Multivariate Statistics

Mathias Balling Christiansen

Contents

Multivariate data and multivariate normal distribution

2D bivariate random vector

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Simultaneous pdf $f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$ complete information.

Marginal PDFs

$$f(x_1) = \int f(x_1, x_2)\,\mathrm{d}x_2 \quad \text{and} \quad f(x_2) = \int f(x_1, x_2)\,\mathrm{d}x_1$$

Mean vector

$$\mu = \begin{pmatrix} E[x_1] \\ E[x_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

Marginal variances

$$\sigma_1^2 = \sigma_{11} = E\left[(X_1 - \mu_1)^2\right] \geq 0$$

$$\sigma_2^2 = \sigma_{22} = E\left[(X_2 - \mu_2)^2\right] \geq 0$$

Covariance

$$\sigma_{12} = \mathrm{Cov}(X_1, X_2) = \sigma_{21} = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

Covariance matrix

$$\sum \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

The matrix is symmetric and has positive semi-definite eigenvalues $\lambda_1, \lambda_2 \geq 0$

Correlation coefficient

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

Correlation matrix

$$\rho = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix}$$

$$-1 \leq \rho_{12} \leq 1$$

Generalized variance

$$\left|\sum\right| = \det\left(\sum\right) = \prod_{i=1}^{2} \lambda_i$$

Measure of amount of random variation

pD bivariate random vector

$$X = \left[x_1, x_2, ..., x_p\right]^T$$

Mean vector

$$\mu = \begin{pmatrix} \mu_1 \\ \dots \\ \mu_p \end{pmatrix}$$

Covariance matrix

$$\sum_{p \times p} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{1p} \\ \dots & \sigma_2 & \dots \\ \dots & \dots & \sigma_p \end{bmatrix} = E\left[(X-\mu)(X-\mu)^T\right]$$

Correlation matrix:

$$\rho_{p \times p} = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \dots & 1 & \dots & \dots \\ \dots & \dots & 1 & \dots \\ \dots & \dots & \dots & 1 \end{bmatrix} - 1 \leq \rho_{ij} \leq 1$$

Sampling

From a random sample we want to estimate the moment of the population $(\mu, \sum, \rho)$.

Data Matrix: (Make = over X)

$$X_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \dots \\ x_n^T \end{bmatrix}$$

Columns are the variables and rows are the observations.

Descriptive statistics

Estimation of moments $(\mu, \sum, \rho)$

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

```
% Matlab
mean(X)
```

$$\widehat{\sum}_{p \times p} = S_{p \times p} = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \hat{\mu})(X_j - \hat{\mu})^T = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ \dots & s_2^2 & \dots & \dots \\ \dots & \dots & s_3^2 & \dots \\ \dots & \dots & \dots & s_p^2 \end{pmatrix}$$

```
% Matlab
cov(X)
```

$$\hat{\rho} = R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1p} \\ \dots & 1 & \dots & \dots & \dots \\ \dots & \dots & 1 & \dots & \dots \\ \dots & \dots & \dots & 1 & \dots \\ \dots & \dots & \dots & \dots & 1 \end{pmatrix}$$

```matlab
% Matlab
corrcoef(X)
```

Inference about mean vectors and comparison of multivariate samples

Introduction to non-parametric hypothesis tests

Regression and the General Linear Model

Dimension reduction

Discrimination and classification

Grouping and clustering