More Tree Models of Similarity and Association

Clustering and Classification

Lecture 6

Today's Class

- Hierarchical clustering methods.
  - Algorithm descriptions

- Estimating the fit of a tree model using multiple regression.

Preliminaries

- Our upcoming schedule will look like this:
  - 2/28 – Hierarchical tree algorithms and fit procedures.
  - 3/2 – How to fit trees in R/empirical research article.
    - Our article this week is now available online.
  - 3/7 – K-means clustering algorithms
  - 3/9 – How to do k-means in R/empirical research article.

Hierarchical Clustering Algorithms

Agglomerative Methods

- Next we discuss several different way to complete Agglomerative hierarchical clustering:
  - Single Linkage
  - Complete Linkage
  - Average Linkage
  - Centroid
  - Median
  - Ward Method

Agglomerative Methods

- In describing these…
  - I will give the definition of how we define similar when clusters are combined.
  - And for the first two I will give detailed examples.

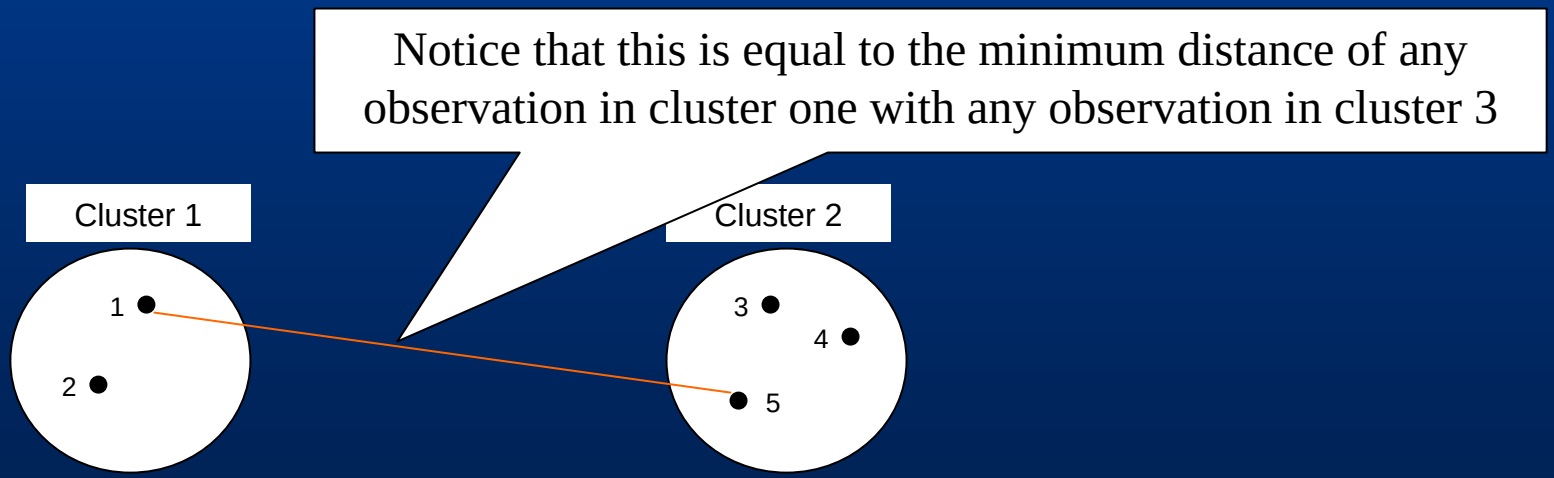
Example Distance Matrix

- The example will be based on the distance matrix below

	1	2	3	4	5
1	0	9	3	6	11
2	9	0	7	5	10
3	3	7	0	9	2
4	6	5	9	0	8
5	11	10	2	8	0

Single Linkage

- The single linkage method of clustering involves combining clusters by finding the "**nearest neighbor**" – the cluster closest to any given observation within the current cluster.



