Introduction to Multivariate Analysis

Lecture 1

July 18, 2011
Advanced Multivariate Statistical Methods
ICPSR Summer Session #2
Today’s Lecture

- Introductions
- Syllabus and course overview
- Chapter 1 (a brief review, really):
  - Data organization/notation
  - Graphical techniques
  - Distance measures
- Introduction to SAS
Multivariate Statistics and Thinking

- Although titled “Advanced Multivariate Statistical Methods” this course is an overview of thinking about data and methods from a multivariate lens:
  - Many methods fall under the label “multivariate statistics” (e.g., Multivariate ANOVA, Discriminant Analysis, Principal Component Analysis)
  - Many multivariate statistical distributions exist (e.g., Multivariate Normal, Wishart)
  - Many modern (univariate) statistical methods rely on these multivariate distributions, especially the multivariate normal distribution

- This course will focus on multivariate thinking, not just about methods, but also about the foundations of multivariate statistical analysis
Course Structure

The course is organized around a central topic each week:

1. **Foundations of Multivariate Thinking and The Multivariate Normal Distribution**
   - Matrix algebra
   - Multivariate normal distribution

2. **Multivariate Normal and Linear Mixed Models**
   - Multivariate ANOVA
   - Discrimination/classification
   - Linear models

3. **Multivariate Data Reduction Procedures**
   - Principal components analysis
   - Factor analysis and structural equation modeling

4. **Generalized Multivariate Techniques**
   - Distance methods
   - Finite mixture models
   - Categorical distributions
Multivariate Statistics

A taxonomy of multivariate statistical analyses shows that most techniques fall into one of the following categories:

1. Data reduction or structural simplification
2. Sorting and grouping
3. Investigation of the dependence among variables
4. Prediction
5. Hypothesis construction and testing
Data Organization

- As a precursor of things to come, here is a preview of the ways data are organized in this book/course

- Multivariate data are a collection of observations (or measurements) of:
  - $p$ variables ($k = 1, \ldots, p$)
  - $n$ “items” ($j = 1, \ldots, n$)

- “items” can also be thought of as subjects/examinees/individuals or entities (when people are not under study)

- In some disciplines (such as educational measurement), “items” are considered the variables collected per individual
Data Organization

- \( x_{jk} \) = measurement of the \( k^{th} \) variable on the \( j^{th} \) entity

<table>
<thead>
<tr>
<th></th>
<th>Variable 1</th>
<th>Variable 2</th>
<th>\ldots</th>
<th>Variable ( k )</th>
<th>\ldots</th>
<th>Variable ( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1:</td>
<td>( x_{11} )</td>
<td>( x_{12} )</td>
<td>\ldots</td>
<td>( x_{1k} )</td>
<td>\ldots</td>
<td>( x_{1p} )</td>
</tr>
<tr>
<td>Item 2:</td>
<td>( x_{21} )</td>
<td>( x_{22} )</td>
<td>\ldots</td>
<td>( x_{2k} )</td>
<td>\ldots</td>
<td>( x_{2p} )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>Item ( j ):</td>
<td>( x_{j1} )</td>
<td>( x_{j2} )</td>
<td>\ldots</td>
<td>( x_{jk} )</td>
<td>\ldots</td>
<td>( x_{jp} )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>Item ( n ):</td>
<td>( x_{n1} )</td>
<td>( x_{n2} )</td>
<td>\ldots</td>
<td>( x_{nk} )</td>
<td>\ldots</td>
<td>( x_{np} )</td>
</tr>
</tbody>
</table>
To represent the entire collection of items and entities, a rectangular array can be constructed:

\[
X = \begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\
  x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np}
\end{bmatrix}
\]

In the next class, we will learn about how arrays like this have an algebra that makes life somewhat easier.

All arrays will be symbolized by boldfaced font.
Array Example

- So, putting things all together, envision standing outside of the bookstore, asking people for receipts

- You are interested in looking at two variables:
  - Variable 1: the total amount of the purchase
  - Variable 2: the number of books purchased

- You find four people, and here is what you see observe (with notation:

\[
\begin{align*}
  x_{11} &= 42 & x_{21} &= 52 & x_{31} &= 48 & x_{41} &= 58 \\
  x_{12} &= 4 & x_{22} &= 5 & x_{32} &= 4 & x_{42} &= 3
\end{align*}
\]
Array Example (Continued)

- The data array would look like:

\[ X = \begin{bmatrix}
    x_{11} & x_{12} \\
    x_{21} & x_{22} \\
    x_{31} & x_{32} \\
    x_{41} & x_{42}
\end{bmatrix} = \begin{bmatrix}
    42 & 4 \\
    52 & 5 \\
    48 & 4 \\
    58 & 3
\end{bmatrix} \]

- Notice for any variable, \( x_{jk} \):
  - The first subscript \( (j) \) represents the ROW location in the data array
  - The second subscript \( (k) \) represents the COLUMN location in the data array
Descriptive Statistics Review

- When we have a large amount of data, it is often hard to get a manageable description of the nature of the variables under study.

