Multivariate (MV) data consist of observations on two or more variables from two or more individuals (objects, human subjects, etc.).

Multivariate analysis (MVA) is an increasingly important area of statistics for several reasons:

1. The vast majority of data actually collected is multivariate.

2. MV data can have a complex structure, requiring a rich set of models and offering a great variety of analytic approaches.

3. Many statistical problems in MVA remain unsolved.

4. There is a synergism between MVA and modern statistical computing.

Hypothetical example of MV data:

Subject	P. Hudson	S. Atkins	J. Abraham	D. Zimmerman	M. Hartman	D. Mercier	Average
History							51.7
Math							43.7
Physics							50.2
Literature							47.2
Geography							47.8
Average	48.0	52.1	41.7	48.5	43.9	54.5	

Questions of interest:

- Which subjects were most difficult? (Look at row averages.)
- Which individuals had better academic performance than others? (Look at column averages.)
- Are results for different subjects correlated? For example, do students who perform above average in physics and math tend to perform below average in literature? (Need table entries to answer this.)

Notation for data:

- $p = \#$ variables, $n = \#$ individuals.
- Let x_{ij} be the value of the jth variable for the ith individual.
- Data matrix is the $n \times p$ matrix

$$\mathbf{X} = (x_{ij}) = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

- Note: some authors define $\mathbf{X}$ as a $p \times n$ matrix.
- Some measurements could be missing.

MVA techniques can be classified as:

- variable-directed, i.e. primarily concerned with relationships between variables (pooled over individuals); or
- individual-directed, i.e. primarily concerned with relationships between individuals (pooled over variables).

This distinction is not always clearcut, however.

Major Specific Objectives of MVA:

1. *Analysis of Interdependence Among Variables.* Applicable to variables that "arise on an equal footing." The strength of the interdependence may range from complete independence to exact collinearity. Includes such things as:
   - Inference on correlation coefficient ($p = 2$) or correlation/covariance matrix ($p > 2$)
   - Determining whether the variables fall into groups such that the correlation is high between variables in the same group but low between variables in different groups (Clustering of variables); and then trying to explain what's responsible for this (FA)
   - Transforming the variables to a new set of uncorrelated variables, ranked in order of importance in describing variability (PCA)

   The above methods generally are variable-directed.

2. *Analysis of Dependence.* Applicable when the variables do not "arise on an equal footing," i.e. one or more of the variables are sensibly regarded as dependent variables and the others are sensibly regarded as explanatory variables. Then we want to study the functional relationship between the dependent variable(s) and the explanatory variable(s), and possibly predict the values of unobserved dependent variables given values of the explanatory variables. Includes:
   - Hotelling's T^2 test
   - Multivariate ANOVA (MANOVA)
   - Multivariate Regression

   These methods also are generally variable-directed.

3. *Classification.* Generally individual-directed. Includes such things as:
   - Techniques for grouping individuals that are similar in some sense (Cluster Analysis)
   - Techniques for assigning an individual to one of two or more pre-established groups (Discriminant Analysis)

Some final introductory remarks:

1. Most MVA methods in this course assume that all the variables are continuous (or discrete with many levels). For discrete MV data, many objectives can be accomplished via methods of categorical data analysis. The mixed case is trickier and has received relatively less attention.
2. Most MVA methods in this course assume that the data have a multivariate normal (MVN) distribution. Why?
   - Mathematical beauty and simplicity
   - CLT for large samples
   - Nonparametric procedures not as well-developed as in univariate case

   In practice, we should examine the data to check the assumption, consider how robust our inferences are to departures from normality, and consider making a transformation to more nearly attain normality.

3. Principles of good sampling design or experimental design (randomization, blocking, etc.) are just as important in MVA as in UVA, but we focus on analysis rather than design in this course.

Some Basic Summary Statistics

Analysis of UV data usually begins with construction of a histogram and computation of two numerical summary statistics: the mean and standard deviation.

Analysis of MV data usually begins with construction of some graphical displays (to be described shortly) and the calculation of three sets of numerical summary statistics: means, standard deviations, and correlation coefficients.

- Sample means:

$$\bar{x}_k = \frac{1}{n} \sum_{j=1}^{n} x_{jk} \qquad (k = 1, \ldots, p)$$

- Sample variances:

$$s_k^2 = \frac{1}{n} \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 \equiv s_{kk} \qquad (k = 1, \ldots, p)$$

  (note the divisor of n rather than $n - 1$)

- Sample standard deviations:

$$s_k = \sqrt{s_k^2} \qquad (k = 1, \ldots, p)$$

- Sample covariances:

$$s_{ik} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \qquad (i, k = 1, \ldots, p)$$

- Sample correlation coefficients:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii} s_{kk}}} \qquad (i, k = 1, \ldots, p)$$

These quantities can be arranged in vectors or matrices:

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}, \quad \mathbf{S}_n = \qquad ,$$

$$\mathbf{R} = \qquad .$$

Interpretations of summary statistics:

- Means and standard deviations — straightforward measures of central tendency and variability
- Covariances — not straightforward as measures of association between pairs of variables
- Correlation coefficients
  - Measures strength and sign of linear association
  - $-1 \leq r \leq 1$
  - $r \begin{cases} < 0 \Rightarrow \text{ negative linear association} \\ = 0 \Rightarrow \text{ no linear association} \\ > 0 \Rightarrow \text{ positive linear association} \end{cases}$
  - Invariant to affine linear transformations with same-sign multiplicative coefficients
  - Very sensitive to outliers

Interpretation of the correlation matrix:

- The $\mathbf{R}$ produced by a computer may often be in a rather unsuitable format for interpretation.
  - too many significant digits are given
  - if $p > 8$ or so, then some rows of $\mathbf{R}$ may be spread over 2 or more rows of computer output

  Example: 100K Race Data (correlations among "split times" only)

  (graphic shown in lecture)

- Some ways to improve the presentation of $\mathbf{R}$:
  - Round coefficients to ≤ 2 decimal places
  - Suppress the ones on the main diagonal
  - Suppress 0's before the decimal point

Revisit the 100K race data:

$$\mathbf{R} = \begin{bmatrix}
 & .95 & .84 & .79 & .62 & .62 & .53 & .48 & .54 & .41 \\
.95 & & .89 & .83 & .64 & .63 & .54 & .51 & .53 & .44 \\
.84 & .89 & & .92 & .76 & .73 & .61 & .62 & .58 & .47 \\
.79 & .83 & .92 & & .89 & .84 & .69 & .70 & .67 & .51 \\
.62 & .64 & .76 & .89 & & .94 & .75 & .79 & .74 & .54 \\
.62 & .63 & .73 & .84 & .94 & & .84 & .84 & .77 & .66 \\
.53 & .54 & .61 & .69 & .75 & .84 & & .78 & .70 & .72 \\
.48 & .51 & .62 & .70 & .79 & .84 & .78 & & .76 & .66 \\
.54 & .53 & .58 & .67 & .74 & .77 & .70 & .76 & & .78 \\
.41 & .44 & .47 & .51 & .54 & .66 & .72 & .66 & .78 &
\end{bmatrix}$$

- In some cases the pattern of the correlations may become clearer if the variables are reordered. However, any natural ordering (e.g. with respect to time of measurement) should be respected.

Another example of a correlation matrix: The Women's National Track Records Data (see Table 1.9 of text).

(graphic shown in lecture)

Computing a correlation matrix (and other simple numerical summary statistics):

- In SAS, can use PROC CORR or

  Solutions→Analysis→Interactive Data Analysis →Analyze→Multivariate

Some Elementary Graphical Summaries

- Univariate histograms for each variable
  - Histograms for Women's National Track Records Data:

    (graphic shown in lecture)

    In SAS, can use

    Solutions→Analysis→Interactive Data Analysis →Analyze→Histogram/Bar Chart

- Matrix scatterplot — a collection of all two-variable scatterplots, laid out in a two-dimensional array. Plots on the main diagonal can be left empty, or a boxplot can be placed there instead.
  - Matrix scatterplot for Women's National Track Records Data:

    (graphic shown in lecture)

    In SAS, can use

    Solutions→Analysis→Interactive Data Analysis →Analyze→ScatterPlot

- 3-D scatterplot — mostly useless unless the positions of the points are represented unambiguously.
  - 3-D scatterplot of lizard data from Table 1.3:

    (graphic shown in lecture)

- Profile plot (growth curve plots) — Applicable generally, but most useful in a situation where the data represent a single characteristic measured on several occasions (repeated measurements, longitudinal data). Plots the characteristic versus time, with line segments connecting consecutive measurements from the same individual.
  - Profile plot of 100K race data:

    (graphic shown in lecture)

    In SAS,

    Solutions→Analysis→Interactive Data Analysis →Analyze→Line Plot

    with some tweaking will create individual-specific profile plots.

- Stars and Chernoff faces — never mind
- Dynamic graphics
  - Brushing linked plots (e.g. within a matrix scatterplot)
  - Rotating a 3-D scatterplot

<u>Statistical Distance</u>

Consider two points, $P = (x_1, \ldots, x_p)$ and $Q = (y_1, \ldots, y_p)$ in p-dimensional Euclidean space. (Can think of P and Q as representing two of the n individuals in the data.)

Euclidean distance between P and Q:

$$d_E(P, Q) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2}.$$

- If Q is fixed, points P satisfying $d_E(P, Q) = c$ (a constant) lie on a hypersphere centered at Q. E.g., if $p = 2$:

- Unsatisfactory for most statistical purposes because it weights each coordinate equally. Coordinates represent variables that often have different variabilities.

Modified distance measure that allows for different variabilities:

$$d_U(P,Q) = \sqrt{\frac{(x_1-y_1)^2}{s_{11}} + \cdots + \frac{(x_p-y_p)^2}{s_{pp}}}$$

- If Q is fixed, points P satisfying $d_U(P,Q) = c$ lie on a hyperellipsoid centered at Q and aligned with the coordinate axes. E.g., if $p = 2$:

- Points which appear farther away from Q visually (i.e. in a Euclidean sense) may be closer using the modified distance.

$d_U(P,Q)$ is still not suitable for general use because it doesn't account for possible correlation among variables. E.g. if $p = 2$:

In such a case we first rotate the coordinate axes by an angle θ of our choice, thereby transforming the coordinate system, and then we measure distance by $d_U(P,Q)$ in the new coordinate system. After some algebra, the **statistical distance** measure is

$$\begin{aligned} d_S(P,Q) = [&a_{11}(x_1-y_1)^2 + \cdots + a_{pp}(x_p-y_p)^2 \\ &+2a_{12}(x_1-y_1)(x_2-y_2) \\ &+2a_{13}(x_1-y_1)(x_3-y_3) + \cdots \\ &+2a_{p-1,p}(x_{p-1}-y_{p-1})(x_p-y_p)]^{1/2} \end{aligned}$$

where the a_{ij}'s are messy functions of θ and the s_{ij}'s.

- If Q is fixed, points P satisfying $d_S(P,Q) = c$ lie on a hyperellipsoid centered at Q, but not necessarily aligned with the original coordinate axes.
- A tidier way to represent $d_S(P,Q)$ will be possible using matrix notation (it's an example of a *quadratic form*).

Other distance measures are possible. $d(P,Q)$ is a distance measure if and only if:

1. $d(P,Q) = d(Q,P)$
2. $d(P,Q) > 0$ if $P \neq Q$
3. $d(P,Q) = 0$ if $P = Q$
4. $d(P,Q) \leq d(P,R) + d(R,Q)$ (triangle inequality)

Notation for vectors and matrices

Column vector: $\mathbf{x}$
Row vector: $\mathbf{x}'$
Matrix: $\mathbf{X}$
Scalar: x

Students' kKnowledge of vector and matrix dimensionality, addition, multiplication, and transposition is assumed for this course.

Some properties of vectors

- Length of vector $\mathbf{x}$: $L_{\mathbf{x}} =$
- Angle between two vectors $\mathbf{x}$ and $\mathbf{y}$:

$$\cos\theta =$$

Example:

- Vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* iff
  Examples:

- A set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is *linearly dependent* if constants $c_1, \ldots, c_k$ exist which are not all equal to 0 and for which

$$\sum_{i=1}^{k} c_i \mathbf{x}_i = \mathbf{0}.$$

  Note that this implies that at least one vector in the set can be expressed as a linear combination of the others.

  If a set of vectors is not linearly dependent it is *linearly independent*.

Examples:

- Projection of a vector $\mathbf{x}$ onto another vector $\mathbf{y}$:

Some properties, types, facts, and functions of matrice

- A matrix $\mathbf{A}$ is symmetric if $\mathbf{A}' = \mathbf{A}$.

  Fact: If $\mathbf{AB}$ is well-defined, then $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

- Identity matrix $\mathbf{I}$ and null matrix $\mathbf{0}$:

- A square matrix $\mathbf{A}$ is *nonsingular* if there exists a matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{I}$. In this case, $\mathbf{B}$ is called the *inverse* of $\mathbf{A}$ and written as $\mathbf{A}^{-1}$.

  Fact: If $\mathbf{A}$ and $\mathbf{C}$ are both nonsingular and of the same dimensions, then $(\mathbf{AC})^{-1} = \mathbf{C}^{-1}\mathbf{A}^{-1}$.

  Fact: A square matrix is nonsingular iff its columns are linearly independent.

- A square matrix $\mathbf{Q}$ is *orthogonal* if $\mathbf{Q}' = \mathbf{Q}^{-1}$. Example:

- If $\mathbf{A}$ is square, its *determinant* is well-defined (definition unnecessary for our purposes) and written as $|\mathbf{A}|$. The determinant measures the magnitude of the matrix, in some sense.

  Fact: $|\mathbf{A}| = 0$ iff $\mathbf{A}$ is singular.

- If $\mathbf{A} = (a_{ij})$ is $k \times k$, its *trace* is defined as $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{k} a_{ii}$. The trace also measures the magnitude of the matrix, in a different sense than the determinant.

Eigenvalues and eigenvectors

Let $\mathbf{A}$ be $k \times k$ and let $\mathbf{x} \neq \mathbf{0}$ be $k \times 1$.

- $\mathbf{A}$ has an *eigenvalue* λ, with corresponding *eigenvector* $\mathbf{x}$, if λ and $\mathbf{x}$ satisfy

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}.$$

- There are k such eigenvalue/eigenvector pairs (not necessarily all distinct).
- If $\mathbf{A}$ is symmetric, then the eigenvectors can be chosen to be orthogonal to each other and normalized to be of length 1.

$$\text{Notation: } (\lambda_1, \mathbf{e}_1), \ldots, (\lambda_k, \mathbf{e}_k)$$
$$\mathbf{e}_i'\mathbf{e}_j = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise} \end{cases}$$

Example:

Quadratic forms

Let $\mathbf{x} = (x_i)$ be $k \times 1$. A *quadratic form* is a function $f(\mathbf{x})$ of the form

$$f(\mathbf{x}) = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} x_i x_j = \mathbf{x}'\mathbf{A}\mathbf{x}$$

where $\mathbf{A} = (a_{ij})$ is a specified symmetric matrix. $\mathbf{A}$ is called the matrix of the quadratic form.

Example:

If $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$ then $\mathbf{x}'\mathbf{A}\mathbf{x}$ and $\mathbf{A}$ are *nonnegative definite* (= *positive semidefinite* for some authors).

If $\mathbf{x}'\mathbf{A}\mathbf{x}$ is n.n.d. and $\mathbf{x}'\mathbf{A}\mathbf{x} = 0$ iff $\mathbf{x} = \mathbf{0}$, then $\mathbf{x}'\mathbf{A}\mathbf{x}$ and $\mathbf{A}$ are *positive definite*. Examples:

Note: $d_S(P, Q)$ can be written as $\sqrt{(\mathbf{x} - \mathbf{y})'\mathbf{A}(\mathbf{x} - \mathbf{y})}$. It satisfies the properties of a distance measure iff $\mathbf{A}$ is p.d.

The spectral decomposition of a matrix

Any $k \times k$ symmetric matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \sum_{i=1}^{k} \lambda_i \mathbf{e}_i \mathbf{e}_i'$$

where $\lambda_1, \ldots, \lambda_k$ are the eigenvalues of $\mathbf{A}$ and $\mathbf{e}_1, \ldots, \mathbf{e}_k$ are the corresponding normalized eigenvectors.

Consequences of the spectral decomposition theorem:

1. A symmetric matrix $\mathbf{A}$ is n.n.d. (or p.d.) iff all of its eigenvalues are nonnegative (or positive).
2. If $\mathbf{A}$ is defined such that $\mathbf{x}'\mathbf{A}\mathbf{x}$ is a squared distance of $\mathbf{x}$ from the origin, then $\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{k} \lambda_i y_i^2$ where $\mathbf{y}_i = \mathbf{e}_i'\mathbf{x}$.
   Can display for case $k = 2$:

3. If $\mathbf{A}$ is p.d. and symmetric, then $\mathbf{A}^{-1} = \sum_{i=1}^{k} \frac{1}{\lambda_i} \mathbf{e}_i \mathbf{e}_i'$.
4. If $\mathbf{A}$ is p.d. and symmetric, then a p.d. symmetric matrix $\mathbf{B}$ exists such that $\mathbf{A} = \mathbf{B}\mathbf{B}$. $\mathbf{B}$ is the *square root matrix* of $\mathbf{A}$, denoted by $\mathbf{A}^{1/2}$, and $\mathbf{A}^{1/2} = \sum_{i=1}^{k} \sqrt{\lambda_i} \mathbf{e}_i \mathbf{e}_i'$.