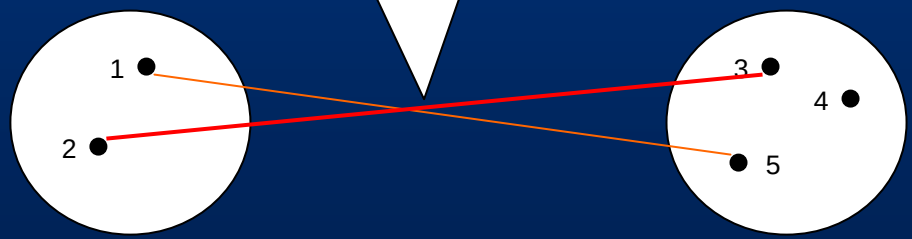
Single Linkage

- So the distance between any two clusters is:

$$D(A,B) = \min\{d(\mathbf{y}_i, \mathbf{y}_j)\}$$

For all y_i in A and y_j in B

Notice any other distance is longer



So how would we do this using our distance matrix?

Single Linkage Example

- The first step in the process is to determine the two elements with the smallest distance, and combine them into a single cluster.
- Here, the two objects that are most similar are objects 3 and 5…we will now combine these into a new cluster, and compute the distance from that cluster to the remaining clusters (objects) via the single linkage rule.

	1	2	3	4	5
1	0	9	3	6	11
2	9	0	7	5	10
3	3	7	0	9	2
4	6	5	9	0	8
5	11	10	2	8	0

Single Linkage Example

- The shaded rows/columns are the portions of the table w
  o
  5
- T
  cluster to the remain
  objects is given below:

Thes
5 wit
distar
1 is

These are the distances of 3 and 5 with 2. Our rule says that the distance of our new cluster with 2 is equal to the minimum of theses two values…7

	1	2	3	4	5
	0	9	3	6	11
	9	0	7	5	10
3	3	7	0	9	2
4	6	5	9	0	8
5	11	10	2	8	0

$d_{(35),1} = \min\{d_{31},d_{51}\} = \min\{3,11\} = 3$

$d_{(35),2} = \min\{d_{32},d_{52}\} = \min\{7,10\} = 7$

$d_{(35),4} = \min\{d_{34},d_{54}\} = \min\{9,8\} = 8$

In equation form our new distances are

Single Linkage Example

- Using the distance values, we now consolidate our table so that (35) is now a single row/column.
- The distance from the (35) cluster to the remaining objects is given below:

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

	(35)	1	2	4
(35)	0			
1	3	0		
2	7	9	0	
4	8	6	5	0

Single Linkage Example

- We now repeat the process, by finding the smallest distance between within the set of remaining clusters.
- The smallest distance is between object 1 and cluster (35).
- Therefore, object 1 joins cluster (35), creating cluster (135).

	(35)	1	2	4
(35)	0	3	7	8
1	3	0	9	6
2	7	9	0	5
4	8	6	5	0

The distance from cluster (135) to the other clusters is then computed:

d(135)2 = min{d(35)2,d12} = min{7,9} = 7

d(135)4 = min{d(35)4,d14} = min{8,6} = 6

Single Linkage Example

- Using the distance values, we now consolidate our table so that (135) is now a single row/column.
- The distance from the (135) cluster to the remaining objects is given below:

	(135)	2	4
(135)	0		
2	7	0	
4	6	5	0

$$d(135)2 = \min\{d(35)2, d12\} = \min\{7,9\} = 7$$
$$d(135)4 = \min\{d(35)4, d14\} = \min\{8,6\} = 6$$

