Multivariate Data Analysis Section 1

Qingyang Li

1 Covariance and Correlation of Bivariate R.V.

Population Covariance of X_i and X_j:

$$Cov(X_i, X_j) = \sigma_{X_i X_j} = E[(X_i - \mu_i)(X_j - \mu_j)]$$

Sample Covaraince of X_i and X_j:

$$S_{X_i X_j} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \mu_i)(X_j - \mu_j)$$

Population Correlation of X_i and X_j:

$$Corr(X_i, X_j) = \rho_{X_i X_j} = \frac{\sigma_{X_i X_j}}{\sigma_{X_i} \sigma_{X_j}} = \frac{E[(X_i - \mu_i)(X_j - \mu_j)]}{\sqrt{E(X_i - \mu_i)^2}\sqrt{E(X_j - \mu_j)^2}}$$

Sample Correlation of X_i and X_j:

$$r_{X_i X_j} = \frac{S_{X_i X_j}}{S_{X_i} S_{X_j}} = \frac{\sum_{i=1}^{n} (X_i - \mu_i)(X_j - \mu_j)}{\sqrt{\sum_{i=1}^{n} (x_i - \mu_i)^2}\sqrt{\sum_{j=1}^{n} (x_j - \mu_j)^2}}$$

2 Sample Mean Vector

Let $\mathbf{x}$ be a random vector of p variables measure on a sampling unit. If there are n individuals in the sample, the n observation vectors are $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$, where $\mathbf{x_i} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$.

How do we find the sample mean vector?

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x_i} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{bmatrix} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_p \end{bmatrix}$$

How do we find the sample mean vector in a data matrix?

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1}^\intercal \\ \mathbf{x_2}^\intercal \\ \vdots \\ \mathbf{x_i}^\intercal \\ \vdots \\ \mathbf{x_n}^\intercal \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

Assume $\mathbf{j}^\intercal$ is $1 \times n$ size of 1's vector.

$$\hat{\mu}^\intercal = \frac{1}{n}\mathbf{j}^\intercal\mathbf{X} \Rightarrow \hat{\mu} = \frac{1}{n}\mathbf{X}^\intercal\mathbf{j}$$

3 Sample Covariance Matrix

$$\begin{aligned}
\widehat{\mathbf{\Sigma}} &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x_i}-\hat{\mu})(\mathbf{x_i}-\hat{\mu})^\intercal \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\begin{bmatrix} x_{i1}-\hat{\mu}_1 \\ x_{i2}-\hat{\mu}_2 \\ \vdots \\ x_{ip}-\hat{\mu}_p \end{bmatrix} \times \begin{bmatrix} x_{i1}-\hat{\mu}_1 & x_{i2}-\hat{\mu}_2 & \cdots & x_{ip}-\hat{\mu}_p \end{bmatrix} \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\begin{bmatrix} (x_{i1}-\hat{\mu}_1)(x_{i1}-\hat{\mu}_1) & (x_{i1}-\hat{\mu}_1)(x_{i2}-\hat{\mu}_2) & \cdots & (x_{i1}-\hat{\mu}_1)(x_{ip}-\hat{\mu}_p) \\ (x_{i2}-\hat{\mu}_2)(x_{i1}-\hat{\mu}_1) & (x_{i2}-\hat{\mu}_2)(x_{i2}-\hat{\mu}_2) & \cdots & (x_{i2}-\hat{\mu}_2)(x_{ip}-\hat{\mu}_p) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{ip}-\hat{\mu}_p)(x_{i1}-\hat{\mu}_1) & (x_{ip}-\hat{\mu}_p)(x_{i2}-\hat{\mu}_2) & \cdots & (x_{ip}-\hat{\mu}_p)(x_{ip}-\hat{\mu}_p) \end{bmatrix} \\
&= \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1p} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{p1} & \hat{\sigma}_{p2} & \cdots & \hat{\sigma}_{pp} \end{bmatrix}
\end{aligned}$$

How do we find sample covariance matrix from data matrix?

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n-1}\mathbf{X}^\intercal(\mathbf{I}-\frac{1}{n}\mathbf{J})\mathbf{X}$$

,where $\mathbf{I}-\frac{1}{n}\mathbf{J}$ is a $n \times n$ matrix.

4 Sample Correlation Matrix

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

The sample correlation between j^{th} and k^{th} varaiables is defined as $r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}}$, where $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, p$.

5 Standardized Data

Sometimes it is easier to work with data which are on the same scale. Standardized data can be used to convert the data to an unitless scale. $z_{jk} = \frac{x_{jk}-\hat{\mu}_k}{\sqrt{s_{jj}}}$, where $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, p$.

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

- After standardizing data, we will have a sample mean of 0 and a sample variance of 1 for each of the variables.
- After standardizing data, the sample correlation matrix will be equal to the sample covariance matrix. Using an example to illustrate,

$$\begin{aligned} \rho_{12} &= \frac{cov(z_1, z_2)}{\sqrt{var(z_1)var(z_2)}} \\ &= \frac{cov(z_1, z_2)}{\sqrt{1 \times 1}} \\ &= cov(z_1, z_2) \end{aligned}$$

- The sample correlation matrix is always the same as you are working on original data or standardized data.