Multivariate Analysis (Slides 5)

- Rather than dimension reduction, which we will return to later, today we will examine cluster analysis.
- The aim of cluster analysis is to establish if there is a group structure in the data set.
- If we establish that there is group structure, then we are interested in knowing how many groups are present and their particular structures.
- **Example:** Consider the data for US arrests previously presented at the beginning of the course.

  For each state in the US, the number of arrests per 100,000 residents for assault, murder, and rape was recorded.

  The percentage of the population living in urban areas was also recorded.
- Of interest is whether there are groups of similar states? If so, how many, and what characterizes the groups?

Similarity/Dissimilarity

- We want to place observations in groups according to their similarity.
- To do this we need to decide what is meant by "similarity" or "dissimilarity" (or distance).
- Many proposals have been made, depending on the type of data, *e.g.*, for measurement, categorical, or binary data.
- Some authors talk about similarities, but others talk about dissimilarities. The two approaches are of course equivalent.
- Some common properties of a dissimilarity measure $d(\mathbf{x}, \mathbf{y})$ are that:

  1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ and $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$
  2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
  3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ (occasionally ignored)

Example: Protein Mutation Distances

- Fitch and Margoliash (Science, 1967) recorded the dissimilarity between species by recording the number of positions in the protein molecule cytochrome-c where the two species have different amino acids.
- A subset of the dissimilarity that they recorded is:

	Man	Monkey	Horse	Pig	Pigeon	Tuna	Mould	Fungus
Man	0	1	17	13	16	31	63	66
Monkey	1	0	16	12	15	32	62	65
Horse	17	16	0	5	16	27	64	68
Pig	13	12	5	0	13	25	64	67
Pigeon	16	15	16	13	0	27	59	66
Tuna	31	32	27	25	27	0	72	69
Mould	63	62	64	64	59	72	0	61
Fungus	66	65	68	67	66	69	61	0

Dissimilarities

- There have been many proposed dissimilarity measures for measurement data, *i.e.*, for data such that observation i is of the form $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{im})$ with $x_{ik} \in \mathbb{R}$. For example, the following have all been suggested for $d(\mathbf{x}_i, \mathbf{x}_j)$.
- Euclidean: $\sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$
- Absolute Distance (Manhattan): $\sum_{k=1}^m |x_{ik} - x_{jk}|$
- Maximum: $\max_{k \in \{1,2,\ldots,m\}} |x_{ik} - x_{jk}|$
- Minkowski: $[\sum_{k=1}^m |x_{ik} - x_{jk}|^p]^{\frac{1}{p}}$, $(p \geq 1)$
- There are many such possibilities.
- **Exercise:** Determine whether or not these satisfy the three suggested properties of a dissimilarity measure.

Standardization

- When constructing a dissimilarity matrix we need to be aware of how our data are scaled.
- It is important that different variables are comparably scaled, as if they are not, the variable with the greatest variance will figure most prominently in the clustering solution.
- Hence variables are generally standardized by dividing through by their standard deviation before being used to calculate a dissimilarity matrix.
- Each variable will then have variance equal to 1.
- Sometimes this may not be advisable as we may wish to give less weight to a variable that carries less information, *e.g.*, if it has very small variance.

Example: US Arrests

- Consider measuring the dissimilarity between states Alabama and Wyoming in the US arrests data.
- The data values are $\mathbf{x}_1^T = (13.2, 236, 58, 21.2)$ and $\mathbf{x}_{50}^T = (6.8, 161, 60, 15.6)$.
- We get the following results:

DISSIMILARITY	VALUE
Euclidean	75.5
Manhattan	89.0
Maximum	75.0
Minkowski (p=3)	75.0
Minkowski (p=1.3)	79.8

Binary Data

- For data that consists of binary variables, that is, $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{im})$ with $x_{ik} \in \{0, 1\}$, we can consider dissimilarity by looking at a cross tabulation of the number of 0's and 1's for each data point.