Random vectors and random matrices

Example:

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \end{pmatrix}$$

Definition: $E(\mathbf{X}) = (E(X_{ij}))$ (if all the expectations exist)

Two results on expectations of random matrices:

1. $E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$
2. $E(\mathbf{A}\mathbf{X}\mathbf{B}) = \mathbf{A}E(\mathbf{X})\mathbf{B}$

First-order and second-order central moments of a random vector $\mathbf{X} = (X_i)$:

1. Mean vector $E(\mathbf{X}) \equiv \boldsymbol{\mu}$

2. Covariance matrix $E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] \equiv \boldsymbol{\Sigma}$
   - $\boldsymbol{\Sigma} = (\sigma_{ik}) = (\mathrm{cov}(X_i, X_k))$
   - Thus

     $\boldsymbol{\Sigma} =$

   - If X_i and X_k are independent, then $\mathrm{cov}(X_i, X_k) = 0$. Converse not true in general, but is true if $(X_i, X_k)'$ have a bivariate normal distribution.
   - If all p variables are independent, then $\boldsymbol{\Sigma}$ is a *diagonal* matrix, i.e. $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_{ii})$.
   - Fact: $\boldsymbol{\Sigma} = E(\mathbf{X}\mathbf{X}') - \boldsymbol{\mu}\boldsymbol{\mu}'$

3. Correlation matrix

$$\boldsymbol{\rho} =$$

where $\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}$.

If we define $\mathbf{V}^{1/2} = \text{diag}(\sqrt{\sigma_{ii}})$, and if $\mathbf{V}$ is positive definite, then it can be shown that

$$\boldsymbol{\rho} = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^{-1/2},$$

where $\mathbf{V}^{-1/2} = (\mathbf{V}^{1/2})^{-1}$.

Partitioning a $p \times 1$ random vector into two subvectors:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \\ X_{q+1} \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}.$$

The mean vector and covariance matrix can be partitioned accordingly:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix},$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \begin{bmatrix} \text{cov}(\mathbf{X}^{(1)}) & \text{cov}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) \\ \text{cov}(\mathbf{X}^{(2)}, \mathbf{X}^{(1)}) & \text{cov}(\mathbf{X}^{(2)}) \end{bmatrix}$$

Partitioning into more than two groups is straightforward.

Results on first-order and second-order moments of linear transformations, $\mathbf{CX}$, of a random vector:

1. $E(\mathbf{CX}) = \mathbf{C}\boldsymbol{\mu}$
2. $\text{cov}(\mathbf{CX}) = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'$
3. $\text{cov}(\mathbf{CX}, \mathbf{DX}) = \mathbf{C}\boldsymbol{\Sigma}\mathbf{D}'$

Examples:

<u>Matrix inequalities and maximization</u>

1. Cauchy-Schwartz inequality: For any two $p \times 1$ vectors $\mathbf{b}$ and $\mathbf{d}$,

   $$(\mathbf{b}'\mathbf{d})^2 \leq (\mathbf{b}'\mathbf{b})(\mathbf{d}'\mathbf{d}),$$

   with equality iff $\mathbf{b} = c\mathbf{d}$ for some constant c.

   Application to sample correlation coefficient:

2. Extended Cauchy-Schwartz inequality: Let $\mathbf{B}$ be any $p \times p$ p.d. matrix. Then for any two $p \times 1$ vectors $\mathbf{b}$ and $\mathbf{d}$,

   $$(\mathbf{b}'\mathbf{d})^2 \leq (\mathbf{b}'\mathbf{B}\mathbf{b})(\mathbf{d}'\mathbf{B}^{-1}\mathbf{d}),$$

   with equality iff $\mathbf{b} = c\mathbf{B}^{-1}\mathbf{d}$ for some constant c.

3. Maximization lemma (immediate consequence of extended C-S inequality): Let $\mathbf{B}$ be any $p \times p$ p.d. matrix and let $\mathbf{d}$ be any $p \times 1$ vector. Then

$$\max_{\mathbf{x} \neq \mathbf{0}} \frac{(\mathbf{x}'\mathbf{d})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{d}'\mathbf{B}^{-1}\mathbf{d},$$

and the maximum is attained iff $\mathbf{x} = c\mathbf{B}^{-1}\mathbf{d}$ for some $c \neq 0$.

An application of this lemma to simultaneous confidence intervals for the components of a mean vector will be given in Chapter 5.

4. Maximization of quadratic forms (for points on the unit sphere): Let $\mathbf{B}$ be a $p \times p$ p.d. matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ and corresponding normalized eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_p$. Then:

(a) $\max_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_1$, and the maximum is attained iff $\mathbf{x} = c\mathbf{e}_1$ for some $c \neq 0$.

(b) $\min_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_p$, and the minimum is attained iff $\mathbf{x} = c\mathbf{e}_p$ for some $c \neq 0$.

(c) For any $k = 1, 2, \ldots, p-1$,

$$\max_{\mathbf{x} \neq \mathbf{0}, \mathbf{x} \perp \mathbf{e}_1, \ldots, \mathbf{e}_k} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_{k+1},$$

and the maximum is attained iff $\mathbf{x} = c\mathbf{e}_{k+1}$ for some $c \neq 0$.

Thus, the smallest and largest eigenvalues of $\mathbf{B}$ represent extreme values of the quadratic form $\mathbf{x}'\mathbf{B}\mathbf{x}$ for points $\mathbf{x}$ on the unit sphere.

We will apply this result to principal component analysis in Chapter 8.

The Geometry of the Sample

Recall that the values of p variables measured on n individuals can be arranged in an $n \times p$ data matrix:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

Two ways to view $\mathbf{X}$:

1. The rows of $\mathbf{X}$ represent n points in $\mathcal{R}^p$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

2. The columns of $\mathbf{X}$ represent p points (or vectors) in $\mathcal{R}^n$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p \end{bmatrix}$$

There is merit to each point of view.

The first point of view provides information on the sample's central location and variation.

Example:

- $\bar{\mathbf{x}}$ measures central location of the n points in $\mathcal{R}^p$.
- $\mathbf{S}_n$ measures spread of the n points in $\mathcal{R}^p$.
- A scalar measure of spread, called the generalized sample variance, will be described soon.

From the second point of view, it is more useful to regard $\mathbf{y}_1, \ldots, \mathbf{y}_p$ as vectors (rather than points) in $\mathcal{R}^n$.

Example:

To interpret the sample mean vector $\bar{\mathbf{x}} = (\bar{x}_i)$:

1. Consider the vector $\frac{1}{\sqrt{n}}\mathbf{1}_n$, and display it in $\mathcal{R}^n$.
2. Next consider the projection of $\mathbf{y}_i$ on $\frac{1}{\sqrt{n}}\mathbf{1}_n$:
3. Thus, $\bar{x}_i$ is the multiple of $\mathbf{1}_n$ required to yield the projection of $\mathbf{y}_i$ onto the line determined by $\mathbf{1}_n$.

Further, for each $\mathbf{y}_i$ we can define a *residual vector*

$$\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1}_n = \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix}$$

and display it (from an appropriate two-dimensional perspective) as follows:

Note that $\mathbf{d}_i$ is orthogonal to $\bar{x}_i\mathbf{1}$.

Two more interpretations of the residual vectors:

1. Lengths of the $\mathbf{d}_i$'s are proportional to the corresponding sample standard deviations, since

   Similarly, $\mathbf{d}_i'\mathbf{d}_k = ns_{ik}$.
2. The cosine of the angle between $\mathbf{d}_i$ and $\mathbf{d}_k$ is the sample correlation coefficient between variables i and k.

Random Samples

Conceptual foundation for MVA: Regard the data matrix $\mathbf{X}$ as a realization of a random matrix, in which the rows $\mathbf{X}_1', \ldots, \mathbf{X}_n'$ represent independent observations from a common p-variate distribution.

Note:

- The variables in any individual $\mathbf{X}_j$ may be correlated.
- If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are not independent, the statistical inferences we make will be of dubious validity.

Important Result: If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are a random sample from any MV distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then

$$E(\bar{\mathbf{X}}) = \boldsymbol{\mu}, \qquad \text{cov}(\bar{\mathbf{X}}) = \frac{1}{n}\boldsymbol{\Sigma},$$

$$E(\mathbf{S}_n) = \left(\frac{n-1}{n}\right)\boldsymbol{\Sigma}.$$

Proof:

Note:

- MV normality not needed for any of these results.
- $\bar{\mathbf{X}}$ is unbiased for $\boldsymbol{\mu}$, but $\mathbf{S}_n$ is biased for $\boldsymbol{\Sigma}$. However, $\mathbf{S} \equiv (\frac{n}{n-1})\mathbf{S}_n$ is unbiased.

Generalized Variance

Motivation: May want a more compact description (in fact, a scalar description) of sample variation than that provided by $\mathbf{S}_n$ or $\mathbf{S}$.

Definition: The *generalized sample variance* of a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is $|\mathbf{S}|$.

Two geometrical interpretations of $|\mathbf{S}|$ (don't dwell on these):

1. In $\mathcal{R}^p$, $|\mathbf{S}| = \text{constant} \times (\text{volume of ellipsoid})^2$, where the ellipsoid is $\{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \leq c^2\}$ (measures statistical distance from $\mathbf{x}$ to $\bar{\mathbf{x}}$).
2. In $\mathcal{R}^n$, $|\mathbf{S}| = \left(\frac{1}{n-1}\right)^p (\text{volume})^2$, where "volume" is the volume of the hypertrapezoidal region generated by the p residual vectors $\mathbf{d}_1, \ldots, \mathbf{d}_p$.

Some facts about the generalized sample variance:

- Very different sample covariance matrices can give the same generalized sample variance, so its usefulness is limited.
  Example:

- $|\mathbf{S}| = 0 \Leftrightarrow \mathbf{d}_1, \ldots, \mathbf{d}_p$ are linearly dependent.
- If $n \leq p$, then $|\mathbf{S}| = 0$ (no matter what $\mathbf{X}$ is).
- If a $p \times 1$ vector $\mathbf{a} \neq 0$ exists for which $\mathbf{a}'\mathbf{X}_j$ is constant for all j, then $|\mathbf{S}| = 0$. If not, and if $n > p$, then $|\mathbf{S}| > 0$.

Definition: The *total sample variance* of a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is $s_{11} + s_{22} + \cdots + s_{pp} = \text{tr}(\mathbf{S})$.

Two geometrical interpretations of total sample variance:

1. In $\mathcal{R}^p$, TSV = sum of squared distances from the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to the point $\bar{\mathbf{x}}$, divided by $n - 1$.

2. In $\mathcal{R}^n$, TSV = sum of squared lengths of $\mathbf{d}_1, \ldots, \mathbf{d}_p$, divided by $n - 1$.

Sample mean, covariance, and correlation as matrix operations on data matrix $\mathbf{X}$

- $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}$

- $\mathbf{S} = \frac{1}{n-1}\mathbf{X}'(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}$
- $\mathbf{R} = [\text{diag}(1/\sqrt{s_{11}}, \ldots, 1/\sqrt{s_{pp}})]\mathbf{S}$ $\times[\text{diag}(1/\sqrt{s_{11}}, \ldots, 1/\sqrt{s_{pp}})]$

Sample moments for linear combinations of variables

Let $\mathbf{x}_j'$ be an arbitrary row of the data matrix $\mathbf{X}$. Consider

$$\mathbf{c}'\mathbf{x}_j = c_1 x_{j1} + \cdots + c_p x_{jp},$$
$$\mathbf{d}'\mathbf{x}_j = d_1 x_{j1} + \cdots + d_p x_{jp}.$$

Facts:

- Sample mean of the n $\mathbf{c}'\mathbf{x}_j$'s: $\mathbf{c}'\bar{\mathbf{x}}$
- Sample variance of the n $\mathbf{c}'\mathbf{x}_j$'s: $\mathbf{c}'\mathbf{S}\mathbf{c}$
- Sample covariance of the n $\mathbf{c}'\mathbf{x}_j$'s and $\mathbf{d}'\mathbf{x}_j$'s: $\mathbf{c}'\mathbf{S}\mathbf{d}$

Example:

The multivariate normal density function

Recall the univariate normal density function:

Note that the exponential part can be written as follows:

Generalization to MV. Here $\mathbf{X}$ represents a $p \times 1$ random vector, and we write $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Assume $\boldsymbol{\Sigma}$ is p.d. Then the density function for $\mathbf{X}$ is:

Note that contours of constant density are ellipsoids defined by the $\mathbf{x}$-values that satisfy

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2.$$

These ellipsoids are centered at $\boldsymbol{\mu}$ and have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$, where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\boldsymbol{\Sigma}$ and $\mathbf{e}_1, \ldots, \mathbf{e}_p$ are the corresponding eigenvectors.

The bivariate normal density function, in detail

Two examples of contours of constant density:

Properties of the Multivariate Normal Distribution

Until noted otherwise, let $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

I. *Linear combinations of the components of* $\mathbf{X}$ *are normally distributed.*

More precise: If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then every linear combination $\mathbf{a}'\mathbf{X} + d$ is distributed as $N(\mathbf{a}'\boldsymbol{\mu} + d, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$. Incidentally, the converse is also true.

Can extend to a vector of linear combinations, as follows: If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{A}$ and $\mathbf{d}$ are a $q \times p$ matrix and $q \times 1$ vector, respectively, then $\mathbf{AX} + \mathbf{d}$ is distributed as $N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{d}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

Examples:

II. *All subsets of the components of* $\mathbf{X}$ *have a MVN distribution.*

More precise: If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and we partition $\mathbf{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

where $\mathbf{X}_1$ is $q \times 1$, then $\mathbf{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

(It is also true that $\mathbf{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.)

III. *Zero covariance implies that the corresponding components of* $\mathbf{X}$ *are distributed independently.*

More precise: Suppose $\mathbf{X}_1$ and $\mathbf{X}_2$ are $q_1 \times 1$ and $q_2 \times 1$, respectively, and

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

Then $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

A related result: If $\mathbf{X}_1 \sim N_{q_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{X}_2 \sim N_{q_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, and $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent, then

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

IV. *The conditional distribution of any subset of components of* $\mathbf{X}$ *given any other disjoint subset of components is MVN.*

More precise: With $\mathbf{X}$ partitioned as in III, the conditional distribution of $\mathbf{X}_1$, given $\mathbf{X}_2 = \mathbf{x}_2$, is MVN, with

- mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ and
- covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Note:

- The conditional mean is a linear function of the values of the variables being conditioned upon.
- The conditional covariance matrix does not depend on the values of variables being conditioned upon.

Examples:

V. *Certain quadratic forms in a MVN vector have a chi-square distribution.*

More precise: If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_p.$$

A special case you're familiar with:

Another example:

A consequence of this result is that the region in $\mathcal{R}^p$ enclosed by the contour

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \chi^2_{p,\alpha}$$

has probability $1 - \alpha$.

VI. *Linear combinations of independent MVN random vectors, each having the same covariance matrix, are MVN.*

More precise: Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent where $\mathbf{X}_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for every j. Define

$$\mathbf{V}_1 = c_1\mathbf{X}_1 + \cdots + c_n\mathbf{X}_n, \quad \mathbf{V}_2 = b_1\mathbf{X}_1 + \cdots + b_n\mathbf{X}_n.$$

Then the joint distribution of $\mathbf{V}_1$ and $\mathbf{V}_2$ is

$$N_{2p}\left(\begin{bmatrix} \sum_{j=1}^n c_j\boldsymbol{\mu}_j \\ \sum_{j=1}^n b_j\boldsymbol{\mu}_j \end{bmatrix}, \begin{bmatrix} (\mathbf{c}'\mathbf{c})\boldsymbol{\Sigma} & (\mathbf{b}'\mathbf{c})\boldsymbol{\Sigma} \\ (\mathbf{b}'\mathbf{c})\boldsymbol{\Sigma} & (\mathbf{b}'\mathbf{b})\boldsymbol{\Sigma} \end{bmatrix}\right).$$

Example:

Maximum Likelihood Estimators (MLEs) of the Parameters of $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and Their Properties

Suppose we have a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. How should we estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$? A well-established method of estimation is MLE, which consists of maximizing the joint density function of the random sample with respect to the parameters. Here, this joint density (also called the likelihood function) is as follows:

Fact 1: The unique MLEs, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, in this setting are

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}, \qquad \hat{\boldsymbol{\Sigma}} = \mathbf{S}_n = \frac{n-1}{n}\mathbf{S}.$$

A nice property of MLE's is the following: Suppose we have obtained MLEs of some parameters, but would like to have the MLE of some function of those parameters. It turns out that the MLE of that function of the parameters is the same function of the MLEs of those parameters. Examples:

Fact 2: In this setting, $\bar{\mathbf{X}}$ and $\mathbf{S}$ are *sufficient* statistics, i.e. the conditional distribution of the random sample, given $\bar{\mathbf{X}}$ and $\mathbf{S}$, does not depend on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Our next two facts give distributional results for the MLEs. Before giving them, recall the following UV result:

Back to the MV case (taking $\mathbf{X}_1, \ldots, \mathbf{X}_n$ to be a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$), we have the following.

