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 $d(x, y)$ are that:
  1. $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$
  2. $d(x, y) = d(y, x)$
  3. $d(x, z) \leq d(x, y) + d(y, z)$ (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:

<table>
<thead>
<tr>
<th></th>
<th>Man</th>
<th>Monkey</th>
<th>Horse</th>
<th>Pig</th>
<th>Pigeon</th>
<th>Tuna</th>
<th>Mould</th>
<th>Fungus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man</td>
<td>0</td>
<td>1</td>
<td>17</td>
<td>13</td>
<td>16</td>
<td>31</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td>Monkey</td>
<td>1</td>
<td>0</td>
<td>16</td>
<td>12</td>
<td>15</td>
<td>32</td>
<td>62</td>
<td>65</td>
</tr>
<tr>
<td>Horse</td>
<td>17</td>
<td>16</td>
<td>0</td>
<td>5</td>
<td>16</td>
<td>27</td>
<td>64</td>
<td>68</td>
</tr>
<tr>
<td>Pig</td>
<td>13</td>
<td>12</td>
<td>5</td>
<td>0</td>
<td>13</td>
<td>25</td>
<td>64</td>
<td>67</td>
</tr>
<tr>
<td>Pigeon</td>
<td>16</td>
<td>15</td>
<td>16</td>
<td>13</td>
<td>0</td>
<td>27</td>
<td>59</td>
<td>66</td>
</tr>
<tr>
<td>Tuna</td>
<td>31</td>
<td>32</td>
<td>27</td>
<td>25</td>
<td>27</td>
<td>0</td>
<td>72</td>
<td>69</td>
</tr>
<tr>
<td>Mould</td>
<td>63</td>
<td>62</td>
<td>64</td>
<td>64</td>
<td>59</td>
<td>72</td>
<td>0</td>
<td>61</td>
</tr>
<tr>
<td>Fungus</td>
<td>66</td>
<td>65</td>
<td>68</td>
<td>67</td>
<td>66</td>
<td>69</td>
<td>61</td>
<td>0</td>
</tr>
</tbody>
</table>
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 $x_1^T = (13.2, 236, 58, 21.2)$ and $x_{50}^T = (6.8, 161, 60, 15.6)$.

- We get the following results:

<table>
<thead>
<tr>
<th>Dissimilarity</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>75.5</td>
</tr>
<tr>
<td>Manhattan</td>
<td>89.0</td>
</tr>
<tr>
<td>Maximum</td>
<td>75.0</td>
</tr>
<tr>
<td>Minkowski (p=3)</td>
<td>75.0</td>
</tr>
<tr>
<td>Minkowski (p=1.3)</td>
<td>79.8</td>
</tr>
</tbody>
</table>
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.

\[
\begin{array}{ccc}
\text{POINT } i & 1 & 0 \\
1 & a & b & a + b \\
0 & c & d & c + d \\
\hline
a + c & b + d & m = a + b + c + d \\
\end{array}
\]
Example: Binary Data

- **Example:** Suppose \( x_i^T = (1, 1, 0, 0, 0, 0, 1) \) and \( x_j^T = (1, 0, 1, 1, 1, 0, 0) \).

### Point \( j \) and Point \( i \)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( a = 1 )</td>
<td>( b = 2 )</td>
</tr>
<tr>
<td>0</td>
<td>( c = 3 )</td>
<td>( d = 1 )</td>
</tr>
<tr>
<td></td>
<td>( a + c = 4 )</td>
<td>( b + d = 3 )</td>
</tr>
</tbody>
</table>
Binary Dissimilarities

- Again many dissimilarities measures have been proposed for binary data (certainly over 40). In particular, there are suggestions that $d(x_i, x_j)$ 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 $x_i = (1, 1, 0, 0, 0, 0, 1)$ and $x_j = (1, 0, 1, 1, 1, 0, 0)$ are,

<table>
<thead>
<tr>
<th>Dissimilarity</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Matching</td>
<td>0.714</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.833</td>
</tr>
<tr>
<td>Kulczynski</td>
<td>0.708</td>
</tr>
<tr>
<td>Czekanowski</td>
<td>0.714</td>
</tr>
</tbody>
</table>

• 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.

<table>
<thead>
<tr>
<th>Alabama</th>
<th>Alaska</th>
<th>Arizona</th>
<th>Arkansas</th>
<th>California</th>
<th>Colorado</th>
<th>Connecticut</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alaska</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arizona</td>
<td>63</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arkansas</td>
<td>47</td>
<td>77</td>
<td>109</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>California</td>
<td>56</td>
<td>45</td>
<td>23</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colorado</td>
<td>42</td>
<td>66</td>
<td>90</td>
<td>37</td>
<td>73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connecticut</td>
<td>128</td>
<td>159</td>
<td>185</td>
<td>85</td>
<td>169</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td></td>
</tr>
<tr>
<td>Texas</td>
<td>42</td>
<td>72</td>
<td>93</td>
<td>33</td>
<td>77</td>
<td>15</td>
<td>93</td>
</tr>
<tr>
<td>Utah</td>
<td>119</td>
<td>148</td>
<td>174</td>
<td>76</td>
<td>157</td>
<td>86</td>
<td>16</td>
</tr>
<tr>
<td>Vermont</td>
<td>190</td>
<td>218</td>
<td>251</td>
<td>144</td>
<td>237</td>
<td>165</td>
<td>77</td>
</tr>
<tr>
<td>Virginia</td>
<td>80</td>
<td>111</td>
<td>139</td>
<td>36</td>
<td>125</td>
<td>53</td>
<td>49</td>
</tr>
<tr>
<td>Washington</td>
<td>93</td>
<td>122</td>
<td>149</td>
<td>51</td>
<td>133</td>
<td>61</td>
<td>38</td>
</tr>
<tr>
<td>West Virginia</td>
<td>157</td>
<td>186</td>
<td>218</td>
<td>110</td>
<td>204</td>
<td>132</td>
<td>48</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>184</td>
<td>214</td>
<td>242</td>
<td>138</td>
<td>226</td>
<td>154</td>
<td>58</td>
</tr>
<tr>
<td>Wyoming</td>
<td>76</td>
<td>107</td>
<td>135</td>
<td>31</td>
<td>122</td>
<td>52</td>
<td>54</td>
</tr>
</tbody>
</table>
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 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 \( d(\{x_{\text{Illinois}}, x_{\text{New York}}\}, \{x_{\text{Michigan}}, x_{\text{Nevada}}\}) \)?

• 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, \( A = \{ x_{a_1}, x_{a_2}, \ldots, x_{a_k} \} \) and \( B = \{ x_{b_1}, x_{b_2}, \ldots, x_{b_l} \} \).

• The following methods have been proposed for measuring the dissimilarity between \( A \) and \( B \):

  • **Single Linkage:** \( d(A, B) = \min_{x \in A, y \in B} d(x, y) \)

  • **Complete Linkage:** \( d(A, B) = \max_{x \in A, y \in B} d(x, y) \)

  • **Average Linkage:** \( d(A, B) = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d(x, y) \)

• 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 $x^T = (1, 0, 2)$, $y^T = (1, 1, 2)$, $z^T = (2, 2, 2)$ (choose any given measure).

- **Exercise:** Create a dissimilarity matrix for the binary data $x^T = (1, 0, 1)$, $y^T = (1, 1, 1)$, $z^T = (0, 1, 1)$ (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).