Chapter 12

An Introduction to Multivariate Methods

Contents

Multivariate statistical methods are used to display, analyze, and describe data on two or more features or variables simultaneously. I will discuss multivariate methods for measurement data. Methods for multi-dimensional count data, or mixtures of counts and measurements are available, but are beyond the scope of what I can do here. I will give a brief overview of the type of problems where multivariate methods are appropriate.

Example: Turtle shells Jolicouer and Mosimann provided data on the height, length, and width of the carapace (shell) for a sample of female painted turtles. **Cluster analysis** is used to identify which shells are similar on the three features. **Principal component analysis** is used to identify the linear combinations of the measurements that account for most of the variation in size and shape of the shells.

Cluster analysis and principal component analysis are primarily descriptive techniques.

Example: Fisher's Iris data Random samples of 50 flowers were selected from three iris species: Setosa, Virginica, and Versicolor. Four measurements were made on each flower: sepal length, sepal width, petal length, and petal width. Suppose the sample means on each feature are computed within the three species. Are the means on the four traits significantly different across species? This question can be answered using four separate one-way ANOVAs. A more powerful **MANOVA** (multivariate analysis of variance) method compares species on the four features simultaneously.

Discriminant analysis is a technique for comparing groups on multidimensional data. Discriminant analysis can be used with Fisher's Iris data to find the linear combinations of the flower features that best distinguish species. The linear combinations are optimally selected, so insignificant differences on one or all features may be significant (or better yet, important) when the features are considered simultaneously! Furthermore, the discriminant analysis could be used to classify flowers into one of these three species when their species is unknown.

MANOVA, discriminant analysis, and **classification** are primarily inferential techniques.

12.1 Linear Combinations

Suppose data are collected on p measurements or features $X_1, X_2, \ldots, X_p$. Most multivariate methods use **linear combinations** of the features as the basis for analysis. A linear combination has the form

$$Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p,$$

where the coefficients $a_1, a_2, \ldots, a_p$ are known constants. Y is evaluated for each observation in the data set, keeping the coefficients constant.

For example, three linear combinations of $X_1, X_2, \ldots, X_p$ are:

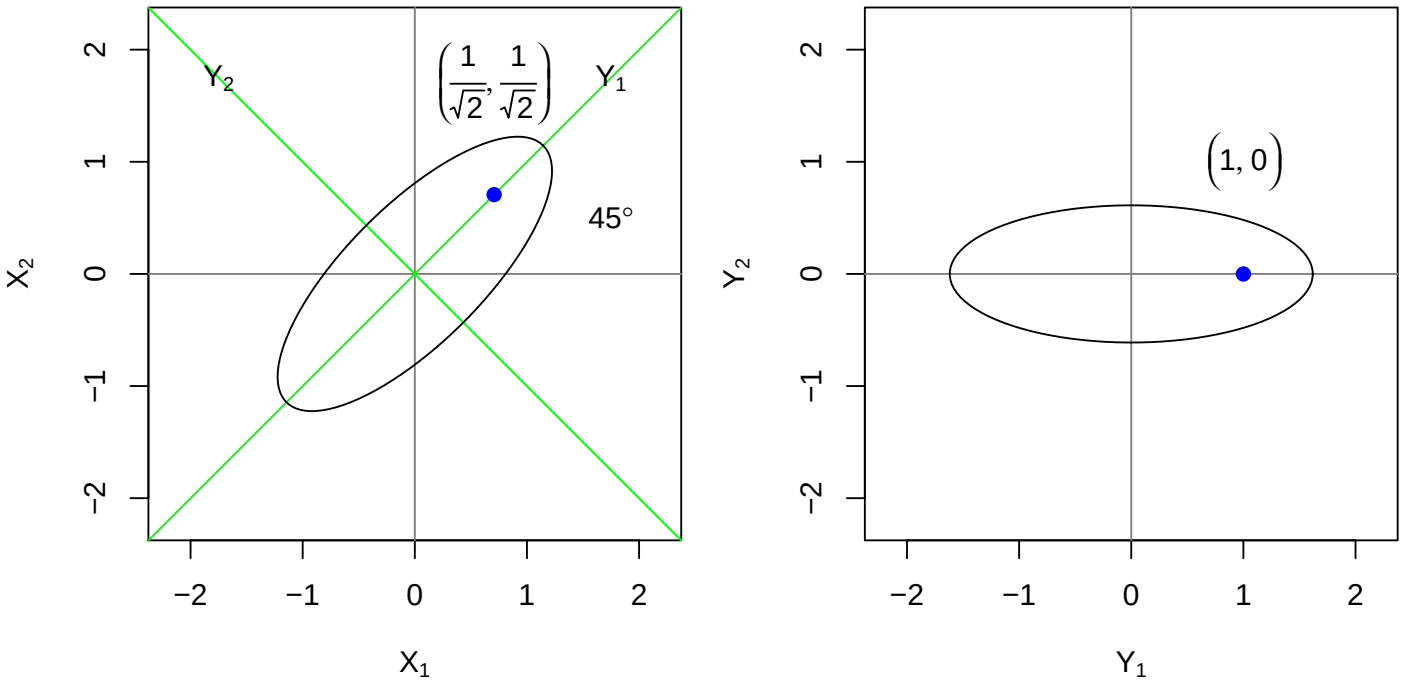
$$\begin{aligned} Y &= 1X_1 + 0X_2 + 0X_3 + \cdots + 0X_p = X_1, \\ Y &= \frac{1}{p}(X_1 + X_2 + \cdots + X_p), \text{ and} \\ Y &= 2X_1 - 4X_2 + 55X_3 - 1954X_4 + \cdots + 44X_p. \end{aligned}$$

Vector and **matrix** notation are useful for representing and summarizing multivariate data. Before introducing this notation, let us try to understand linear combinations geometrically when $p = 2$.

Example: $-45°$ **rotation** A plot of data on two features X_1 and X_2 is given below. Also included is a plot for the two linear combinations

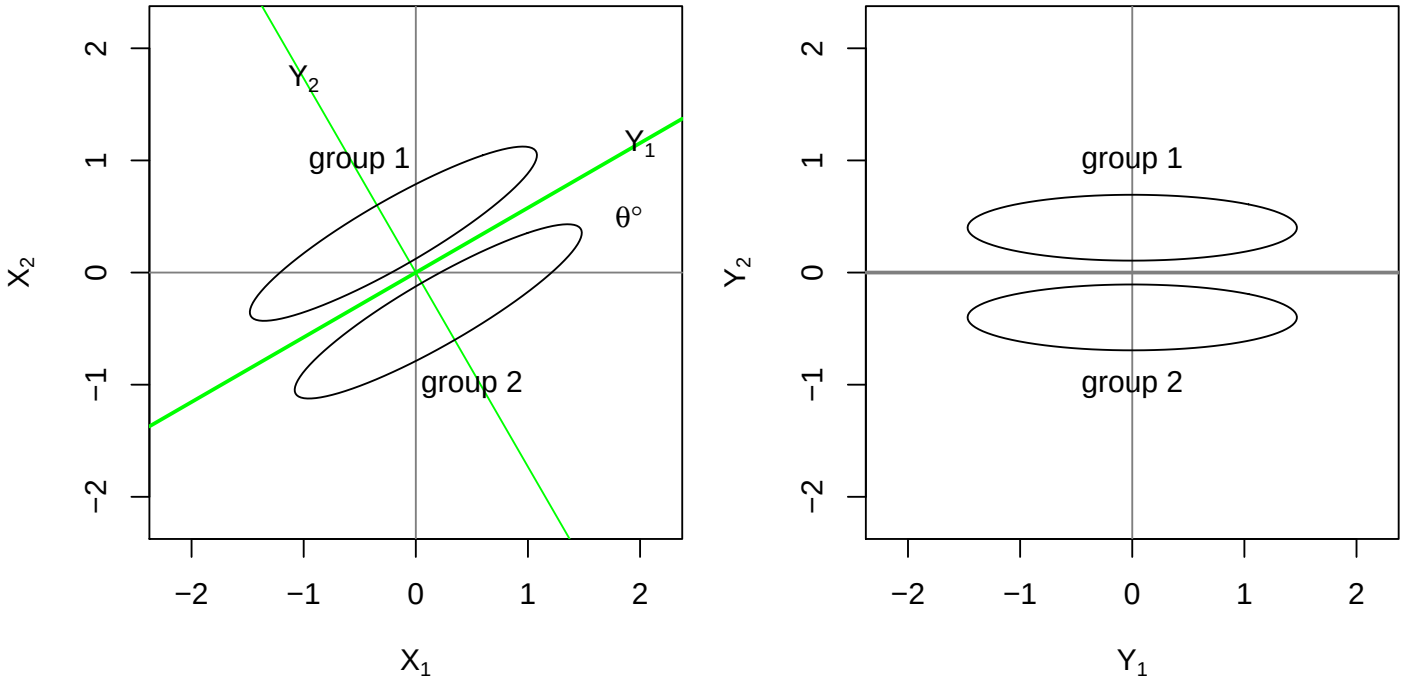
$$Y_1 = \frac{1}{\sqrt{2}}(X_1 + X_2) \qquad \text{and}$$
$$Y_2 = \frac{1}{\sqrt{2}}(X_2 - X_1).$$

This transformation creates two (roughly) uncorrelated linear combinations Y_1 and Y_2 from two highly correlated features X_1 and X_2. The transformation corresponds to a **rotation** of the original coordinate axes by -45 degrees. Each data point is then expressed relative to the new axes. The new features are uncorrelated!



The $\sqrt{2}$ divisor in Y_1 and Y_2 does not alter the interpretation of these linear combinations: Y_1 is essentially the sum of X_1 and X_2, whereas Y_2 is essentially the difference between X_2 and X_1.

Example: Two groups The plot below shows data on two features X_1 and X_2 from two distinct groups.



If you compare the groups on X_1 and X_2 separately, you may find no significant differences because the groups overlap substantially on each feature. The plot on the right was obtained by rotating the coordinate axes $-\theta$ degrees, and then plotting the data relative to the new coordinate axes. The rotation corresponds to creating two linear combinations:

$$\begin{aligned} Y_1 &= \cos(\theta)X_1 + \sin(\theta)X_2 \\ Y_2 &= -\sin(\theta)X_1 + \cos(\theta)X_2. \end{aligned}$$

The two groups differ substantially on Y_2. This linear combination is used with discriminant analysis and MANOVA to distinguish between the groups.

The linear combinations used in certain multivariate methods do not correspond to a rotation of the original coordinate axes. However, the pictures given above should provide some insight into the motivation for the creating linear combinations of two features. The ideas extend to three or more features, but are more difficult to represent visually.

12.2 Vector and Matrix Notation

A **vector** is a string of numbers or variables that is stored in either a row or in a column. For example, the collection $X_1, X_2, \ldots, X_p$ of features can be represented as a **column-vector** with p rows, using the notation

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}.$$

The entry in the j^{th} row is X_j. The transpose of X, represented by X', is a **row-vector** with p columns: $X' = (X_1, X_2, \ldots, X_p)$. The j^{th} column of X' contains X_j.

Suppose you collect data on p features $X_1, X_2, \ldots, X_p$ for a sample of n individuals. The data for the i^{th} individual can be represented as the column-vector:

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{bmatrix}.$$

or as the row-vector $X_i' = (X_{i1}, X_{i2}, \cdots, X_{ip})$. Here X_{ij} is the value on the j^{th} variable. Two subscripts are needed for the data values. One subscript identifies the individual and the other subscript identifies the feature.

A **matrix** is a rectangular array of numbers or variables. A data set can be viewed as a matrix with n rows and p columns, where n is the sample size. Each row contains data for a given individual:

$$\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}.$$

Vector and matrix notation are used for summarizing multivariate data. For example, the **sample mean vector** is

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix},$$

where $\bar{X}_j$ is the sample average on the j^{th} feature. Using matrix algebra, $\bar{X}$ is defined using a familiar formula:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This mathematical operation is well-defined because vectors are added elementwise.

The sample variances and covariances on the p variables can be grouped together in a $p \times p$ **sample variance-covariance matrix** S (i.e., p rows and p columns)

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix},$$

where

$$s_{ii} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ki} - \bar{X}_i)^2$$

is the sample variance for the i^{th} feature, and

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)$$

is the sample covariance between the i^{th} and j^{th} features. The subscripts on the elements in S identify where the element is found in the matrix: s_{ij} is

stored in the i^{th} row and the j^{th} column. The variances are found on the **main diagonal** of the matrix. The covariances are **off-diagonal** elements. S is symmetric, meaning that the elements above the main diagonal are a reflection of the entries below the main diagonal. More formally, $s_{ij} = s_{ji}$.

Matrix algebra allows you to express S using a formula analogous to the sample variance for a single feature:

$$S = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})'.$$

Here $(X_k - \bar{X})(X_k - \bar{X})'$ is the matrix product of a column vector with p entries times a row vector with p entries. This matrix product is a $p \times p$ matrix with $(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)$ in the i^{th} row and j^{th} column. The matrix products are added up over all n observations and then divided by $n - 1$.

The interpretation of covariances is enhanced by standardizing them to give correlations. The **sample correlation matrix** is denoted by the $p \times p$ symmetric matrix

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}.$$

The i^{th} row and j^{th} column element of R is the correlation between the i^{th} and j^{th} features. The diagonal elements are one: $r_{ii} = 1$. The off-diagonal elements satisfy

$$r_{ij} = r_{ji} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}.$$

In many applications the data are standardized to have mean 0 and variance 1 on each feature. The data are standardized through the so-called **Z-score transformation**: $(X_{ki} - \bar{X}_i)/s_{ii}$ which, on each feature, subtracts the mean from each observation and divides by the corresponding standard deviation. The sample variance-covariance matrix for the standardized data is the correlation matrix R for the raw data.

Example: Let X_1, X_2, and X_3 be the reaction times for three visual stimuli named A, B and C, respectively. Suppose you are given the following summaries based on a sample of 30 individuals:

$$\bar{X} = \begin{bmatrix} 4 \\ 5 \\ 4.7 \end{bmatrix},$$

$$S = \begin{bmatrix} 2.26 & 2.18 & 1.63 \\ 2.18 & 2.66 & 1.82 \\ 1.63 & 1.82 & 2.47 \end{bmatrix},$$

$$R = \begin{bmatrix} 1.00 & 0.89 & 0.69 \\ 0.89 & 1.00 & 0.71 \\ 0.69 & 0.71 & 1.00 \end{bmatrix}.$$

The average response time on B is 5. The sample variance of response times on A is 2.26. The sample covariance between response times on A and C is 1.63. The sample correlation between response times on B and C is 0.71.

12.3 Matrix Notation to Summarize Linear Combinations

Matrix algebra is useful for computing sample summaries for linear combinations of the features $X' = (X_1, X_2, \ldots, X_p)$ from the sample summaries on these features. For example, suppose you define the linear combination

$$Y_1 = a_1X_1 + a_2X_2 + \cdots + a_pX_p.$$

Using matrix algebra, Y_1 is the matrix product $Y_1 = a'X$, where $a' = (a_1, a_2, \ldots, a_p)$. The sample mean and variance of Y_1 are

$$\bar{Y} = a_1\bar{X}_1 + a_2\bar{X}_2 + \cdots + a_p\bar{X}_p = a'\bar{X}$$

and

$$s_Y^2 = \sum_{ij} a_i a_j s_{ij} = a'Sa,$$

where $\bar{X}$ and S are the sample mean vector and sample variance-covariance matrix for $X' = (X_1, X_2, \ldots, X_p)$.

Similarly, the sample covariance between Y_1 and

$$Y_2 = b'X = b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

is

$$s_{Y_1,Y_2} = \sum_{ij} a_i b_j s_{ij} = a'Sb = b'Sa.$$

Example: In the stimuli example, the total reaction time per individual is

$$Y \;=\; [1\ 1\ 1]\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \;=\; X_1 + X_2 + X_3.$$

The mean reaction time is

$$\begin{aligned} \bar{Y} \;&=\; [1\ 1\ 1] \\ \bar{X} \;&=\; [1\ 1\ 1]\begin{bmatrix} 4 \\ 5 \\ 4.7 \end{bmatrix} \;=\; 4+5+4.7 \;=\; 13.7. \end{aligned}$$

The variance of Y is the sum of the elements in the variance-covariance matrix:

$$s^2_Y \;=\; [1\ 1\ 1]\, S \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \;=\; \sum_{ij} s_{ij} \;=\; 2.26 + 2.18 + \cdots + 2.47 \;=\; 18.65.$$