Lecture 2

Thomas Woodson

Stony Brook University

Warm Up

- Load dataset about lizards. The dataset has 25 observations and 3 variables. The variables are mass, snout vent length (SVL), and hind limb span (HLS).

A. Find the descriptive statistics of the 3 variables **Trouvez les statistiques descriptives des 3 variables**

C. Find the covariance and correlation matrices of the data. Can you discuss any of the relationships between the variables? **Trouvez les matrices de covariance et de corrélation des données. Pouvez-vous discuter des relations entre les variables?**

Plotting

- R is a powerful program because it is easy to make very nice plots.

Example

- Example with strength of paper

Practice

- Create a scatter plot of the lizards data.

Grouping vs Classification

- Grouping (clustering). Unknown set of groups
- No assumptions made about data
- Based on (dis)similarity or distance measures
- Regroupement (clustering). Ensemble inconnu de groupes
- Aucune hypothèse faite sur les données
- Basé sur des mesures de (dis) similarité ou de distance

- Assign something to know group
- More assumptions about data and structure
- Attribuer quelque chose à connaître au groupe
- Plus d'hypothèses sur les données et la structure

Distance and clustering

Distance: $d(O,P) = \sqrt{x_1^2 + \cdots + x_p^2}$

$$d(\boldsymbol{x}, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2}$$

Standardized coordinates:

$$d(O,P) = \sqrt{(\frac{x_1}{\sqrt{s_{11}}})^2 + \cdots + (\frac{x_2}{\sqrt{s_{22}}})^2} = \sqrt{(\frac{x_1}{s_{11}})^2 + \cdots + (\frac{x_2}{s_{22}})^2}$$

$$d(\boldsymbol{x}, y) = \sqrt{(x - y)' S^{-1} (x - y)}$$

Other Distance Formulas

Canberra Metric

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

Minkowski Metric

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

Binary distance (Hamming Distance)

- Measuring distance between categorical data
- Mesurer la distance entre les données catégoriques

	V1 (hair)	V2 (job)	V3 (color)	V4 (state)	V5 (belief)
ITEM 1	1	0	0	1	1
ITEM 2	1	1	0	1	0

Contingency Table

		Item 1		
		1	**0**	**total**
Item 2	**1**	2	1	3
	0	1	1	2
	total	3	2	

	1	0	
1	a	B	A+B
0	c	D	C+D
	A+C	B+D	

Similarity Coefficients

How do we weight the similarities and differences?

Most common is equal weights for matches

$$\frac{a+d}{p}$$

	1	0	
1	a	B	A+B
0	c	D	C+D
	A+C	B+D	P=a+b+c+d

Example Human Characteristics

	Height	Weight	Eye	Hair	Handness	Gender
Individual 1	68	140	Green	Blond	Right	Female
Individual 2	73	185	Brown	Brown	Right	Male
Individual 3	67	165	Blue	Blond	Right	Male
Individual 4	64	120	Brown	Brown	Right	Female
Individual 5	76	210	brown	Brown	left	male

Example Human Characteristics

Define six binary variables $X_1, X_2, X_3, X_4, X_5, X_6$ as

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

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

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

The scores for individuals 1 and 2 on the $p = 6$ binary variables are

Example Human Characteristics

		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

$$\frac{a+d}{p} = \frac{1+0}{6} = \frac{1}{6}$$

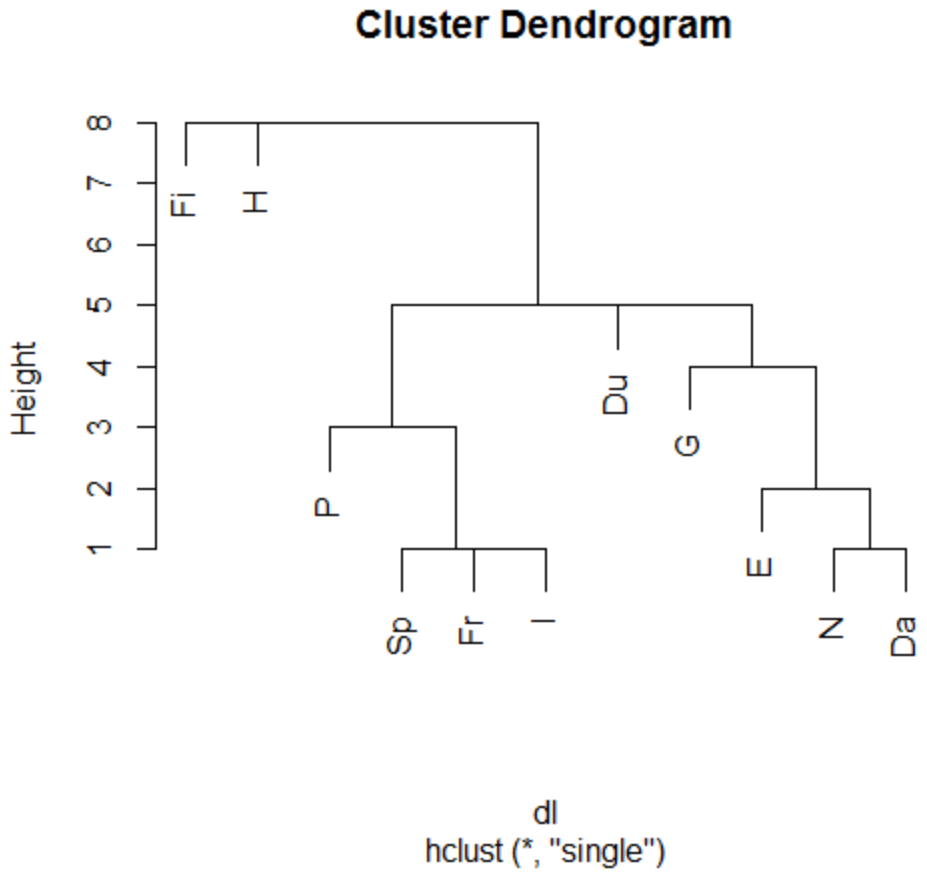
		Individual 2		
		1	0	Total
Individual 1	1	1	2	3
	0	3	0	3
	Totals	4	2	6

