다변량통계방법론

2021년 2학기

고려대학교 통계학과 대학원

Ch 1. Aspects of Multivariate Analysis

- Multivariate analysis: statistical analysis for data with simultaneous measurements on $p > 1$ variables.

  관심있는 변수가 여러개

- Examples of multivariate analysis

  - Data reduction and simplification PCA . FA ... linear comb. non-linear comb.
  - Sorting and grouping observation들에 대해, cluster analysis등. or 분류과정
  - Investigation of the dependence among variables 변수사이의관계.
  - Prediction
  - Hypothesis testing ANOVA multivariate

1.3 The Organization of Data

- Arrays

  - x_{jk} : measurement of the kth variable on the jth item
  - Data with n measurements of p variables

obs

	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}

  - Data is expressed as a matrix X with n rows and p columns:

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

p개 평균, 분산

1.3 The Organization of Data

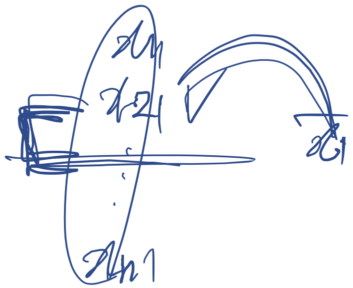
- Descriptive statistics 요약 통계량.
  - Sample mean: $\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}, \quad k = 1,2,\ldots,p$
  - Sample variance: $s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{jk} - \bar{x}_k\right)^2, \quad k = 1,2,\ldots,p$
  - Sample covariance: $s_{ik} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right), \quad i = 1,2,\ldots,p,\ k = 1,2,\ldots,p$ $\frac{p(p-1)}{2}$ 개 공분산
  - 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:

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

각각의 변동치



  - 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 from n observations on p variables are summarized as

  - Sample means: $\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$ $p \times 1$

  - Sample variances and covariances: $S_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$ symm.

    Sample covariance!

    mle 할때 n이 쓰임.

    $\boxed{S : \text{불편}}$

  - Sample correlations: $R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$ symm.

1.3 The Organization of Data

변수의 분포 읽는편이지만 그리기 어려움.

- *p*-dimensional scatterplot (*n* points in *p* dimensions)
  - The *p* measurements (x_{j1}, x_{j2}, …, x_{jp}) on the *j*th item represent the coordinates of a point in *p*-dimensional space.
  - 하나하나의 축
  - The coordinate axes correspond to the variables.

data 더 많으니까 훨씬 그리기 어려워..

- *n*-dimensional scatterplot (*p* points in *n* dimensions)
  - The *n* observations of the *p* variables can be regarded as *p* points in *n*-dimensional space.
  - The coordinate axes correspond to the observations.

그러니 작은편이지. 하지만 변수간 연관성 보기엔 좋다.

1.5 Distance

- Euclidean distance
  - The straight-line distance from point $P = (x_1, x_2, \ldots x_p)$ to the origin $O = (0,0,\ldots 0)$

$$d(O,P) = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$$

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

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

통계계산에 있어서
분산이 큰 변수에 영향을 많이 받다.

표준화

1.5 Distance

표준화 마친 거리 euclidean

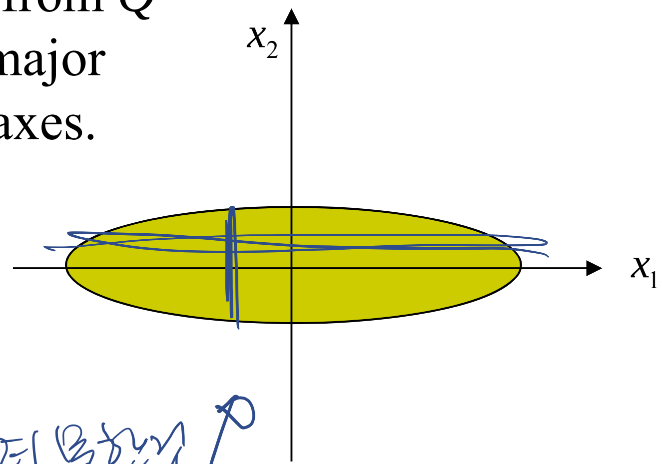
- Statistical distance (when the variables are not correlated)
  - Statistical distance for the point $P = (x_1, x_2, \ldots x_p)$ to the origin $O = (0,0,\ldots0)$

