22}}{N}$	weights for matches
2. $\frac{2(n_{11}+n_{22})}{N+n_{11}+n_{22}}$	double weight for matches
3. $\frac{n_{11}+n_{22}}{N+n_{12}+n_{21}}$	double weight for mismatches
4. $\frac{n_{11}}{N}$	ratio of 1-1 matches
5. $\frac{n_{11}}{N-n_{22}}$	0-0 matches are treated as irrelevant
6. $\frac{2n_{11}}{n_{1.}+n_{.1}}$	Homework Explain coefficients 6,
7. $\frac{n_{11}}{n_{11}+2(n_{12}+n_{21})}$	7 and 8. And prove that coefficients
8. $\frac{n_{11}}{n_{12}+n_{21}}$	1-3 (5-7) are monotonically related.

Table 4: Similarity coefficients for clustering items

Pearson's χ^2 Test

Theorem 1 (Pearson) By the null hypothesis $H_0 : P(X|Y) = P(X)$,

$$\chi^2 = \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}} \sim \chi^2(1) \quad (13)$$

$Y \setminus X$	1	0	marginal distribution of X
1	n_{11}	n_{12}	$n_{1.}$
0	n_{21}	n_{22}	$n_{2.}$
marginal distribution of Y	$n_{.1}$	$n_{.2}$	N

Table 5: Use Pearson's χ^2 to test whether X and Y are independent

Hierarchical Clustering

- Agglomerative Hierarchical Method (AHM):
The most similar objects are grouped, and these initial groups are merged according to their similarities. In the opposite direction, we have
- Divisive Hierarchical Method

To measure the distance between clusters, we need linkage methods:

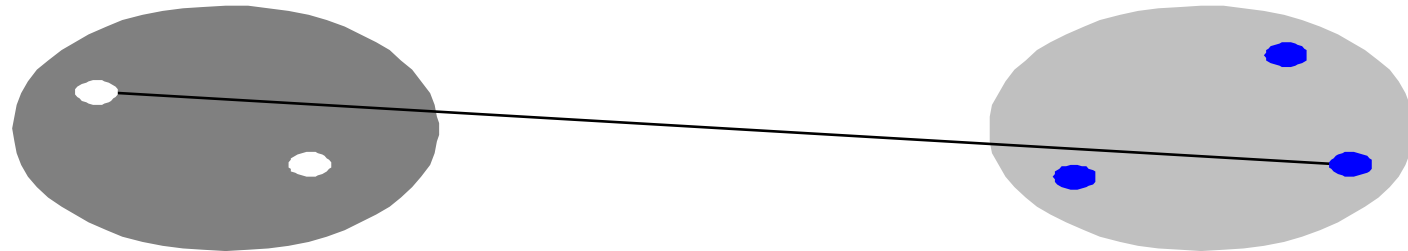
1. Single Linkage
2. Complete Linkage
3. Average Linkage

Linkage method works well for clustering items, as well as variables. We will focus on this method. 

