Multivariate Statistics

Sudipta Das

Assistant Professor,
Department of Data Science,
Ramakrishna Mission Vivekananda University, Kolkata
Slides adapted from Jhonson & Winchern

Outline I

1. Random Vectors & Random Sample
   - Random Vectors
   - Random Samples
   - Generalized Sample Variance
   - Statistical Distance

Random Vectors I

- Random vector: Vector of random variables

$$\underline{X} = [X_1, X_2, \ldots, X_p]'$$

- Mean vector

$$E(\underline{X}) = [\mu_1, \mu_2, \ldots, \mu_p]' = \underline{\mu}$$

- Covariance matrix

$$Cov(\underline{X}) = E(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix} = \Sigma.$$

- Example 2.13 (Page 70)

- Correlation matrix

$$
\begin{aligned}
Cor(\underline{X}) &= \begin{bmatrix}
\frac{\sigma_{11}}{\sqrt{\sigma_{11}\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}\sigma_{pp}}} \\
\frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}\sigma_{pp}}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\sigma_{1p}}{\sqrt{\sigma_{11}\sigma_{pp}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}\sigma_{pp}}}
\end{bmatrix} \\
&= \begin{bmatrix}
1 & \rho_{12} & \cdots & \rho_{1p} \\
\rho_{12} & 1 & \cdots & \rho_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{1p} & \rho_{2p} & \cdots & 1
\end{bmatrix} = \rho.
\end{aligned}
$$

Random Vectors III

- Standard deviation matrix

$$V^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{\sigma_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\sigma_{pp}} \end{bmatrix}.$$

- Relation between Σ and ρ through V.

$$\Sigma = V^{\frac{1}{2}} \rho V^{\frac{1}{2}}$$

and

$$\rho = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}}.$$

- Example 2.14 (Page 72)

- Result: For any real constant vector $\underline{c} = [c_1, c_2, \ldots, c_p]'$, the linear combination $\underline{c}'\underline{X} = c_1X_1 + c_2X_2 + \ldots + c_pX_p$ has mean

$$E(\underline{c}'\underline{X}) = \underline{c}'\underline{\mu}$$

and variance

$$Var(\underline{c}'\underline{X}) = \underline{c}'\Sigma\underline{c}.$$

- Result: For any real matrix

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{q1} & c_{q2} & \cdots & c_{qp} \end{bmatrix}$$

the linear combination $\underline{Z} = C\underline{X}$ has mean

$$\underline{\mu}_Z = E(\underline{Z}) = E(C\underline{X}) = C\underline{\mu}_X$$

and variance

$$\Sigma_Z = Cov(\underline{Z}) = Cov(C\underline{X}) = C\Sigma_X C'.$$

- Result: For any two random vectors $\underline{X}_1$ and $\underline{X}_2$ of same order, let $\underline{Z} = \underline{X}_1 + \underline{X}_2$

$$
\begin{aligned}
\mu_Z &= E[\underline{X}_1 + \underline{X}_2] \\
&= E[\underline{X}_1] + E[\underline{X}_2] \\
&= \mu_1 + \mu_2.
\end{aligned}
$$

and

$$
\begin{aligned}
\Sigma_Z &= Var[\underline{X}_1 + \underline{X}_2] \\
&= Var[\underline{X}_1] + Var[\underline{X}_2] + Cov[\underline{X}_1, \underline{X}_2] + Cov[\underline{X}_2, \underline{X}_1] \\
&= \Sigma_{11} + \Sigma_{22} + \Sigma_{12} + \Sigma_{21}.
\end{aligned}
$$

Random Samples I

- Let $\mathbf{X}_{1\cdot}, \mathbf{X}_{2\cdot}, \ldots, \mathbf{X}_{n\cdot}$ be n samples drawn from a random distribution. Then the sample mean $\bar{\mathbf{X}}$ is calculated as

$$\begin{aligned}
\bar{\mathbf{X}} &= \frac{1}{n}[\mathbf{X}_{1\cdot} + \mathbf{X}_{2\cdot} + \cdots + \mathbf{X}_{n\cdot}]' \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} X_{i1}, \sum_{i=1}^{n} X_{i2}, \ldots, \sum_{i=1}^{n} X_{ip}\right]' \\
&= [\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p]'
\end{aligned}$$

and the sample variance is calculated as

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_{i\cdot} - \bar{\mathbf{X}})(\mathbf{X}_{i\cdot} - \bar{\mathbf{X}})'$$

$$= \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} X_{i1} - \bar{X}_1 \\ X_{i2} - \bar{X}_2 \\ \vdots \\ X_{ip} - \bar{X}_p \end{bmatrix}\begin{bmatrix} X_{i1} - \bar{X}_1 & X_{i2} - \bar{X}_2 & \cdots & X_{ip} - \bar{X}_p \end{bmatrix}$$

$$= \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n}(X_{i1}-\bar{X}_1)^2 & \sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2) & \cdots & \sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{ip}-\bar{X}_p) \\ \sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2) & \sum_{i=1}^{n}(X_{i2}-\bar{X}_2)^2 & \cdots & \sum_{i=1}^{n}(X_{i2}-\bar{X}_2)(X_{ip}-\bar{X}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{ip}-\bar{X}_p) & \sum_{i=1}^{n}(X_{i2}-\bar{X}_2)(X_{ip}-\bar{X}_p) & \cdots & \sum_{i=1}^{n}(X_{ip}-\bar{X}_p)^2 \end{bmatrix}$$

- Example 1.2 (Page 10)

Random Samples III

- Let $\mathbf{X}_{1\cdot}, \mathbf{X}_{2\cdot}, \ldots, \mathbf{X}_{n\cdot}$ be random samples from a joint distribution that has mean vector $\underline{\mu}$ and covariance matrix Σ. Then for the sample mean $\bar{\mathbf{X}}$,

$$E(\bar{\mathbf{X}}) = \underline{\mu} \quad \text{and} \quad Cov(\bar{\mathbf{X}}) = \frac{1}{n}\Sigma$$

and for the sample variance S_n,

$$E(S_n) = \frac{n-1}{n}\Sigma$$

Random Samples IV

- Therefore an unbiased estimator of Σ is

$$\begin{aligned} S &= \left(\frac{n}{n-1}\right) S_n \\ &= \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_{i.} - \bar{\mathbf{X}})(\mathbf{X}_{i.} - \bar{\mathbf{X}})' \\ &= \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{12} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1p} & S_{2p} & \dots & S_{pp} \end{bmatrix} \end{aligned}$$

Generalized Sample Variance I

- Determinant of S is called as generalized sample variance.
- One can show that for a $p-$variate data set

$$\text{Generalized Sample Variance} = |S| = (n-1)^{-p}(\textit{hyper volume})^2$$

  by induction.
  - Geometrical interpretation for bivariate data:
    Example 3.7 (Page 124)
- For highly correlated data generalized sample variance will be smaller.

Statistical Distance I

- Statistical distance (d) between any two sample points

$$P = X_{i.} \text{ and } Q = X_{j.}$$

in a sample set $\{X_{1.}, X_{2.}, \ldots, X_{n.}\}$ is defined as

$$d^2(P, Q) = (X_{i.} - X_{j.})' S^{-1} (X_{i.} - X_{j.})$$

- Figure 1.25 (Page 37)