Individual

Individual	1	2	3	4	5
1	1				
2	$\frac{1}{6}$	1			
3	$\frac{4}{6}$	$\frac{2}{6}$	1		
4	$\frac{4}{6}$	$\frac{3}{6}$	$\frac{2}{6}$	1	
5	0	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	1

Hierarchical Clustering

- Single Linkage
  - Closest
  - le plus proche
- Complete Linkage
  - Most distance
  - Plus de distance
- Average Linkage
  - Average distance
  - Distance moyenne
- Ward's Linkage
  - Minimize information loss
  - Minimiser la perte d'informations



Single Linkage

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

Algorithm for Single linkage

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

$$d_{(35)1} = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3$$
$$d_{(35)2} = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7$$
$$d_{(35)4} = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8$$

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

$$d_{(135)2} = \min\{d_{(35)2}, d_{12}\} = \min\{7, 9\} = 7$$
$$d_{(135)4} = \min\{d_{(35)4}, d_{14}\} = \min\{8, 6\} = 6$$

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

$$d_{(135)(24)} = \min\{d_{(135)2}, d_{(135)4}\} = \min\{7, 6\} = 6$$

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

Single Linkage

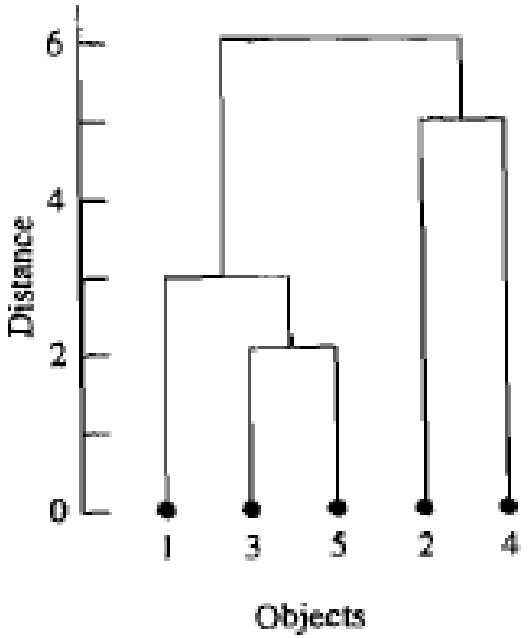
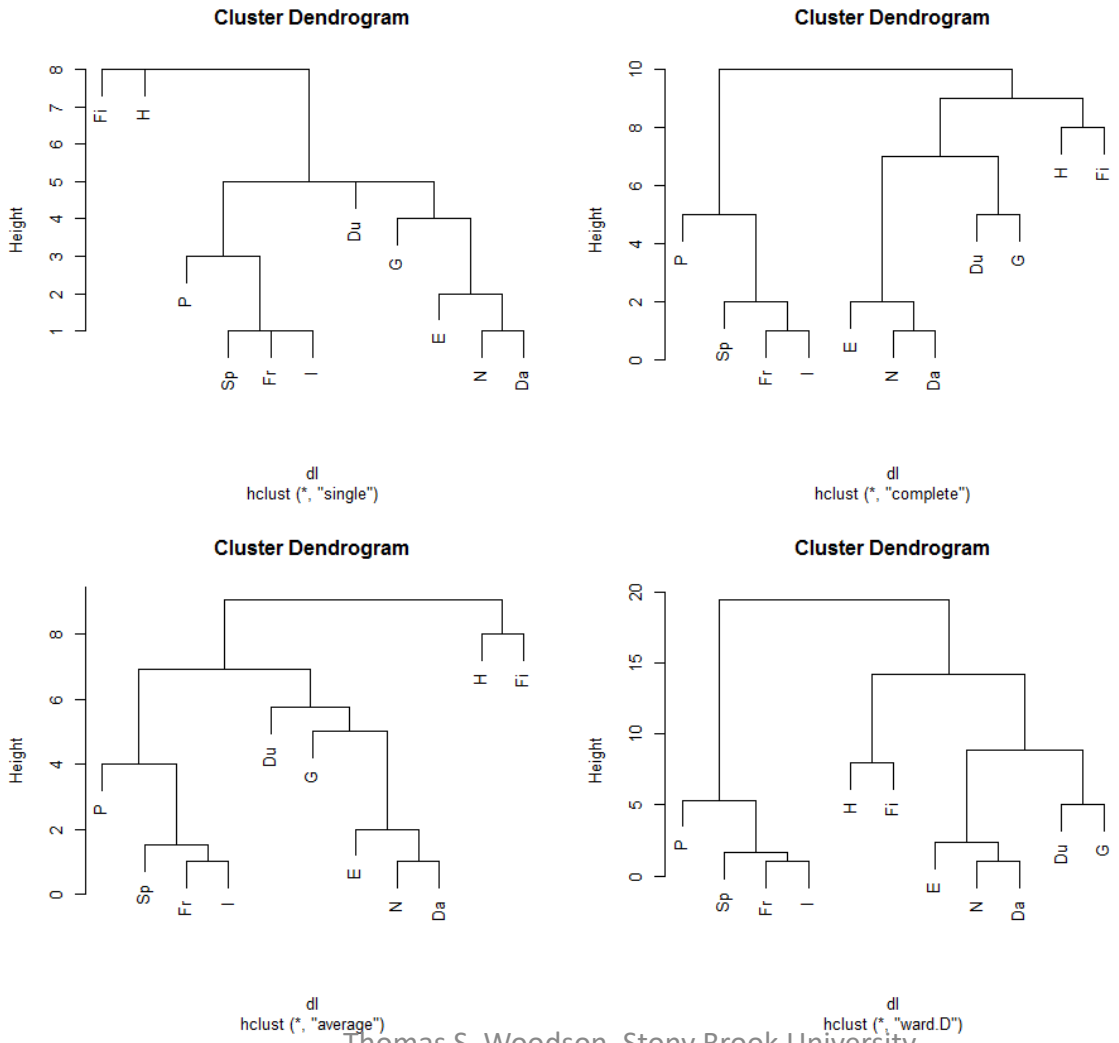


Figure 12.3 Single linkage dendrogram for distances between five objects.

R-Examples of clustering

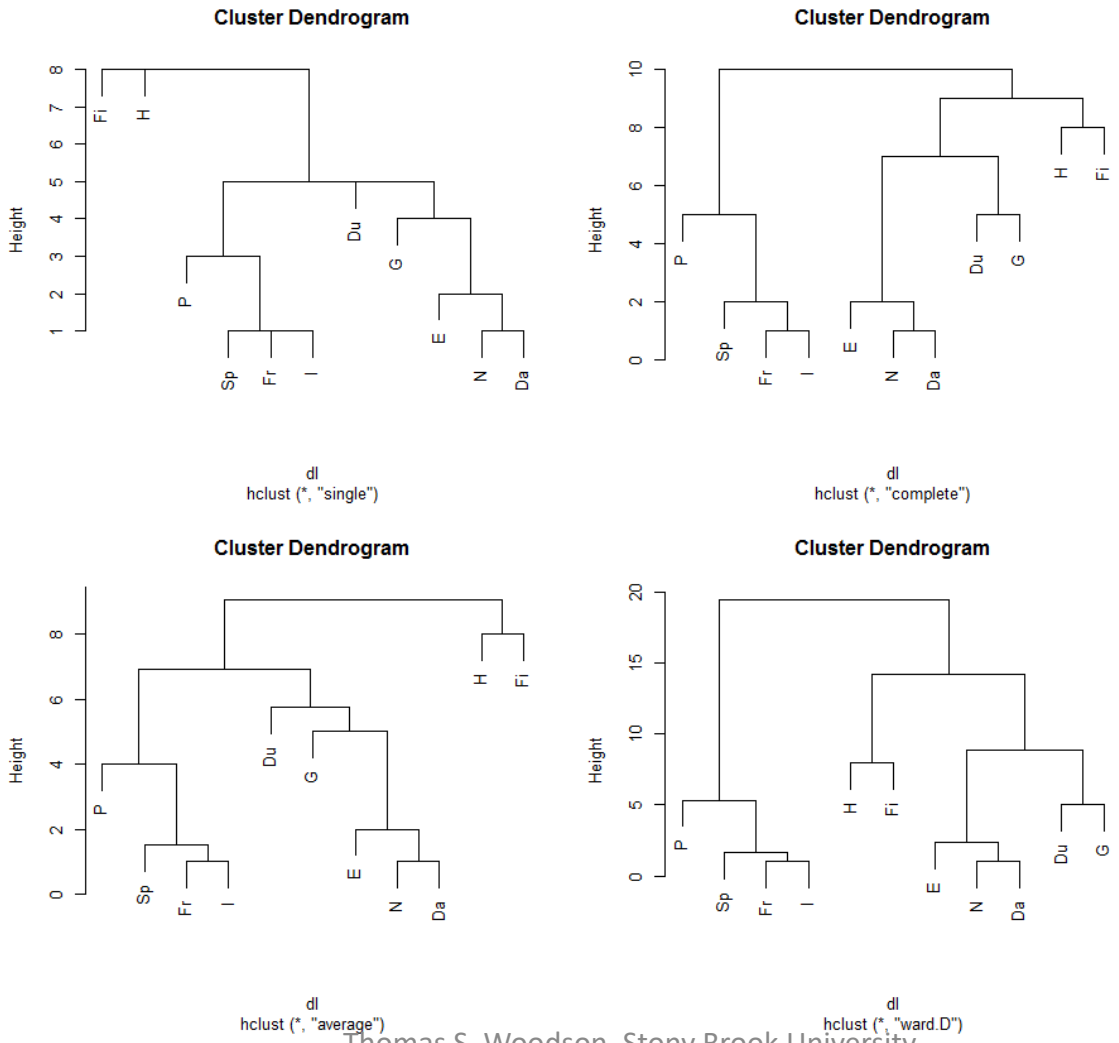


Counting to 10…

Table 12.2 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

R-Examples of clustering



R Code Demonstration

Practique

Using data called electricity

a. Find the correlation matrix of the variables

b. Create a dendrogram of the data using the single linkage method

c. Create a dendrogram of the data using the complete linkage method

d. Create a dendrogram of the data using the average linkage method

e. Create a dendrogram of the data using the Ward linkage method

a Trouver la matrice de corrélation (tableau 12.5)

b. Créer un dendrogramme des données en utilisant la méthode de liaison unique

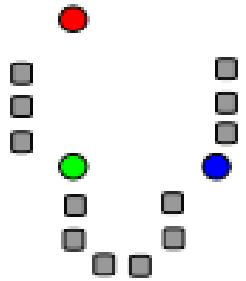
c. Créer un dendrogramme des données en utilisant la méthode de couplage complète

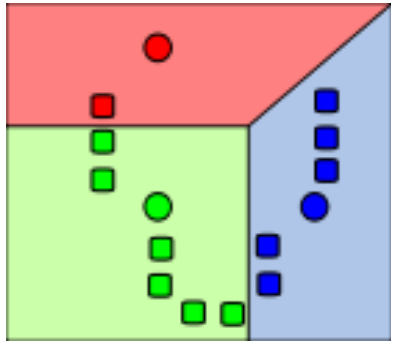
