Review of Basic Concepts in Multivairate Data

1. Representations of Data Matrix

$n \times p$ *Data Matrix* :

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

$$= (x_{ij}),\ \ i = 1,\ldots,n,\ j = 1,\ldots,p$$

1.1 Rows representation

- $X = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_i^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$ *with* $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ip})^T \in \mathbb{Q}^p,\ \ i = 1,\ldots,n$

1.2 Columns representation

- $X = [\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(j)},\ldots,\boldsymbol{x}_{(p)}]$ *with* $\boldsymbol{x}_{(j)} = (x_{1j},\ldots,x_{nj})^T \in \mathbb{R}^n,\ \ j = 1,\ldots,p$
- $\boldsymbol{x}_{(j)} = (x_{1j},\ldots,x_{nj})^T \in \mathbb{R}^n \Leftrightarrow$ *data points in observational space*

Review of Basic Concepts in Multivairate Data

2. Scaling

- *Data point* : $\boldsymbol{x}_{(j)} = (x_{1j}, \dots, x_{nj})^T \in \mathbb{R}^n$
- *Mean* : $\bar{x}_j = \dfrac{x_{1j} + \cdots + x_{nj}}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_{ij}$
- *Variance* : $s_j^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$

2.1 Centring

- $x_{ij} - \bar{x}_j = y_{ij}$: *centring*

$\Rightarrow \quad Y = (y_{ij}), \quad i = 1, \dots, n \ ; \ j = 1, \dots, p.$: *centered data matrix*

2.2 Standardizing

- $(x_{ij} - \bar{x}_j)/s_j = z_{ij}$: *standardizing*

$\Rightarrow \quad Z = (z_{ij}), \quad i = 1, \dots, n \ ; \ j = 1, \dots, p.$: *standardized data matrix*

Review of Basic Concepts in Multivairate Data

3. Algebraic Calculations of Y and Z from X

3.1 Centring

- $Y = HX$ with $H = I - J/n$: $n \times n$ centering matrix

3.2 Standardizing

- $Z = HXD_s^{-1/2}$ with $D_s^{1/2} = diag(\sqrt{s_{11}}, \cdots, \sqrt{s_{pp}})$: standard deviation matrix

Review of Basic Concepts in Multivairate Data

4. Summary Multivariate Statistics

- $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{Q}^p, \ i = 1, \ldots, n$

4.1 Mean Vector

- $\overline{\boldsymbol{x}} = (\overline{x}_1, \ldots, \overline{x}_j, \ldots, \overline{x}_p)^T = \dfrac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \dfrac{X^T 1_n}{n}$

4.2 Covariance Matrix S from Y

- $S = \dfrac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = \dfrac{1}{n-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T - n\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T\right)$

  $\Leftrightarrow \ S = \dfrac{1}{n-1} Y^T Y = \dfrac{1}{n-1} X^T H X$

4.3 Correlation Matrix R from Z

- $R = \dfrac{1}{n-1} Z^T Z = D_s^{-1/2} S \, D_s^{-1/2}$

Review of Basic Concepts in Multivairate Data

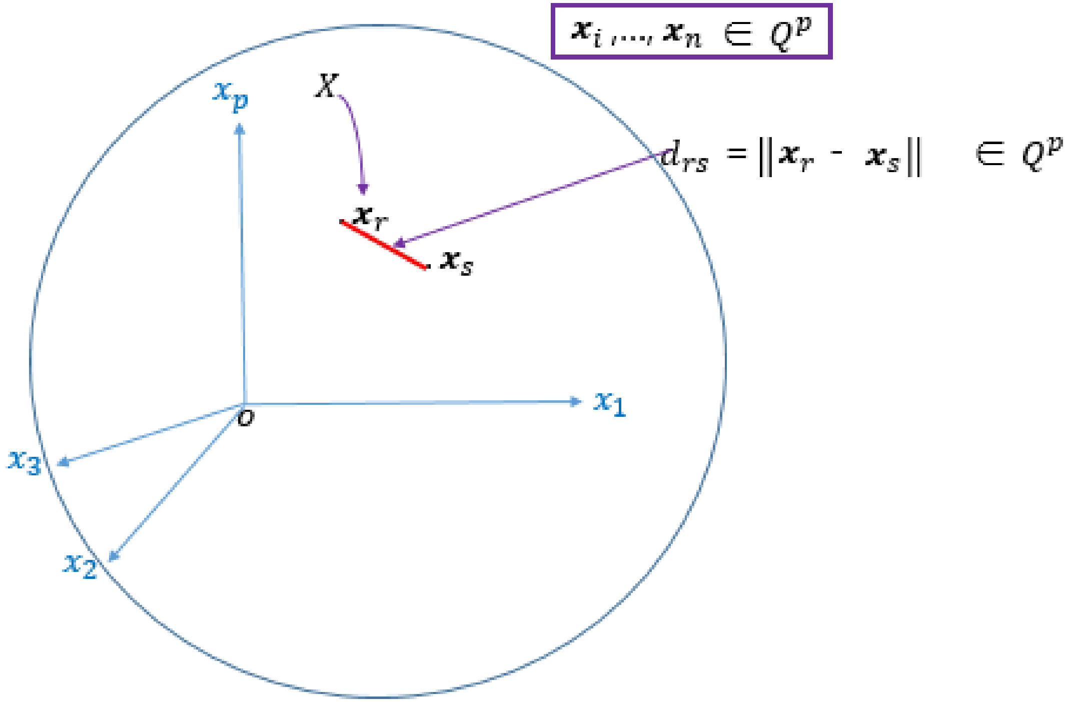
5. *Distance and Correlation of* X

5.1 *Norm : Length of vector*

- $\forall\, \boldsymbol{y} \in \mathbb{R}^n, \qquad \|\boldsymbol{y}\| = (\boldsymbol{y} \bullet \boldsymbol{y})^{1/2} = (\boldsymbol{y}^T \boldsymbol{y})^{1/2} = \left[\sum_{j=1}^{p} y_j^2\right]^{1/2}$

