Discriminant Analysis

Clustering and Classification

Lecture 3

2/14/06
Today’s Class

• Introduction to Discriminant Analysis
  – From Johnson & Wichern, Chapter 11.

• Assumptions of DA.

• How DA works.
  – How to arrive at discriminant functions.
  – How many discriminant functions to use.
  – How to interpret the results.

• How to classify objects using DA.
General Introduction
General Introduction

• DA is a statistical technique that allows the user to investigate the differences between multiple sets of objects across several variables simultaneously.

• DA works off of matrices used in Multivariate Analysis of Variance (MANOVA).
When to Use Discriminant Analysis

- Data should be from distinct groups.
  - Group membership must already be known prior to initial analysis.
- DA is used to interpret group differences.
- DA is used to classify new objects.
Assumptions

• Data must not have linear dependencies.
  – Must be able to invert matrices.

• Population covariance must be equal for each group.

• Each group must be drawn from a population where the variables are multivariate normal (MVN).
Notation

- $g = \text{number of groups}$
- $p = \text{number of discriminating variables}$
- $n_i = \text{number of cases in group } i$
- $n_\cdot = \text{number of cases over all the groups}$
More Assumptions

1. two or more groups $g \geq 2$
2. at least two cases per group $n_i \geq 2$
3. any number of discriminating variables, provided that they are less than the total number of cases minus two: $0 < p < (n - 2)$
4. discriminating variables are measured at the interval level
5. no discriminating variable may be a linear combination of the other discriminating variables
More Assumptions

5. no discriminating variable may be a linear combination of the other discriminating variables

6. the covariance matrices for each group must be (approximately) equal, unless special formulas are used

7. each group has been drawn from a population with a MVN distribution on the discriminating variables.
Example from Kleckka

• To demonstrate DA, Kleckka (1980) uses an example of data taken from senatorial factions (citing Bardes, 1975 and 1976).

• Bardes wanted to know how US Senate voting factions changed over time
  – How stable they were from year to year
  – How much they were influenced by other issues.
Groups of Senators

- Known Groups of Senators:
  1. Generally favoring foreign aid (9)
  2. Generally opposing foreign aid (2)
  3. Opposed to foreign involvements (5)
  4. Anti-Communists (3)
Variables

• Six variables (from roll call votes):
  1. CUTAID – cut aid funds
  2. RESTRICT – add restrictions to the aid program
  3. CUTASIAN – cut funds for Asian nations
  4. MIXED – Mixed issues: liberal aid v. no aid to communists
  5. ANTIYUGO – Anti-aid to Yugoslavia
  6. ANTINUET – Anti-aid to neutral countries
### Univariate Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUTAID</td>
<td>1.422</td>
<td>3.000</td>
<td>2.200</td>
<td>2.100</td>
<td>1.900</td>
</tr>
<tr>
<td>RESTRICT</td>
<td>1.944</td>
<td>1.000</td>
<td>2.000</td>
<td>2.333</td>
<td>1.921</td>
</tr>
<tr>
<td>Cutasian</td>
<td>1.000</td>
<td>3.000</td>
<td>2.000</td>
<td>1.333</td>
<td>1.526</td>
</tr>
<tr>
<td>Mixed</td>
<td>2.667</td>
<td>2.000</td>
<td>1.800</td>
<td>1.667</td>
<td>2.211</td>
</tr>
<tr>
<td>Antiyugo</td>
<td>1.556</td>
<td>2.500</td>
<td>2.600</td>
<td>3.000</td>
<td>2.158</td>
</tr>
<tr>
<td>Antineut</td>
<td>1.259</td>
<td>1.667</td>
<td>2.133</td>
<td>2.444</td>
<td>1.719</td>
</tr>
</tbody>
</table>

**TABLE 1**

Means for "Known" Senators
How Discriminant Analysis Works
Canonical Discriminant Analysis

- The canonical discriminant function looks like this:
  \[ f_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + \ldots + u_p X_{pkm} \]

- Here:
  - \( f_{km} \) = the value (score) on the canonical discriminant function for case \( m \) in the group \( k \)
  - \( X_{ikm} \) = the value on discriminating variable \( X_i \) for case \( m \) in group \( k \)
  - \( u_i \) = coefficients which produce the desired characteristics of the function.
Number of Functions

• Because Canonical DA makes use of methods similar to Canonical Correlations, a set of discriminant functions are derived.
  – The first function is built to maximize group differences.
  – The next functions are built to be orthogonal to the first, and still maximize group differences.

• The number of functions derived is equal to max(g-1,p)
  – In the example, this would be max(4-1,6)=6.
Deriving the Canonical Discriminant Functions

• To get at the canonical discriminant functions, we must first construct a set of sums of squares and crossproducts (SSCP) matrices.
  – A total covariance matrix
  – A within group covariance matrix
  – A between group covariance matrix

• Once we have the between and within matrices, we take the eigenvalues and eigenvectors of each.
Total SSCP Matrix

Each element of the total SSCP matrix:

\[ t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (X_{ikm} - X_{i..}) (X_{jkm} - X_{j..}) \]

- \( g \) = number of groups
- \( n_k \) = number of cases in group \( k \)
- \( n_\cdot \) = total number of cases over all groups
- \( X_{ikm} \) = the value of variable \( i \) for case \( m \) in group \( k \)
- \( X_{ik..} \) = mean value of variable \( i \) for cases in group \( k \)
- \( X_{i..} \) = mean value of variable \( i \) for all cases
Within SSCP Matrix

- Each element of the within SSCP matrix:

\[ w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (X_{ikm} - X_{ik.})(X_{jkm} - X_{jk.}) \]

- \( g \) = number of groups
- \( n_k \) = number of cases in group \( k \)
- \( n_\cdot \) = total number of cases over all groups
- \( X_{ikm} \) = the value of variable \( i \) for case \( m \) in group \( k \)
- \( X_{ik.} \) = mean value of variable \( i \) for cases in group \( k \)
- \( X_{i..} \) = mean value of variable \( i \) for all cases
Between SSCP Matrix

• Once we have $W$ and $T$, we can compute $B$ by the following formula:

$$B = T - W$$

• When there are no differences between the group centroids (the mean vectors of each group), $W = T$.

• The extent they differ will define the distinctions among the observed variables.
Obtaining Discriminant Functions

• Once we have $\mathbf{B}$ and $\mathbf{W}$, we then find the solutions ($v_i$) to the following equations:

\[
\begin{align*}
\sum b_{1i}v_i &= \lambda \sum w_{1i}v_i \\
\sum b_{2i}v_i &= \lambda \sum w_{2i}v_i \\
&\vdots \\
\sum b_{pi}v_i &= \lambda \sum w_{pi}v_i
\end{align*}
\]

• There is also a constraint that the sum of the squared $v_i$ equal one (as typical in PCA).
Step 2: Converting to Functions

• Once the $\lambda$ and $v_i$ parameters are found, one then converts these into the weights for the discriminant functions:

$$u_i = v_i \sqrt{n_i - g}$$

$$u_0 = - \sum_{i=1}^{p} u_i X_i.$$
Interpreting the Discriminant Functions
Example Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Function 1</th>
<th>Function 2</th>
<th>Function 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant (u_0)</td>
<td>5.4243</td>
<td>3.5685</td>
<td>-4.3773</td>
</tr>
<tr>
<td>CUTAID</td>
<td>.8078</td>
<td>-.5225</td>
<td>1.6209</td>
</tr>
<tr>
<td>RESTRICT</td>
<td>.7940</td>
<td>-1.1177</td>
<td>-.3339</td>
</tr>
<tr>
<td>CUTASIAN</td>
<td>-4.6004</td>
<td>-1.1228</td>
<td>-1.1431</td>
</tr>
<tr>
<td>MIXED</td>
<td>-.6957</td>
<td>-1.3160</td>
<td>1.1418</td>
</tr>
<tr>
<td>ANTIYUGO</td>
<td>-1.1114</td>
<td>1.1132</td>
<td>.3781</td>
</tr>
<tr>
<td>ANTINEUT</td>
<td>1.4387</td>
<td>1.0422</td>
<td>.2000</td>
</tr>
</tbody>
</table>
### Example Function Scores for an Observation

<table>
<thead>
<tr>
<th>Variable</th>
<th>FUNCTION 1</th>
<th></th>
<th>FUNCTION 2</th>
<th></th>
<th>FUNCTION 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff. x Value = Contribution</td>
<td></td>
<td>Coeff. x Value = Contribution</td>
<td></td>
<td>Coeff. x Value = Contribution</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>5.4243</td>
<td></td>
<td>3.5685</td>
<td></td>
<td>−4.3773</td>
<td></td>
</tr>
<tr>
<td>CUTAID</td>
<td>.8078</td>
<td>1.0</td>
<td>−.5225</td>
<td>1.0</td>
<td>−.5225</td>
<td>1.6209</td>
</tr>
<tr>
<td>RESTRICT</td>
<td>.7940</td>
<td>3.0</td>
<td>−1.1177</td>
<td>3.0</td>
<td>−3.3531</td>
<td>−.3339</td>
</tr>
<tr>
<td>CUTASIAN</td>
<td>−4.6004</td>
<td>1.0</td>
<td>−1.1228</td>
<td>1.0</td>
<td>−1.1228</td>
<td>−1.1431</td>
</tr>
<tr>
<td>MIXED</td>
<td>−.6957</td>
<td>3.0</td>
<td>−1.3160</td>
<td>3.0</td>
<td>−3.9480</td>
<td>1.1418</td>
</tr>
<tr>
<td>ANTIYUGO</td>
<td>−1.114</td>
<td>1.0</td>
<td>1.1132</td>
<td>1.0</td>
<td>1.1132</td>
<td>.3781</td>
</tr>
<tr>
<td>ANTINEUT</td>
<td>1.4387</td>
<td>1.0</td>
<td>1.4387</td>
<td>1.0</td>
<td>1.0422</td>
<td>.2000</td>
</tr>
<tr>
<td>discriminant score</td>
<td>2.2539</td>
<td></td>
<td>−3.2225</td>
<td></td>
<td>−.8977</td>
<td></td>
</tr>
</tbody>
</table>
Example Interpretation