Fact 3:

- $\bar{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$
- $(n-1)\mathbf{S} \sim$ Wishart with $n-1$ df and covariance parameter $\boldsymbol{\Sigma}$, i.e. $W_{n-1}(\boldsymbol{\Sigma})$
- $\bar{\mathbf{X}}$ and $\mathbf{S}$ are independent.

The Wishart distribution is defined as follows: If $\mathbf{Z}_1, \ldots, \mathbf{Z}_m$ are iid $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, then the distribution of $\Sigma_{j=1}^{m} \mathbf{Z}_j \mathbf{Z}_j'$ is called the Wishart distribution with m df and covariance (matrix) parameter $\boldsymbol{\Sigma}$.

Fact 4 (A CLT): Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from ANY p-variate distribution with mean $\boldsymbol{\mu}$ and p.d. covariance matrix $\boldsymbol{\Sigma}$. Then for large $n - p$,

$$\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})$$

is distributed approximately as $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, and

$$n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$$

is distributed approximately as χ^2_p.

Assessing the MVN Assumption

- Important because many MV inferences depend on the extent to which the data are MVN (or at least approximately so).
- A single, all-encompassing diagnostic is elusive.
- The text emphasizes diagnostics that assess UV and BV normality.
- UV and BV assessments not foolproof, but the risk of there existing some higher-dimensional non-normality is regarded as small in most practical situations.

A. Assessments of UVN (done separately for each of the p variables)

1. Dot diagrams or histograms
2. Chi-square test on numbers of observations within intervals
3. Kolmogorov-Smirnov and Cramer-von Mises tests based on the sample cumulative distribution function
4. Normal probability (or Q-Q) plots — a plot of the sample quantiles versus the quantiles of a standard normal distribution. "Straightness" of the points in the plot indicate conformance to a MVN assumption.

Details of construction:

Example (see handout).

A Q-Q plot is a graphical diagnostic for UVN. It would be good to have an objective measure of "straightness" to accompany the plot. Such a measure is the correlation coefficient between the $x_{(j)}$'s and $q_{(j)}$'s:

$$r_Q = \frac{\Sigma_{j=1}^{n}(x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\Sigma_{j=1}^{n}(x_{(j)} - \bar{x})^2 \Sigma_{j=1}^{n}(q_{(j)} - \bar{q})^2}}$$

Reject UVN at significance level α if $r_Q < r_{Q,\alpha,n}$, where the critical values $r_Q < r_{Q,\alpha,n}$ are found in Table 4.2 of the textbook. For example, when $n = 100$, the critical values are as follows:

.01	.05	.10
.9822	.9873	.9895

Return to example on handout.

B. Assessments of BVN (done separately for each pair of variables)

1. Two-dimensional scatterplots of each pair of variables
   - BVN $\Rightarrow$ the conditional mean appears roughly linear.
   - BVN $\Rightarrow$ the points' convex hull appears roughly elliptical.

2. Quadratic forms in each pair of variables

   Partition $\mathbf{X}$ as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

   where $\mathbf{X}_1$ is 2×1. Partition $\boldsymbol{\mu}$, $\bar{\mathbf{X}}$, $\boldsymbol{\Sigma}$, and $\mathbf{S}$ accordingly. Then

$$\begin{aligned} \mathbf{X} &\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\Rightarrow \mathbf{X}_1 \sim N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ &\Rightarrow (\mathbf{X}_1 - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{X}_1 - \boldsymbol{\mu}_1) \sim \chi^2_2. \end{aligned}$$

   Thus, BVN $\Rightarrow$ approximately $100(1-\alpha)\%$ of the bivariate observations should lie in the ellipse

$$(\mathbf{X}_1 - \bar{\mathbf{X}}_1)' \mathbf{S}_{11}^{-1} (\mathbf{X}_1 - \bar{\mathbf{X}}_1) \leq \chi^2_{2,\alpha}.$$

   Can take $\alpha = 0.5$.

C. A more MV approach: Chi-square plots

Uses the same idea as in method B2, but applies it to the entire $p \times 1$ vector of observed variables (and incorporates the Q-Q plot idea).

Three-step procedure:

1. Compute $d_j^2 = (\mathbf{X}_j - \bar{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_j - \bar{\mathbf{X}})$ for $j = 1, \ldots, n$ and rank them to obtain $d^2_{(1)} \leq d^2_{(2)} \leq \cdots \leq d^2_{(n)}$.
2. Obtain quantiles of the χ^2_p distribution, i.e. $\chi^2_{p,1-\frac{j-0.5}{n}}$ $(j = 1, \ldots, n)$.
3. Plot the $d^2_{(j)}$'s versus the $\chi^2_{p,1-\frac{j-0.5}{n}}$'s.

$\mathbf{X}$ MVN $\Rightarrow$ the points in the plot should lie approximately on a straight line.

Can also check whether approximately 50% of the d_j^2's are less than $\chi^2_{p,.50}$.

See radiotherapy data example once more.

Transformations to Normality

Suppose we find that the random sample does not appear to have been drawn from a MVN distribution. One way to proceed is to seek transformations of the variables such that the joint distribution of the transformed variables is consistent with a MVN assumption (or more nearly so).

How do we select a transformation?

1. On the basis of theoretical considerations
   - known physical laws
   - common statistical experience
2. On the basis of what the data tell us

Consider *power transformations*, i.e. x^λ. Special cases:

Power transformations generally "work" when the histogram departs from normality by being skewed either right or left.

How does a power transformation affect symmetry about the mean?

- If $\lambda < 1$, large values of x are pulled in.

- If $\lambda > 1$, large values of x are pushed out.

Some trial and error with different choices of λ may lead to a satisfactory result.

A more formal way to select a power transformation is by the Box-Cox procedure. Recall that the Box-Cox family of power transformations is

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log x & \text{if } \lambda = 0 \end{cases}$$

In this non-regression context, the "best" value of λ is the one that minimizes the sample variance of the $y_j^{(\lambda)}$'s, where

$$y_j^{(\lambda)} = \begin{cases} \frac{x_j^\lambda - 1}{\lambda[(\Pi_{i=1}^n x_i)^{1/n}]^{\lambda-1}} & \text{if } \lambda \neq 0 \\ (\Pi_{i=1}^n x_i)^{1/n} \log x & \text{if } \lambda = 0 \end{cases}$$

Remarks:

- Minimizing the sample variance of the $y_j^{(\lambda)}$'s can be shown to be equivalent to maximizing

  $$\log L(\lambda) = -\frac{n}{2}\log(S(\lambda)) + (\lambda - 1)\sum_{j=1}^n \log x_j$$

  where $S(\lambda)$ is the sample variance of the $x_j^{(\lambda)}$'s.
- The required calculations can be performed using PROC TRANSREG of SAS.
- There is no guarantee that the chosen value of λ transforms the data to normality, so we need to examine Q-Q plots of the transformed data.

Refer to the radiotherapy data example, handed out in class.

The application of the Box-Cox method described above focuses on making the UV marginals more nearly normal. We can also apply it to the entire p-variate distribution, by seeking the vector $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_p]'$ that maximizes

$$\log L(\boldsymbol{\lambda}) = -\frac{n}{2}\log|\mathbf{S}(\boldsymbol{\lambda})| + (\lambda_1 - 1)\sum_{j=1}^n \log x_{j1} + \cdots + (\lambda_p - 1)\sum_{j=1}^n \log x_{jp}.$$

SAS PROC TRANSREG does not do this, so we have to program it ourselves. See the radiotheraphy example for details.

Hotelling's T^2-test for $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$

Recall the following UV problem:

> Suppose $X_1, \ldots, X_n$ are a random sample from $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown (but σ^2 is positive). Suppose further that we want to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Then a size-α test is to reject H_0 if
>
> $$t \equiv \frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} > t_{\alpha/2, n-1}.$$

This is the well-known *one-sample t-test*.

The MV analogue of this problem is as follows:

> Suppose $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown (but $\boldsymbol{\Sigma}$ is p.d.). Suppose further that we want to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.

The MV analogue of the UV t-statistic is called Hotelling's T^2-statistic:

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0).$$

Note:

- T^2 is a measure of statistical distance from $\bar{\mathbf{X}}$ to $\boldsymbol{\mu}_0$.

- T^2 reduces to t^2 when $p = 1$:

What is (are) the critical value(s) for T^2? It turns out that under H_0,

$$T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p}.$$

So a size-α test is to reject H_0 if $T^2 > \frac{(n-1)p}{n-p} F_{\alpha,p,n-p}$.

Now refer to the sweat data example, on handout provided in class.

Issues regarding Hotelling's T^2 test:

- The test is based on an assumption of MVN, so we should check this assumption (at least when $n - p < 25$ or 30).
  - Radiotherapy data: $n = 98$, so test is OK even though the data aren't MVN.
  - Sweat data: $n = 20$, so outcome of our check for MVN is important.
- It's instructive to compare the structure of T^2 with t^2 in terms of the random variables, vectors, or matrices involved:

- T^2 is invariant under nonsingular linear transformations of the $p \times 1$ vector of measurements on an individual. That is, if we let

  $$\mathbf{Y}_j = \mathbf{C}\mathbf{X}_j + \mathbf{d} \qquad (j = 1, \ldots, n)$$

  where $\mathbf{C}$ is nonsingular, then T^2 for the transformed data will be identical to T^2 for the original data. Thus T^2 does not depend on the units of measurement.
- Hotelling's T^2 test is (equivalent to) the likelihood ratio test of $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.
- Usually data analysts are (or should be) more interested in a confidence region for $\boldsymbol{\mu}$ than in hypothesis testing. That is, they want to determine a region of $\boldsymbol{\mu}$-values within which the true value of $\boldsymbol{\mu}$ is likely to lie.

Confidence Regions for $\boldsymbol{\mu}$ (one-sample situation)

General definition: Let $\mathbf{X}$ be the data matrix, let $\alpha \in (0, 1)$, and let $\boldsymbol{\theta}$ be a vector of parameters belonging to a parameter space Θ. A $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$ is a set $R(\mathbf{X})$ of $\boldsymbol{\theta}$-values in Θ such that, prior to collecting the data,

$$P[\boldsymbol{\theta} \in R(\mathbf{X})] = 1 - \alpha.$$

Application of definition to the problem at hand: We seek a subset $R(\mathbf{X})$ of $\mathcal{R}^p$ such that $P[\boldsymbol{\mu} \in R(\mathbf{X})] = 1 - \alpha$.

Derivation of such a confidence region:

Example: 95% confidence region for $\boldsymbol{\mu}$ in the sweat data example is:

Remarks:

- This confidence region is a p-dimensional ellipsoid. E.g., if $p = 2$ the region is as follows:

- The orientation and lengths of the axes of the ellipsoid can be determined from the eigenvalues λ_i and eigenvectors $\mathbf{e}_i$ of $\mathbf{S}$.

- There is an equivalence between the T^2 test and this confidence region. Specifically, the size-α T^2 test of $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ is rejected $\Leftrightarrow$ $\boldsymbol{\mu}_0$ lies outside the aforementioned confidence region.
- Because this confidence region is "curved," from its algebraic specification it is not easy to determine at a glance if a specified $\boldsymbol{\mu}$-value belongs to it. This is a deficiency relative to a univariate confidence interval. For ease of interpretation it would be better for the confidence region to be "rectangular," i.e. the Cartesian product of intervals (one for each scalar component of $\boldsymbol{\mu}$). This motivates the idea of *simultaneous confidence intervals*.

One-at-a-time Confidence Intervals for $\mu_1, \ldots, \mu_p$

Fact: For any $p \times 1$ vector $\boldsymbol{\ell} \neq \mathbf{0}$,

$$\frac{\boldsymbol{\ell}'\bar{\mathbf{X}} - \boldsymbol{\ell}'\boldsymbol{\mu}}{\sqrt{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}/n}} \sim t_{n-1}$$

and therefore a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{\ell}'\boldsymbol{\mu}$ is given by

$$\boldsymbol{\ell}'\bar{\mathbf{X}} \pm t_{\alpha/2,n-1}\sqrt{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}/n}.$$

For the specific case of $\boldsymbol{\ell}_i'\boldsymbol{\mu} = \mu_i$ $(i = 1, \ldots, p)$ we obtain the following set of intervals, each of which is a $100(1-\alpha)\%$ confidence interval for the respective μ_i:

$$\begin{aligned}
\bar{X}_1 &\pm t_{\alpha/2,n-1}\sqrt{s_{11}/n} \\
\bar{X}_2 &\pm t_{\alpha/2,n-1}\sqrt{s_{22}/n} \\
&\vdots \\
\bar{X}_p &\pm t_{\alpha/2,n-1}\sqrt{s_{pp}/n}
\end{aligned}$$

Now think in terms of events and probabilities. Let C_i represent the event that the ith $100(1-\alpha)\%$ one-at-a-time confidence interval for μ_i contains μ_i. Three questions:

1. What is $P(C_i)$?
2. The simultaneous coverage probability (SCP) is defined as $P(C_1 \cap C_2 \cap \cdots \cap C_p)$. How does the SCP compare to $P(C_i)$?
3. If the C_i's are independent, what is the SCP?

Since the C_i's are generally not independent, we need some alternative strategies to obtain a desired SCP.

Bonferroni-based Simultaneous Confidence Intervals for $\mu_1, \ldots, \mu_p$

The Bonferroni Inequality says that if $A_1, \ldots, A_m$ are events, then

$$P(A_1 \cap A_2 \cap \cdots \cap A_m) \geq 1 - \sum_{i=1}^{m} (1 - P(A_i)).$$

Applied to the setting we're currently interested in, we have

$$P(C_1 \cap C_2 \cap \cdots \cap C_p) \geq 1 - \sum_{i=1}^{p} \alpha = 1 - p\alpha.$$

Example calculation: If we have a collection of 95% one-at-a-time confidence intervals for $\mu_1, \ldots, \mu_7$, their SCP will be at least what?

Consequently, the Bonferroni-based intervals

$$\begin{aligned} \bar{X}_1 &\pm t_{\alpha/(2p),n-1}\sqrt{s_{11}/n} \\ \bar{X}_2 &\pm t_{\alpha/(2p),n-1}\sqrt{s_{22}/n} \\ &\vdots \\ \bar{X}_p &\pm t_{\alpha/(2p),n-1}\sqrt{s_{pp}/n} \end{aligned}$$

have SCP at least $1 - \alpha$.

Remarks:

- The Bonferroni-based intervals result from a simple modification of the expressions for the one-at-a-time intervals.
- As the number of components of $\boldsymbol{\mu}$ increase, the Bonferroni-based intervals get wider.
- The Bonferroni-based intervals are conservative (i.e. they may be wider than necessary).
- Comparison of widths of Bonferroni-based intervals and one-at-a-time intervals:

- The Bonferroni method can also be applied to variations of the problem currently under consideration. E.g., simultaneous confidence intervals for a subset of the components of $\boldsymbol{\mu}$, or for all pairwise contrasts $\mu_i - \mu_j$ $(i \neq j)$.

T^2-based Simultaneous Confidence Intervals for $\mu_1, \ldots, \mu_p$

The distributional result at the top of page 36 implies that for every $\boldsymbol{\ell} \neq \mathbf{0}$,

$$P\left(\frac{n[\boldsymbol{\ell}'(\bar{\mathbf{X}} - \boldsymbol{\mu})]^2}{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}} \leq t^2_{\alpha/2,n-1}\right) = 1 - \alpha.$$

In fact, for each $\boldsymbol{\ell} \neq \mathbf{0}$ the interval

$$\boldsymbol{\ell}'\bar{\mathbf{X}} \pm t_{\alpha/2,n-1}\sqrt{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}/n}$$

contains $\boldsymbol{\ell}'\boldsymbol{\mu}$ iff the inequality in the probability statement above is satisfied.

Take $c > 0$ and let $C_{\boldsymbol{\ell}}$ be the event

$$\left\{\frac{n[\boldsymbol{\ell}'(\bar{\mathbf{X}} - \boldsymbol{\mu})]^2}{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}} \leq c\right\}.$$

Consider the infinite collection of intervals of the form $\boldsymbol{\ell}'\bar{\mathbf{X}} \pm \sqrt{c}\sqrt{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}/n}$ (one interval for every $\boldsymbol{\ell} \neq \mathbf{0}$). The event that every interval in this collection contains its respective $\boldsymbol{\ell}'\boldsymbol{\mu}$ is $\cap_{\boldsymbol{\ell} \neq \mathbf{0}} C_{\boldsymbol{\ell}}$.

But note that

$$\begin{aligned}\cap_{\boldsymbol{\ell}\neq\mathbf{0}} C_{\boldsymbol{\ell}} &= \left\{\max_{\boldsymbol{\ell}\neq\mathbf{0}}\left(\frac{n[\boldsymbol{\ell}'(\bar{\mathbf{X}}-\boldsymbol{\mu})]^2}{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}}\right)\leq c\right\}\\ &= \left\{n(\bar{\mathbf{X}}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu})\leq c\right\}\\ &= \left\{T^2\leq c\right\}.\end{aligned}$$

Thus, if we want the probability to be $1-\alpha$ that the collection of all intervals of the form $\boldsymbol{\ell}'\bar{\mathbf{X}}\pm\sqrt{c}\sqrt{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}/n}$ contain their respective $\boldsymbol{\ell}'\boldsymbol{\mu}$, we can take the cut-off point c to be $\frac{(n-1)p}{n-p}F_{\alpha,p,n-p}$.