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

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

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

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



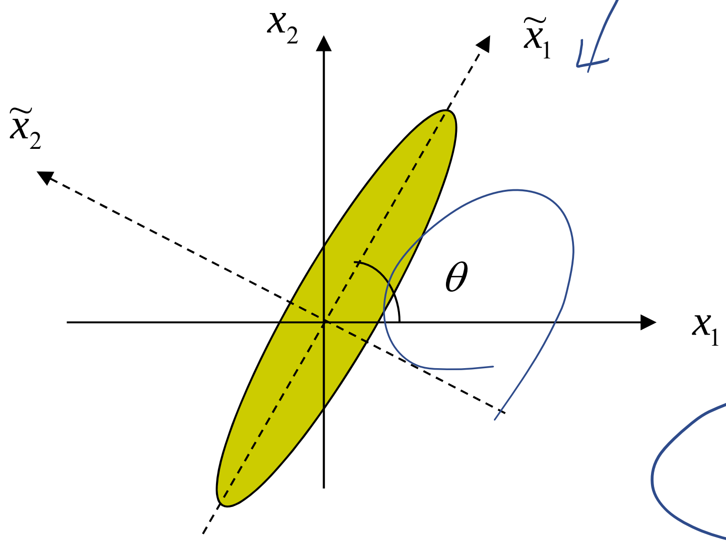
hyper-타원인거로 볼수있다

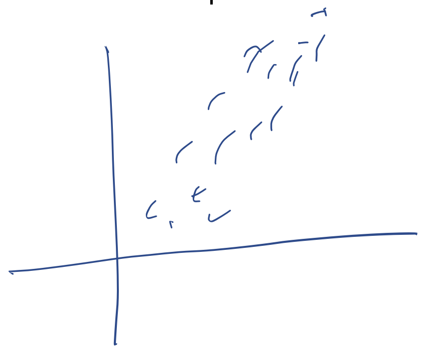
하지만 corr 고려 못한것 ρ

1.5 Distance

corr 있다(x)

- Statistical distance (when the variables are correlated)





rotate

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

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

rotate하니 corr가 없어짐

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

1.5 Distance

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

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

    $=\sqrt{a_{11}x_1^2+a_{22}x_2^2+2a_{12}x_1x_2}.$

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

  - This is called the Mahalanobis distance.

Mahalanobis distance !!

1.5 Distance

변수 여러개로 확장

cross product ?

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

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

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

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

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

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

  - 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

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

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:

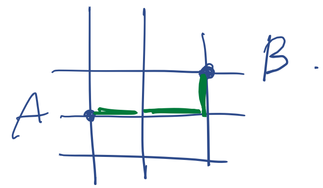
  - $d(P,Q) = d(Q,P)$;
  - $d(P,Q) > 0$ if $P \neq Q$;
  - $d(P,Q) = 0$ if $P = Q$;
  - $d(P,Q) \leq d(P,R) + d(R,Q)$ (triangle inequality).

바람직한 distance measure 조건.

1.5 Distance

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

$$d(P,Q) = \left[\sum_{i=1}^{p} |x_i - y_i|^m \right]^{\frac{1}{m}}$$



  - For $m = 1$, it is called the "city-block" distance.
  - For $m = 2$, it is the Euclidean distance.

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

$$d(P,Q) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}$$

1.5 Distance

범주형 변수, 2가지면

- For p binary variables and two observations P and Q with values $P = (x_1, x_2, \ldots x_p)$ and $Q = (y_1, y_2, \ldots y_p)$,

Euclidean

$$(x_i - y_i)^2 = \begin{cases} 0 & \text{if } x_i = y_i. \\ 1 & \text{if } x_i \neq y_i. \end{cases}$$

- $d(P,Q) = \sum_{i=1}^{p} (x_i - y_i)^2$

- This Euclidean distance measures the number of discordance.

다른 개수

1.5 Distance

- Gower distance between two arbitrary points P and Q with values $P = (x_1, x_2, \ldots x_p)$ and $Q = (y_1, y_2, \ldots y_p)$

  - For categorical variables,

$$d_i = \begin{cases} 0 & \text{if } x_i = y_i. \\ 1 & \text{if } x_i \neq y_i. \end{cases}$$

  - For numeric variables,

$$d_i = \frac{|x_i - y_i|}{R_i},$$

  where R_i is the range of the ith variable.

  - $$d(P,Q) = \frac{\sum_{i=1}^{p} \delta_i d_i w_i}{\sum_{i=1}^{p} \delta_i w_i}$$

// 연구과제 //

유명 distance 中

2개 골라서.

4가지 만족해

안하나 확인하기

안하기

Eucli X

Mahal X

~화 (14)

목(16) 발표준비