Introduction

- In many observational or designed studies, observations are collected simultaneously on more than one variable on each experimental unit.

- Multivariate analysis is the collection of methods that can be used to analyze these multiple measurements.

- Idea is to exploit potential correlations among the multiple measurements to improve inference.

- Some multivariate techniques rely on an underlying probability model, typically the multivariate normal distribution. Others are 'model-free'.
Objectives of Multivariate Analysis

- *Dimensionality reduction*: Can we reduce the dimensionality of the problem by considering a small number of (linear) combinations of a large number of measurements without losing important information?

- *Grouping*: Identify groups of 'similar' units using a common set of measured traits.

- *Classification*: Classify units into previously defined groups using a common set of measured traits.
Objectives of Multivariate Analysis

- Dependence among variables: What is the nature of associations among variables?

- Prediction: If variables are associated, then we might be able to predict the value of some of them given information on the others.

- Hypothesis testing: Are differences in sets of response means for two or more groups large enough to be distinguished from sampling variation?
Examples: Classification and Grouping

- The US IRS uses data collected from tax returns (income, amount withheld, deductions, contributions to charity, age) to classify taxpayers into two groups: those who will be audited and those who will not.

- Using the concentration of elements (copper, silver, tin, antimony) in the lead alloy used in bullets, the FBI identifies 'compositional groups' of 'similar' bullets that may be used to infer whether bullets were produced from the same batch of lead.
Examples: Classification and Grouping

- An insurance company wishes to group customers with respect to products purchased and demographic variables to target marketing efforts at different groups.

- A marketing company examines traits of people who respond or fail to respond to a mass mailing.

- An entomologist uses measurements on various physical characteristics to group insects into categories representing subspecies.
Examples: Hypothesis Testing

- A transportation company wants to know if means for gasoline mileage, repair costs, downtime due to repairs differ for different truck models.

- An insurance company wants to know if changing case management practices leads to changes in mean length of hospital stay, mean infection rates, mean costs, measures of patient satisfaction, ...

- Water quality monitors want to know if different tillage practices lead to different patterns of nitrate concentrations in nearby waterways?
Examples: Dimensionality Reduction

- An index of consumer satisfaction with new car ownership can be constructed from dozens of questions on a survey.

- A single index of patient reaction to radiotherapy can be constructed from measurements on several response variables.

- Wildlife ecologists can construct a few indices of habitat preference from measurements of dozens of features of nesting sites selected by a certain bird species.
Organization of Data and Notation

- We will use $n$ to denote the number of individuals or units in our sample and use $p$ to denote the number of variables measured on each unit.

- If $p = 1$, then we are back in the usual univariate setting.

- $x_{ik}$ is the value of the $k$-th measurement on the $i$-th unit. For the $i$-th unit we have measurements $x_{i1}, x_{i2}, \ldots, x_{ip}$
Organization of Data and Notation

- We often collect all measurements taken on the \( i \)-th unit into a column vector. If five measurements are taken on the \( i \)-th unit, we would have

\[
x_i = \begin{bmatrix}
x_{i1} \\
x_{i2} \\
x_{i3} \\
x_{i4} \\
x_{i5}
\end{bmatrix}
\]
Organization of Data and Notation

- We often display measurements from a sample of $n$ units in matrix form:

$$X_{n \times p} = \begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1p} \\
  x_{21} & x_{22} & \cdots & x_{2p} \\
  \vdots & \vdots & & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{np}
\end{bmatrix}$$

is a matrix with $n$ rows (one for each unit) and $p$ columns (one for each measured trait or variable).
Descriptive Statistics

- The sample mean of the $k$th variable ($k = 1, ..., p$) is computed as

$$
\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik}
$$

- The sample variance of the $k$th variable is usually computed as

$$
s_k^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2
$$

and the sample standard deviation is given by

$$
s_k = \sqrt{s_k^2}
$$
Descriptive Statistics

- Sometimes the variance is defined with a denominator of $n$ instead of $n - 1$, and this will be clear from the notation.
- We often use $s_{kk}$ to denote the sample variance for the $k$-th variable. Thus,
  $$s_k^2 = s_{kk}$$
- The sample covariance between variable $k$ and variable $j$ is computed as
  $$s_{jk} = \frac{1}{n - 1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$
- If variables $k$ and $j$ are independent, the population covariance will be exactly zero, but the sample covariance will vary about zero.
Descriptive Statistics

- The sample correlation between variables \( k \) and \( j \) is defined as
  
  \[ r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj} s_{kk}}} \]

- \( r_{jk} \) is between -1 and 1.

