다변량통계방법론

2019년 2학기

고려대학교 통계학과 대학원

Ch 1. Aspects of Multivariate Analysis

- Multivariate analysis: statistical method for data with simultaneous measurements on many variables

  - There are $p > 1$ variables.

  - The values of these variables are all recorded for each distinct item, individual, or experimental unit.

- Examples of multivariate analysis

  - Data reduction and simplification

  - Sorting and grouping

  - Investigation of the dependence among variables

  - Prediction

  - Hypothesis testing

1.3 The Organization of Data

- Arrays

  - x_{jk} : measurement of the kth variable on the jth item

  - n measurements of p variables can be displayed as follows:

	Variable 1	Variable 2	$\cdots$	Variable k	$\cdots$	Variable p
Item 1:	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
Item 2:	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item j:	x_{j1}	x_{j2}	$\cdots$	x_{jk}	$\cdots$	x_{jp}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item n:	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

  - These data can be displayed as a rectangular array X of n rows and p columns:

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

1.3 The Organization of Data

- Descriptive statistics

  - Sample mean: $$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}, \quad k = 1,2,\ldots,p$$

  - Sample variance: $$s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{jk} - \bar{x}_k\right)^2, \quad k = 1,2,\ldots,p$$

  - Sample covariance: $$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right), \quad i = 1,2,\ldots,p,\ k = 1,2,\ldots,p$$

  - Sample correlation coefficient (or Pearson's product-moment correlation coefficient):

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right)}{\sqrt{\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)^2}\sqrt{\sum_{j=1}^{n}\left(x_{jk} - \bar{x}_k\right)^2}}, \quad i = 1,\ldots,p, k = 1,\ldots,p$$

  - Sum of squares of the deviations from the mean:

$$w_{kk} = \sum_{j=1}^{n}\left(x_{jk} - \bar{x}_k\right)^2, \quad k = 1,2,\ldots,p$$

  - Sum of cross-product deviations:

$$w_{ik} = \sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right), \quad i = 1,2,\ldots,p,\ k = 1,2,\ldots,p$$

1.3 The Organization of Data

- Descriptive statistics (continued)

  The descriptive statistics computed from n measurements on p variables can be organized into arrays:

  - Sample means:

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

  - Sample variances and covariances:

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

  - Sample correlations:

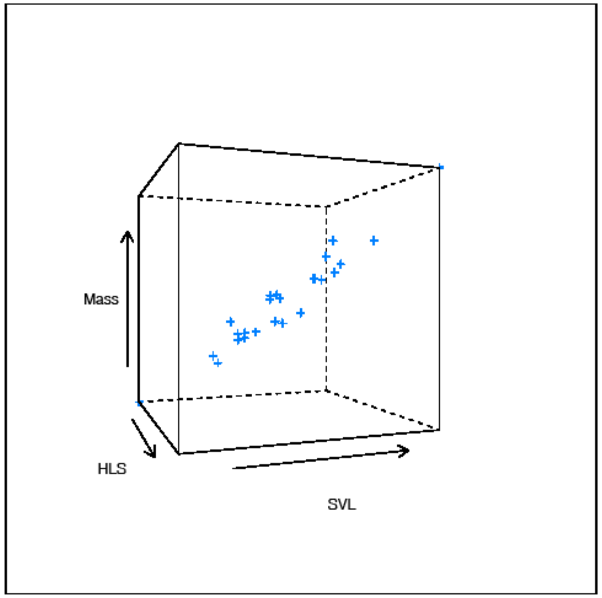
$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

1.3 The Organization of Data

- n points in p dimensions (p-dimensional scatterplot)
  - The p measurements ($x_{j1}, x_{j2}, \ldots, x_{jp}$) on the jth item represent the coordinates of a point in p-dimensional space.
  - The coordinate axes are taken to correspond to the variables, so that the jth point is x_{j1} units along the first axis, x_{j2} units along the second, …, x_{jp} units along the pth axis.
  - The resulting plot with n points not only exhibits the overall pattern of variability, but also will show similarities among the n items.

- 3D scatter plot
  - Example 1.6. Measurements on 25 lizards

  Variables: Mass
  Snout vent length (SVL)
  hind limb span (HLS)



1.3 The Organization of Data

- p points in n dimensions
  - The n observations of the p variables can be regarded as p points in n-dimensional space.
  - Each column of X determines one of the points. That is, the ith column,

$$\begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{bmatrix},$$

  consisting of all n measurements on the ith variable, determines the ith point.
  - Closeness of points in n dimensions can be related to measures of association between the corresponding variables.

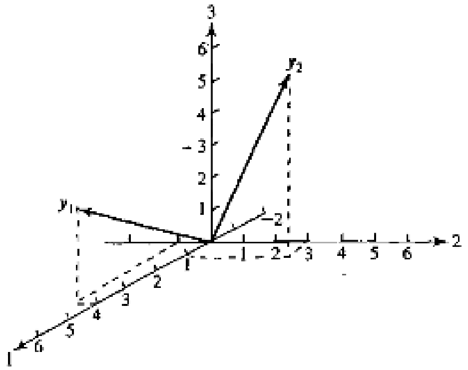


Figure 3.2 A plot of the data matrix $\mathbf{X}$ as $p = 2$ vectors in $n = 3$ space.

1.5 Distance

- Most multivariate techniques are based upon the simple concept of distance.
- Euclidean distance

 - Consider the straight-line distance from point $P = (x_1, x_2, \ldots x_p)$ to the origin $O = (0,0,\ldots 0)$.
 - The Euclidean distance is defined as $d(O,P) = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$.
 - Note that all points that lie a constant squared distance, such as c^2, from the origin satisfy the equation

$$d^2(O,P) = x_1^2 + x_2^2 + \cdots + x_p^2 = c^2.$$

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

$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}.$$

 - Euclidean distance is unsatisfactory for most statistical purposes, because each coordinate contributes equally to the calculations of Euclidean distance.
 - It is often desirable to weight coordinates subject to a great deal of variability less heavier than those that are not highly variable.

