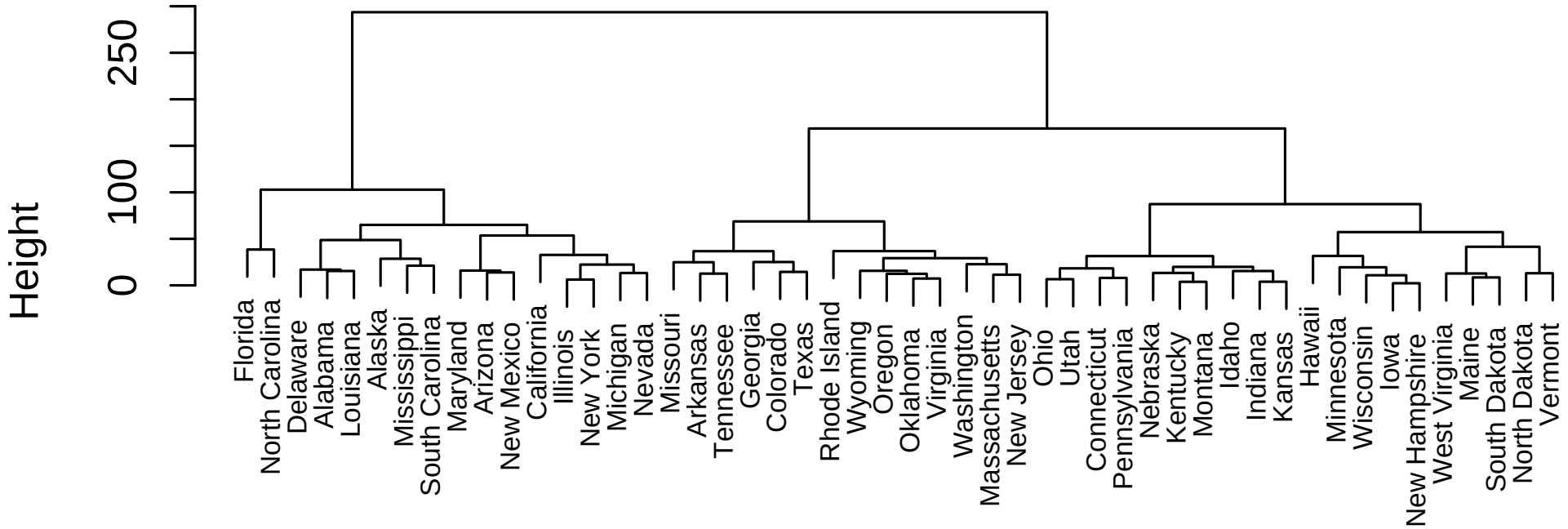
Multivariate Analysis (Slides 6)

- We will work through a hierarchical clustering to test our understanding and to determine the effect that different dissimilarity measures or linkages have.
- We will then discuss the phenomenon of "chaining" and how we may determine the appropriate number of groups within the data.
- Remember the objective is to cluster the data so that there is little dissimilarity within groups, but large dissimilarity between groups.