		Point j		
		1	**0**	
Point i	**1**	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	$m = a+b+c+d$

Example: Binary Data

- **Example:** Suppose $\mathbf{x}_i^T = (1, 1, 0, 0, 0, 0, 1)$ and $\mathbf{x}_j^T = (1, 0, 1, 1, 1, 0, 0)$.

		Point j		
		1	**0**	
Point i	**1**	$a = 1$	$b = 2$	$a + b = 3$
	0	$c = 3$	$d = 1$	$c + d = 4$
		$a + c = 4$	$b + d = 3$	$m = a + b + c + d = 7$

Binary Dissimilarities

- Again many dissimilarities measures have been proposed for binary data (certainly over 40). In particular, there are suggestions that $d(\mathbf{x}_i, \mathbf{x}_j)$ be given by:
- Simple Matching (Hamming): $1 - \frac{a+d}{a+b+c+d}$
  - Proportion of variables in agreement.
- Jaccard: $1 - \frac{a}{a+b+c}$
  - Ignore double absence, as may be a redundant variable.
- Kulczynski: $1 - \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right)$
  - Average of ratios of agreement from two samples.
- Czekanowski: $1 - \frac{2a}{2a+b+c}$
  - More emphasis on double presence than double absence.
- **Exercise:** Determine whether or not these satisfy the three suggested properties of a dissimilarity measure.

Example: Binary Data

- The dissimilarities between the binary data values $\mathbf{x}_i = (1, 1, 0, 0, 0, 0, 1)$ and $\mathbf{x}_j = (1, 0, 1, 1, 1, 0, 0)$ are,

DISSIMILARITY	VALUE
Simple Matching	0.714
Jaccard	0.833
Kulczynski	0.708
Czekanowski	0.714

- To determine which dissimilarity measure is relevant we have to consider the application it is to be used for.
- Think about the choice between Simple Matching and Jaccard for example. What is the key difference between these?

Categorical Data

- When we have categorical data, then we tend to just use a simple matching-type measure of dissimilarity.
- In this respect we can just count the number of terms that differ.
- For example, suppose that we have recorded the following categorical variables for two subjects: Gender, Hair Colour, Eye Colour, and Education Level.
- If we compare subjects with values (Male, Brown, Brown, Secondary) and (Female, Brown, Green, Third), then we notice the data points differ in three of the variables, and so we may assign a dissimilarity value of three.

Mixed Data

- If we have mixed data, then we can work out a dissimilarity for the measurement variables, for the binary variables, and for the categorical variables.
- A weighted combination of the dissimilarities can then be used to give an overall dissimilarity value between data points.
- Again alternative possibilities for computing dissimilarities in this situation have been proposed.

US Arrests: Dissimilarities

- We can produce a table of the dissimilarities between the US arrest data (we used Euclidean below). This is called a dissimilarity matrix.

```
               Alabama Alaska Arizona Arkansas California Colorado Connecticut ...
Alaska              37
Arizona             63     47
Arkansas            47     77     109
California          56     45      23       98
Colorado            42     66      90       37         73
Connecticut        128    159     185       85        169       98
.                    .      .       .        .          .        .           .
:                    :      :       :        :          :        :           :
Texas               42     72      93       33         77       15          93 ...
Utah               119    148     174       76        157       86          16 ...
Vermont            190    218     251      144        237      165          77 ...
Virginia            80    111     139       36        125       53          49 ...
Washington          93    122     149       51        133       61          38 ...
West Virginia      157    186     218      110        204      132          48 ...
Wisconsin          184    214     242      138        226      154          58 ...
Wyoming             76    107     135       31        122       52          54 ...
```