1.5 Distance

- Statistical distance (when the variables are not correlated)

  - Consider the point $P = (x_1, x_2, \ldots x_p)$ to the origin $O = (0,0,\ldots 0)$.

  - The statistical distance is defined as

$$d(O,P) = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} + \cdots + \frac{x_p^2}{s_{pp}}}.$$

  - Note that all points that lie a constant squared distance, such as c^2, from the origin satisfy the equation

$$d^2(O,P) = \frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} + \cdots + \frac{x_p^2}{s_{pp}} = c^2.$$

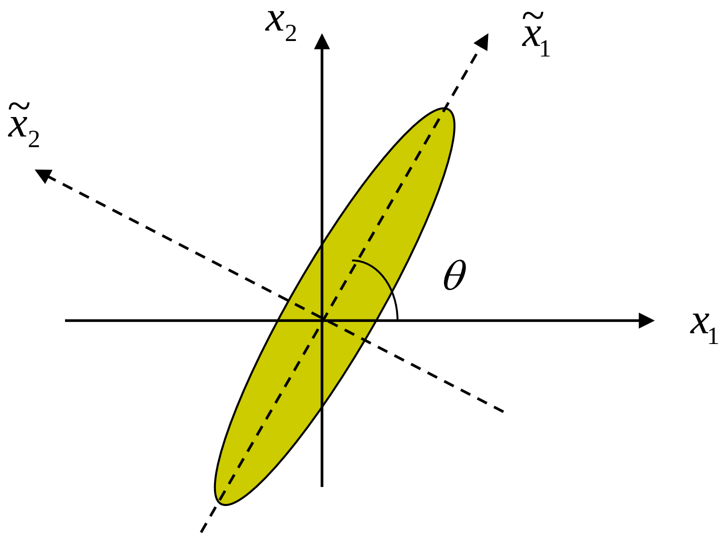
  - The statistical distance between two arbitrary points P and Q with coordinates $P = (x_1, x_2, \ldots x_p)$ and $Q = (y_1, y_2, \ldots y_p)$ is

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + \cdots + \frac{(x_p - y_p)^2}{s_{pp}}}.$$

  - All points P that are a constant squared distance from Q lie on a hyperellipsoid centered at Q whose major and minor axes are parallel to the coordinate axes.

1.5 Distance

- Statistical distance (when the variables are correlated)



- Rotate the original coordinate system through the angle θ while keeping the scatter fixed and label the rotated axes $\tilde{x}_1$ and $\tilde{x}_2$.

- This suggests that we calculate the sample variances using the $\tilde{x}_1$ and $\tilde{x}_2$ coordinates and measure distance as

$$d(O,P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}}.$$

1.5 Distance

- Statistical distance (when the variables are correlated) (continued)

  - The relation between the original coordinates (x_1, x_2) and the rotated coordinates $(\tilde{x}_1, \tilde{x}_2)$ is provided by

$$\tilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta),$$
$$\tilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta).$$

  - After some algebraic manipulations, the distance from $P = (\tilde{x}_1, \tilde{x}_2)$ to the origin O = (0,0) can be written in terms of the original coordinates (x_1, x_2) as

$$d(O,P) = \sqrt{a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2}.$$

  - Note that these distances are completely determined by the coefficients (weights) a_{ij}, i=1,2, j=1,2, shown as the array

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$$

  so that a constant squared distance, such as c^2, from the origin satisfy

$$d^2(O,P) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underline{x'Ax = c^2}.$$

1.5 Distance

- Statistical distance (when the variables are correlated) (continued)

  - The statistical distance between the point $P = (x_1, x_2, \dots x_p)$ to the origin $O = (0,0,\dots 0)$ is expressed by

$$d(O,P) = \sqrt{a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{pp}x_p^2 + 2a_{12}x_1x_2 + +2a_{13}x_1x_3 + \cdots + 2a_{p-1,p}x_{p-1}x_p}.$$

  - The statistical distance between two arbitrary points P and Q with coordinates $P = (x_1, x_2, \dots x_p)$ and $Q = (y_1, y_2, \dots y_p)$ is expressed by

$$d(P,Q) = \sqrt{\begin{aligned} &a_{11}(x_1 - y_1)^2 + a_{22}(x_2 - y_2)^2 + \cdots + a_{pp}(x_p - y_p)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) \\ &+ 2a_{13}(x_1 - y_1)(x_3 - y_3) + \cdots + 2a_{p-1,p}(x_{p-1} - y_{p-1})(x_p - y_p) \end{aligned}},$$

  where the a's are numbers such that the distances are always nonnegative.

  - Note that these distances are completely determined by the coefficients (weights) a_{ik}, i=1,2,…,p, k=1,2,…,p, shown as a rectangular array

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{12} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp} \end{bmatrix}$$ so that $d^2(O,P) = x'Ax.$

1.5 Distance

- The entries in the array specify the distance functions.

  - The a_{ik}'s cannot be arbitrary numbers; they must be such that the computed distance is nonnegative for every pair of points.

- Other measures of distance is also possible. Any distance measure $d(P,Q)$ between two points P and Q is valid provided that it satisfies the following properties, where R is any other intermediate point:

  - $d(P,Q) = d(Q,P)$;
  - $d(P,Q) > 0$ if $P \neq Q$;
  - $d(P,Q) = 0$ if $P = Q$;
  - $d(P,Q) \leq d(P,R) + d(R,Q)$ (triangle inequality).