5.2 *Distance between* $\boldsymbol{x}_r = (x_{r1}, \ldots, x_{rp})^T$ *and* $\boldsymbol{x}_s = (x_{s1}, \ldots, x_{sp})^T$

- $d_{rs} = \|\boldsymbol{x}_r - \boldsymbol{x}_s\| = \left[(\boldsymbol{x}_r - \boldsymbol{x}_s)^T (\boldsymbol{x}_r - \boldsymbol{x}_s)\right]^{1/2} = \left[\sum_{j=1}^{p} (x_{rj} - x_{sj})^2\right]^{1/2}$: *Euclidean distance*
- $d_{rs} = \|\boldsymbol{x}_r - \boldsymbol{x}_s\|_{S^{-1}} = \left[(\boldsymbol{x}_r - \boldsymbol{x}_s)^T S^{-1} (\boldsymbol{x}_r - \boldsymbol{x}_s)\right]^{1/2}$: *Mahalanobis distance*



Geometric Concepts of Multivariate Data

Review of Basic Concepts in Multivairate Data

5.3 Correlation coefficient from Y

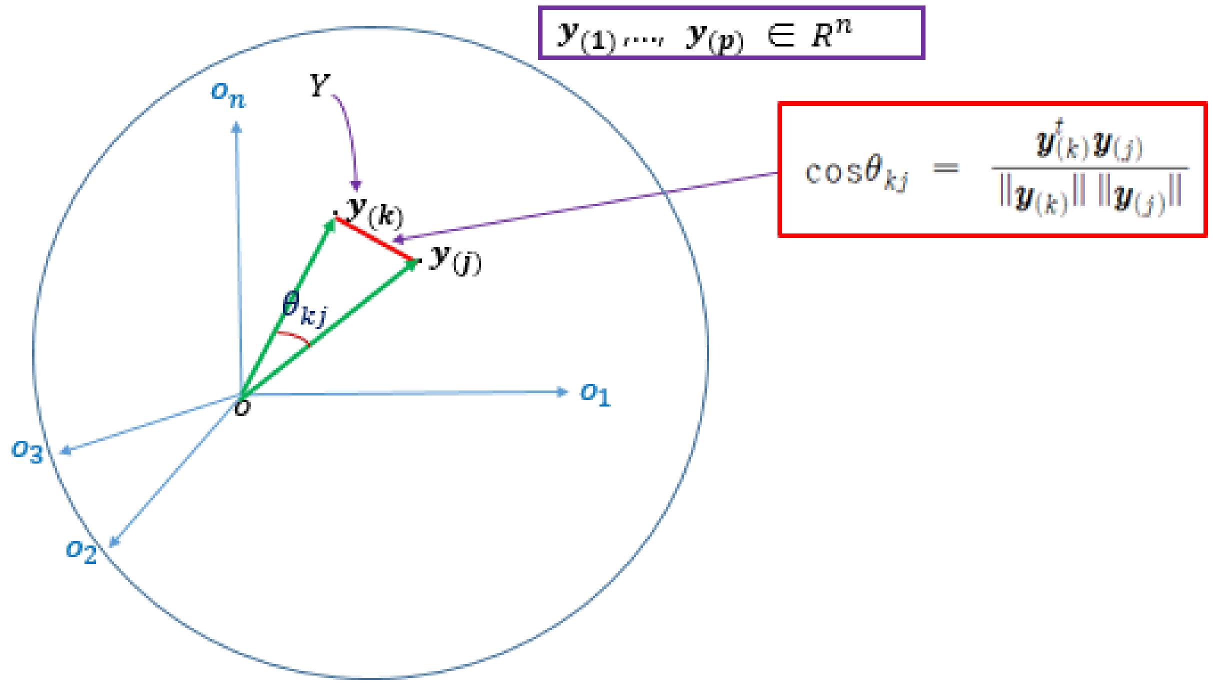
- $\boldsymbol{y}_{(j)} = (y_{1j}, \dots, y_{nj})^T = (x_{1j} - \overline{x}_j, \dots, x_{nj} - \overline{x}_j)^T, j = 1, \dots, p$

- $\boldsymbol{y}_{(k)} = (y_{1k}, \dots, y_{nk})^T$ *and* $\boldsymbol{y}_{(j)} = (y_{1j}, \dots, y_{nj})^T$

$$\Rightarrow \quad \cos\theta_{kj} = \frac{\boldsymbol{y}_{(k)}^T \boldsymbol{y}_{(j)}}{\|\boldsymbol{y}_{(k)}\| \, \|\boldsymbol{y}_{(j)}\|},$$

$$= \frac{\sum_{i=1}^{n} (x_{ik} - \overline{x}_k)(x_{ij} - \overline{x}_j)}{\sqrt{\sum_{i=1}^{n} (x_{ik} - \overline{x}_k)^2} \sqrt{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}},$$

$$= r_{kj}$$



Geometric Concepts of Multivariate Data