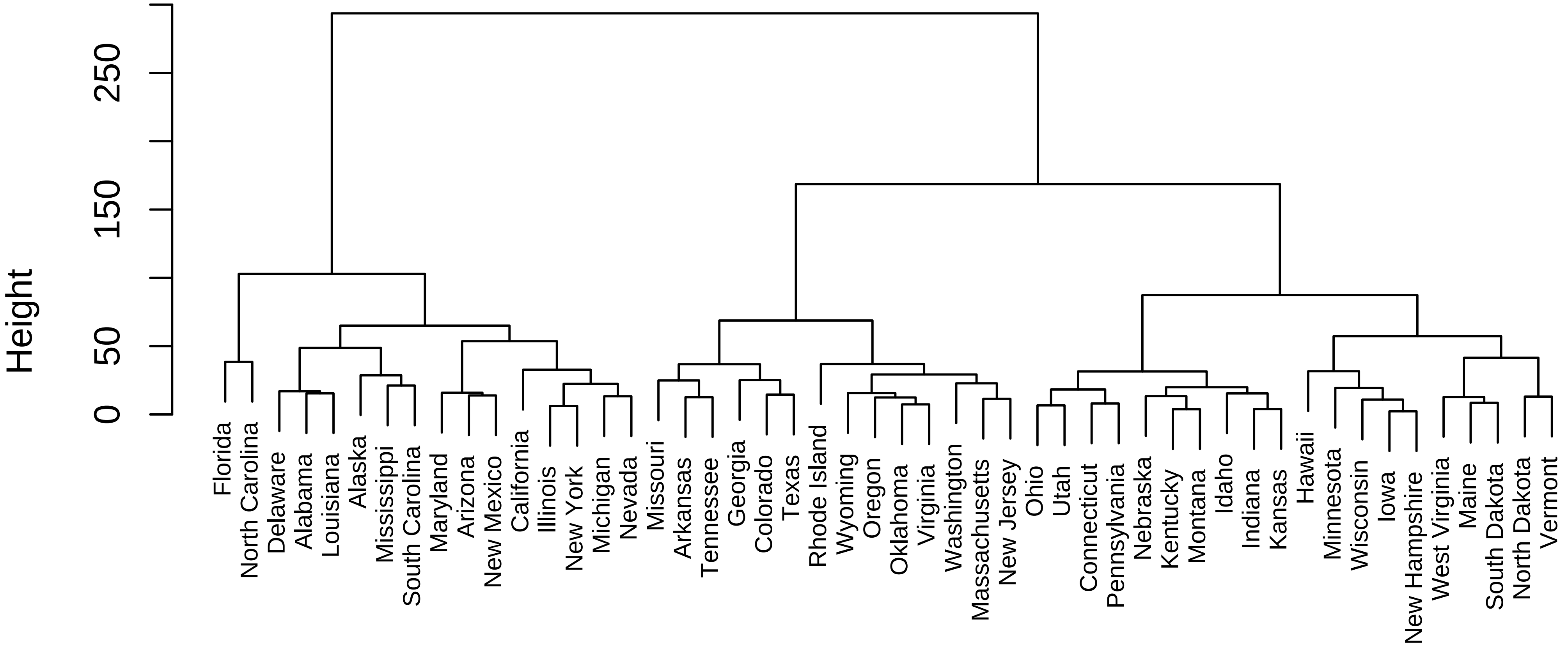
Finding Groups of Similarity

- Many methods for finding groups of similar observations have been proposed.
- These methods fall under the title of “Cluster Analysis”.
- The aim of cluster analysis is to find groups of observations such that observations within a group are very similar, and such that different groups are very dissimilar.
- Two types of cluster analysis methods commonly used are:
  - **Hierarchical:** These methods construct a tree-like structure to show groups of observations. The clustering is built up over a series of steps in which similar observations are joined together.
  - **Iterative:** These methods start with an initial clustering of observations and iteratively update the clustering until the “best” clustering is found.

Hierarchical Clustering

- One method of hierarchical clustering starts by assigning each observation to a group on its own.
  - For the US arrests data this would mean that each state is in a group on its own.
- The two closest groups are found and combined into a single group.
  - For the US arrests data, with Euclidean dissimilarity, Iowa and New Hampshire are closest.
- This leaves one fewer group.
- The process is repeated until only one group is left.