Therefore, the infinite collection of intervals

$$\{\boldsymbol{\ell}'\bar{\mathbf{X}}\pm\sqrt{\frac{(n-1)p}{n-p}F_{\alpha,p,n-p}}\sqrt{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}/n}:\boldsymbol{\ell}\in\mathcal{R}^p\}$$

has SCP $1-\alpha$. Choosing the $\boldsymbol{\ell}$'s to give us the components of $\boldsymbol{\mu}$ yields the $100(1-\alpha)\%$ T^2-intervals

$$\begin{aligned}\bar{X}_1 &\pm \sqrt{\frac{(n-1)p}{n-p}F_{\alpha,p,n-p}}\sqrt{s_{11}/n}\\ \bar{X}_2 &\pm \sqrt{\frac{(n-1)p}{n-p}F_{\alpha,p,n-p}}\sqrt{s_{22}/n}\\ &\vdots\\ \bar{X}_p &\pm \sqrt{\frac{(n-1)p}{n-p}F_{\alpha,p,n-p}}\sqrt{s_{pp}/n}\end{aligned}$$

Remarks:

- T^2-based intervals are wider than the corresponding Bonferroni-based intervals when you're only interested in the p components of $\boldsymbol{\mu}$.
- T^2-based intervals are ideal for "data snooping."

Sweat data example again:

Inference when some values are missing: The EM Algorithm

In the UV context, missing values are easily handled if they are "missing at random."

But in the MV context, some observations (which are vectors) may be only partially missing. For example, the observed data matrix could be

$$\mathbf{X}=\begin{bmatrix}— & 0 & 3\\ 7 & 2 & 6\\ 5 & 1 & 2\\ — & — & 5\end{bmatrix}.$$

Most MV techniques require all components of the data matrix $\mathbf{X}$ to be observed.

If $\mathbf{X}$ is not complete, we can use the EM algorithm to impute the missing values, and then proceed as though the imputed values were actually observed.

The EM algorithm has two steps:

1. Prediction step — Given an estimate $\tilde{\boldsymbol{\theta}}$ of the unknown parameters, predict the contribution of any missing observation to the sufficient statistics for the complete data set.
2. Estimation step — Use the predicted sufficient statistics to compute a revised estimate of the parameters.

In our setting $\mathbf{X}_1,\ldots,\mathbf{X}_n$ is a random sample from $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and the complete-data sufficient statistics are

$$\mathbf{T}_1=\sum_{j=1}^{n}\mathbf{X}_j=n\bar{\mathbf{X}},$$

$$\mathbf{T}_2=\sum_{j=1}^{n}\mathbf{X}_j\mathbf{X}_j'=(n-1)\mathbf{S}+n\bar{\mathbf{X}}\bar{\mathbf{X}}'.$$

Notation: Let $\mathbf{x}_j^{(1)}$ and $\mathbf{x}_j^{(2)}$ represent the missing components and observed components, respectively, of the "complete" observation $\mathbf{x}_j$. Let $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ be estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, with components arranged to correspond to the ordering of $\mathbf{x}_j = \begin{bmatrix} \mathbf{x}_j^{(1)} \\ \mathbf{x}_j^{(2)} \end{bmatrix}$.

Predicted contribution of $\mathbf{x}_j^{(1)}$ to $\mathbf{T}_1$:

$$\begin{aligned} \tilde{\mathbf{x}}_j^{(1)} &= E(\mathbf{X}_j^{(1)} | \mathbf{X}_j^{(2)} = \mathbf{x}_j^{(2)}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \\ &= \tilde{\boldsymbol{\mu}}^{(1)} + \tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}(\mathbf{x}_j^{(2)} - \tilde{\boldsymbol{\mu}}^{(2)}) \end{aligned}$$

Predicted contribution of $\mathbf{x}_j^{(1)}$ to $\mathbf{T}_2$:

$$\begin{aligned} \widetilde{\mathbf{x}_j^{(1)}\mathbf{x}_j^{(1)\prime}} &= E(\mathbf{X}_j^{(1)}\mathbf{X}_j^{(1)\prime} | \mathbf{X}_j^{(2)} = \mathbf{x}_j^{(2)}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \\ &= \tilde{\boldsymbol{\Sigma}}_{11} - \tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\tilde{\boldsymbol{\Sigma}}_{21} + \tilde{\mathbf{x}}_j^{(1)}\tilde{\mathbf{x}}_j^{(1)\prime}, \\ \widetilde{\mathbf{x}_j^{(1)}\mathbf{x}_j^{(2)\prime}} &= E(\mathbf{X}_j^{(1)}\mathbf{X}_j^{(2)\prime} | \mathbf{X}_j^{(2)} = \mathbf{x}_j^{(2)}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \\ &= E(\mathbf{X}_j^{(1)} | \mathbf{X}_j^{(2)} = \mathbf{x}_j^{(2)}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})\mathbf{x}_j^{(2)\prime} = \tilde{\mathbf{x}}_j^{(1)}\mathbf{x}_j^{(2)\prime} \end{aligned}$$

Estimation step:

$$\tilde{\boldsymbol{\mu}} = \frac{1}{n}\tilde{\mathbf{T}}_1, \qquad \tilde{\boldsymbol{\Sigma}} = \frac{1}{n}\tilde{\mathbf{T}}_2 - \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}'.$$

Continue cycling between the steps until "convergence."

Once final estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are obtained, we can, for example, test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ by comparing $n(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)$ to $\chi^2_{\alpha,p}$. This is approximately a size-α test.

See section 5.7 of text for a numerical example.

The validity of this approach requires the data to be "missing at random."

Large-Sample Inference for $\mu_1, \ldots, \mu_p$

The exact validity of Hotelling's T^2 test, the associated ellipsoidal confidence region, and the associated Bonferroni-based and T^2-based simultaneous confidence intervals rests on the assumption that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

However, as a result of the CLT, these inference procedures are approximately valid for a dataset of sufficiently large sample size, even when the normality assumption is not satisfied.

For large-sample inference, the authors suggest replacing $t_{\alpha/2,n-1}$ with $z_{\alpha/2}$ and replacing $\frac{(n-1)p}{n-p}F_{\alpha,p,n-p}$ with $\chi^2_{\alpha,p}$ in the expressions for test statistics and confidence regions/intervals. For example, the large-sample Hotelling's T^2 test:

But it's OK (it's a little more conservative) to just leave the expressions alone.

An application to categorical data:

- Suppose that a nominal attribute variable with $q+1$ levels or categories is measured on every individual in a population.
- An individual from category k will be assigned the following value:
- So the possible values are the vectors

  and the corresponding probabilities are $p_1, p_2, \ldots, p_q, p_{q+1}$ where $p_{q+1} = 1 - (p_1 + \cdots + p_q)$.
- Regard the observations as a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from this distribution. Let X_{ij} be the ith component of $\mathbf{X}_j$.

- It can be shown that

$$E(\mathbf{X}_j) = \mathbf{p} = (p_i), \qquad \text{var}(X_{ij}) = p_i(1-p_i),$$
$$\text{cov}(X_{ij}, X_{kj}) = -p_i p_k.$$

- Sample mean vector is $\bar{\mathbf{X}} = \frac{1}{n}\Sigma_{j=1}^{n}\mathbf{X}_j$ = sample proportion vector = $\hat{\mathbf{p}} = (\hat{p}_i)$.
- Covariance matrix of sample mean vector is $\frac{1}{n}\mathbf{\Sigma}$ where $\mathbf{\Sigma}$ is the $(q+1)\times(q+1)$ matrix with elements given two bullets ago.
- By CLT, $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \sim N_{q+1}(\mathbf{0}, \mathbf{\Sigma})$ approximately, for large n.
- Thus, can do large-sample Hotelling's T^2-test of $H_0 : \mathbf{p} = \mathbf{p}_0$ versus $H_1 : \mathbf{p} \neq \mathbf{p}_0$ as follows:

  where $\hat{\mathbf{\Sigma}}$ is obtained by replacing the p_i's in $\mathbf{\Sigma}$ with the $\hat{p}_i$'s.
- How large must n be for approximately valid inference? A rule of thumb is: $n\hat{p}_i \geq 20$ for all $i = 1, \ldots, q+1$.

Example:

Multivariate Two-Sample Inference

In the UV situation, the simplest kind of 2-sample situation is *paired data*, i.e. a case where 2 "treatments" are applied to the same units or to units that are probably much more similar than randomly selected units. Examples:

Review of inference for UV paired data:

- Let X_{1j} = response to Treatment 1 for jth replicate; let X_{2j} = response to Treatment 2 for jth replicate; and let $D_j = X_{1j} - X_{2j}$.
- If D_j's are iid $N(\delta, \sigma_d^2)$, then

$$\frac{\bar{D} - \delta}{s_d/\sqrt{n}} \sim t_{n-1},$$

  where $\bar{D}$ and s_d are the sample mean and standard deviation of the D_j's.

- Thus, we can test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ using the test statistic $\sqrt{n}\bar{D}/s_d$, and a confidence interval for δ can be obtained easily.
- In essence, the two-sample inference problem is changed to a one-sample problem by reducing the data to within-pair differences.

There are two important ways in which this UV situation can be extended to a MV one:

1. Still have 2 treatments, but measure $p > 1$ responses on each replicate (= MV paired comparisons)
2. Still measure 1 response variable, but have $q > 2$ treatments for each replicate (= Repeated measures)

Multivariate Paired Comparisons

Notation:

Result: Assuming that the $\mathbf{D}_j$'s are a random sample from $N_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_d)$,

$$n(\bar{\mathbf{D}}-\boldsymbol{\delta})'\mathbf{S}_d^{-1}(\bar{\mathbf{D}}-\boldsymbol{\delta}) \sim T^2 \sim \frac{(n-1)p}{n-p}F_{p,n-p}$$

where $\bar{\mathbf{D}}$ and $\mathbf{S}_d$ are the sample mean vector and covariance matrix of the $\mathbf{D}_j$'s.

Ramification: Can proceed with inference on $\boldsymbol{\delta}$ as though it were $\boldsymbol{\mu}$ in a one-sample situation.

Two features not generally shared with the one-sample situation:

- The scientific hypothesis of interest is usually whether there is any difference between the average effects of the two treatments, i.e. the null hypothesis is $H_0 : \boldsymbol{\delta} = \mathbf{0}$. So our test statistic is usually $n\bar{\mathbf{D}}'\mathbf{S}_d^{-1}\bar{\mathbf{D}}$.
- It is possible for the original data to be non-normal and yet the $\mathbf{D}_j$'s be MVN. Example:

  Consequently, we should check the $\mathbf{D}_j$'s, not the $\mathbf{X}_{ij}$'s, for consistency with MVN.

Repeated Measures

Each individual receives each treatment once, on successive occasions. There are q treatments and n individuals. The data on the jth individual are thus $\mathbf{X}_j' = [X_{j1}, \ldots, X_{jq}]$.

The scientific hypothesis of interest usually is whether there is any difference between the average effects of any two treatments. I.e.,

$$H_0 : \mu_1-\mu_2 = 0, \quad \mu_1-\mu_3 = 0, \quad \ldots, \quad \mu_1-\mu_q = 0$$

i.e.,

$$H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0} \quad \text{versus } H_1 : \mathbf{C}\boldsymbol{\mu} \neq \mathbf{0}$$

where

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix}.$$

$\mathbf{C}$ is a *contrast matrix*.

Assuming that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a random sample from $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then:

- $\mathbf{C}\mathbf{X}_1, \ldots, \mathbf{C}\mathbf{X}_n$ is a random sample from $N_{q-1}(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')$.
- The sample mean and covariance matrix of the $\mathbf{C}\mathbf{X}_j$'s are $\mathbf{C}\bar{\mathbf{X}}$ and $\mathbf{C}\mathbf{S}\mathbf{C}'$.

Consequently,

$$\begin{aligned} T^2 &= n(\mathbf{C}\bar{\mathbf{X}} - \mathbf{C}\boldsymbol{\mu})'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}(\mathbf{C}\bar{\mathbf{X}} - \mathbf{C}\boldsymbol{\mu}) \\ &\sim \frac{(n-1)(q-1)}{n-q+1}F_{q-1,n-q+1} \end{aligned}$$

and a size-α test of the above hypotheses is to reject H_0 if

$$n(\mathbf{C}\bar{\mathbf{X}})'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}\mathbf{C}\bar{\mathbf{X}} > \frac{(n-1)(q-1)}{n-q+1}F_{\alpha,q-1,n-q+1}$$

$100(1-\alpha)\%$ simultaneous confidence intervals for all contrasts $\mathbf{c}'\boldsymbol{\mu}$ are the set of all intervals

$$\mathbf{c}'\bar{\mathbf{X}} \pm \sqrt{\frac{(n-1)(q-1)}{n-q+1}F_{\alpha,q-1,n-q}}\sqrt{\frac{\mathbf{c}'\mathbf{S}\mathbf{c}}{n}} :$$

$\mathbf{c}$ is a contrast vector

Remarks:

- $n(\mathbf{C}\bar{\mathbf{X}})'(\mathbf{CSC}')^{-1}\mathbf{C}\bar{\mathbf{X}}$ is invariant to the choice of $\mathbf{C}$, provided that $\mathbf{C}$ is a $(q-1)\times q$ contrast matrix with linearly independent rows.
- In essence, the q-treatment UV problem is changed to a one-sample MV problem in which the $\mathbf{CX}_j$'s are the observations.

See hypothetical headache medicine example, handed out in class.

<u>Inference for the Difference of Two Means Based on Independent Samples</u>

Recall the following UV problem. If $X_{11}, \ldots, X_{1,n_1}$ and $X_{21}, \ldots, X_{2,n_2}$ are independent random samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively, then a size-α test of $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ is to reject H_0 if

$$\frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} > t_{\alpha/2, n_1+n_2-2}.$$

Distinguishing feature (compared to paired data): There is no defensible rationale for pairing individuals, even if the two sample sizes are equal.

Notation for MV problem:

- Samples $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1,n_1}$ and $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2,n_2}$
- Sample means $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$
- Sample covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$.

Assumption: The two samples are independent random samples from $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively. Initially we shall also assume that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

Hypotheses of interest:

- $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ (primary)
- $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ versus $H_1 : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ (secondary)

How the first pair of hypotheses are tested depends on the outcome of a test of the second pair of hypotheses. Formal tests of equality of the covariance matrices exist, but here we will use the following informal rule of thumb:

If, on the basis of this informal check (or a formal test), we decide to assume $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, then we pool the two sample covariance matrices to obtain an estimate of the common covariance matrix:

$$\mathbf{S}_{pooled} = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1+n_2-2}$$

We have the following distributional result:

$$\begin{aligned}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))'[(\frac{1}{n_1} + \frac{1}{n_2})\mathbf{S}_{pooled}]^{-1} \\ \times[\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \\ \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p, n_1+n_2-p-1}.\end{aligned}$$

From this distributional result, we can do the following inference procedures:

- Two-sample size-α Hotelling's T^2 test for $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1: \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ is to reject H_0 if

$$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}]^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) > \frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{\alpha,p,n_1+n_2-p-1} \equiv c_\alpha.$$

- A $100(1-\alpha)\%$ ellipsoidal confidence region for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is given by

$$\{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 : [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]' \times [\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}]^{-1}[\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \leq c_\alpha\}.$$

The directions and lengths of the confidence ellipsoid's axes are determined from the eigenvectors and eigenvalues of $\mathbf{S}_{pooled}$.

- Bonferroni-based $100(1-\alpha)\%$ simultaneous confidence intervals for the p population mean differences are given by

$$(\bar{X}_{1i} - \bar{X}_{2i}) \pm t_{\alpha/(2p),n_1+n_2-2}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s_{ii,pooled}} \quad (i = 1, \ldots, p).$$

- T^2-based $100(1-\alpha)\%$ simultaneous confidence intervals for all linear combinations of the components of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ are given by

$$\left\{\boldsymbol{\ell}'(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \pm \sqrt{c_\alpha}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\ell}'\mathbf{S}_{pooled}\boldsymbol{\ell}} : \boldsymbol{\ell} \in \mathcal{R}^p\right\}.$$

Large-sample inference: If the assumption of equal covariance matrices is dubious, and/or the assumption of MVN is dubious for either population, but $n_1 - p$ and $n_2 - p$ are both "large," then we can test the equality of means hypothesis at approximate size α by rejecting H_0 if

$$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \geq \chi^2_{\alpha,p}.$$

The corresponding ellipsoidal confidence region and simultaneous confidence intervals are straightforward.

Now see an example using the effluent data, on a class handout.

<u>Inference for the Mean Vectors of More Than Two Populations: The One-Way MANOVA</u>

Recall the following UV problem. Suppose we observe independent random samples from g normal populations with means $\mu_1, \ldots, \mu_g$ and common variance σ^2, i.e.