Single Linkage Example

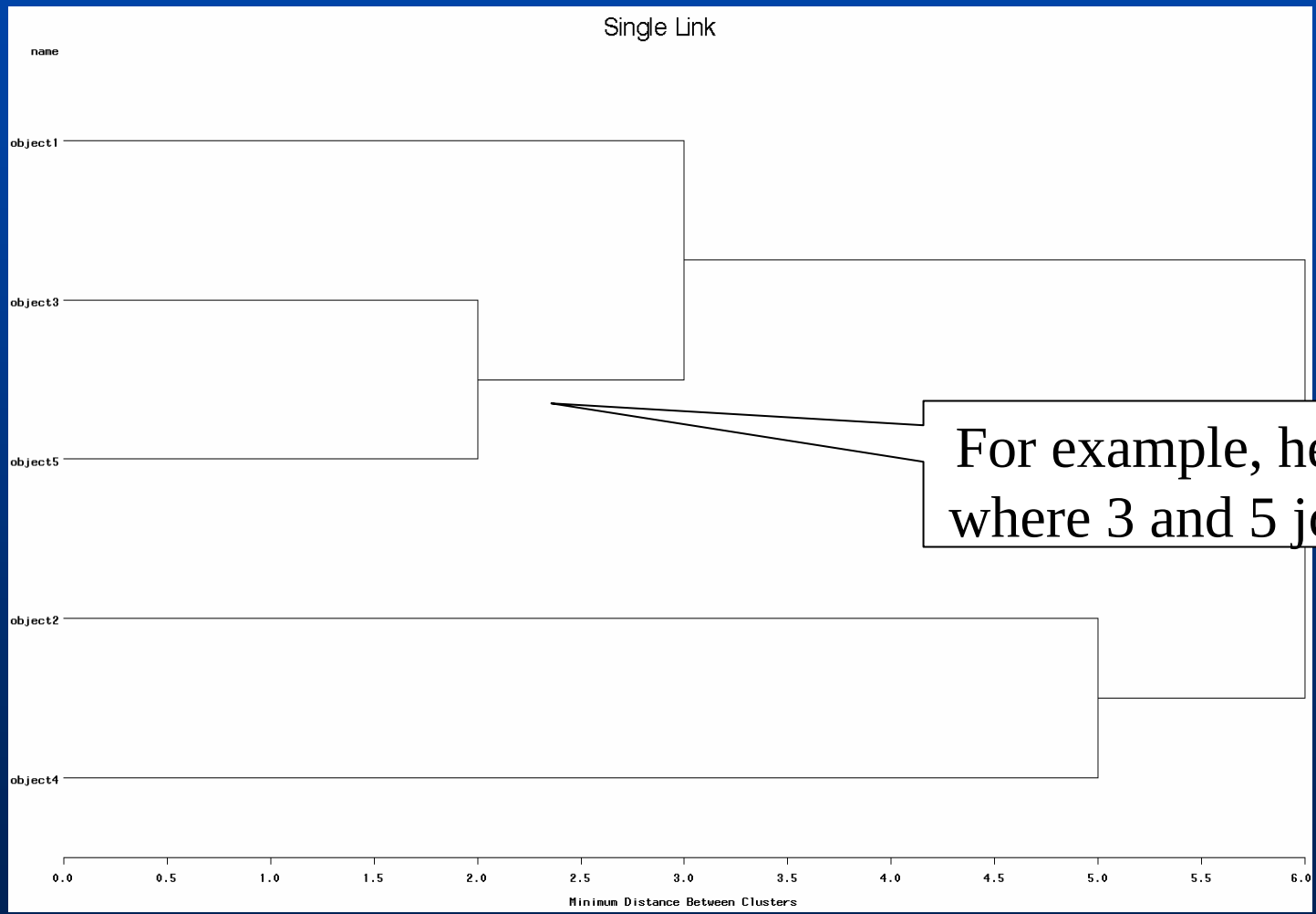
- We now repeat the process, by finding the smallest distance between within the set of remaining clusters.
- The smallest distance is between object 2 and object 4.
- These two objects will be joined to from cluster (24).
- The distance from (24) to (135) is then computed.

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

- The final cluster is formed (12345) with a distance of 6.

	(135)	2	4
(135)	0		
2	7	0	
4	6	5	0

The Dendogram



For example, here is where 3 and 5 joined

Complete Linkage

- The complete linkage method of clustering involves combining clusters by finding the “farthest neighbor” – the cluster farthest to any given observation within the current cluster.
- This ensures that all objects in a cluster are within some maximum distance of each other.