Dendogram


Height
0
50
150
250
Florida
North Carolina
Delaware
Alabama
Louisiana
Alaska
Mississippi
South Carolina
Maryland
Arizona
New Mexico
California
Illinois
New York
Michigan
Nevada
Missouri
Arkansas
Tennessee
Georgia
Colorado
Texas
Rhode Island
Wyoming
Oregon
Oklahoma
Virginia
Washington
Massachusetts
New Jersey
Ohio
Utah
Connecticut
Pennsylvania
Nebraska
Kentucky
Montana
Idaho
Indiana
Kansas
Hawaii
Minnesota
Wisconsin
Iowa
New Hampshire
West Virginia
Maine
South Dakota
North Dakota
Vermont

Dendogram Explained

- The tree-like structure used to summarize hierarchical clustering results is called a dendogram.
- The groups joined at the bottom of the graph are close together, whereas the groups at the top of the graph are far apart.
- The vertical scale of the dendogram shows the value of the distance between the groups when they were joined.
- The first two groups were joined at a distance of 2.29, whereas the last two groups joined at distance 293.6.

Questions

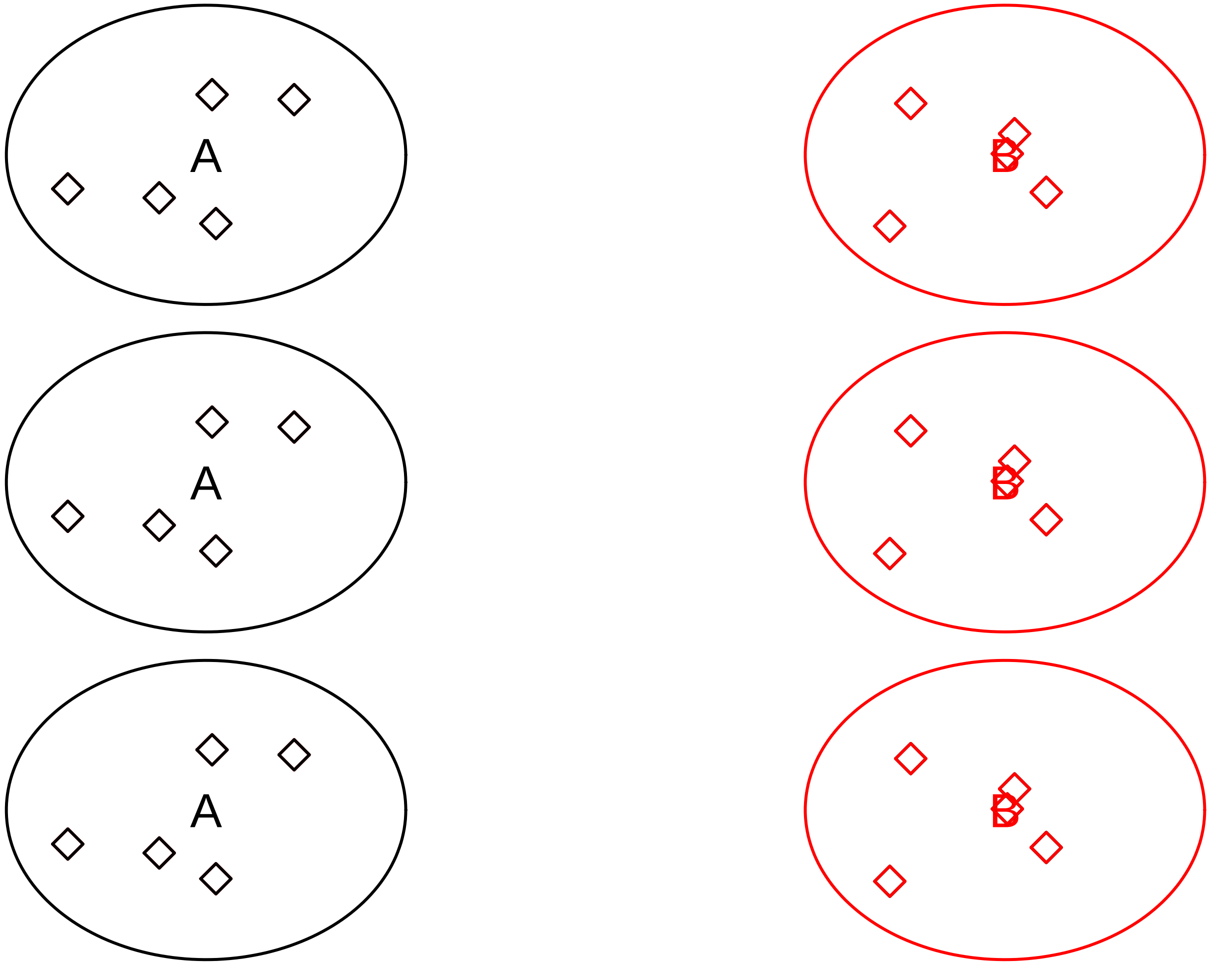
- How many groups are there?
- How do we measure the dissimilarity between two groups?
- For example, what is the value of $d(\{\mathbf{x}_{\text{Illinois}}, \mathbf{x}_{\text{New York}}\}, \{\mathbf{x}_{\text{Michigan}}, \mathbf{x}_{\text{Nevada}}\})$?
- The dendogram reports this to be 22.4.