• In the example, we saw that Senator Aiken had discriminant scores of 2.25, -3.22, and -0.90.
  – These scores are in standard deviation units…of the discriminant space
• Positive values shows an object being high on a dimension.
• Negative values shows an object being low on a dimension.
• We will come to learn how to interpret the dimensions.
Group Centroids

• What we are really after is the group means for each of the discriminant functions.

• The means in this case are:
  1. 1.74, -0.94, 0.02
  2. -6.93, -0.60, 0.28
  3. -1.48, 0.69, -0.30
  4. 1.86, 2.06, 0.25

• These will be used to classify our observations.
Standardized Coefficients

- To interpret each dimension, we look at the standardized coefficients.

- Standardized coefficients are created by:

\[ c_i = u_i \sqrt{\frac{w_{ii}}{n - g}} \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Standardized Coefficient</th>
<th>Function 1</th>
<th>Function 2</th>
<th>Function 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUTAID</td>
<td>.6094</td>
<td>-3942</td>
<td>1.2227</td>
<td></td>
</tr>
<tr>
<td>RESTRICT</td>
<td>.7068</td>
<td>-.9950</td>
<td>-.2973</td>
<td></td>
</tr>
<tr>
<td>CUTASIAN</td>
<td>-2.1859</td>
<td>-.5335</td>
<td>-.5432</td>
<td></td>
</tr>
<tr>
<td>MIXED</td>
<td>-.4760</td>
<td>-.9004</td>
<td>.7812</td>
<td></td>
</tr>
<tr>
<td>ANTIYUGO</td>
<td>-.6077</td>
<td>.8090</td>
<td>.2748</td>
<td></td>
</tr>
<tr>
<td>ANTINEUT</td>
<td>1.0168</td>
<td>.7365</td>
<td>.1414</td>
<td></td>
</tr>
</tbody>
</table>
How Many Significant Functions?

• To see how many functions are needed to describe group differences, we need to look at the eigenvalues, $\lambda$, for each dimension.

• We will have a test statistic based on the eigenvalue.

• The statistic provides the result of a hypothesis test testing that the dimension (and all subsequent dimensions) are not significant.
Example Test Statistics

<table>
<thead>
<tr>
<th>Functions Derived, k</th>
<th>Wilks's Lambda</th>
<th>Chi-Square</th>
<th>Degrees of Freedom</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.0345</td>
<td>43.760</td>
<td>18</td>
<td>.001</td>
</tr>
<tr>
<td>1</td>
<td>.3680</td>
<td>12.996</td>
<td>10</td>
<td>.224</td>
</tr>
<tr>
<td>2</td>
<td>.9492</td>
<td>.678</td>
<td>4</td>
<td>.954</td>
</tr>
</tbody>
</table>
Classifying Objects
Classifying Objects

- Several methods exist for classifying objects.
- Each is based on the distance of an object from each group’s centroid.
  - The object is then classified into the group with the smallest distance.
- Many classification methods use the raw data.
- The canonical discriminant functions can be used as well.
Validation of Classification

• We will show more about classification in the next class.

• Basically, once we classify objects, we need to see how good we are at putting our objects into groups.

• There are multiple ways to test whether or not we do a good job.
  – Most easy is to just classify all of our objects and see how good we recover our original groups.
# Classification Matrix Example

<table>
<thead>
<tr>
<th>Original Group</th>
<th>Predicted Group</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Unknown</td>
<td>33</td>
<td>10</td>
<td>27</td>
<td>4</td>
</tr>
</tbody>
</table>
Wrapping Up

• Discriminant Analysis is a long-standing method for deriving the dimensions along which groups differ.

• We will see that it is often the first method used when approaching a classification problem.

• We must have a training data set in place to be able to use this method.  
  – All of our other methods will not require this.
Next Time

• How to do discriminant analysis in R

• Presentation of Anderson (2005) article.