Suppose we want to test whether the means are all equal, i.e. $H_0: \mu_1 = \mu_2 = \cdots = \mu_g$ versus H_1 : at least one of the μ_l's is different from the others. By writing

$$\mu_l = \mu + (\mu_l - \mu) \equiv \mu + \tau_l,$$

for any number μ, we see that the H_0 above is equivalent to $H_0: \tau_1 = \tau_2 = \cdots = \tau_g$.

It is useful to cast the problem in terms of a linear model for the data:

$$X_{lj} = \mu + \tau_l + e_{lj}, \quad e_{lj} \sim \text{iid } N(0, \sigma^2)$$
$$(j = 1, \ldots, n_l;\ l = 1, \ldots, g).$$

This model is not "identified," i.e. one or more of its parameters are not uniquely determined. The constraint $\sum_{l=1}^{g} n_l \tau_l = 0$ identifies the model, so we add it to the model specification. (Note: this is not the only constraint that identifies the model.)

Least squares estimates: $\hat{\mu} = \bar{X}$, $\hat{\tau}_l = \bar{X}_l - \bar{X}$.

Fitted residual: $\hat{e}_{lj} = X_{lj} - (\hat{\mu} + \hat{\tau}_l) = X_{lj} - [\bar{X} + (\bar{X}_l - \bar{X})] = X_{lj} - \bar{X}_l$.

So we have the decomposition

$$X_{lj} = \bar{X} + (\bar{X}_l - \bar{X}) + (X_{lj} - \bar{X}_l) = \hat{\mu} + \hat{\tau}_l + \hat{e}_{lj}.$$

Partitioning the variability in the data accordingly yields the following decomposition of the total sum of squares (corrected for the mean), also called the one-way analysis of variance (ANOVA):

$$\sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X})^2 = \sum_{l=1}^{g} n_l (\bar{X}_l - \bar{X})^2 + \sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X}_l)^2,$$

i.e.

$$SS_{cor} = SS_{tr} + SS_{res}.$$

This decomposition plus some additional information can be laid out in a table, called the ANOVA table, as follows:

The hypotheses of interest can be tested based on this table. Specifically, we reject $H_0 : \tau_1 = \tau_2 = \cdots = \tau_g$ in favor of H_1 : not H_0 at the α level of significance if

$$F \equiv \frac{SS_{tr}/(g-1)}{SS_{res}/(\sum_{l=1}^{g} n_l - g)} > F_{\alpha, g-1, \sum_{l=1}^{g} n_l - g}.$$

Note, for future reference, that rejecting H_0 for large F is equivalent to rejecting H_0 for large values of SS_{tr}/SS_{res}, which in turn is equivalent to rejecting H_0 for small values of $SS_{res}/(SS_{res} + SS_{tr})$.

Now for the MV extension of these ideas.

An identified model for $\mathbf{X}_{lj}$:

$$\mathbf{X}_{lj} = \boldsymbol{\mu} + \boldsymbol{\tau}_l + \mathbf{e}_{lj}, \quad \sum_{l=1}^{g} n_l \boldsymbol{\tau}_l = \mathbf{0},$$
$$\mathbf{e}_{lj} \sim \text{iid } N_p(\mathbf{0}, \boldsymbol{\Sigma}) \qquad (j = 1, \ldots, n_l;\ l = 1, \ldots, g).$$

Least squares estimates of parameters:

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}, \quad \hat{\boldsymbol{\tau}}_l = \bar{\mathbf{X}}_l - \bar{\mathbf{X}}.$$

Decomposition of an observation in terms of least squares estimates:

$$\mathbf{X}_{lj} = \bar{\mathbf{X}} + (\bar{\mathbf{X}}_l - \bar{\mathbf{X}}) + (\mathbf{X}_{lj} - \bar{\mathbf{X}}_l) = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\tau}}_l + \hat{\mathbf{e}}_{lj}.$$

The decomposition above gives rise to the following decomposition of the total (mean-corrected) *sum of squares and cross-products matrix*:

$$\sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{X}_{lj} - \bar{\mathbf{X}})(\mathbf{X}_{lj} - \bar{\mathbf{X}})' = \sum_{l=1}^{g} n_l (\bar{\mathbf{X}}_l - \bar{\mathbf{X}})(\bar{\mathbf{X}}_l - \bar{\mathbf{X}})' + \sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{X}_{lj} - \bar{\mathbf{X}}_l)(\mathbf{X}_{lj} - \bar{\mathbf{X}}_l)',$$

i.e.

Total SS+CP matrix = Between SS+CP matrix
+Within SS+CP matrix.

We can lay this out in a table called the multivariate analysis of variance (MANOVA) table, as follows:

Suppose that we wish to use a scalar-valued statistic computable from this MANOVA as a test statistic for testing $H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\tau}_g$ versus H_1 : not H_0. What statistic is a reasonable extension of the univariate F-test?

Although several test statistics are possible, we will feature *Wilks' lambda* statistic,

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|}.$$

We reject H_0 if Λ^* is too small. How small is too small?

- If $p = 1$ or $g = 2$, a multiple of $(1 - \Lambda^*)/\Lambda^*$ has a particular F distribution under H_0, so we reject H_0 if this statistic is larger than an appropriate size-α F cutoff point. See Table 6.3 for details.

- If $p = 2$ or $g = 3$, a multiple of $(1-\sqrt{\Lambda^*})/\sqrt{\Lambda^*}$ has a particular F distribution under H_0, so we reject H_0 if this statistic is larger than an appropriate size-α F cutoff point. See Table 6.3 for details.
- If $p > 2$ and $g > 3$, the following result can be used if $n \equiv \Sigma_{l=1}^{g} n_l$ is large: Reject H_0 at level α if

$$-(n - 1 - \frac{p+g}{2}) \log \Lambda^* > \chi^2_{\alpha, p(g-1)}.$$

- In most applications we just use the P-value provided in the computer output to judge statistical significance.

Now refer to the supplemental class handouts for some examples, analyzed using SAS.

Simultaneous Confidence Intervals for Components of Differences of Mean Vectors

Suppose that a MANOVA has indicated that not all the population means are equal, or equivalently that there are significant differences due to "treatments." Then we'd like to know which variable(s) differ significantly, and for which treatments they do so.

We will learn the following Bonferroni-based approach:

- The treatment difference vectors are $\boldsymbol{\tau}_k - \boldsymbol{\tau}_l$ $(l < k = 2, \ldots, g)$.
- The ith component of $\boldsymbol{\tau}_k - \boldsymbol{\tau}_l$ is
- So the number of scalar differences that we want confidence intervals for, such that their SCP is at least $1 - \alpha$, is
- Our point estimate of $\tau_{ki} - \tau_{li}$ is

- The variance of our point estimate is $\text{var}(\bar{X}_{ki} - \bar{X}_{li}) =$
- A good estimate of the variance of our point estimate is $\hat{\text{var}}(\bar{X}_{ki} - \bar{X}_{li}) =$
- Thus, a set of confidence intervals for all differences $\tau_{ki} - \tau_{li}$ $(l < k = 2, \ldots, g;\ i = 1, \ldots, p)$ whose SCP is at least $1 - \alpha$ is given by

$$(\bar{X}_{ki} - \bar{X}_{li}) \pm t_{\alpha/[pg(g-1)],n-g} \sqrt{\frac{w_{ii}}{n-g}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)}.$$

Application to the turtle example:

The Two-Way MANOVA

The MANOVA described previously is appropriate when a model with only one factor of classification is appropriate.

We want to extend the same ideas to a setting where there are two factors of classification, such as:

- A randomized blocks experiment
- A completely randomized experiment, with treatments that have a factorial structure
- An observational study in which individuals are cross-classified by two factors (e.g. socioeconomic class and region of the country)

The UV model and ANOVA:

- Notation: X_{lkr} is the rth observation at level l of Factor 1 and level k of Factor 2 $(l = 1, \ldots, g;\ k = 1, \ldots, b;\ r = 1, \ldots, n)$. Note that for simplicity we consider only the "balanced" case.
- Model:

$$\begin{aligned} X_{lkr} &= \mu + \tau_l + \beta_k + \gamma_{lk} + e_{lkr}, \\ & e_{lkr} \text{ iid } N(0, \sigma^2), \\ & \sum_{l=1}^{g} \tau_l = \sum_{k=1}^{b} \beta_k = \sum_{l=1}^{g} \gamma_{lk} = \sum_{k=1}^{b} \gamma_{lk} = 0. \end{aligned}$$

  The γ_{lk}'s are "interaction" terms. Note: This model cannot be fit unless $n \geq 2$; if $n = 1$, however, we can fit the model without interaction terms.
- ANOVA:

- Using F-tests, we first test for interaction. If we find no significant interaction, then we test for main effects of each of the two factors.

Now for the MV analogue:

- Model:
- Assume $n \geq 2$.

• MANOVA:

As in the UV case, begin by testing for no interaction, i.e. for testing $H_0 : \boldsymbol{\gamma}_{11} = \boldsymbol{\gamma}_{12} = \cdots = \boldsymbol{\gamma}_{gb} = \mathbf{0}$ versus H_1: not H_0.

• Test statistic is

• Reject H_0 at level α if

$$-[gb(n-1) - \frac{p+1-(g-1)(b-1)}{2}] \log \Lambda^* > \chi^2_{\alpha,(g-1)(b-1)p}$$

If the test for no interaction is rejected, it is inadvisable to proceed with tests on the other effects. Instead, do UV two-way ANOVA's on each variable to see which of the variables have significant interaction effects. For those variables, do interaction plots. The variables that have no significant interaction may be used for further tests.

Test for main effect of Factor 1, i.e. $H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\tau}_g = \mathbf{0}$ versus H_1: not H_0.

• Test statistic is

• Reject H_0 at level α if

$$-[gb(n-1) - \frac{p+1-(g-1)}{2}] \log \Lambda^* > \chi^2_{\alpha,(g-1)p}$$

Test for main effect of Factor 2, i.e. $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\beta}_b = \mathbf{0}$ versus H_1: not H_0.

• Test statistic is

• Reject H_0 at level α if

$$-[gb(n-1) - \frac{p+1-(b-1)}{2}] \log \Lambda^* > \chi^2_{\alpha,(b-1)p}$$

A very nice example of a two-way MANOVA, replete with SAS code and output, can be found on pp. 312-317 of text.

Simultaneous Confidence Intervals for Components of Differences in Main Effect Vectors

Notation:

• $\mathbf{E} = SSP_{res}$, and E_{ii} is ith diagonal element of $\mathbf{E}$.

• $\bar{X}_{l\cdot i}$ is ith component of $\bar{X}_l$.

• $\bar{X}_{\cdot ki}$ is ith component of $\bar{X}_{\cdot k}$

Bonferroni-based $100(1-\alpha)\%$ simultaneous confidence intervals for all Factor 1 differences $\tau_{li} - \tau_{mi}$ $(i = 1, \ldots, p;\ l < m = 2, \ldots, g)$:

$$(\bar{X}_{l\cdot i} - \bar{X}_{m\cdot i}) \pm t_{\alpha/[pg(g-1)],gb(n-1)} \sqrt{\frac{E_{ii}}{gb(n-1)} \cdot \frac{2}{nb}}$$

Bonferroni-based $100(1-\alpha)\%$ simultaneous confidence intervals for all Factor 2 differences $\beta_{ki} - \beta_{qi}$ $(i = 1, \ldots, p;\ k < q = 2, \ldots, b)$:

$$(\bar{X}_{\cdot ki} - \bar{X}_{\cdot qi}) \pm t_{\alpha/[pb(b-1)],gb(n-1)} \sqrt{\frac{E_{ii}}{gb(n-1)} \cdot \frac{2}{ng}}$$

Note: Don't compute these intervals if the interaction effect is significant.

Profile Analysis (for 2 groups)

Suppose that we have obtained independent random samples from two populations, and that the p measurement variables are *commensurable*, i.e. they are expressed in the same units. Then, we may not only be interested in testing the equality of mean vectors hypothesis (which we already know how to do), but also in testing whether certain other linear functions of components of the mean vectors are equal.

Assume subsequently that all p variables are commensurable, that the two populations are MVN, and that the two groups have a common covariance matrix. In this case it is very useful to graph the mean vector for each group as a *profile*, as follows:

Relevant hypotheses to be tested:

1. Are the two groups' profiles equal?

2. Are the two groups' profiles parallel?

3. Assuming that the profiles are parallel, are they coincident?

4. Assuming that the profiles are coincident, are they level?

Methodology for addressing these questions:

1. Been there, done that. A size-α test is to reject H_0 if
$$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}]^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$$
$$> \quad \frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{\alpha,p,n_1+n_2-p-1}$$
(or equivalently can use Wilks' lambda).

2. To address the parallelism hypothesis, define the contrast matrix
$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix}.$$
Then, parallelism holds iff $H_{01} : \mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2$, or equivalently $\mathbf{C}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$. A size-$\alpha$ test is to reject H_{01} if
$$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\mathbf{C}'[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{C}\mathbf{S}_{pooled}\mathbf{C}']^{-1}\mathbf{C}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$$
$$> \quad \frac{(n_1+n_2-2)(p-1)}{n_1+n_2-p}F_{\alpha,p-1,n_1+n_2-p}.$$

Can interpret this as a test for the equality of the means of the $\mathbf{C}\mathbf{X}_{1j}$'s and the $\mathbf{C}\mathbf{X}_{2j}$'s.

3. To test for coincidence given parallelism, first note that two parallel profiles never cross, and thus they are coincident iff the "total heights,"

   are equal. So the hypothesis is equivalent to $H_{02} : \mathbf{1}'\boldsymbol{\mu}_1 = \mathbf{1}'\boldsymbol{\mu}_2$.

   We can test this hypothesis by a UV t-test for the equality of means of the $\mathbf{1}'\mathbf{X}_{1j}$'s and the $\mathbf{1}'\mathbf{X}_{2j}$'s. Equivalently, a size-α test is to reject H_{02} if
$$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\mathbf{1}[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{1}'\mathbf{S}_{pooled}\mathbf{1}]^{-1}\mathbf{1}'(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$$
$$> \quad t^2_{\alpha/2,n_1+n_2-2} = F_{\alpha,1,n_1+n_2-2}.$$

4. To test for level profiles given coincidence, first note that:

   (a) Coincidence occurs iff the two populations have the same mean vector $\boldsymbol{\mu} = (\mu_i)$, the UMVUE of which is

   $$\bar{\mathbf{X}} =$$

   (b) Level profiles occur iff every consecutive difference, $\mu_i - \mu_{i-1}$, is 0.

   Thus, the null hypothesis is $H_{03} : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$. A size-$\alpha$ test is to reject H_{03} if

   $$(n_1 + n_2)\bar{\mathbf{X}}'\mathbf{C}'[\mathbf{CSC}']^{-1}\mathbf{C}\bar{\mathbf{X}}$$
   $$> \frac{(n_1 + n_2 - 1)(p-1)}{n_1 + n_2 - p + 1} F_{\alpha, p-1, n_1+n_2-p+1},$$

   where $\mathbf{S}$ is the sample covariance matrix of the $\mathbf{X}_{1j}$'s and $\mathbf{X}_{2j}$'s assuming they have a common mean.

Each of the hypotheses of interest can be tested equivalently using Wilks' lambda in the context of a particular MANOVA.

Now see the maternal attitudes example on the class handout.

Three questions and their answers:

1. If we determine that the profiles are parallel, what is the advantage of the test for coincidence (given under Point 3 on the previous page) over the standard test for equality of two mean vectors (given under Point 1 on the previous page)?

2. If we determine that the profiles are parallel but not coincident, can we test for levelness of the profiles, and if so, how?

3. If we reject parallelism, it is possible that one (but not both) of the profiles is level. How can we test for levelness of each profile individually?

Univariate Regression Analysis

Our next major topic is MV regression. To warm up, let's briefly review the UV regression model and assumptions, and mention some of the standard UV regression methodology.

UV multiple linear regression model:

$$y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \cdots + \beta_r z_{jr} + \epsilon_j,$$
$$\epsilon_j\text{'s uncorrelated}, \quad E(\epsilon_j) = 0, \quad \text{var}(\epsilon_j) = \sigma^2$$
$$(j = 1, \ldots, n).$$

Can write in vector/matrix notation as

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad E(\boldsymbol{\epsilon}) = \mathbf{0}, \quad \text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}.$$

Note: The above assumptions are sufficient to do least squares point estimation for the model parameters. To do interval estimation and/or hypothesis testing, we generally also assume normality of the ϵ_j's.

Least squares estimates:

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y},$$
$$S^2 = (\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})/(n - r - 1)$$

provided that $\mathbf{Z}$ has full rank $r + 1$.