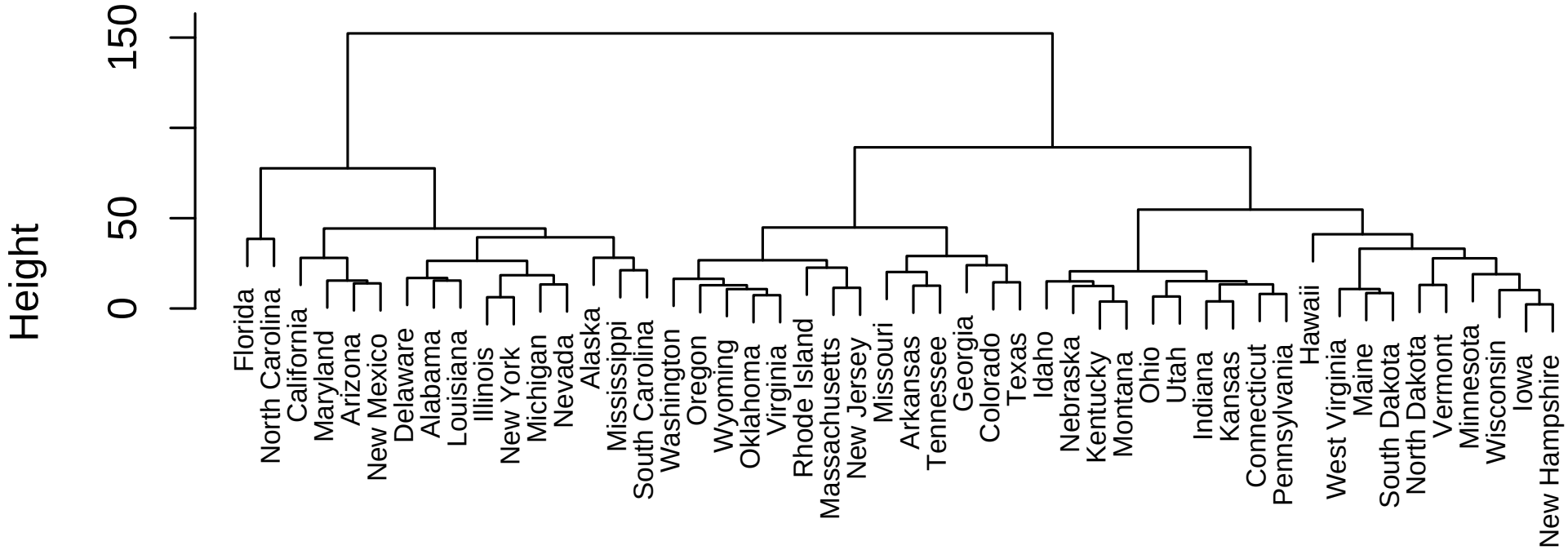
US Arrests: Complete Linkage

- Using complete linkage and Euclidean dissimilarity:



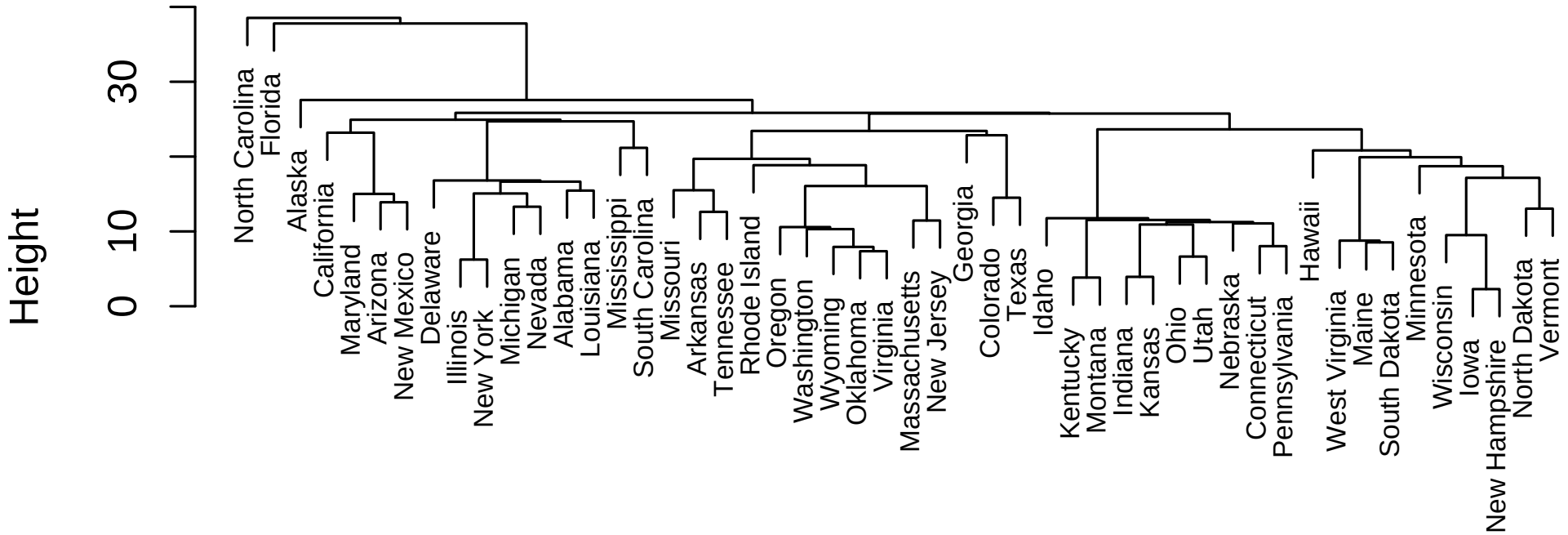
US Arrests: Average Linkage

- Using average linkage and Euclidean dissimilarity:



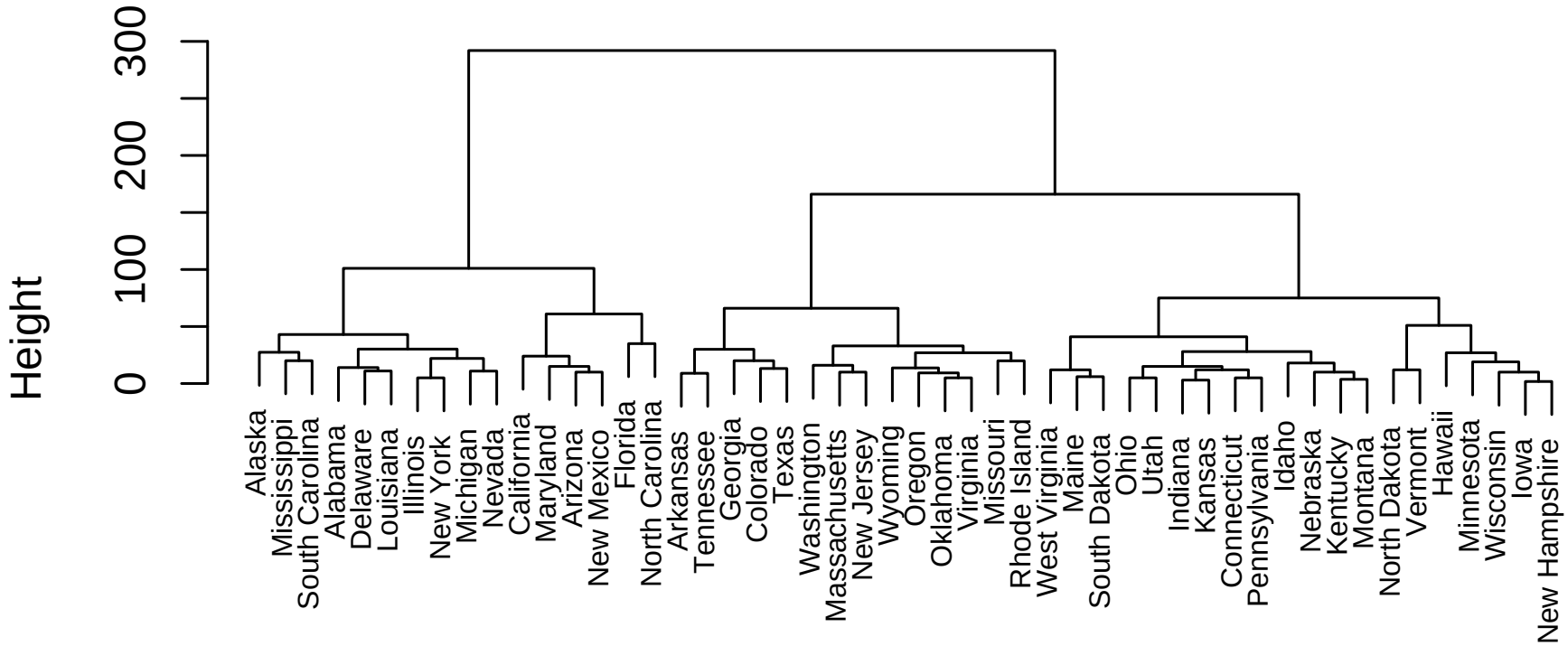
US Arrests: Single Linkage

- Using single linkage and Euclidean dissimilarity:



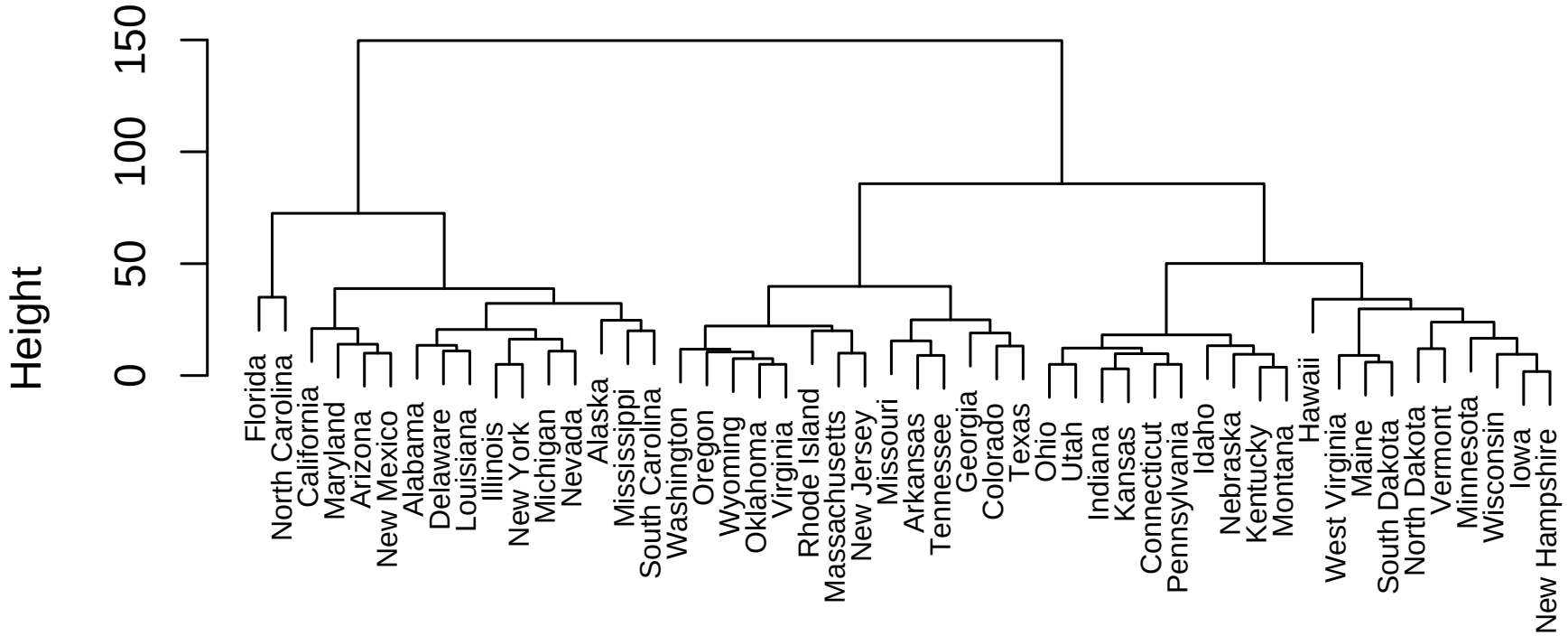
US Arrests: Complete Linkage

- Using complete linkage and maximum dissimilarity:



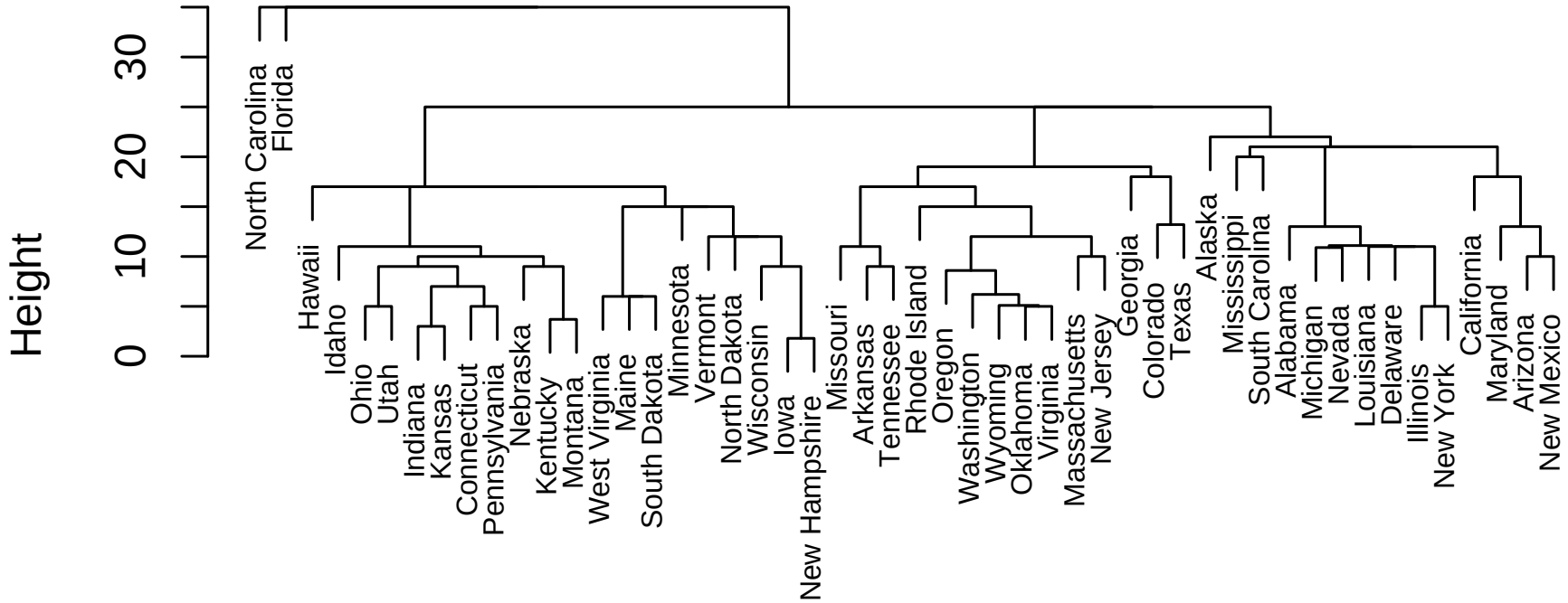
US Arrests: Average Linkage

- Using average linkage and maximum dissimilarity:



US Arrests: Single Linkage

- Using single linkage and maximum dissimilarity:



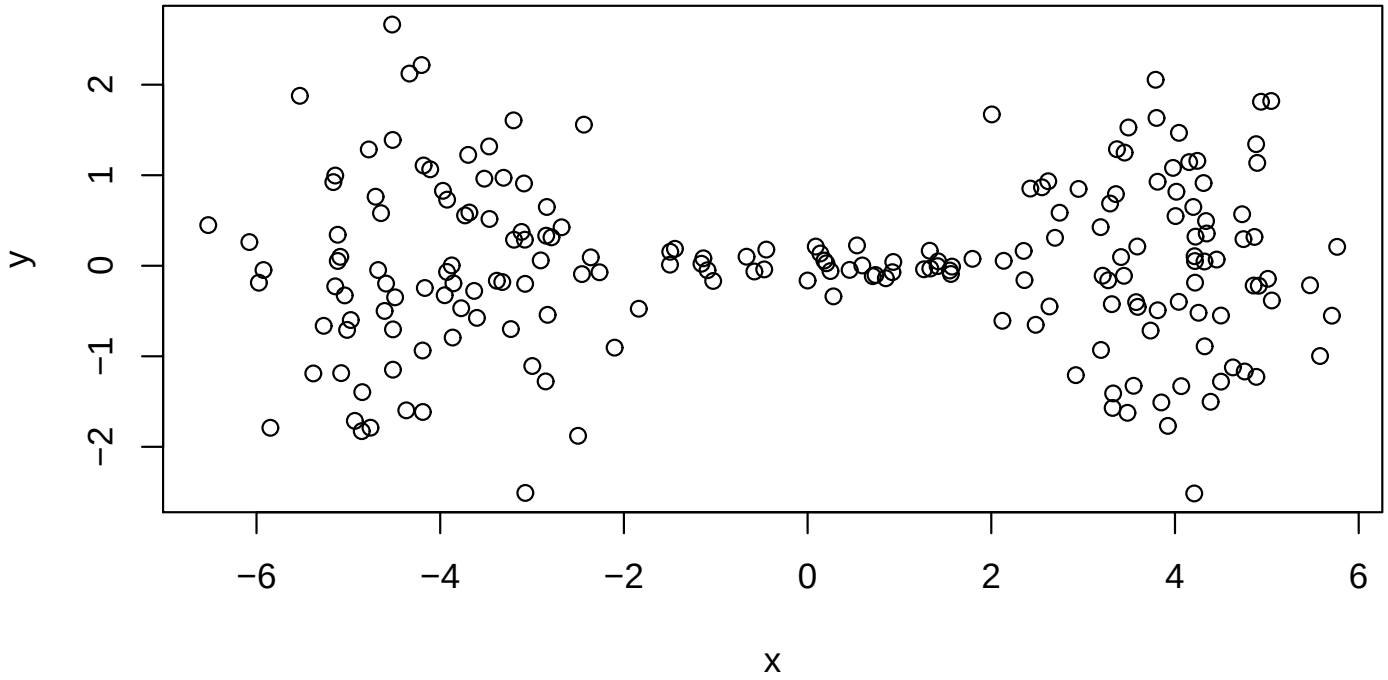
Linkage Effects

- Complete linkage joins the final clusters at a much larger measure of dissimilarity.
- Complete and average linkage result in 'spherical clusters', with good internal similarity.
- Single linkage displays outliers, whilst these are often hidden in complete linkage.
- Complete and single linkage are invariant under monotonic transformations, *e.g.*, taking logs, whilst average linkage is not.
- Complete linkage is likely to suggest a smaller number of large clusters with roughly equal size.
- Which linkage (or even dissimilarity measure) should be used requires careful thought, ideally before attempting the analysis. To try multiple versions and decide which output 'looks best' will return to using pre-conceptions, turning an 'objective' solution into a 'subjective' solution.

Chaining

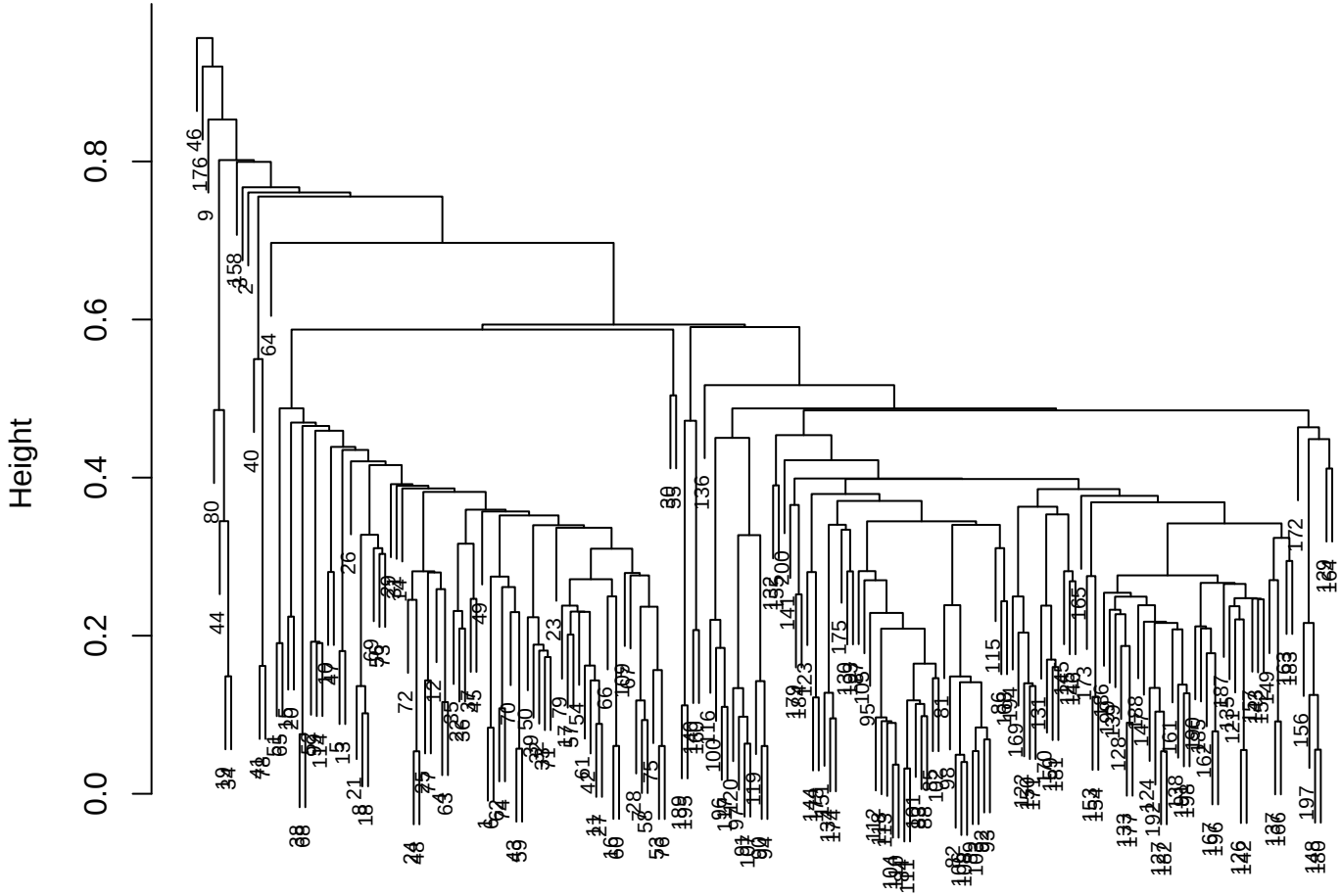
- Consider the dendrogram for single linkage with Euclidean dissimilarity. The tree shows a tendency to add a single observation to the same group that continues to get larger and larger.
- This phenomenon is called "chaining".
- Chaining occurs because a unit joins a group based on similarity with just one member of that group.
- Single Linkage is susceptible to this, resulting in elongated clusters that may include quite dissimilar points.
- This is not always bad, *e.g.*, evolutionary chain mechanisms.

Chaining



Chaining Example

- The previous plot of bivariate data appears to show two clusters at either end of the graph, with a few points scattered in between.

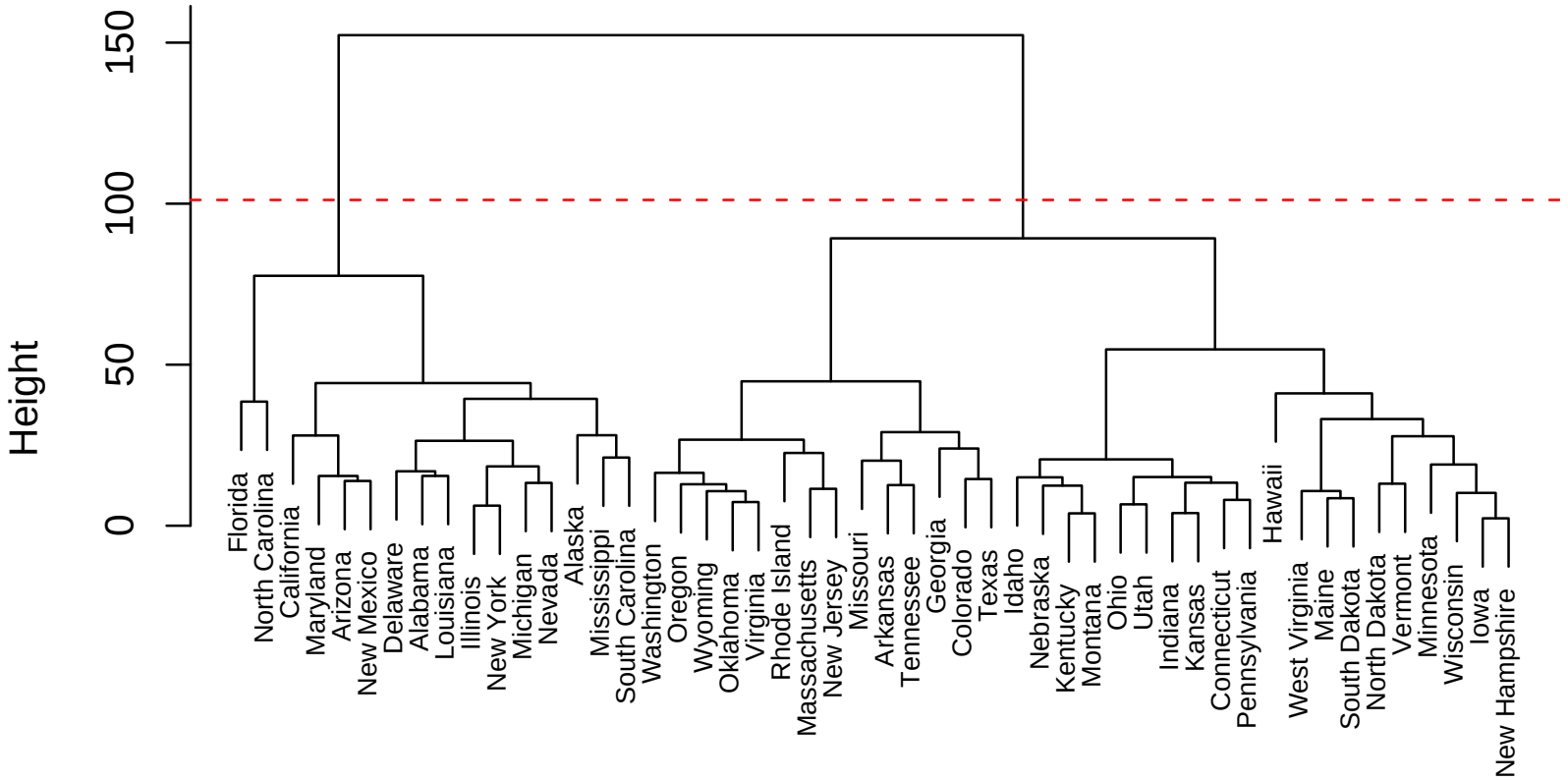


How Many Groups?

- How do we decide if groups are present, and if so, how many?
- One possibility is to look in the dendrogram for joints that happen at very large height values.
- This is because the height on the dendrogram is interpretable through the method of linkage used.
- We could make use of background knowledge for the problem, but an often suggested rule is to cut the tree at $\overline{h} + 3s_h$, where $\overline{h}$ is the average height of the joints and s_h is the standard deviation of the heights.

US Arrests: How Many Groups?

- A horizontal line is drawn at $\overline{h} + 3s_h$, suggesting two groups.



Performance

- A good way to assess the performance of the various clustering choices is to look at their performance on artificial data that has been created to include specific and known group structure.
- This allows a test to determine if the method does indeed find the correct structure when it is known to exist.
- After all, hierarchical clustering is an exploratory tool rather than a specialist clustering tool. If you want specific cluster memberships for each observation there are better clustering techniques available.

Cluster Agreement

- Suppose that two different clustering methods are applied to the same data.
- **Example:** Using both single-linkage hierarchical clustering and complete-linkage clustering.
- **Example:** An expert may cluster the observations using his expert knowledge and a statistician may cluster the data not knowing what it represents.
- Can we quantify the level of agreement between the two approaches?

Cross Tabulation

- A cross tabulation of the cluster memberships from the two methods permits a comparison of results.
- For example,

Table 1

		Method B		
		Cluster 1	Cluster 2	
Method A	Cluster 1	30	10	40
	Cluster 2	15	60	75
		45	70	115

- Do the results of the two clustering methods agree?

Examples

- Do we have agreement here?

Table 2

		Method B		
		Cluster 1	Cluster 2	
Method A	Cluster 1	10	30	40
	Cluster 2	60	15	75
		70	45	115

- What about here?

Table 3

		Method B		
		Cluster 1	Cluster 2	
Method A	Cluster 1	24	16	40
	Cluster 2	46	29	75
		70	45	115

Trickier Example

- What about when the methods suggest a different number of groups?

Table 4

		Method B		
		Cluster 1	Cluster 2	
Method A	Cluster 1	10	30	40
	Cluster 2	20	5	25
	Cluster 3	40	10	50
		70	45	115

Rand Index

- Rand (1971) proposed an index for measuring the agreement between two clusterings.
- The Rand Index is a number between 0 and 1, with 0 representing little agreement and 1 representing strong agreement.
- The formula for the Rand Index is given by,

$$\text{Rand Index} = \frac{\binom{n}{2} + 2\sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{c_1}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{c_2}\binom{n_{\cdot j}}{2}\right]}{\binom{n}{2}}$$

- Here n_{ij} is the number of points that are in cluster i for Method A and cluster j for Method B, c_1 is the number of clusters for Method A and c_2 the number of clusters for Method B. Also, $n_{\cdot j} = \sum_{i=1}^{c_1} n_{ij}$, $n_{i\cdot} = \sum_{j=1}^{c_2} n_{ij}$, and $n = n_{\cdot\cdot} = \sum_{i=1}^{c_1}\sum_{j=1}^{c_2} n_{ij}$.

Rand Index Quantities

- The quantities in the Rand Index are found through the cross tabulation entries:

		Method B				
		Cluster 1	Cluster 2	$\cdots$	Cluster c_2	
Method A	Cluster 1	n_{11}	n_{12}	$\cdots$	n_{1c_2}	$n_{1\cdot}$
	Cluster 2	n_{21}	n_{22}	$\cdots$	n_{2c_2}	$n_{2\cdot}$
		$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	Cluster c_1	$n_{c_1 1}$	$n_{c_1 2}$	$\cdots$	$n_{c_1 c_2}$	$n_{c_1\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c_2}$	$n_{\cdot\cdot}$

Rand Index Explained

- Let us break down the formula for the Rand Index:
- Let α be the number of pairs of data points so that both data points fall in the same cluster under method A and also fall in the same cluster under method B.
- Similarly let β be the number of pairs so that both do not fall in the same cluster under method A, and also do not fall in the same cluster under method B.
- Let γ be the number of pairs so that both fall in the same cluster under method A, but do not fall in the same cluster under method B.
- Let δ be the number of pairs so that both do not fall in the same cluster under method A, but do fall in the same cluster under method B.
- Finally, note that the number of possible pairs of data points is:

$$\binom{n}{2} = \alpha + \beta + \gamma + \delta$$

.

Rand Index Explained

- Hence:

$$R = \frac{\alpha + \beta}{\alpha + \beta + \gamma + \delta} = \frac{\binom{n}{2} - \gamma - \delta}{\binom{n}{2}}$$

- Here:

$$\gamma = \sum_{i=1}^{c_1} \binom{n_{i.}}{2} - \sum_{i=1}^{c_1} \sum_{j=1}^{c_2} \binom{n_{ij}}{2}$$

- And:

$$\delta = \sum_{j=1}^{c_2} \binom{n_{.j}}{2} - \sum_{i=1}^{c_1} \sum_{j=1}^{c_2} \binom{n_{ij}}{2}$$

- In determining γ and δ we effectively find the number of pairs of points that share a cluster under one method, and subtract the number of pairs of points that share a cluster under both methods.

Rand Index: Values

- The Rand Index can be computed quite easily.
- For Tables 1 to 4 previously shown we find the following values:

Table	Rand Index
1	0.657
2	0.657
3	0.499
4	0.588

- The value for Table 3 is still quite large, but there appeared to be little agreement between the clustering methods used to create it.

Example

- **Exercise:** Calculate the Rand Index value for the following:

Table		Method B		
		Cluster 1	Cluster 2	
Method A	Cluster 1	5	25	30
	Cluster 2	20	5	25
	Cluster 3	40	5	45
		65	35	100

Rand Index: Problems

- The Rand Index tends to give quite large values even when clustering methods are in substantial disagreement.
- Even a random assignment of points to clusters can lead to large Rand Index values.
- Hubert and Arabie (1985) proposed an adjustment to the Rand Index in order to account for agreement by chance. They did this by considering a distribution for assigning points to clusters under the condition that cluster sizes remained unchanged.

Adjusted Rand Index

- The Adjusted Rand Index is calculated as follows,

$$\text{Adjusted Rand} = \frac{\binom{n}{2}\sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\binom{n_{ij}}{2} - \sum_{i=1}^{c_1}\binom{n_{i\cdot}}{2}\sum_{j=1}^{c_2}\binom{n_{j\cdot}}{2}}{\frac{1}{2}\binom{n}{2}\left[\sum_{i=1}^{c_1}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{c_2}\binom{n_{\cdot j}}{2}\right] - \sum_{i=1}^{c_1}\binom{n_{i\cdot}}{2}\sum_{j=1}^{c_2}\binom{n_{\cdot j}}{2}}$$

- It arises from:

$$\text{Adjusted Rand} = \frac{\text{Rand Index} - \text{Expected Rand Index}}{\text{Max Rand Index} - \text{Expected Rand Index}}$$

- Details omitted.

Adjusted Rand Index: Values

- This index can be negative (corresponding to very low agreement), but it can't be greater than 1.
- The Adjusted Rand Index can also be computed quite easily.
- For Tables 1 to 4 previously shown we find the following values:

Table	Rand Index
1	0.311
2	0.311
3	-0.006
4	0.185

- Now we have a small value for Table 3 as we would expect.