Linkage

- There are at least three proposed methods for measuring dissimilarity between two groups.
- Consider two groups, $\mathcal{A} = \{\mathbf{x}_{a_1}, \mathbf{x}_{a_2}, \ldots, \mathbf{x}_{a_k}\}$ and $\mathcal{B} = \{\mathbf{x}_{b_1}, \mathbf{x}_{b_2}, \ldots, \mathbf{x}_{b_l}\}$.
- The following methods have been proposed for measuring the dissimilarity between $\mathcal{A}$ and $\mathcal{B}$:
- **Single Linkage:** $d(\mathcal{A}, \mathcal{B}) = \min_{\mathbf{x} \in \mathcal{A}, \mathbf{y} \in \mathcal{B}} d(\mathbf{x}, \mathbf{y})$
- **Complete Linkage:** $d(\mathcal{A}, \mathcal{B}) = \max_{\mathbf{x} \in \mathcal{A}, \mathbf{y} \in \mathcal{B}} d(\mathbf{x}, \mathbf{y})$
- **Average Linkage:** $d(\mathcal{A}, \mathcal{B}) = \frac{1}{|\mathcal{A}||\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{A}} \sum_{\mathbf{y} \in \mathcal{B}} d(\mathbf{x}, \mathbf{y})$
- How can we interpret these?
- **Exercise:** Determine whether or not these satisfy the three suggested properties of a dissimilarity measure.

Linkage Interpretation?

- What do single, complete and average linkage measure?



Examples

- **Exercise:** Create a dissimilarity matrix for the data $\mathbf{x}^T = (1, 0, 2)$, $\mathbf{y}^T = (1, 1, 2)$, $\mathbf{z}^T = (2, 2, 2)$ (choose any given measure).

- **Exercise:** Create a dissimilarity matrix for the binary data $\mathbf{x}^T = (1, 0, 1)$, $\mathbf{y}^T = (1, 1, 1)$, $\mathbf{z}^T = (0, 1, 1)$ (choose any given binary measure).

- **Exercise:** Create a dissimilarity matrix for the groups $\mathcal{A} = \{(1, 2), (0, 3)\}$, $\mathcal{B} = \{(0, 4), (2, 1)\}$, $\mathcal{C} = \{(2, 1), (3, 3)\}$ (choose any given measure and any linkage).