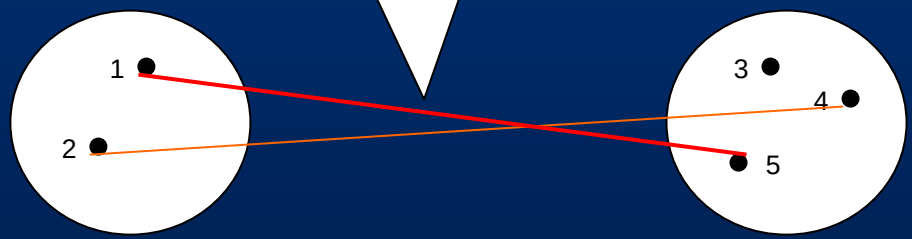
Complete Linkage

- So the distance between any two clusters is:

$$D(A,B) = \max\left\{d(\mathbf{y}_i, \mathbf{y}_j)\right\}$$

For all y_i in A and y_j in B

Notice any other distance is shorter



So how would we do this using our distance matrix?

Complete Linkage Example

- To demonstrate complete linkage in action, consider the five-object distance matrix.
- The first step in the process is to determine the two elements with the smallest distance, and combine them into a single cluster.
- Here, the two objects that are most similar are objects 3 and 5.
- We will now combine these into a new cluster, and compute the distance from that cluster to the remaining clusters (objects) via the complete linkage rule.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Complete Linkage Example

- The shaded rows/columns are the portions of the table with the distances from an object to object 3 or object 5.

- The distance from the (35) cluster to the remaining objects is given below:

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

	1	2	3	4	5
1	0	9	3	6	11
2	9	0	7	5	10
3	3	7	0	9	2
4	6	5	9	0	8
5	11	10	2	8	0

Complete Linkage Example

- Using the distance values, we now consolidate our table so that (35) is now a single row/column.

- The distance from the (35) cluster to the remaining objects is given below:

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

	(35)	1	2	4
(35)	0			
1	11	0		
2	10	9	0	
4	9	6	5	0

Notice our new computed distances with (35)

Complete Linkage Example

- We now repeat the process, by finding the smallest distance between within the set of remaining clusters.
- The smallest distance is between object 2 and object 4. Therefore, they form cluster (24).
- The distance from cluster (24) to the other clusters is then computed:

$d_{(24)(135)} = \max\{d_{2(35)}, d_{4(35)}\} = \max\{10,9\} = 10$

$d_{(24)1} = \max\{d_{21}, d_{41}\} = \max\{9,6\} = 9$

	(35)	1	2	4
(35)	0	11	10	9
1	11	0	9	6
2	10	9	0	5
4	9	6	5	0

So now we use our rule to combine 2 and 4

Complete Linkage Example

- Using the distance values, we now consolidate our table so that (24) is now a single row/column.

- The distance from the (24) cluster to the remaining objects is given below:

$d_{(24)(135)} = \max\{d_{2(35)}, d_{4(35)}\} = \max\{10,9\} = 10$

$d_{(24)1} = \max\{d_{21}, d_{41}\} = \max\{9,6\} = 9$

	(35)	(24)	1
(35)	0	10	11
(24)	10	0	9
1	11	9	0

Notice our 10 and 9

Complete Linkage Example

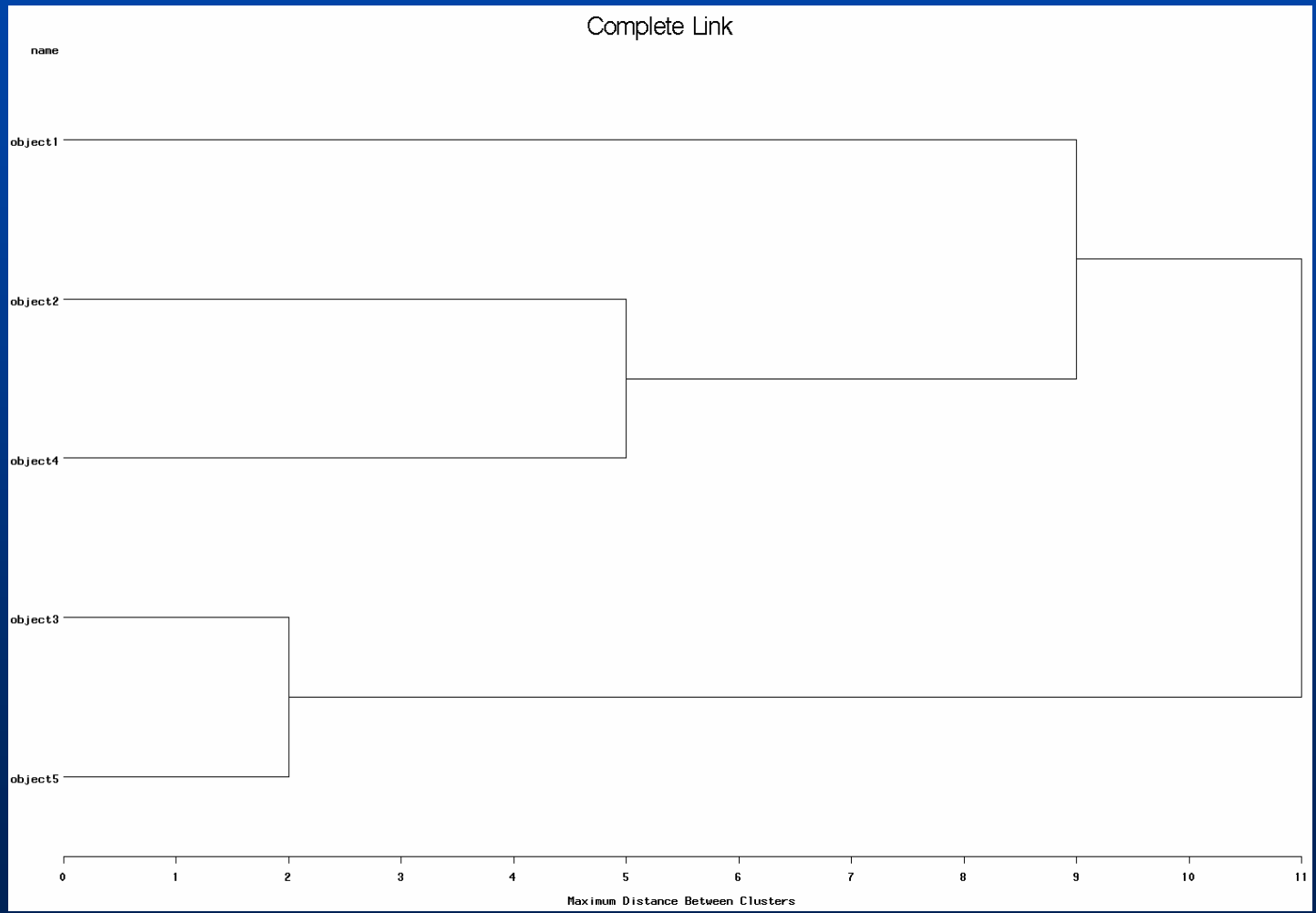
- We now repeat the process, by finding the smallest distance between within the set of remaining clusters.
- The smallest distance is between cluster (24) and object 1.
- These two objects will be joined to from cluster (124).
- The distance from (124) to (35) is then computed.

$d_{(35)(124)} = \max\{d_{1(35)}, d_{(24)(35)}\} = \max\{11,10\} = 11$

- The final cluster is formed (12345) with a distance of 11.

	(35)	(24)	1
(35)	0	10	
(24)	10	0	
1	11	9	0

The Dendogram



Average Linkage

- The average linkage method proceeds similarly to the single and complete linkage methods, with the exception that at the end of each agglomeration step, the distance between clusters is now represented by the average distance of all objects within each cluster.

- In reality, the average linkage method will produce very similar results to the complete linkage method.