- For this reason (and as a way of introducing a review topics from previous courses), descriptive statistics are used.

- Such descriptive statistics include:
  - Means
  - Variances
  - Covariances
  - Correlations
Sample Mean

- For the \( k^{th} \) variable, the sample mean is:

\[
\bar{x}_k = \frac{1}{n} \sum_{j=1}^{n} x_{jk}
\]

- An array of the means for all \( p \) variables then looks like this (which we will come to know as the mean vector):

\[
\bar{x} = \begin{bmatrix}
\bar{x}_1 \\
\bar{x}_2 \\
\bar{x}_3 \\
\bar{x}_4
\end{bmatrix}
\]
Sample Variance

- For the $k^{th}$ variable, the sample variance is:

$$s_k^2 = s_{kk} = \frac{1}{n} \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2$$

- Note the “kk” subscript, this will be important because the equation that produces the variance for a single variable is a derivation of the equation of the covariance for a pair of variables.

- Also note the division by $n$
  - Reasons for this will become apparent in the near future (hint: it’s a type of estimate).

- For a pair of variables, $i$ and $k$, the sample covariance is:

$$s_{ik} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$
Sample Covariance Matrix

- Making an array of all sample covariances give us:

\[
S_n = \begin{bmatrix} 
S_{11} & S_{12} & \ldots & S_{1p} \\
S_{21} & S_{22} & \ldots & S_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
S_{p1} & S_{p2} & \ldots & S_{pp} 
\end{bmatrix}
\]
Sample Correlation

- Sample covariances are dependent upon the scale of the variables under study

- For this reason, the correlation is often used to describe the association between two variables

- For a pair of variables, $i$ and $k$, the sample correlation is found by dividing the sample covariance by the product of the standard deviation of the variables:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii} s_{kk}}}$$

- The sample correlation:
  - Ranges from -1 to 1
  - Measures linear association
  - Is invariant under linear transformations of $i$ and $k$
  - Is a biased statistic
Sample Correlation Matrix

- Making an array of all sample correlations give us:

$$R = \begin{bmatrix}
1 & r_{12} & \cdots & r_{1p} \\
 r_{21} & 1 & \cdots & r_{2p} \\
 \vdots & \vdots & \ddots & \vdots \\
 r_{p1} & r_{p2} & \cdots & 1
\end{bmatrix}$$
Graphical Techniques

- Displaying multivariate data can be difficult due to our natural limitations of seeing the world in three dimensions.

- Several simple ways of displaying data include:
  - Bivariate scatterplots
  - Three-dimensional scatterplots
Bivariate Scatterplots

Density
Max 0.97
Med 0.81
Min 0.76

Strength (MD)
Max 135.4
Med 121.4
Min 102.5

Strength (CD)
Max 80.33
Med 70.70
Min 48.93
Trivariate Scatterplots
Graphical Techniques

- But you likely already have seen those plots
- Some plots that can be achieved by multivariate methods include:
  - “Stars”
  - Chernoff faces
  - Dendrograms
  - Bivariate plots, but of the variable space
  - Network graphs
Stars

Arizona Public Service Co. (1)

Boston Edison Co. (2)

Central Louisiana Electric Co. (3)

Commonwealth Edison Co. (4)
Chernoff Faces

Cluster 1

Cluster 2

Cluster 3

Cluster 5

Cluster 7

Cluster 4

Cluster 6
Dendrograms

Figure 12.12 A dendrogram for similarities between 109 pure malt Scotch whiskies.
Figure 12.18 A two-dimensional representation of universities produced by metric multidimensional scaling.
Distance Measures

- A great number of multivariate techniques revolve around the computation of distances:
  - Distances between variables
  - Distances between entities (people, objects, etc.)

- The formula for the Euclidean distance formula between the coordinate pair \( P = (x_1, x_2) \) and the origin \( O = (0, 0) \):

\[
d(O, P) = \sqrt{(x_1 - 0)^2 + (x_2 - 0)^2}
\]
Distance Measures

- Elaborate discussions of distance measures will be found later in the class

- There are also some statistical analogs to distance measures, taking the variability of variables into account

- Also be aware that there are literally an infinite number of distance measures!

- To be considered an actual “distance”, a distance measure must satisfy the following:
  - \( d(P, Q) = d(Q, P) \)
  - \( d(P, Q) > 0 \) if \( P \neq Q \)
  - \( d(P, Q) = 0 \) if \( P = Q \)
  - \( d(P, Q) \leq d(P, R) + d(R, Q) \) (known as the triangle inequality)
Final Thoughts

- We introduced what this course will be about - the wild world of multivariate statistics

- Things will become increasingly relevant as time progresses...but do not hesitate to ask “why?”

- We will now head down to the lab for a SAS introduction session

- Tomorrow’s Class: Matrix algebra (Chapter 2 and Supplement 2A)