Fitted values: $\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$

Fitted residuals: $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = [\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{y}$

Additional topics to review:

- Partition of total variation of $\mathbf{y}$ (the ANOVA)
- Sampling properties of least squares estimators
- Confidence ellipsoid for $\boldsymbol{\beta}$; simultaneous confidence intervals for elements of $\boldsymbol{\beta}$
- Tests for the general linear hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$; full-versus-reduced model testing approach
- Inference for $E(y|\mathbf{z})$ and for prediction of y

- Diagnostics: residual plots, testing for outliers, influence and leverage, collinearity
- Variable selection: R^2, Mallow's C_p, PRESS, stepwise techniques

The Multivariate Regression Model

Now suppose we have more than one dependent variable, say m of them. Suppose further that each dependent variable follows its own regression model, with the stipulation that the explanatory variables in each dependent variable's regression equation are identical. An example and non-example:

More generally, the model is

For each fixed i, the ϵ_{ji}'s are assumed to be uncorrelated and such that $E(\epsilon_{ji}) = 0$, $\text{var}(\epsilon_{ji}) = \sigma_{ii}$ for $j = 1, \ldots, n$. However, the errors associated with different measurements on the same individual may be correlated. Thus,

- $\text{cov}(\epsilon_{ji}, \epsilon_{jk}) = \sigma_{ik}$ for all $j = 1, \ldots, n$
- $\text{cov}(\epsilon_{ji}, \epsilon_{lk}) = 0$ if $j \neq l$

The model is more convenient to represent using matrix notation. Define

Then we can write the MV regression model as

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad E(\boldsymbol{\epsilon}) = \mathbf{0}, \quad \text{cov}(\boldsymbol{\epsilon}_{(i)}, \boldsymbol{\epsilon}_{(k)}) = \sigma_{ik}\mathbf{I}$$
$$(i, k = 1, \ldots, m).$$

We will assume that $\mathbf{Z}$ has full rank, so that $\mathbf{Z}'\mathbf{Z}$ is invertible.

In terms of a model for each dependent variable, we have

$$\begin{aligned} \mathbf{y}_{(i)} &= \mathbf{Z}\boldsymbol{\beta}_{(i)} + \boldsymbol{\epsilon}_{(i)}, \qquad (i = 1, \ldots, m) \\ E(\boldsymbol{\epsilon}_{(i)}) &= \mathbf{0}, \qquad \text{cov}(\boldsymbol{\epsilon}_{(i)}, \boldsymbol{\epsilon}_{(k)}) = \sigma_{ik}\mathbf{I}. \end{aligned}$$

Point Estimation of Parameters

Because of the possible dependence between $\boldsymbol{\epsilon}_{(i)}$ and $\boldsymbol{\epsilon}_{(k)}$, it would seem inappropriate to perform inference on the elements of $\boldsymbol{\beta}$ by running m separate UV regression analyses.

In fact, this is only partly true. It turns out that despite the possible dependence, the least squares estimate (LSE) of $\boldsymbol{\beta}_{(i)}$ can be obtained exclusively from the ith UV regression. That is, the LSE of $\boldsymbol{\beta}_{(i)}$ is

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}_{(i)}.$$

Can put these all into one big matrix:

Matrix of predicted (or fitted) values: $\hat{\mathbf{Y}} =$

Matrix of fitted residuals: $\hat{\boldsymbol{\epsilon}} =$

The residuals are orthogonal to the columns of $\mathbf{Z}$, and thus to the predicted values, i.e.

Consequently, the Total SS&CP matrix, $\mathbf{Y}'\mathbf{Y}$, can be decomposed into 2 parts, as follows:

Likewise the mean-corrected Total SS&CP matrix, $(\mathbf{Y} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{Y})'(\mathbf{Y} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{Y})$, can be decomposed as follows, to yield the so-called *overall MANOVA*:

Can estimate $\boldsymbol{\Sigma} = (\sigma_{ik})$ by the Residual SS&CP matrix divided by its df, i.e.,

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{n - r - 1}\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}.$$

Some sampling properties of the LSE's and related quantities:

1. $\hat{\boldsymbol{\beta}}_{(i)}$ is unbiased, i.e., $E(\hat{\boldsymbol{\beta}}_{(i)}) = \boldsymbol{\beta}_{(i)} \quad (i = 1, \ldots, m)$
2. $\text{cov}(\hat{\boldsymbol{\beta}}_{(i)}, \hat{\boldsymbol{\beta}}_{(k)}) = \sigma_{ik}(\mathbf{Z}'\mathbf{Z})^{-1} \quad (i, k = 1, \ldots, m)$
3. $E(\hat{\boldsymbol{\epsilon}}_{(i)}) = \mathbf{0} \quad (i = 1, \ldots, m)$
4. $E(\tilde{\boldsymbol{\Sigma}}) = \boldsymbol{\Sigma}$.
5. $\hat{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{\beta}}$ are uncorrelated, i.e. every column of $\hat{\boldsymbol{\epsilon}}$ is uncorrelated with every column of $\hat{\boldsymbol{\beta}}$.

Now suppose that we are also willing to assume that the ϵ_{ji}'s have a MVN distribution. More precisely, assume that

Then:

- $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$
- $\hat{\boldsymbol{\beta}}$, when its columns are strung out as one long vector, has a MVN distribution
- $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = \left(\frac{n-r-1}{n}\right)\tilde{\boldsymbol{\Sigma}}$ is the MLE of $\boldsymbol{\Sigma}$
- $n\hat{\boldsymbol{\Sigma}} \sim W_{n-r-1}(\boldsymbol{\Sigma})$

Hypothesis Testing

The first general type of hypothesis we shall test is of the form

$$H_0 : \boldsymbol{\beta}^{(2)} = \mathbf{0} \quad \text{where } \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \end{bmatrix}$$

Examples:

- An *overall* test of the regression, i.e. testing whether all of the coefficients corresponding to the explanatory variables (other than the intercept) are equal to 0. For this case, $q = 0$.

- Testing for the importance of any single explanatory variable in the absence of any others (except for the intercept). For this case, $q = 0$ and $r = 1$.

- Testing for the importance of any single explanatory variable in the presence of t others (plus the intercept). For this case, $q = t$ and $r = t + 1$.

- The previous two could be strung together to yield a sequence of "single-variable hypotheses."

- Testing for the importance of s explanatory variables in the presence of t others (plus the intercept). For this case, $q = t$ and $r = t+s$.

To test these hypotheses, consider a "reduced-model versus full-model" testing approach. Under $H_0 : \boldsymbol{\beta}^{(2)} = \mathbf{0}$, the model is

$$\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\beta}^{(1)} + \boldsymbol{\epsilon}$$

and the Error SS&CP matrix is

Under $H_1 : \boldsymbol{\beta}^{(2)} \neq \mathbf{0}$, the model is

$$\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\beta}^{(1)} + \mathbf{Z}_2\boldsymbol{\beta}^{(2)} + \boldsymbol{\epsilon}$$

and the Error SS&CP matrix is

Now, the reduced model cannot fit the data any better than full model (in the sense of minimizing the residual SS&CP matrix). If the reduced model fits almost as well as the full model, then the two Error SS&CP matrices should be about equal. But if the reduced model fits substantially worse than the full model, then $n\hat{\boldsymbol{\Sigma}}$ will be "smaller" (in some sense) than $n\hat{\boldsymbol{\Sigma}}_0$.

Specifically, we will reject H_0 if

$$\Lambda^* = \frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_0|},$$

is too small. This coincides with the likelihood ratio test.

Equivalently, we could reject H_0 if

$$-n \log\left(\frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_0|}\right)$$

is too large.

To obtain an actual cut-off value, we will use a large-sample result. Specifically, we will reject H_0 at approximate level α if

$$-[n-r-1-\frac{1}{2}(m-r+q+1)] \log\left(\frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_0|}\right) > \chi^2_{m(r-q),\alpha}.$$

To yield a good approximation, both $n-r$ and $n-m$ should be "large."

Now see the first two handouts on the analysis of the "girls9" data.

Hypothesis Testing (Continued)

Next, consider testing hypotheses of the form $H_0 : \mathbf{C}\boldsymbol{\beta}\mathbf{M} = \mathbf{0}$ versus $H_1 : \mathbf{C}\boldsymbol{\beta}\mathbf{M} \neq \mathbf{0}$.

- $\mathbf{C}$ picks out or compares coefficients corresponding to different explanatory variables.
- $\mathbf{M}$ picks out or compares coefficients corresponding to different dependent variables.
- The general hypothesis we considered previously, $H_0 : \boldsymbol{\beta}^{(2)} = \mathbf{0}$, is the special case of this hypothesis obtained by putting $\mathbf{C} = [\mathbf{0}_{(r-q)\times(q+1)}, \mathbf{I}_{r-q}]$ and $\mathbf{M} = \mathbf{I}_m$.

Examples, in the particular case of the "girls9" data:

- Overall test of significance of the regression

- Test of importance of *ht* in a model with *age* in it

- Test of difference in slope coefficients of *ht* in a model with *age* in it

All of these hypotheses can be tested by an appropriate Wilk's lambda statistic.

Now see the third handout of the analysis of the "girls9" data.

Confidence ellipsoids and simultaneous confidence intervals

1. For the vector of mean responses (and its elements) at a given $\mathbf{z}_0$
   - Point estimate of $(\mathbf{z}_0'\boldsymbol{\beta})' = \boldsymbol{\beta}'\mathbf{z}_0$ is $\hat{\boldsymbol{\beta}}'\mathbf{z}_0$.
   - $100(1-\alpha)\%$ confidence ellipsoid for $\boldsymbol{\beta}'\mathbf{z}_0$:
   $$(\boldsymbol{\beta}'\mathbf{z}_0 - \hat{\boldsymbol{\beta}}'\mathbf{z}_0)'[\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0\tilde{\mathbf{\Sigma}}]^{-1}(\boldsymbol{\beta}'\mathbf{z}_0 - \hat{\boldsymbol{\beta}}'\mathbf{z}_0) \leq \frac{m(n-r-1)}{n-r-m}F_{\alpha,m,n-r-m}$$
   - $100(1-\alpha)\%$ Bonferroni-based simultaneous confidence intervals for $\mathbf{z}_0'\boldsymbol{\beta}_{(i)}$:
   $$\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)} \pm t_{\alpha/(2m),n-r-1}\sqrt{\tilde{\sigma}_{ii}\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0} \quad (i = 1, \ldots, m)$$
   - $100(1-\alpha)\%$ T^2-based simultaneous confidence intervals for $\mathbf{z}_0'\boldsymbol{\beta}_{(i)}$:
   $$\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)} \pm \sqrt{\frac{m(n-r-1)}{n-r-m}F_{\alpha,m,n-r-m}} \times\sqrt{\tilde{\sigma}_{ii}\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0} \quad (i = 1, \ldots, m)$$

2. For the vector of new responses (and its elements) at a given $\mathbf{z}_0$
   - Point estimate of $\mathbf{y}_0 \equiv (\mathbf{z}_0'\boldsymbol{\beta} + \boldsymbol{\epsilon}_0')' = \boldsymbol{\beta}'\mathbf{z}_0 + \boldsymbol{\epsilon}_0$ is $\hat{\boldsymbol{\beta}}'\mathbf{z}_0$.
   - $100(1-\alpha)\%$ confidence ellipsoid for $\mathbf{y}_0$:
   $$(\mathbf{y}_0 - \hat{\boldsymbol{\beta}}'\mathbf{z}_0)'[(1+\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0)\tilde{\boldsymbol{\Sigma}}]^{-1}(\mathbf{y}_0 - \hat{\boldsymbol{\beta}}'\mathbf{z}_0) \leq \frac{m(n-r-1)}{n-r-m}F_{\alpha,m,n-r-m}$$
   - $100(1-\alpha)\%$ Bonferroni-based simultaneous confidence intervals for y_{0i}:
   $$\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)} \pm t_{\alpha/(2m),n-r-1}\sqrt{\tilde{\sigma}_{ii}(1+\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0)} \quad (i = 1, \ldots, m)$$
   - $100(1-\alpha)\%$ T^2-based simultaneous confidence intervals for y_{0i}:
   $$\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)} \pm \sqrt{\frac{m(n-r-1)}{n-r-m}F_{\alpha,m,n-r-m}} \times\sqrt{\tilde{\sigma}_{ii}(1+\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0)} \quad (i = 1, \ldots, m)$$

3. $100(1-\alpha)\%$ Bonferroni-based simultaneous confidence intervals for the elements of $\boldsymbol{\beta}$:
$$\hat{\beta}_{qi} \pm t_{\alpha/[2m(r+1)],n-r-1}\sqrt{\tilde{\sigma}_{ii}c_{qq}} \quad (q = 1, \ldots, r+1;\ i = 1, \ldots, m)$$
where c_{qq} is the qth diagonal element of $(\mathbf{Z}'\mathbf{Z})^{-1}$.

Since normality is assumed for hypothesis testing and interval estimation, this assumption should be checked if these inference procedures are to be used. The techniques of Chapter 4 can be applied to the rows of $\hat{\boldsymbol{\epsilon}}$.

The Concept of Linear Regression

Suppose we take independent observations of a $p \times 1$ random vector $\mathbf{X}$, which we partition into two subvectors $\mathbf{Y}$ and $\mathbf{Z}$ as follows (note that we're using some symbols to represent different things now than they represented last time):
$$\mathbf{X} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix}.$$
Recall from Chapter 4 that if $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where
$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_Z \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_{ZZ} \end{bmatrix},$$
then the conditional distribution of $\mathbf{Y}|\mathbf{Z} = \mathbf{z}$ is

The two quantities

$$\boldsymbol{\mu}_{Y \cdot Z} \equiv$$

and

$$\boldsymbol{\Sigma}_{YY \cdot Z} \equiv$$

are useful when $\mathbf{X}$ has any distribution, not just the MVN distribution (though they are not necessarily the conditional mean and conditional covariance matrix of $\mathbf{Y}$ given $\mathbf{Z}$ for other distributions).

- $\boldsymbol{\mu}_{Y \cdot Z}$ is called *the (population) regression function of* $\mathbf{Y}$ *on* $\mathbf{Z}$. It has certain optimality properties as a predictor of $\mathbf{Y}$, regardless of the distribution of $\mathbf{X}$.
- $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_{ZZ}^{-1}$ is called *the matrix of (population) regression coefficients.*
- $\boldsymbol{\Sigma}_{YY \cdot Z}$ is called the *partial covariance matrix of* $\mathbf{Y}$ (*adjusted for* $\mathbf{Z}$). Its diagonal elements are called the *partial variances* and its off-diagonal elements are called the *partial covariances* of $\mathbf{Y}$ (adjusted for $\mathbf{Z}$). They convey how much variation and covariation there is among the elements of $\mathbf{Y}$ after adjusting for any linear relationship existing between $\mathbf{Y}$ and $\mathbf{Z}$.

Example:

Now suppose that we have observed a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of the random vector denoted by $\mathbf{X}$ on the previous page. Then:

- A "good" estimate of the regression function is

$$\hat{\boldsymbol{\mu}}_{Y \cdot Z}(\mathbf{z}) = \bar{\mathbf{Y}} + \mathbf{S}_{YZ}\mathbf{S}_{ZZ}^{-1}(\mathbf{z} - \bar{\mathbf{Z}})$$

where

$$\mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})' = \begin{bmatrix} \mathbf{S}_{YY} & \mathbf{S}_{YZ} \\ \mathbf{S}_{ZY} & \mathbf{S}_{ZZ} \end{bmatrix}$$

$=$

Note that $\hat{\boldsymbol{\mu}}_{Y \cdot Z}(\mathbf{z})$ is identical to the fitted value at $\mathbf{Z} = \mathbf{z}$ from a MV regression analysis.

Note also that $\hat{\boldsymbol{\mu}}_{Y \cdot Z}(\mathbf{z})$ is equal to the MLE if the original distribution is MVN.

- A "good" estimate of the partial covariance matrix is

$$\mathbf{S}_{YY \cdot Z} = \mathbf{S}_{YY} - \mathbf{S}_{YZ}\mathbf{S}_{ZZ}^{-1}\mathbf{S}_{ZY}.$$

Example (continued):

Let

$$\boldsymbol{\Sigma}_{YY \cdot Z} = (\sigma_{ij \cdot Z}), \qquad \mathbf{S}_{YY \cdot Z} = (S_{ij \cdot Z}).$$

Also let

$$\boldsymbol{\sigma}'_{Y_i Z} = [\mathrm{cov}(Y_i, Z_1), \ldots, \mathrm{cov}(Y_i, Z_r)],$$

$\mathbf{S}'_{Y_i Z}$ = the corresponding vector within $\mathbf{S}$.

Certain functions of these quantities describe certain types of linear association:

1. The partial correlation coefficient

$$\rho_{ij \cdot Z} = \frac{\sigma_{ij \cdot Z}}{\sqrt{\sigma_{ii \cdot Z}\sigma_{jj \cdot Z}}}$$

measures the linear association between Y_i and Y_j after eliminating (adjusting for) the effects of $Z_1, \ldots, Z_r$. It can be estimated as follows:

$$r_{ij \cdot Z} = \frac{S_{ij \cdot Z}}{\sqrt{S_{ii \cdot Z}S_{jj \cdot Z}}}.$$

2. The multiple correlation coefficient

$$\rho_{Y_i(Z)} = \sqrt{\frac{\boldsymbol{\sigma}'_{Y_i Z}\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\sigma}_{Y_i Z}}{\sigma_{ii}}}$$

measures the linear association between Y_i and $\mathbf{Z}$. Specifically, for each i it equals the maximum correlation between Y_i and all possible linear combinations of $\mathbf{Z}$. It can be estimated as follows:

$$R_{Y_i(Z)} = \sqrt{\frac{\mathbf{S}'_{Y_i Z}\mathbf{S}_{ZZ}^{-1}\mathbf{S}_{Y_i Z}}{S_{ii}}}.$$

Example (continued):

Another example (Problem 7.14 from textbook):

- Y = rate of return on investment
- Z_1 = manager's attitude toward risk (1=very conservative, ..., 5=very aggressive)
- Z_2 = years of experience
- Observed correlation matrix based on random sample of size 25:

$$\mathbf{R} = \begin{bmatrix} 1.0 & -.35 & .82 \\ -.35 & 1.0 & -.60 \\ .82 & -.60 & 1.0 \end{bmatrix}$$

Principal Components Analysis (PCA)

Introductory remarks:

- PCA is a technique for examining relationships among a set of p (generally correlated) variables.
- PCA linearly transforms the original set of p variables to a new set of p uncorrelated variables, and orders the variables in decreasing order of "importance." The new variables are called the principal components.
- The usual objective of PCA is to see if the first few PC's account for most of the variation in the original data. If they do, then it is argued that the effective dimensionality of the variables is less than p.