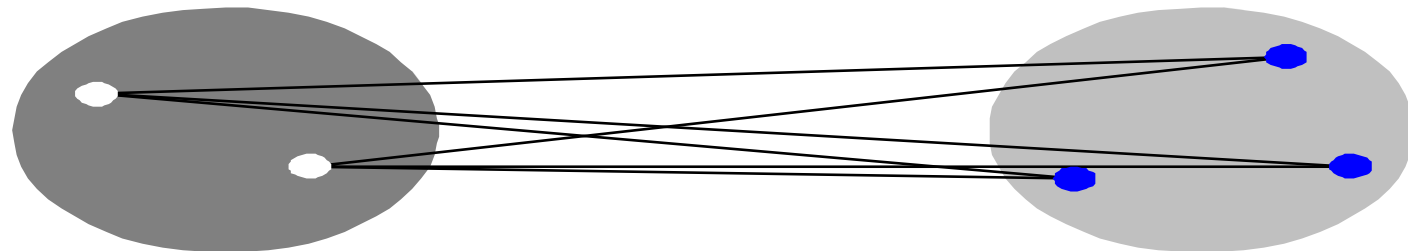
Distances between Clusters



Single linkage



Complete linkage



Average linkage

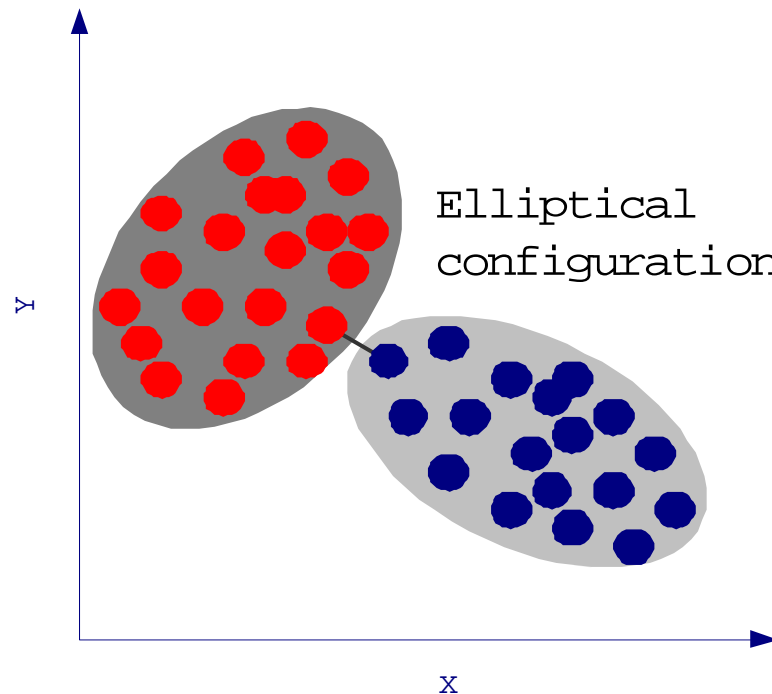
Algorithm of AHM

1. Start with n clusters, each containing a single entity. Let the distance matrix be $D_{n \times n}$.
2. Search D for the nearest pair of clusters by some linkage, merge them and update D .
3. Repeat Step 2 a total of $n - 1$ times to get a dendrogram (or tree diagram).

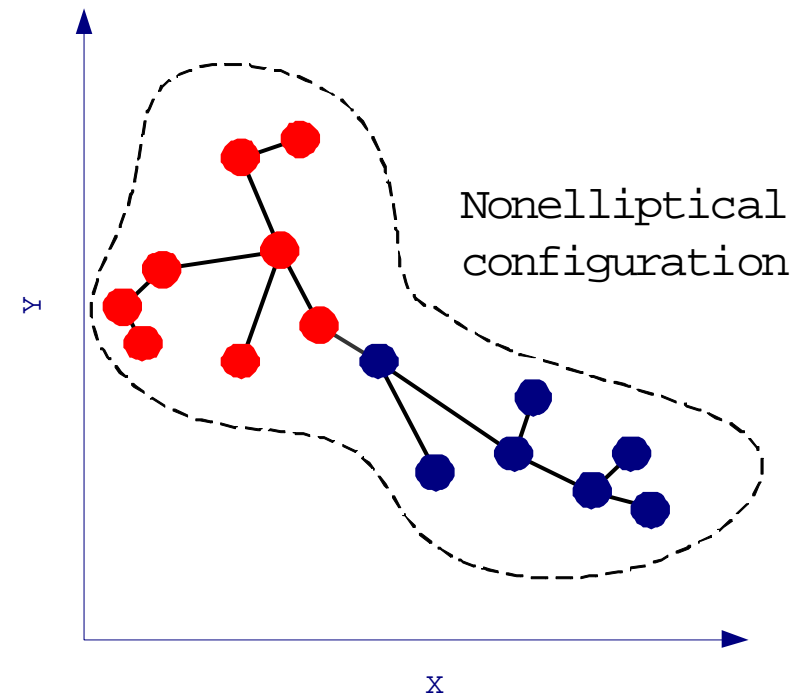
	x_1	x_2	x_3	x_4	x_5		x_3	x_4	x_5	CL4	
x_1	0	1	3.5	5	7	\Rightarrow	x_3	0	1.5	3.5	2.5
x_2		0	2.5	4	6		x_4		0	2	4
x_3			0	1.5	3.5		x_5			0	6
x_4				0	2		CL4				0
x_5					0						

Disadvantage of Single Linkage

Homework 1 Calculate the clustering in the last slide by single, complete and average linkages. And compare the three results.



single linkage confused
by near overlap



chaining effect

Clustering Variables

- Calculate the pairwise correlation between m variables in Table 1:

	X_1	X_2	\dots	X_m
X_1	1			
X_2	ρ_{21}	1		
\vdots	\vdots			
X_m	ρ_{m1}	ρ_{m2}	\dots	1

Table 6: Correlation matrix

- Cluster the variables by the correlation matrix, using some linkage.

Ward's Hierarchical Clustering

Ward considered hierarchical clustering based on minimizing the **error sum of squares** (ESS) from joining two groups.

Definition 1 For a given cluster k , let ESS_k be the sum of the squared deviations for each item in the cluster from the cluster mean. Assume there are

currently l categories, define $ESS = \sum_{i=1}^l ESS_i$.

- Initial: $ESS_i = 0$, where $i = 1, \dots, n$.
- Join two clusters with the minimum ESS.

- $ESS = \sum_{i=1}^K (\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}})$.

Usual Linkages with Scores

1. **SIN**gle linkage (☒☒)
2. **COM**plete method (☒☒)
3. **MED**ian method (☒☒☒)
4. **CEN**troid method (☒☒☒)
5. **AVE**rage linkage (☒☒☒☒)
6. **FLE**xible-beta method (☒☒☒☒)
7. **MCQ** method (☒☒☒☒)
8. **WARD** method (☒☒☒☒)
9. **EML** method (☒☒☒☒)
10. **DEN** method (☒☒☒☒)
11. **TWO** method (☒☒☒☒)

Nonhierarchical Clustering

➔ Algorithm of K -means method:^a

1. Partition the items into K initial clusters.
2. Proceed through the list of items, assigning an item to the cluster whose centroid is nearest.
Update the centroids.
3. Repeat step 2 until no more reassignments.

➔ Understanding K -means method:

- The result depends on the initial assignment.
- K -means method works well in the case that partial items are suitably clustered.

^aNonhierarchical clustering techniques are designed to group items, rather than variables, into a collection of K clusters.

Example of K -means Method

Example 2 The basic idea of K -means method is that the centroid of partial near points is not far from that of actual cluster.

Item	X_1	X_2		Cluster	\bar{X}_1	\bar{X}_2
A	5	3		(AB)	2	2
B	-1	1	→	(CD)	-1	-2
C	1	-2				
D	-3	-2				

$d(A, (AB)) < d(A, (CD))$
 $d(B, (AB)) > d(B, (CD))$

Cluster	\bar{X}_1	\bar{X}_2	A	B	C	D
A	5	3	0	40	41	89
(BCD)	-1	-1	52	4	5	5

Disadvantage of K -means

1. If two or more seeds points inadvertently lie within a single cluster, the result will be poor.
2. The existence of an outlier might produce at least one group with very disperse items.
3. K -means method works poorly on the rare data.
4. Initial partition may lead to poor result.

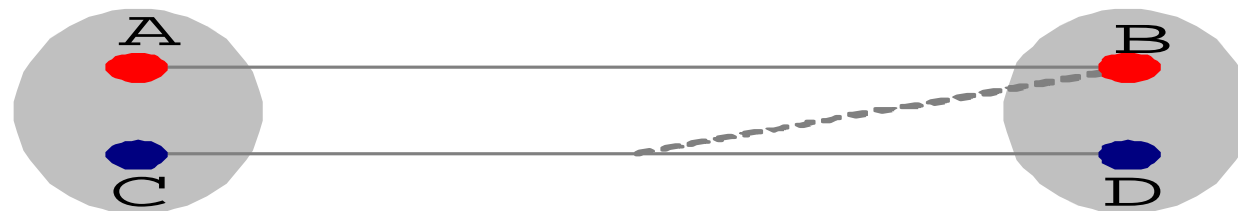


Figure 2: Counterexample of K -means method
 $d(B, (CD)) > d(B, (AB))$

Multidimensional Scaling

Task: Represent the items in low-dimensional space.

Algorithm: There are $N = n(n - 1)/2$ similarities between items.

1. Sort the similarities to $s_{i_1 j_1} < \dots < s_{i_N j_N}$
2. Project the feature vectors on a q -dimensional space, making **Kruskal's Stress(q)** as small as possible:

$$\text{KStress}(q) = \sqrt{\frac{\sum_{i < j} \left[d_{ij}^{(q)} - \hat{d}_{ij}^{(q)} \right]^2}{\sum_{i < j} \left[d_{ij}^{(q)} \right]^2}} \quad (14)$$

where $d_{i_k j_k}^{(q)}$ denote the distance between item i_k and item j_k in q dimensions, and $\hat{d}_{i_1 j_1}^{(q)} > \dots > \hat{d}_{i_N j_N}^{(q)}$.

Tankane's Stress

Definition 2 For a given dimension q , a more preferred stress is defined by Tankane,

$$\text{TStress}(q) = \sqrt{\frac{\sum_{i < j} \sum [d_{ij}^2 - \hat{d}_{ij}^2]^2}{\sum_{i < j} \sum d_{ij}^4}} \quad (15)$$

➔ Multidimensional scaling makes the plotting of items available in 2 or 3 dimensional space, which provides us a intuitive clustering.

Stress	Goodness of fit
20%	Poor
10%	Fair
5%	Good
2.5%	Excellent
0%	Perfect

Correspondence Matrix

Definition 3 From the observation matrix \mathbf{X} , we get the **correspondence matrix**^a $\mathbf{P} = (p_{ij})_{n \times m} = \frac{1}{t} \mathbf{X}$, where $t = \sum_{i,j} x_{ij}$.

- The vector of row sums $\mathbf{r} = \mathbf{P}_{n \times m} \mathbf{1}_{m \times 1}$
- The vector of column sums $\mathbf{c} = \mathbf{P}^T \mathbf{1}_{n \times 1}$

Definition 4 $\mathbf{D}_r^{1/2} = \text{diag}(\sqrt{r_1}, \dots, \sqrt{r_n})$, $\mathbf{D}_c^{1/2} = \text{diag}(\sqrt{c_1}, \dots, \sqrt{c_m})$, $\mathbf{D}_r^{-1/2} = \text{diag}(\frac{1}{\sqrt{r_1}}, \dots, \frac{1}{\sqrt{r_n}})$ and $\mathbf{D}_c^{-1/2} = \text{diag}(\frac{1}{\sqrt{c_1}}, \dots, \frac{1}{\sqrt{c_m}})$.

^aMore details can be found in [3].

Correspondence Analysis (CA)

CA can be formulated as the weighted least squares problem to select $\hat{P} = (\hat{p}_{ij})_{n \times m}$ with a specified reduced rank, to minimize

$$\begin{aligned} & \sum_{i,j} \frac{(p_{ij} - \hat{p}_{ij})^2}{r_i c_j} \\ &= \text{tr}[(D_r^{-1/2}(P - \hat{P})D_c^{-1/2})(D_r^{-1/2}(P - \hat{P})D_c^{-1/2})^\top] \end{aligned} \quad (16)$$

Theorem 2 (Generalized SVD) The optimal \hat{P} with rank s is

$$\begin{aligned} \hat{P} &= \sum_{k=1}^s \tilde{\lambda}_k (D_r^{1/2} \tilde{u}_k)(D_c^{1/2} \tilde{v}_k)^\top \\ &= r c^\top + \sum_{k=2}^s \tilde{\lambda}_k (D_r^{1/2} \tilde{u}_k)(D_c^{1/2} \tilde{v}_k)^\top \end{aligned} \quad (17)$$

where $\tilde{\lambda}_k$ are singular values of $D_r^{-1/2} P D_c^{-1/2}$ and \tilde{u}_k, \tilde{v}_k are corresponding singular vectors.

CA (Cont'd)

Corollary 1 The minimum of (16) is $\sum_{k=s+1}^m \tilde{\lambda}_k^2$.

Theorem 3 The optimal approximation to $\mathbf{P} - \mathbf{r}\mathbf{c}^\top$ with reduced rank s is

$$\sum_{k=1}^s \lambda_k (\mathbf{D}_r^{1/2} \mathbf{u}_k) (\mathbf{D}_c^{1/2} \mathbf{v}_k)^\top \quad (18)$$

where λ_k are singular values of $\mathbf{D}_r^{-1/2} (\mathbf{P} - \mathbf{r}\mathbf{c}^\top) \mathbf{D}_c^{-1/2}$ and $\mathbf{u}_k, \mathbf{v}_k$ are corresponding singular vectors.^a

Homework 2 Prove that $\lambda_k = \tilde{\lambda}_{k+1}$, $\mathbf{u}_k = \tilde{\mathbf{u}}_{k+1}$ and $\mathbf{v}_k = \tilde{\mathbf{v}}_{k+1}$ for $k = 1, \dots, m - 1$.

^aThe proof is done by the properties of SVD, that can be found in pp713-714 of [4].

Symmetric Map

The vectors $\mathbf{D}_r^{1/2}\mathbf{u}_k$ and $\mathbf{D}_c^{1/2}\mathbf{v}_k$ need not have length 1, but satisfying the scaling

$$\begin{aligned}(\mathbf{D}_r^{1/2}\mathbf{u}_k)^\top \mathbf{D}_r^{-1} (\mathbf{D}_r^{1/2}\mathbf{v}_k)^\top &= 1 \\(\mathbf{D}_c^{1/2}\mathbf{u}_k)^\top \mathbf{D}_c^{-1} (\mathbf{D}_c^{1/2}\mathbf{v}_k)^\top &= 1\end{aligned}$$

Let $\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^\top)\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ be an SVD, it is usual in CA to plot $\lambda_k\mathbf{D}_r^{-1/2}\mathbf{u}_k$ and $\lambda_k\mathbf{D}_c^{-1/2}\mathbf{v}_k$ for $k = 1, 2$, and maybe 3.

Definition 5 The joint plot of the coordinates in $\lambda_k\mathbf{D}_r^{-1/2}\mathbf{u}_k$ and $\lambda_k\mathbf{D}_c^{-1/2}\mathbf{v}_k$ is called a **symmetric map**, in which the geometry for the row points is identical to that for the column points.

Profile Approximation Method

Definition 6 Algebraically, the **row profiles** are the rows of $D_r^{-1}P$. The CA can be defined as the approximation of the row profiles (denoted by P^*) by points in a low-dimensional space.

To minimize

$$\begin{aligned} \sum_{i,j} \frac{(p_{ij} - \hat{p}_{ij})^2}{r_i c_j} &= \sum_i r_i \sum_j \frac{(p_{ij}/r_i - p_{ij}^*)^2}{c_j} \\ &= \text{tr}[AA^T] \end{aligned} \quad (19)$$

where $A = (D_r^{-1/2}P - D_r^{1/2}P^*)D_c^{-1/2}$, we have

Theorem 4 $P^* = \mathbf{1}_n \mathbf{c}^T + \sum_{k=2}^s \tilde{\lambda}_k (D_r^{-1/2} \tilde{\mathbf{u}}_k) (D_c^{1/2} \tilde{\mathbf{v}}_k)^T$ with

rank $s < m$, where $D_r^{-1/2} P D_c^{-1/2} = \sum_{k=1}^m \tilde{\lambda}_k \tilde{\mathbf{u}}_k \tilde{\mathbf{v}}_k^T$ is an SVD.

Inertia

Homework 3 Suppose λ_k , \mathbf{u}_k and \mathbf{v}_k are from the SVD of $B = D_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^\top)D_c^{-1/2}$, we have

$$\mathbf{P}^* - \mathbf{1}_n\mathbf{c}^\top \approx \sum_{k=1}^{s-1} \lambda_k (\mathbf{D}_r^{-1/2}\mathbf{u}_k)(\mathbf{D}_c^{1/2}\mathbf{v}_k)^\top \quad (20)$$

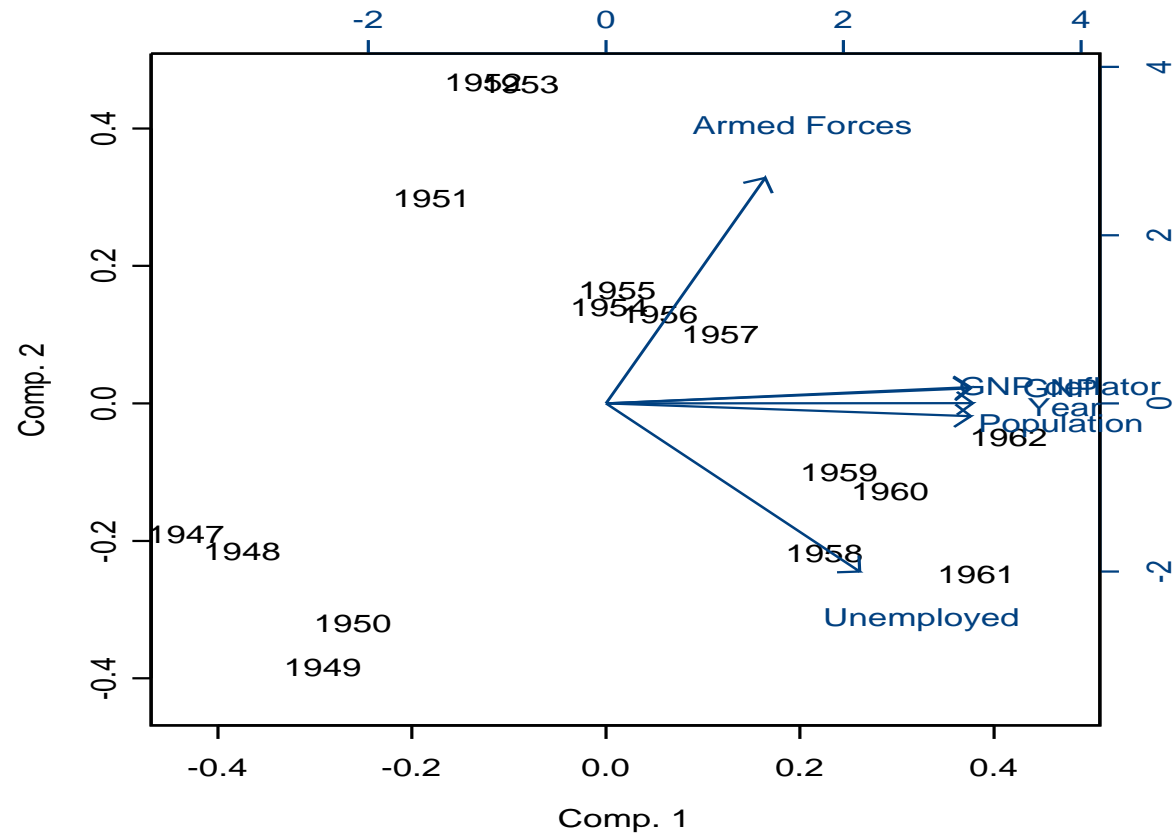
Total inertia^a is a measure of the variation in the count data and is defined as the weighted sum of squares

$$\begin{aligned} & \text{tr}[\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^\top)\mathbf{D}_c^{-1/2}(\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^\top)\mathbf{D}_c^{-1/2})^\top] \\ &= \sum_{i,j} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} = \sum_{k=1}^{m-1} \lambda_k^2 \end{aligned} \quad (21)$$

^aIt is just Pearson's χ^2 statistic by the contingency table.

Biplot

Definition 7 A **biplot** is a graphical representation in which both the observations and the variables are represented in a two dimensional space.



How to Make a Biplot?

1. Make the observation matrix $\mathbf{X}_{n \times m}$ to the mean corrected data matrix \mathbf{X}_c with rows $(\mathbf{x}_j - \bar{\mathbf{x}})^\top$.
2. SVD of $\mathbf{X}_c = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{V} = (\mathbf{e}_1, \dots, \mathbf{e}_m)$ are the eigenvectors of $\mathbf{X}_c^\top \mathbf{X}_c$.
3. $\mathbf{X}_c \approx \mathbf{U}\mathbf{\Lambda}^* \mathbf{V}^\top = (\mathbf{y}_1, \mathbf{y}_2) \begin{pmatrix} \mathbf{e}_1^\top \\ \mathbf{e}_2^\top \end{pmatrix}$, where $\mathbf{\Lambda}^* = \text{diag}(\lambda_1, \lambda_2, 0, \dots, 0)$.
4. How to make a biplot?
 - The i th variable is represented by (e_{1i}, e_{2i}) .
 - The n items are represented by matrix $(\mathbf{y}_1, \mathbf{y}_2)_{n \times 2}$.

Procrustes Analysis

Definition 8 A numerical comparison of two configurations, obtained by moving one configuration so that it aligns best with the other, is called **Procrustes analysis**. The comparison steps are

- \mathbf{X} is reduced to $\mathbf{X}_{n \times p}^*$ by method 1, and $\mathbf{Y}_{n \times q}^*$ by method 2, where $q \leq p$.
- By adding columns of zeros to \mathbf{Y}^* to a $n \times p$ matrix. Find a rigid transformation to make the **Procrustes residual sum of squares** minimum

$$\text{PR}^2 = \min_{\mathbf{Q}, \mathbf{b}} \sum_{j=1}^n (\mathbf{x}_j - \mathbf{Q}\mathbf{y}_j - \mathbf{b})^\top (\mathbf{x}_j - \mathbf{Q}\mathbf{y}_j - \mathbf{b}) \quad (22)$$

Measure of Agreement

Theorem 5 Let \mathbf{X}^* and \mathbf{Y}^* both be centered so that all rows have mean zero. Then,

$$PR^2 = \text{tr}[\mathbf{X}^* \mathbf{X}^{*\top}] + \text{tr}[\mathbf{Y}^* \mathbf{Y}^{*\top}] - 2\text{tr}[\mathbf{\Lambda}] \quad (23)$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$ and the rigid transformation is

$$\mathbf{Q} = \mathbf{V}\mathbf{U}^\top, \quad \mathbf{b} = \mathbf{0} \quad (24)$$

Here $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top = \mathbf{Y}^{*\top}\mathbf{X}^*$ is an SVD.

Definition 9 (Sibson Measure, 1978) Another measure for the agreement between two configurations is

$$\gamma = 1 - \frac{(\text{tr}[(\mathbf{Y}^{*\top}\mathbf{X}^*\mathbf{X}^{*\top}\mathbf{Y}^*)^{1/2}])^2}{\text{tr}[\mathbf{X}^{*\top}\mathbf{X}^*]\text{tr}[\mathbf{Y}^{*\top}\mathbf{Y}^*]} \quad (25)$$

Clustering of Words

Example 3 Given ten days *People's Daily*, classified in 6 categories. For every word, count the frequency in each category:

	news	economics	culture	politics	synthesis	computer
w_1	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
\vdots	\vdots					
w_n	f_{n1}	f_{n2}	f_{n3}	f_{n4}	f_{n5}	f_{n6}

Note SAS provides several ways to automatic clustering. We find that the random variables of observation are pivotal for the precision.

Example of SAS Code

```
data frequency;
  infile 'f:\data\cluster\frequency.txt';
  input word $1-20 news 21-30 economics 31-40 culture 41-50
  politics 51-60 synthesis 61-70 computer 71-80 sum 81-90;
run;
```

```
data frequency1;
  set frequency;
  IF 10<=sum;
run;
```

```
data frequency2;
  set frequency1;
  f_news=news/sum;f_economics=economics/sum;f_culture=culture/sum;
  f_politics=politics/sum;f_synthesis=synthesis/sum;f_computer=computer/sum;
run;
```

```
proc cluster data=work.frequency2 method=average
  outtree=otree pseudo ccc;
  var f_news f_economics f_culture f_politics f_synthesis f_computer;
  copy word;
run;
```

```
proc tree data=otree graphics
  horizontal nclusters=10 out=oclust;
  copy word;
run;
```

Class Average Method

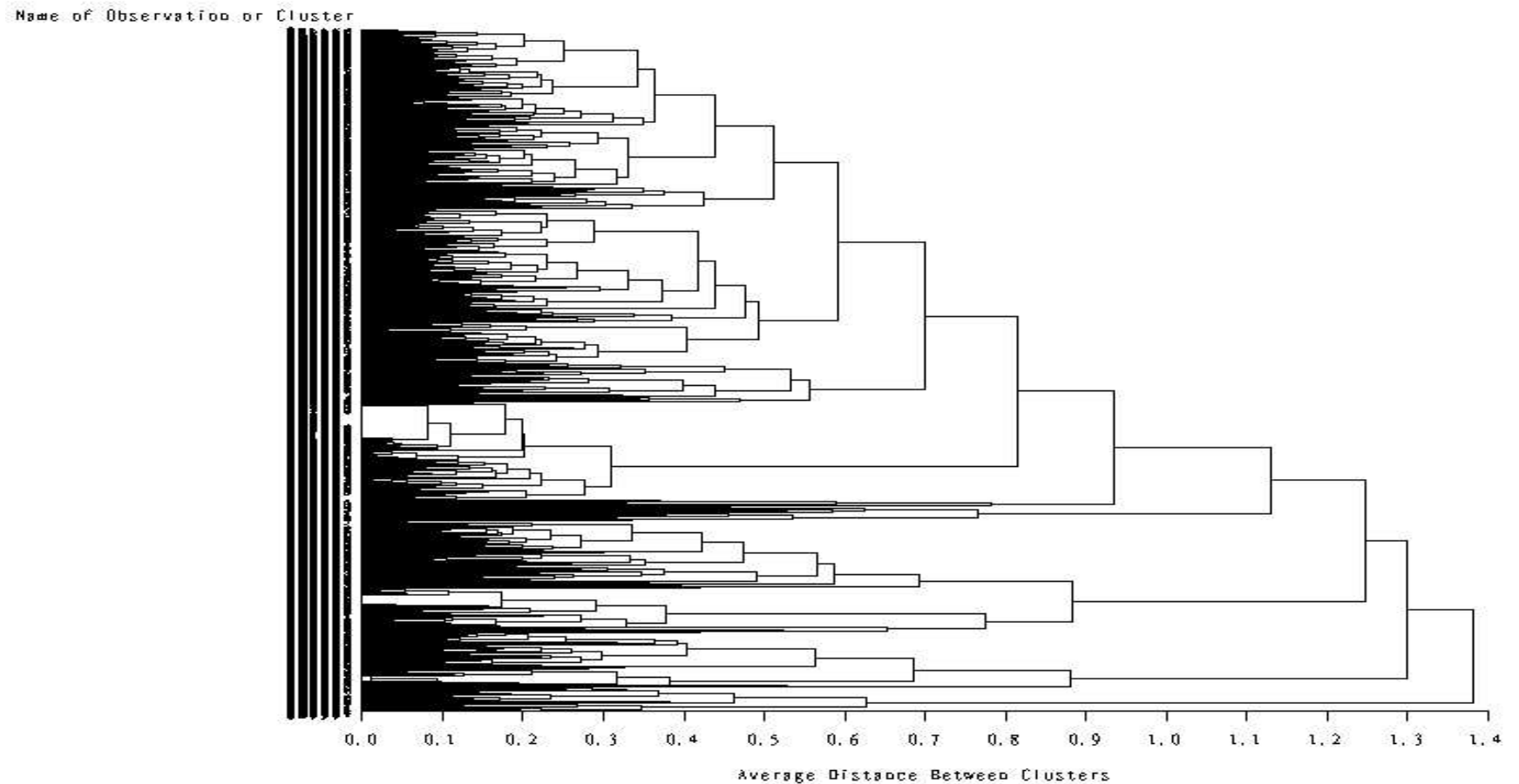


Figure 3: cluster tree by class average method

Ward Method

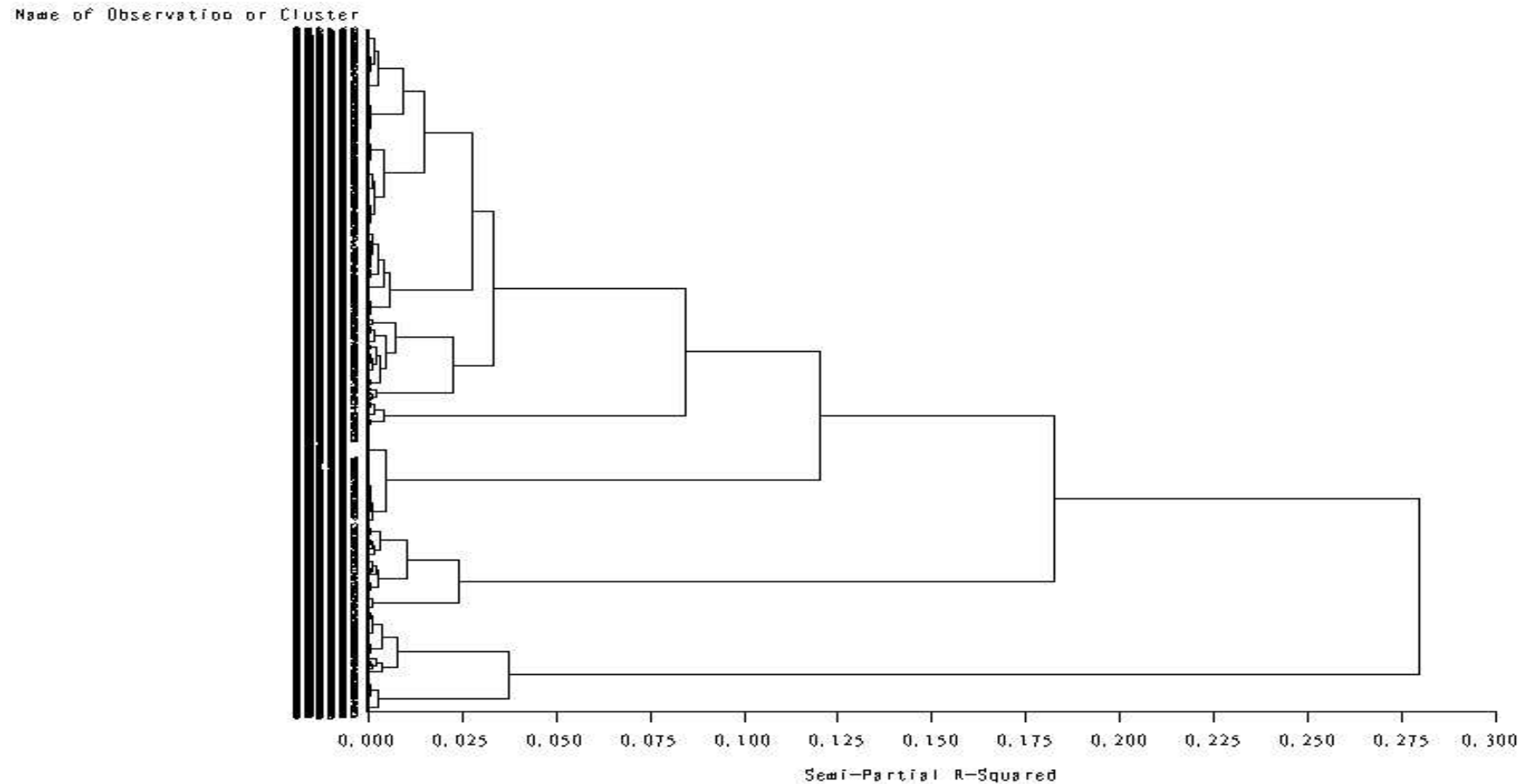


Figure 4: cluster tree by Ward method

What's a Class?

It is difficult to define *class* mathematically. In 1977, C. R. Rao once gave three definitions of class: Given a threshold value $d_0 > 0$, a set of C is a class if

1. $\forall \mathbf{x}_i, \mathbf{x}_j \in C$, we have $d(\mathbf{x}_i, \mathbf{x}_j) \leq d_0$.
2. $\forall \mathbf{x}_i \in C$, we have

$$\frac{1}{n-1} \sum_{\mathbf{x}_j \in C} d(\mathbf{x}_i, \mathbf{x}_j) \leq d_0 \quad (26)$$

3. given a threshold value $t_0 > d_0$ s.t. $d(\mathbf{x}_i, \mathbf{x}_j) \leq t_0$ and

$$\frac{1}{n(n-1)} \sum_{\mathbf{x}_i, \mathbf{x}_j \in C} d(\mathbf{x}_i, \mathbf{x}_j) \leq d_0 \quad (27)$$

References

1. P. J. Bickel and K. A. Doksum (2001), *Mathematical Statistics — Basic Ideas and Selected Topics* (Second Edition). Prentice-Hall, Inc.
2. Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland (1975), *Discrete Multivariate Analysis: Theory and Practice*. MIT.
3. M. J. Greenacre (1984), *Theory and Applications of Correspondence Analysis*. London: Academic Press.
4. R. A. Johnson and D. W. Wichern (1998), *Applied Multivariate Statistical Analysis*, Prentice-Hall, Inc.
5. SAS Institute Inc. (1998), *Introduction to Statistics Using SAS/INSIGHT Software*.
6. S. Theodoridis and K. Koutroumbas (2003), *Pattern Recognition* (Second Edition). Elsevier Science.
7. W. N. Venables and B. D. Ripley (1999), *Modern Applied Statistics with S-PLUS* (Third Edition). Springer-Verlag New York, Inc.



**Thank you
for your attention!**