d. Créer un dendrogramme des données en utilisant la méthode de liaison moyenne

e. Créer un dendrogramme des données en utilisant la méthode de couplage de Ward

K-Means

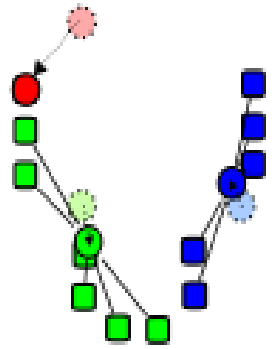
1. Select # of clusters
Randomly place them

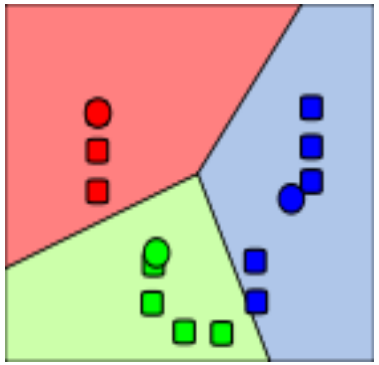




2. Create clusters based on nearest neighbor to random point

3. Find centroid of cluster. That centroid becomes new center





4. Repeat until there is a convergence

R-example

In-Class Problem 3

- Use the track dataset and calculate the Euclidean distances between pairs of countries
- Treating the distance as measures of (dis)similarity, cluster the countries using the single linkage and complete linkage hierarchical procedures and construct dendrogram
- Conduct a k-means analysis with 5 clusters of the track data
    - Use k-means to cluster the variables into K clusters. Compare to the previous part.