- \( r_{jk} = r_{kj} \)

- The sample correlation is the same whether \( n \) or \( n - 1 \) is used as the divisor in evaluating sample variances and covariances.
Descriptive Statistics

- The sample correlation is equal to the sample covariance if measurements are standardized.

- Covariance and correlation measure linear association. Other non-linear dependencies may exist among variables even if $r_{jk} = 0$.

- A population correlation of zero means no linear association but it does not necessarily imply independence.

- The sample correlation ($r_{ij}$) will vary about the value of the population correlation ($\rho_{ij}$).
Descriptive Statistics

• Sums of squares and cross-products:

\[ a_{kk} = \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2, \quad k = 1, \ldots, p \]

\[ a_{jk} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \quad k, j = 1, \ldots, p \]

• Sample statistics can be organized as vectors and matrices:
  – \( \bar{x} \) is the \( p \times 1 \) vector of sample means.
  – \( S \) is the \( p \times p \) symmetric matrix of variances (on the diagonal) and covariances (the off-diagonal elements).
  – \( R \) is the \( p \times p \) symmetric matrix of sample correlations. Diagonal elements are all equal to 1.
Example: Bivariate Data

• Data consist of $n = 5$ receipts from a bookstore. On each receipt we observe the total amount of the sale ($) and the number of books sold ($p = 2$). Then

$$X_{5 \times 2} = \begin{bmatrix}
x_{11} & x_{12} \\
x_{21} & x_{22} \\
x_{31} & x_{32} \\
x_{41} & x_{42} \\
x_{51} & x_{52}
\end{bmatrix} = \begin{bmatrix}
42 & 4 \\
52 & 5 \\
88 & 7 \\
58 & 4 \\
60 & 5
\end{bmatrix}$$

• Sample mean vector is:

$$\bar{x} = \begin{bmatrix}
\bar{x}_1 \\
\bar{x}_2
\end{bmatrix} = \begin{bmatrix}
60 \\
5
\end{bmatrix}$$
Example: Bivariate Data

- Sample covariance matrix is

\[
S = \begin{bmatrix}
    s_{11} & s_{12} \\
    s_{21} & s_{22}
\end{bmatrix} = \begin{bmatrix}
    294.0 & 19.0 \\
    19.0 & 1.5
\end{bmatrix}
\]

- Sample correlation matrix is

\[
R = \begin{bmatrix}
    r_{11} & r_{12} \\
    r_{21} & r_{22}
\end{bmatrix} = \begin{bmatrix}
    1 & 0.90476 \\
    0.90476 & 1
\end{bmatrix}
\]
Distance

• Multivariate methods rely on distances between units.

• Clustering: group units that are 'closest' in some sense.

• Classification: allocate each unit to the 'closest' group.

• Distance can be defined in different ways.
Euclidean Distance

- Straight-line distance from a point $P = \{x_1, x_2, ..., x_p\}'$ in $p$ dimensions to the origin $O$ is
  $$d(O, P) = \sqrt{x_1^2 + x_2^2 + ... + x_p^2}$$

- All points $P$ at an equal squared distance $c^2$ from the origin satisfy:
  $$d^2(O, P) = x_1^2 + x_2^2 + ... + x_p^2 = \sum_{j=1}^{p} x_j^2 = c^2,$$
  which defines a hypersphere centered at $O$. 
Euclidean Distance (cont’d)

- Euclidean distance between two arbitrary points $P$ and $Q$ with $p$ coordinates: $P = (x_1, x_2, \ldots, x_p)'$, $Q = (y_1, y_2, \ldots, y_p)'$ is

$$
    d(P, Q) = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}
$$

- Even though $p$ variables may be observed with different precision, Euclidean distance gives equal weight to all.

- All points $P$ at an equal squared distance $c^2$ from $Q$ satisfy:

$$
    d^2(P, Q) = \sum_{j=1}^{p} (x_j - y_j)^2 = c^2,
$$

a hypersphere centered at $Q$. 
Standardized Distance

• Suppose that the variability in each of the \( p \) dimensions (variables) is different.

• We wish to give more weight in the distance calculation to those dimensions (variables) that are measured more precisely.

• Weights are inversely proportional to the standard deviation in the measurements:

\[
d(P, Q) = \sqrt{\sum_{j=1}^{p} \left( \frac{x_j - y_j}{s_j} \right)^2}
\]
Standardized Distance (cont’d)

• If we define $P = (x_1, x_2, \ldots, x_p)'$, $Q = (y_1, y_2, \ldots, y_p)'$ and $D = diag\{s_{jj}\}$, then

$$d(P, Q) = \sqrt{(P - Q)'D^{-1}(P - Q)}$$

• This measure of distance does not account for correlations among variables: $D$ is a diagonal matrix with all covariances set equal to zero.

• All points $P$ at the same standardized distance $c$ from the origin satisfy:

$$(P - 0)'D^{-1}(P - 0) = P'D^{-1}P = c^2,$$
Standardized Distance (cont’d)

• Any $P$ at a standardized distance $c$ from $Q$ satisfies

$$(P - Q)'D^{-1}(P - Q) = c^2$$

• This defines a hyper-ellipsoid centered at $Q$, with axes parallel to the coordinates axes. The half-length of the axis parallel to the $j$-th coordinate axis is equal to $c\sqrt{s_{jj}}$
Other Distance Measures

• Suppose now that the various measurements do not vary independently.

• What is a reasonable distance measure when the variability in each direction is different and the variables are correlated?
Properties of Distance Measures

- For any three points $P = (x_1, x_2, ..., x_p)'$, $Q = (y_1, y_2, ..., y_p)'$, and $R = (z_1, z_2, ..., z_p)'$, a distance measure must satisfy
  
  \begin{align*}
  &- d(P, Q) = d(Q, P) \\
  &- d(P, Q) > 0 \text{ if } P \neq Q \\
  &- d(P, Q) = 0 \text{ if } P = Q \\
  &- d(P, Q) \leq d(P, R) + d(R, Q) \text{ triangle inequality}
  \end{align*}
General Distance Measure

• A general distance measure is

\[ d(P, Q) = \sqrt{(P - Q)' A (P - Q)} \]

where \( A \) is a symmetric positive definite matrix, a matrix with entries \( a_{jk} = a_{kj} \) such that the distances are always non-negative.

• For \( p = 2 \),

\[ d(P, Q) = \sqrt{[(x_1 - y_1)^2 a_{11}^2 + 2(x_1 - y_1)(x_2 - y_2)a_{12} + (x_2 - y_2)^2 a_{22}]} \]
General Distance (cont’d)

- All points $P = (x_1, x_2, ..., x_p)'$ a constant distance $c^2$ from some fixed point $Q = (y_1, y_2, ..., y_p)'$ satisfy

$$d^2(P, Q) = (P - Q)'A(P - Q) = c^2$$

which is the equation of an ellipse. See figure in next transparency for the case $p = 2$.

- The axes of the ellipse are parallel to the set of new axes $(\tilde{x}_1, \tilde{x}_2)$ obtained by rotating the original axes by an angle $\theta$. 
General Distance (cont’d)
When the measurements are correlated, we can construct a statistical distance that accounts for correlations and unequal variances by:

- First rotating the axes to be parallel to the axes of the ellipsoid

- Then using the expression for a standardized distance

We can re-express any point with respect to the rotated coordinates. For $P = (x_1, x_2)'$, we have

$$
\tilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta) \\
\tilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta).
$$
Statistical Distance (cont’d)
Statistical Distance (cont’d)

- A distance measure that automatically does this is obtained by choosing $A$ as the inverse of the covariance matrix.

- The squared statistical distance between $P = (x_1, x_2, \ldots, x_p)'$ and $Q = (y_1, y_2, \ldots, y_p)'$ is

$$d^2(P, Q) = (P - Q)'S^{-1}(P - Q)$$

- This is also called the squared Mahalanobis distance

- When measurements are uncorrelated this becomes the standardized distance