Stat 428/528: Advanced Data Analysis 2

Chapters 12: An Introduction to Multivariate Methods

April 17, 2019

THE UNIVERSITY OF NEW MEXICO

Multivariate Methods

Multivariate statistical methods are used to display, analyze, and describe data on two or more features or variables simultaneously.

- Methods for measurement data (we will discuss in this class)
- Methods for multi-dimensional count data
- mixtures of counts and measurements

Example: Turtle shells Data on the height, length, and width of the carapace (shell) for a sample of female painted turtles.
——measurements: height, length and width

- 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.
- Both cluster analysis and principal component analysis are primarily descriptive techniques.

Example: Fishers Iris data

consider three iris species: Setosa, Virginica, and Versicolor

random samples of 50 flowers were selected from each of three iris species

four measurements were made on each flower: sepal length, sepal width, petal length, and petal width.

- MANOVA (multivariate analysis of variance): suppose the sample means on each measurements (sepal length, sepal width, petal length, and petal width) 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 com- pares species on the four features simultaneously.

- Discriminant analysis is a technique for comparing groups on multidimensional data.

  —— Discriminant analysis can be used with Fishers 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.

Linear Combinations

Suppose data are collected on p measurements or features $X_1, X_2, \cdots, X_p$.

- Most multivariate methods use linear combinations of the features as the basis for analysis.
- A linear combination has the form

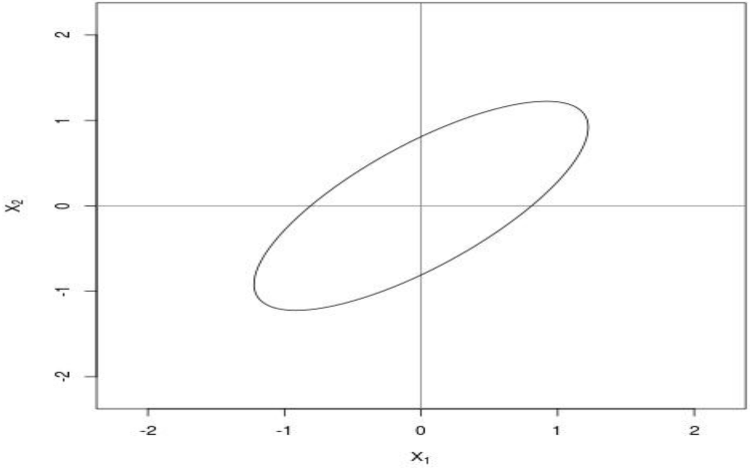
$$Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p,$$

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

Example 1: -45° rotation

A plot of data on two features X_1 and X_2 is given below.

Figure: Plot of data two features X_1 and X_2



Now perform transformations on X_1 and X_2.

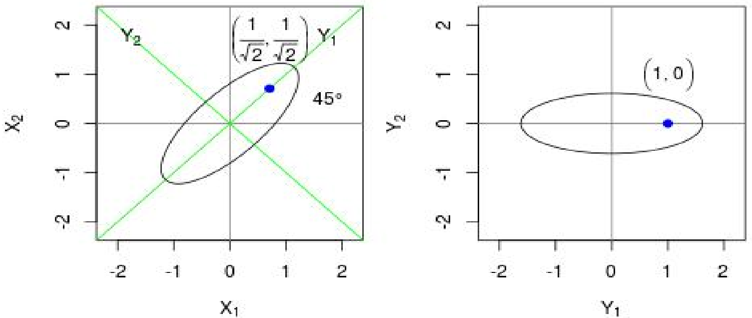
$$Y_1 = \frac{1}{\sqrt{2}}(X_1 + X_2)$$

and

$$Y_2 = \frac{1}{\sqrt{2}}(X_2 - X_1)$$

Plots of data on two features X_1 and X_2 and the transformed features Y_1 and Y_2 are given below.

Figure: Plot of data two features X_1 and X_2 together with Y_1 and Y_2

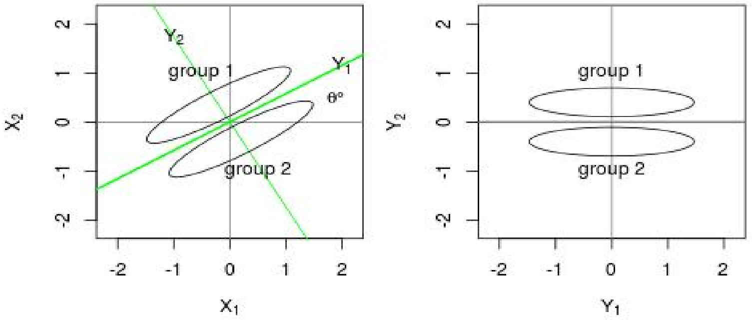


This transformation creates two (roughly) uncorrelated linear combinations Y_1 and Y_2 from two highly correlated features X_1 and X_2 .

- 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!

Example 2: two groups

Figure: Plot of 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 θ degrees, and then plotting the data relative to the new coordinate axes.
- The rotation corresponds to creating two linear combinations:

$$Y_1 = cos(\theta)X_1 + sin(\theta)X_2$$

$$Y_2 = -sin(\theta)X_1 + cos(\theta)X_2$$

The two groups differ substantially on Y_2.

Comments:

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.
- the pictures given above 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.

Vector and Matrix Notation

A vector is a string of numbers or variables that is stored in either a row or in a column.
The collection $X_1, X_2, \cdots, X_p$ of features can be represented as

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

$$\mathbf{X}' = (X_1, X_2, \cdots, X_p)$$

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

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$$

$$\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{ip})$$

Here x_{ij} is the value on the jth variable. Two subscripts are needed for the data values. One subscript identifies the individual and the other subscript identifies the feature.

A data set can be viewed as a matrix with n rows and p columns, where n is the sample size, and p is the number of features. Each row contains data for a given individual:

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

sample mean vector is

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

where $\bar{x}_j$ is the sample average on the j th feature

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

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

where sample variance s_{ii} and sample covariance s_{ij} are defined as follows

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

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

The sample correlation matrix is defined as

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

where sample correlation $r_{ii} = 1$ and

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

Matrix Notation to Summarize Linear Combinations

Let $\mathbf{X}' = (X_1, X_2, \cdots, X_p)$, $\mathbf{a}' = (a_1, a_2, \cdots, a_p)$ and

$$Y_1 = a_1 X_1 + a_2 X_2 + \cdots a_p X_p = \mathbf{a}'\mathbf{X}$$

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

and $s_{Y_1}^2 = \sum_{ij} a_i a_j s_{ij} = \mathbf{a}'\mathbf{S}\mathbf{a}$ where $\bar{\mathbf{X}}$ and $\mathbf{S}$ are the sample mean vector and sample variance-covariance matrix for $\mathbf{X}' = (X_1, X_2, \cdots, X_p)$.

Example: Data collected on features X_1, X_2 and X_3 has

$$\bar{\mathbf{x}} = \begin{bmatrix} 4 \\ 5 \\ 4.7 \end{bmatrix} \text{ and } \mathbf{S} = \begin{bmatrix} 2.26 & 2.18 & 1.63 \\ 2.18 & 2.66 & 1.82 \\ 1.63 & 1.82 & 2.47 \end{bmatrix}$$

Let $Y = (1,1,1)\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = X_1 + X_2 + X_3$

$$\bar{Y} = (1,1,1)\begin{bmatrix} 4 \\ 5 \\ 4.7 \end{bmatrix} = 13.7$$

$$s_Y^2 = (1,1,1)\mathbf{S}(1,1,1)' = \sum_{ij} s_{ij} = 18.65$$