- If the first few PC's account for most of the variation, then two main uses of PCA historically have been to:
  - Give meaningful interpretations to the PC's.
  - Reduce the dimensionality of the data to simplify further MVA.

  Most statisticians now believe their main use should only be the second of these.
- PCA should be applied only to variables that are "on an equal footing," i.e. not to variables that can be separated into dependent and explanatory variables.
- PCA is a mathematical technique; it does not require the user to specify a probability distribution for the data.

Mathematical derivation of (population) principal components (of covariance matrix):

- Suppose $\mathbf{X}$ is a $p \times 1$ random vector with p.d. covariance matrix $\mathbf{\Sigma}$.
- $\mathbf{\Sigma}$ has p (eigenvalue, eigenvector) pairs

  where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and the $\mathbf{e}_i$'s are orthonormal.
- Consider p arbitrary (for now) linear combinations of the elements of $\mathbf{X}$:

The PC's are special choices of the $\boldsymbol{\ell}_i'\mathbf{X}$'s; loosely, they are the choices that maximize the variance, subject to being uncorrelated with all the others. More precisely, the ith PC is the $\boldsymbol{\ell}_i'\mathbf{X}$ that maximizes $\boldsymbol{\ell}_i'\boldsymbol{\Sigma}\boldsymbol{\ell}_i$ subject to the constraints $\boldsymbol{\ell}_i'\boldsymbol{\ell}_i = 1$ and $\boldsymbol{\ell}_i'\boldsymbol{\Sigma}\boldsymbol{\ell}_k = 0$ for $k = 1, \ldots, i-1$.

- It turns out (see Result 8.1 in text) that the $\boldsymbol{\ell}_i$'s that do the job are the eigenvectors of $\boldsymbol{\Sigma}$, i.e. the $\mathbf{e}_i$'s. Thus, the PC's are

  The magnitude of e_{ki} indicates the importance of X_k to the ith PC.

- Furthermore, the maximized value of $\text{var}(\boldsymbol{\ell}_i'\mathbf{X})$ is

  That is, the PC's have variances equal to the eigenvalues of $\boldsymbol{\Sigma}$.

- Now, the sum of $\boldsymbol{\Sigma}$'s eigenvalues is equal to the sum of the variances, i.e.

  Thus, the proportion of the total population variance due to the kth PC is

Example:

PCA of (population) correlation matrix:

- It is quite common to calculate the PC's of a set of variables after they have been standardized to have unit variance.
- This means that we would be finding the PC's of the correlation matrix $\boldsymbol{\rho}$, rather than of $\boldsymbol{\Sigma}$. The eigenvalues and eigenvectors of $\boldsymbol{\rho}$ are generally NOT the same as those of $\boldsymbol{\Sigma}$.
- Choosing to do a PCA of $\boldsymbol{\rho}$ rather than $\boldsymbol{\Sigma}$ involves a definite but arbitrary decision to regard the original variables as equally important. (More discussion to come later.)
- For $\boldsymbol{\rho}$, what do the eigenvalues sum up to?

  Hence the proportion of total variation accounted for by the ith PC is:

Example (continued):

Now examine the class handout of a PCA of the radiotherapy data using PROC PRINCOMP in SAS.

Principal Components Analysis, Part 2

Geometric Interpretation: PCA can be thought of as a rotation of the original coordinate axes in p-dimensional space.

Example:

Some important special cases in PCA:

- *Uncorrelated original variables.* Suppose one original variable, say X_i is uncorrelated with all the other original variables and has variance σ_{ii}. Then the ith column of $\mathbf{\Sigma}$ is

  Moreover,

  Thus, X_i is itself a PC, and σ_{ii} is the eigenvalue corresponding to this PC. Which PC X_i is depends on the value of σ_{ii} relative to the other eigenvalues of $\mathbf{\Sigma}$.

  If all p variables are uncorrelated, the PC's are the same as the original variables, but arranged in decreasing order of the variances. PCA is of no benefit in this case.

- *Repeated eigenvalues.* If, for some q and k,

$$\lambda_{q+1} = \lambda_{q+2} = \cdots = \lambda_{q+k},$$

  then $\lambda \equiv \lambda_{q+1}$ is said to be an eigenvalue with multiplicity k. The eigenvectors corresponding to multiple eigenvalues are NOT unique, so no meaning should be ascribed to the corresponding PC's. If all the eigenvalues are equal, then PCA is of little benefit.

- *Structured covariance matrices.* Some covariance matrices have a certain pattern or structure to them. It is often of interest what the principal components are for such a matrix.

  Examples:

  (a) Compound symmetry

  (b) First-order autoregressive structure

- *Zero eigenvalues.* If the original variables are linearly dependent, then at least one of the eigenvalues will be 0. (The converse is also true.) We excluded this possibility by assuming that $\mathbf{\Sigma}$ is positive definite, but if we relaxed this assumption by allowing $\mathbf{\Sigma}$ to be nonnegative definite, then we would interpret the PC's corresponding to zero eigenvalues as explaining none of the variation.

PCA for a Random Sample (Rather than a Population):

Our discussion to this point has dealt with PCA of a population covariance matrix or correlation matrix, thus we have implicitly assumed that it is known. Generally this is not so, in which case we apply PCA to the sample covariance matrix or sample correlation matrix.

- We can regard the PC's and their estimated variances as estimates of the population PC's and their variances.
- If the population is MVN, some asymptotic theory on the behavior of these estimates has been worked out.
- Modern view: PCA is an exploratory data analytic technique for dimension reduction. Inferential aspects are de-emphasized.

The Thorny Issue of Scaling:

PCA depends critically on the scales used to measure the variables; thus, they are generally not a unique characteristic of the data. Some implications of this are:

- If one variable has a much larger variance than the others, then this variable will dominate the first PC of $\mathbf{\Sigma}$ regardless of the correlation structure.
- If we change the units on this variable so that it has a much smaller variance, then the first PC of $\mathbf{\Sigma}$ may change dramatically.

Therefore, there is little point in doing a PCA unless one of the following is true:

- The variables are unitless (for example, percentages).
- The variables are all measured in the same units.
- The variables have (or are transformed to have) roughly similar variances.

Consequently, unless the situation fits into the first or second of these cases, the PCA should be based on the correlation matrix. Even basing PCA on the correlation matrix is somewhat arbitrary, since it forces the variables to be given equal importance.

Scaling problems do not arise in most of the other kinds of MVA procedures we have considered.

Worthwhile uses of PCA — some examples:

- Checking for MVN by bivariate scatterplots and Q-Q plots of the first few PC's.
- Checking for unusual observations by plots of the same kind as above. Clusters of points may also be revealed. Example of changes in pollen spectrum over time:

- To avoid problems caused by multicollinearity of explanatory variables in MV regression analyses. Instead of regressing $\mathbf{Y}$ on all the explanatory variables, we could regress $\mathbf{Y}$ on the first few PC's of the explanatory variables.
- To avoid singularity problems in discriminant analysis when the number of observations is less than the number of variables. (More to come later.)

None of these uses require us to ascribe any meaning to the PC's.

Factor Analysis (FA)

In FA, we attempt to represent the p variables in $\mathbf{X}$ as linear combinations of a few *latent* variables $F_1, F_2, \ldots, F_m$ ($m < p$) called factors.

The factors:

- have values that vary from individual to individual
- are not observed
- are hypothetical constructs

For example, suppose

$$\mathbf{R} = \begin{bmatrix} 1 & .95 & .05 & .05 & .05 \\ & 1 & .03 & .06 & .02 \\ & & 1 & .91 & .86 \\ & & & 1 & .92 \\ & & & & 1 \end{bmatrix}.$$

Here, X_1 and X_2 may be representable by a single factor; also, X_3, X_4, and X_5 may be representable by a single factor.

FA is similar to PCA in that both seek a simpler structure in a set of variables (i.e., dimension reduction). They have the following differences, however:

1. PC's are linear combinations of the original variables. In FA, the original variables are expressed as linear combinations of the factors.
2. In PCA we seek to explain a large proportion of the total variation. In FA we seek to account for the covariances or correlations among the original variables.

The Orthogonal Factor Model:

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \epsilon_1 \\ &\vdots \\ X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \epsilon_p \end{aligned}$$

or

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\epsilon}.$$

Here $F_1, \ldots, F_m$ are the common *factors*; the ℓ_{ij}'s are the *loadings*. Also assume that

$$E(\mathbf{F}) = \mathbf{0}, \quad E(\boldsymbol{\epsilon}) = \mathbf{0}, \quad \text{cov}(\mathbf{F}) = \mathbf{I},$$
$$\text{cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} \quad \text{(diagonal)},$$

and that $\mathbf{F}$ and $\boldsymbol{\epsilon}$ are independent. Then $\boldsymbol{\Sigma} = \text{cov}(\mathbf{X}) = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$. Note that $\mathbf{L}\mathbf{L}'$ is a simplified configuration for the covariances. FA consists of estimating $\mathbf{L}$ and $\boldsymbol{\Psi}$ from $\mathbf{S}$ (or $\mathbf{R}$).

There are several popular methods of estimation; unfortunately, they often give quite different answers. If $m \geq 2$, the factor loadings are not unique (can post–multiply $\mathbf{L}$ by any orthogonal matrix $\mathbf{Q}$).

Canonical Correlation

Recall that the ordinary correlation coefficient between two scalar variables X_1 and X_2, i.e. $\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$, measures the linear association between X_1 and X_2. Also recall that the multiple correlation coefficient between a variable X_1 and a vector $\mathbf{X}_2$, i.e.

$$\rho_{X_1(\mathbf{X}_2)} = \sqrt{\frac{\boldsymbol{\sigma}'_{X_1\mathbf{X}_2}\boldsymbol{\Sigma}^{-1}_{\mathbf{X}_2\mathbf{X}_2}\boldsymbol{\sigma}_{X_1\mathbf{X}_2}}{\sigma_{X_1X_1}}},$$

measures the linear association between X_1 and $\mathbf{X}_2$, and that it is equal to $\max_{\mathbf{b}} \text{corr}(X_1, \mathbf{b}'\mathbf{X}_2)$.

Suppose we want to measure the correlation between two sets of variables, $\mathbf{X}_1$ and $\mathbf{X}_2$, by a scalar. How should we measure this? By analogy with $\rho_{X_1(\mathbf{X}_2)}$, we might consider something like

$$\left(\frac{|\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{-1}_{22}\boldsymbol{\Sigma}_{21}|}{|\boldsymbol{\Sigma}_{11}|}\right)^{\frac{1}{2}}.$$

There are other possibilities, however, and it turns out that the best measure, from several standpoints, is the largest eigenvalue of

$$\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}.$$

Denote this largest eigenvalue by ρ_1^{*2} and call $\rho_1^* = \sqrt{\rho_1^{*2}}$ the first canonical correlation coefficient.

It turns out that $|\rho_1^*| = \max_{\mathbf{a},\mathbf{b}} \text{corr}(\mathbf{a}'\mathbf{X}_1, \mathbf{b}'\mathbf{X}_2)$ and that this maximum is attained when $\mathbf{a}' = \mathbf{e}_1'\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}$, $\mathbf{b}' = \mathbf{f}_1'\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}$ where $\mathbf{e}_1$ is the eigenvector of $\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}$ associated with its largest eigenvalue ρ_1^{*2}, and $\mathbf{f}_1$ is the eigenvector of $\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}$ associated with its largest eigenvalue (which also happens to be ρ_1^{*2}).

Facts:

- Canonical correlations are invariant to changes of scale on either the elements of $\mathbf{X}_1$ or $\mathbf{X}_2$. (Same for ordinary and multiple correlation coefficients.)
- $|\rho_1^*| \geq \max_i \rho_{X_i(\mathbf{X}_2)} \geq \max_{ij} \rho_{X_iX_j}$ where $i \in (1, 2, \ldots, q)$ and $j \in (q+1, \ldots, p)$.

Example:

$$\mathbf{R} = \begin{bmatrix} 1.0 & .6328 & .2412 & .0586 \\ & 1.0 & -.0553 & .0655 \\ & & 1.0 & .4248 \\ & & & 1.0 \end{bmatrix}$$

$n = 160$ children.

$$\mathbf{X}_1 = \begin{bmatrix} \text{reading speed} \\ \text{reading power} \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} \text{arithmetic speed} \\ \text{arithmetic power} \end{bmatrix}$$

We obtain

$$\begin{aligned} \hat{\rho}_1^* &= .39, \\ \hat{U}_1 &= 1.26 * (\text{reading speed}) - 1.03 * (\text{reading power}), \\ \hat{V}_1 &= 1.1 * (\text{arithmetic speed}) \\ &\quad -.45 * (\text{arithmetic power}). \end{aligned}$$

Thus, reading ability (summarized by $\hat{U}_1$) correlates moderately positively with arithmetic ability (summarized by $\hat{V}_1$).

Discrimination, Classification, and Clustering

These are all methods for grouping individuals based on their values of a multivariate vector of measurements $\mathbf{X}$. They can be distinguished by what one "knows" about the groups.

1. Discrimination — separating individuals into a specified number of groups. The number of groups is specified and we know which individuals belong to which groups, but we need to determine what is the best variable or combination of variables for effectively separating the groups.

   Example: R.A. Fisher's classical iris data set. Data on X_1 = sepal length, X_2 = sepal width, X_3 = petal length, and X_4 = petal width of 50 flowers from each of three closely related species of iris. We may want to determine the linear combination of the 4 variables that best separates the 3 species.

2. Classification (Allocation) — classifying a "new" individual, whose group membership is unknown, into one of two or more groups. The number of groups is known. We want to classify the new individual in such a way that the probability or cost of misclassification is minimized. Often, this is accomplished using the results of a discriminant analysis.

   Example: Good and bad credit risks. A loan officer at a bank could collect data on $X_1 =$ income, $X_2 =$ age, $X_3 =$ length of employment at current job, etc. of people who have previously received loans from the bank, and then try to classify each new loan application as good or bad using these variables.

3. Clustering — forming groups of similar individuals. Neither the number of groups, nor which individuals belong to which groups, is pre-specified.

   Example: 18 different measurements on 8 hominids.

Classification for Two Populations

Problem: We need to assign a new object, with corresponding $p \times 1$ measurement vector $\mathbf{x}_0$, to one of two populations π_1 and π_2.

We assume that π_1 and π_2 are large and that the variables measured are "not too discrete," so that the probability distribution of $\mathbf{X}$ for each population can be modeled by a pdf. Let

$$f_1(\mathbf{x}) = \text{pdf for } \pi_1, \qquad f_2(\mathbf{x}) = \text{pdf for } \pi_2.$$

Also let Ω be the combined sample space, i.e. the set of all possible values of $\mathbf{X}$ regardless of which population it comes from.

In classification, we must partition Ω into two disjoint subsets R_1 and R_2, such that if $\mathbf{x}_0 \in R_1$ then we will classify it as belonging to π_1 and if $\mathbf{x}_0 \in R_2$ then we will classify it as belonging to π_2. Illustration:

If the support of π_1 overlaps with the support of π_2, then errors (misclassification errors) are possible. In fact, there are four possible outcomes of the entire classification process, two of which are errors:

1. Correctly classifying a π_1-object
2. Misclassifying a π_2-object as belonging to π_1
3. Correctly classifying a π_2-object
4. Misclassifying a π_1-object as belonging to π_2

Expressions for the probabilities of each of these possible outcomes can be determined, as follows. Let p_i be the prior probability that a randomly selected object from $\pi_1 \cup \pi_2$ belongs to π_i ($i = 1, 2$). Also define

$$P(i|j) = P(\mathbf{X} \in R_i|\pi_j) = \int_{R_i} f_j(\mathbf{x})\, d\mathbf{x}$$
$$(i = 1, 2;\ j = 1, 2).$$

Thus, for a randomly selected object from $\pi_1 \cup \pi_2$,

1. P(correctly classifying a π_1-object)=
2. P(misclassifying a π_2-object as belonging to π_1)=
3. P(correctly classifying a π_2-object)=
4. P(misclassifying a π_1-object as belonging to π_2)=

A good classification procedure should have small misclassification probabilities. Even more important, however, is having small expected misclassification costs. Often, in practice, the cost of one of the misclassification errors is different from the cost of the other (e.g., making a bad loan may be more costly than not making a good one).