So…

- To summarize…we have explained three of the methods to combine groups.
- Notice that once things are in the same group they cannot be separated
- The agglomeration method used is largely up to the user.

Using Multiple Regression to Assess Fit of Tree Models

Estimating Fit of a Tree Structure

- Corter (1996) describes a method for estimating the fit of a tree structure to a proximity matrix.
  - The fit metric is the familiar R^2 coefficient from Regression.

- R^2 indicates the amount of variance of a dependent variable is accounted for by the regression.

Fit Example

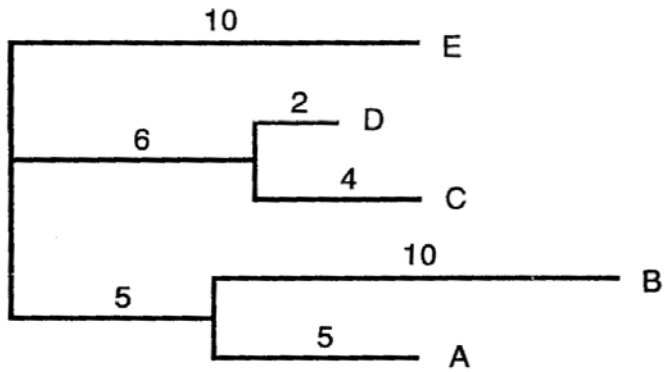
- To demonstrate the fit evaluation process, consider the following dissimilarity matrix (from p. 17):

	A	B	C	D	E
A	-				
B	15	-			
C	20	25	-		
D	18	23	6	-	
E	20	25	20	18	-

a. Dissimiliarities

	A	B	C	D	E
Worker A	--				
Worker B	15	--			
Worker C	20	25	--		
Worker D	18	23	6	--	
Worker E	20	25	20	18	--

b. Additive Tree



re 2.2. Additive Tree of Hypothetical Worker Communication Patterns

Beginning Steps

- We begin by reorganizing our dissimilarities matrix into a column vector:

```
> d=rbind(15,20,25,18,23,6,20,25,20,18)
> d
      [,1]
 [1,]   15
 [2,]   20
 [3,]   25
 [4,]   18
 [5,]   23
 [6,]    6
 [7,]   20
 [8,]   25
 [9,]   20
[10,]   18
```

	A	B	C	D	E
A	-				
B	15	-			
C	20	25	-		
D	18	23	6	-	
E	20	25	20	18	-

Model Matrix

- For the next step, we create a model matrix, **X**, which specifies which parameters enter into which interobject distances in the additive tree.
  - The parameters represent the arc length of the portion of the tree.
- The model matrix has n(n-1)/2 rows, which equals the number of proximity values.
- Each entry of the matrix is set to 1 if the parameter corresponding to that column is included in the path between the pair of objects corresponding to that row.

Model Matrix for Additive Trees

(B,A)	15		1	1	0	0	0	0	0
(C,A)	20		1	0	1	0	0	1	1
(C,B)	25		0	1	1	0	0	1	1
(D,A)	18		1	0	0	1	0	1	1
(D,B)	23	=	0	1	0	1	0	1	1
(D,C)	6		0	0	1	1	0	0	0
(E,A)	20		1	0	0	0	1	1	0
(E,B)	25		0	1	0	0	1	1	0
(E,C)	20		0	0	1	0	1	0	1
(E,D)	18		0	0	0	1	1	0	1

Column 7

Column 6

Column 5

Column 4

Column 3

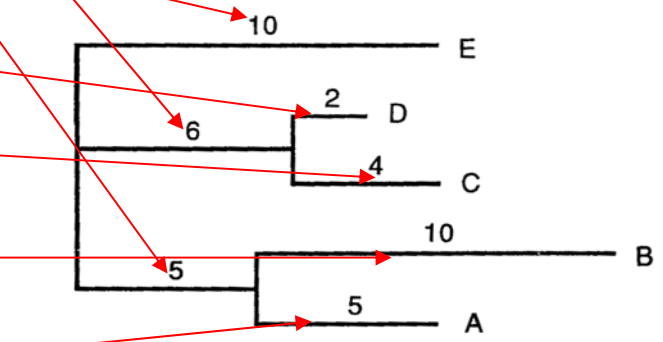
Column 2

Column 1

a. Dissimiliarities

	A	B	C	D	E
Worker A	--				
Worker B	15	--			
Worker C	20	25	--		
Worker D	18	23	6	--	
Worker E	20	25	20	18	--

b. Additive Tree



re 2.2. Additive Tree of Hypothetical Worker Communication Patterns

How Does This Work?

The distance from A to B is the sum of the arc lengths. Therefore:

$$d(B,A) = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + e$$
$$d(B,A) = 1b_1 + 1b_2 + 0b_3 + 0b_4 + 0b_5 + 0b_6 + 0b_7 + e$$
$$d(B,A) = b_1 + b_2$$

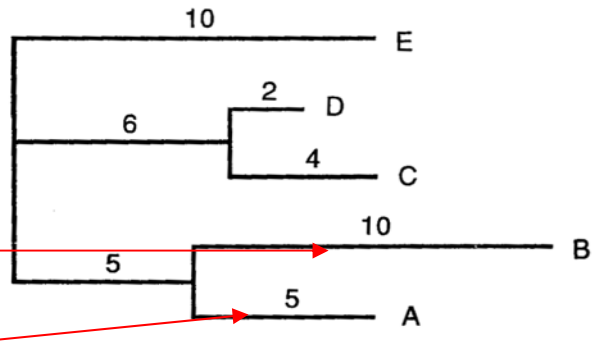
Column 2

Column 1

a. Dissimiliarities

	A	B	C	D	E
Worker A	--				
Worker B	15	--			
Worker C	20	25	--		
Worker D	18	23	6	--	
Worker E	20	25	20	18	--

b. Additive Tree

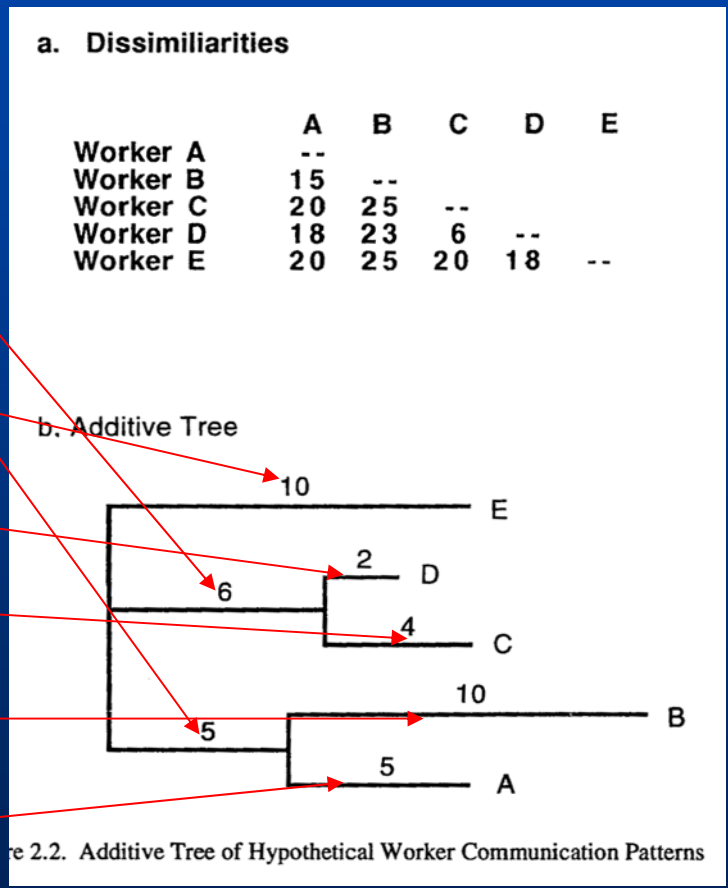


re 2.2. Additive Tree of Hypothetical Worker Communication Patterns

Now What?

- Once we have our model matrix, we simply input the matrix into a standard linear regression program.
  - The model must not be fit with an intercept.
- And for an example…

Column 7
Column 6
Column 5
Column 4
Column 3
Column 2
Column 1



re 2.2. Additive Tree of Hypothetical Worker Communication Patterns

What About Ultrametric Trees?

- The fit of ultrametric trees can also be assessed via linear regression.
- A similar procedure is used.
  - First we vectorize the distance matrix.
  - Next we create the model matrix.

- The model matrix has a more strict structure in the ultrametric tree fitting procedure.

An Ultrametric Example

- Notice that in the ultrametric trees, the paths within a cluster are constrained to be of equal length.

	Ar	Bu	Pe	Sh	Va
Arson	--				
Burglary	8	--			
Perjury	10	10	--		
Shoplifting	8	2	10	--	
Vandalism	4	8	10	8	--

b. Ultrametric Tree

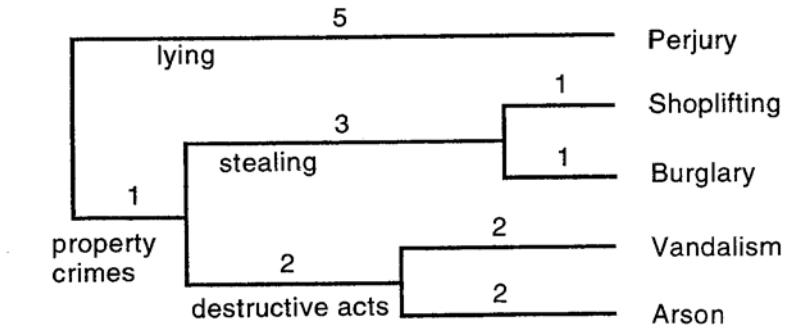


Figure 2.1. Ultrametric Tree of Hypothetical Dissimilarities Among Crimes

An Ultrametric Example

$$\begin{bmatrix} (bu,ar) & 8 \\ (pe,ar) & 10 \\ (pe,bu) & 10 \\ (sh,ar) & 8 \\ (sh,bu) & 2 \\ (sh,pe) & 10 \\ (va,ar) & 4 \\ (va,bu) & 8 \\ (va,pe) & 10 \\ (va,sh) & 8 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 2 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{bmatrix}$$

	Ar	Bu	Pe	Sh	Va
Arson	--				
Burglary	8	--			
Perjury	10	10	--		
Shoplifting	8	2	10	--	
Vandalism	4	8	10	8	--

b. Ultrametric Tree

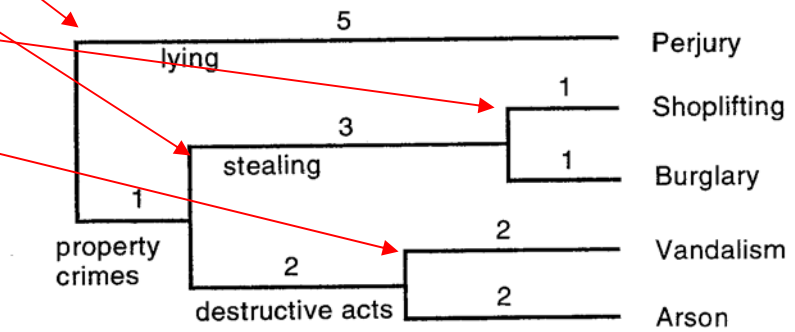


Figure 2.1. Ultrametric Tree of Hypothetical Dissimilarities Among Crimes

The parameters in the ultrametric represent the heights of the tree where nodes are fused.

Wrapping Up

- Agglomerative tree fitting procedures are widely available.
  - We will see that this also presents a problem when it comes to fit.

Next Time

- How to do hierarchical clustering in R.
- Presentation and discussion of an empirical research article featuring hierarchical clustering.