Cost matrix:

Then the expected cost of misclassification, ECM, is:

Minimum ECM Rule:

(a) *General case.* The regions R_1 and R_2 that minimize the ECM are as follows:

$$R_1 = \left\{\mathbf{x} : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right\}$$
$$R_2 = \left\{\mathbf{x} : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right\}$$

Example:

(b) *MVN case with common covariance matrix* $\boldsymbol{\Sigma}$. Suppose

$f_1(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $f_2(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$.

Also, suppose initially that the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are known. Then after some algebra we find that the minimum ECM rule is to classify a new object $\mathbf{x}_0$ as π_1 if

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \geq \log\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right].$$

In practice the parameters are unknown so we take random samples from each population and replace the parameters with sample-based estimates to obtain the following "estimated" ECM rule: Classify a new object $\mathbf{x}_0$ to π_1 if

$$(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'\mathbf{S}_{pooled}^{-1}\mathbf{x}_0-\frac{1}{2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2) \geq \log\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right].$$

Observe that the LHS is a linear function of $\mathbf{x}_0$.

Example:

(c) *MVN case with* $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. The estimated ECM rule is to classify $\mathbf{x}_0$ as π_1 if

$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1}-\mathbf{S}_2^{-1})\mathbf{x}_0+(\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1}-\bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0-k \geq \log\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right]$$

where k, which doesn't depend on $\mathbf{x}_0$, is defined in the textbook. Observe that the LHS is a quadratic function of $\mathbf{x}_0$.

Example:

Discriminant Analysis for Two Populations

Consider again the classification rule given previously:

Classify a new object $\mathbf{x}_0$ in π_1 if

$$(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)'\mathbf{S}_{pooled}^{-1}\mathbf{x}_0-\frac{1}{2}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{X}}_1+\bar{\mathbf{X}}_2) \geq \log\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right].$$

This rule was motivated as an estimated minimum ECM rule for a situation with two normal populations with common covariance matrix.

But there is a completely different motivation that leads to essentially the same rule. This second motivation is *discrimination*, i.e. maximum separation of two populations.

Suppose we have random samples from two MV populations, π_1 and π_2, i.e.

$$\mathbf{X}_{11},\ldots,\mathbf{X}_{1n_1} \sim \text{iid from } \pi_1,$$
$$\mathbf{X}_{21},\ldots,\mathbf{X}_{2n_2} \sim \text{iid from } \pi_2.$$

Suppose further that the two populations have the same covariance matrix, but they need not be MVN. Now consider how we might linearly transform each MV observation to a UV observation in such a way that the UV observations from π_1 are "separated as much as possible" from the UV observations from π_2. (The same linear transformation is to be applied to all observations.) Now, for any given $p \times 1$ vector $\boldsymbol{\ell}$, we could define UV observations

$$Y_{ij} = \boldsymbol{\ell}'\mathbf{X}_{ij},$$

and then

$$Y_{11},\ldots,Y_{1n_1} \sim \text{iid from } \pi_1,$$
$$Y_{21},\ldots,Y_{2n_2} \sim \text{iid from } \pi_2.$$

Now suppose we measure the separation of the Y_{1j}'s from the Y_{2j}'s by the difference in $\bar{Y}_1$ and $\bar{Y}_2$, scaled by the pooled standard deviation of the Y-values. That is, choose $\boldsymbol{\ell}$ to maximize

$$\frac{|\bar{Y}_1-\bar{Y}_2|}{S_Y}$$

where

$$S_Y^2 = \frac{\sum_{j=1}^{n_1}(Y_{1j}-\bar{Y}_1)^2+\sum_{j=1}^{n_2}(Y_{2j}-\bar{Y}_2)^2}{n_1+n_2-2}.$$

Key fact: Maximizing the expression above is equivalent to maximizing

$$\frac{[\boldsymbol{\ell}'(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)]^2}{\boldsymbol{\ell}'\mathbf{S}_{pooled}\boldsymbol{\ell}}.$$

By the extended Cauchy-Schwartz inequality (from Chapter 2), the maximizing value of $\boldsymbol{\ell}$ is

$$\hat{\boldsymbol{\ell}} \equiv \mathbf{S}_{pooled}^{-1}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)$$

and the maximum value is

$$D^2 \equiv (\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2).$$

The function $g(\mathbf{x}) = \hat{\boldsymbol{\ell}}'\mathbf{x} = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\mathbf{S}_{pooled}^{-1}\mathbf{x}$ is called Fisher's discriminant function. Note that it is linear in $\mathbf{x}$, hence the equation obtained by setting the function equal to a constant is a line in $\mathcal{R}^2$, a plane in $\mathcal{R}^3$, etc.

Toy example:

The discriminant function can be used to construct the following classification rule. Simply classify $\mathbf{x}_0$ in π_1 if $y_0 \equiv \hat{\boldsymbol{\ell}}'\mathbf{x}_0$ is closer to $\bar{Y}_1$ than to $\bar{Y}_2$.

It turns out that the midpoint between $\bar{Y}_1$ and $\bar{Y}_2$ is equal to $\frac{1}{2}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2)$ and that numbers larger than this are on the "$\bar{Y}_1$-side" of this midpoint. Thus, the classification rule above can be expressed as follows:

Classify $\mathbf{x}_0$ in π_1 if

$$y_0 \geq \frac{1}{2}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2),$$

or equivalently if

$$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\mathbf{S}_{pooled}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2) \geq 0.$$

Observe that this rule is the same as the special case of the estimated minimum ECM rule for two normal populations with common covariance matrix, in which $c(1|2)p_2 = c(2|1)p_1$.

Now see a worked example, using SAS, on a class handout.

Error Rates for Classification Procedures

The goal of classification is to assign new observations to the populations, making as few misclassifications as possible (perhaps taking costs into account). Thus, it is desirable to have some way to quantify the performance of a classification procedure, i.e. to estimate the probability of misclassifying a new observation.

For simplicity we will assume that the misclassification costs are equal, i.e.

$$c(1|2) = c(2|1) \equiv c.$$

Then the total probability of misclassification, TPM, is equal to ECM/c and is given by

$$\begin{aligned}\text{TPM} &= P(\text{misclassify as } \pi_1) + P(\text{misclassify as } \pi_2) \\ &= P(\text{classify as } \pi_1 | \text{ belongs to } \pi_2)p_2 \\ &\quad + P(\text{classify as } \pi_2 | \text{ belongs to } \pi_1)p_1 \\ &= \end{aligned}$$

The smallest value of TPM, i.e. the value obtained by the best choice of R_1 and R_2, is called the *optimum error rate* (OER).

So for example in the MVN case with common covariance matrix, OER is equal to the expression for TPM above with

$$\begin{aligned}R_1 = \Big\{\mathbf{x} : &(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} \\ &-\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \\ &\geq \quad \log\left[\frac{p_2}{p_1}\right]\Big\}.\end{aligned}$$

In practice, of course, we generally do not know the two probability distributions' parameters so we cannot calculate OER. Instead, we must replace the minimum TPM rule (R_1, R_2) with the estimated minimum TPM rule, say $(\hat{R}_1, \hat{R}_2)$, which is calculated from the data sampled from the two populations (collectively called the *training sample*).

For example, in the MVN case with common covariance matrix,

$$\begin{aligned}\hat{R}_1 = \Big\{\mathbf{x} : &(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\mathbf{S}_{pooled}^{-1}\mathbf{x} \\ &-\frac{1}{2}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2) \\ &\geq \quad \log\left[\frac{p_2}{p_1}\right]\Big\}.\end{aligned}$$

The error rate of the rule $(\hat{R}_1, \hat{R}_2)$ is called the *actual error rate* (AER) and is given by

AER =

But AER, like OER, cannot be calculated in practice because it is still functionally dependent on the probability distributions' parameters (through $f_1(\cdot)$ and $f_2(\cdot)$). Nevertheless, we can estimate AER from the same data used to calculate $(\hat{R}_1, \hat{R}_2)$. Specifically, we estimate AER by the *apparent error rate* (APER) defined by

$$\text{APER} = \frac{\#\text{ misclassifications in the training sample}}{\#\text{ observations in the training sample}}.$$

2×2 table of outcome frequencies in training sample:

APER =

Unfortunately, APER tends to underestimate AER. (This is due to the fact that the same data used to calculate $(\hat{R}_1, \hat{R}_2)$ are used to estimate AER.)

Consequently, some other, more complicated estimates of AER have been devised. One of these is based on a cross-validation idea similar to the case-deletion diagnostic idea in regression analysis. This approach proceeds as follows:

1. Delete the first observation from the training sample.
2. Calculate the estimated minimum TPM classification rule based on the remaining $n_1 + n_2 - 1$ observations.
3. Classify the deleted observation, and compare the result to its known group membership.
4. Repeat steps 1, 2, and 3 for every observation and tabulate the number of misclassifications.

The resulting estimate of AER is called $\hat{E}(\text{AER})$ in the textbook and is given by

$\hat{E}(\text{AER}) =$

Now return to the class handout to see how to obtain APER and $\hat{E}(\text{AER})$ using SAS.

Extension of Classification to More Than Two Populations

We will consider the MVN case and estimated minimum TPM rule only.

Suppose we have g MVN populations $\pi_1, \ldots, \pi_g$ and we take a random sample from each:

$$\begin{aligned}\mathbf{X}_{11} \ldots, \mathbf{X}_{1n_1} &\sim \text{ iid } \text{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ \mathbf{X}_{21} \ldots, \mathbf{X}_{2n_2} &\sim \text{ iid } \text{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \\ &\vdots \\ \mathbf{X}_{g1} \ldots, \mathbf{X}_{gn_g} &\sim \text{ iid } \text{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\end{aligned}$$

The estimated minimum TPM rule is as follows:

(a) If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g$, then assign $\mathbf{x}$ to π_k if $\hat{d}_k(\mathbf{x}) = \max\{\hat{d}_1(\mathbf{x}), \ldots, \hat{d}_g(\mathbf{x})\}$, where

$$\hat{d}_i(\mathbf{x}) = \bar{\mathbf{X}}_i' \mathbf{S}_{pooled}^{-1} \mathbf{x} - \frac{1}{2} \bar{\mathbf{X}}_i' \mathbf{S}_{pooled}^{-1} \bar{\mathbf{X}}_i + \log p_i$$

(linear discriminant scores)

(b) If covariance matrices are unequal, then assign $\mathbf{x}$ to π_k if $\hat{d}_k^Q(\mathbf{x}) = \max\{\hat{d}_1^Q(\mathbf{x}), \ldots, \hat{d}_g^Q(\mathbf{x})\}$, where

$$\hat{d}_i^Q(\mathbf{x}) = -\frac{1}{2} \log |\mathbf{S}_i| - \frac{1}{2} (\mathbf{x} - \bar{\mathbf{X}}_i)' \mathbf{S}_i^{-1} (\mathbf{x} - \bar{\mathbf{X}}_i) + \log p_i$$

(quadratic discriminant scores)

Now see the class handout analyzing the iris data using SAS.

Cluster Analysis: Similarity and Dissimilarity Measures

Clustering is an exploratory technique for discovering relationships among objects (thus it is very individual-directed). Objects are formed into groups on the basis of how "similar" they are (in some sense). Note: the same techniques can be used to form groups of variables too, but we won't consider this.

No prior assumptions are made on how many groups there are, nor on which individuals belong to which groups.

Two important components of a cluster analysis are:

- A similarity (or dissimilarity) measure
- A procedure for grouping similar objects

We begin by considering similarity (and dissimilarity) measures. These may be functions of p-dimensional vectors of measurements taken on two objects, $\mathbf{x}_i$ and $\mathbf{x}_k$. Or, they may arise more directly, as in the following example.

In general, a similarity measure is larger for more similar objects, and a dissimilarity measure is smaller for more similar objects. For any similarity measure s, a dissimilarity measure can be easily defined as constant$-s$.

Some common indirect measures:

- Euclidean distance metric

$$d(\mathbf{x}_i, \mathbf{x}_k) =$$

- Minkowski distance metric

$$d(\mathbf{x}_i, \mathbf{x}_k) = (\sum_{j=1}^{p} |x_{ij} - x_{kj}|^m)^{1/m}$$

- Correlation coefficients (*among individuals*!) Consider

$$r_{ik}^* = \frac{\sum_{j=1}^{p} (x_{ij} - \bar{x}_i^*)(x_{kj} - \bar{x}_k^*)}{\sqrt{\sum_{j=1}^{p} (x_{ij} - \bar{x}_i^*)^2} \sqrt{\sum_{j=1}^{p} (x_{kj} - \bar{x}_k^*)^2}}$$

where

$$\bar{x}_i^* = \frac{1}{p} \sum_{j=1}^{p} x_{ij} \quad .$$

Is this a sensible measure of similarity among individuals?

Better would be the correlation coefficients based on standardized measurements, i.e.

- Euclidean distance when all variables are binary (i.e. presence or absence of a characteristic is coded as 1 or 0, respectively):

$$(x_{ij}-x_{kj})^2 = \begin{cases} 0 \text{ if } x_{ij} = x_{kj} = 0 \text{ or } x_{ij} = x_{kj} = 1 \\ 1 \text{ if } x_{ij} \neq x_{kj} \end{cases}$$

  Note that $\Sigma_{j=1}^{p}(x_{ij} - x_{kj})^2$ equals the number of "mismatches" between objects i and k, and thus it treats 0-0 matches and 1-1 matches equally. In some situations it may be more sensible to give more similarity weight to the 1-1 matches than to the 0-0 matches. See Table 12.2 in text for several possible ways to do this; for example, we could use $s =$# 1-1 matches$/p$.

Once the similarity or dissimilarity measure is determined, and computed for each pair of individuals, we can use the resulting similarity or dissimilarity matrix to form groups. This can be done informally, but less subjective approaches may be preferable.

Cluster Analysis: Hierarchical Agglomerative Clustering

- "Hierarchical" — the smallest groups can be grouped into larger groups, these larger groups can be grouped into still larger groups, etc. (like a family tree).
- "Agglomerative" — the algorithm starts with the distinct objects, then groups some together, then groups these groups together, etc. until all groups are fused into one. (Divisive clustering starts with the entire group and partitions it successively.)

General 4-step procedure (in terms of dissimilarity):

1. Obtain the dissimilarity matrix, $\mathbf{D}$, for the N clusters (at this first stage, the clusters are the individuals).
2. Scan $\mathbf{D}$ for the most similar pair of clusters; call these clusters U and V.

3. Merge clusters U and V to form a new cluster, (UV). Update $\mathbf{D}$ by deleting the rows and columns corresponding to U and V individually and by adding a row and column giving the dissimilarities between (UV) and the remaining clusters. (Several ways to define distance between clusters will be defined very shortly.)
4. Repeat steps 2 and 3 until all objects are in a single cluster. (Requires $N - 1$ repetitions.)

The results of any hierarchical clustering algorithm can be displayed in the form of a *dendrogram*, or tree diagram.

There are 3 common ways to measure distance between clusters. Here, let d_{ik} represent the distance between object i in cluster (UV) and object k in cluster W, and let $N_{(UV)}$ and N_W be the numbers of objects in these two clusters.

- Single linkage (also called minimum distance or nearest-neighbor linkage)

$$d_{(UV)W} = \min\{d_{ik} : i \in (UV), k \in W\}$$

- Complete linkage (also called maximum distance or farthest-neighbor linkage)

$$d_{(UV)W} = \max\{d_{ik} : i \in (UV), k \in W\}$$

- Average linkage

$$d_{(UV)W} = \frac{\Sigma_{i=1}^{N_{(UV)}} \Sigma_{k=1}^{N_W} d_{ik}}{N_{(UV)} N_W}$$

Toy Example:

Cluster Analysis: Wrap-Up of Hierarchical Agglomerative Methods

- There are several agglomerative hierarchical clustering procedures besides single linkage, complete linkage, and average linkage.
- It is sensible to try several of these procedures and several different similarity or dissimilarity measures. If the dendrograms corresponding to all these choices are quite consistent with one another, perhaps we can claim that the grouping is meaningful.
- Some agglomerative hierarchical clustering procedures can produce an *inversion*, which occurs when an object joins an existing cluster at a smaller dissimilarity than that of a previous consolidation. However, this cannot happen with single linkage, complete linkage, and average linkage.

- The *stability* of a clustering procedure can be checked by adding or substracting small positive numbers to the similarities or dissimilarities, and seeing how much the dendrogram changes.
- *Ties* in the similarity or dissimilarity matrix lead to multiple solutions of a hierarchical agglomerative clustering procedure. Single linkage tends to result in fewer solutions than average linkage, which in turn tends to result in fewer solutions than complete linkage. The dendrograms corresponding to multiple solutions should be compared, provided there are not so many solutions that this is infeasible.

Now see the class handout giving an example of a cluster analysis using SAS.

Cluster Analysis: Non-Hierarchical Methods

- These methods group objects into K clusters; K generally must be specified to run the algorithm, but the algorithm should usually be re-run with different values of K to determine its "best" value.
- A popular non-hierarchical clustering method is called the *K-means method*, which is carried out as follows:
  1. Partition the objects *arbitrarily* into K initial clusters.
  2. Proceed through the list of objects, reassigning each object to the cluster whose centroid (p-dimensional mean vector) is closest (in a Euclidean distance sense) to the object. Recalculate the centroid for the cluster receiving the new object and for the cluster losing the object.
  3. Repeat the previous step until no more reassignments take place.

• Toy example: