inner diameter outer diameter

$X_1, X_2, X_3 \rightarrow$ random variables

i	X_1 ID	X_2 OD	X_3 Thickness
1.	x_{11}	x_{12}	x_{13}
2.	x_{21}	x_{22}	x_{23}
⋮	⋮	⋮	⋮
n	x_{n1}	x_{n2}	x_{n3}

Variate $\quad \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}_{p \times 1}$$

↳ all random vectors under study

$$X_{n \times p} = \begin{array}{c|c} i \backslash j & \begin{array}{ccccc} x_1 & x_2 & x_3 \cdots & x_j \cdots & x_p \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ i \\ \vdots \\ n \end{array} & \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{1j} & x_{1p} \\ x_{21} & x_{22} & x_{23} & x_{2j} & x_{2p} \\ x_{31} & x_{32} & x_{33} & x_{3j} & x_{3p} \\ x_{i1} & x_{i2} & x_{i3} & x_{ij} & x_{ip} \\ x_{n1} & x_{n2} & x_{n3} & x_{nj} & x_{np} \end{bmatrix} \end{array}$$

gold X

data matrix

$$X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ x_{i4} \\ \vdots \\ x_{ip} \end{bmatrix}_{p \times 1}$$

column matrix for a particular observation and all random vars w.r.t it.

$$X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ x_{3j} \\ \vdots \\ x_{nj} \end{bmatrix}_{n \times 1}$$

column matrix for a particular random variable and all observations w.r.t it.

Univariate $\qquad\qquad$ Multivariate

$$E(X) = \sum_{\text{all } x_i} x_i \, p(x_i) \rightarrow \text{Discrete} \quad \leftarrow \quad E(X_j) = \sum_{\text{all } x_j} x_j \, p(x_j)$$

$$= \int_{-\infty}^{\infty} x \, f(x) \, dx \rightarrow \text{continuous} \quad \leftarrow \quad = \int_{-\infty}^{\infty} x_j \, f(x_j) \, dx_j$$

Univariate mean

$$\bar{x} = \hat{\mu}$$
$$= \frac{1}{n} \sum_{i=1}^{n} x_i$$

Multivariate mean

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^{n} x_{i1} \\ \frac{1}{n} \sum_{i=1}^{n} x_{i2} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} x_{ip} \end{bmatrix}$$

- univariate
  - central tendency
    - mean
    - median
    - mode
  - Dispersion
    - range
    - IQR
    - variance

- Multivariate
  - Mean vector
  - covariance matrix
  - correlation matrix

Multivariate Statistics

- Multivariate descriptive statistics
  - → Mean vector
  - → covariance matrix
  - → correlation matrix
- Multivariate normal distribution
- Multivariate inferential statistics
  - → Testing of hypothesis MANOVA
  - → Hotellings T^2 distribution.

* Dependence relationship
- → Response variable
  Volume
- → Explanatory variable
  (ID, OD, maybe thickness)

deciding explanatory and response variables depend on the situation, where response variables are derived from explanatory variables

* Interdependence relationship $\rightarrow$ explanatory variables and response variables cannot be precisely distinguished.

Mean Vector

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix} = \begin{bmatrix} 1/n \sum_{i=1}^{n} x_{i1} \\ 1/n \sum_{i=1}^{n} x_{i2} \\ \vdots \\ 1/n \sum_{i=1}^{n} x_{ip} \end{bmatrix}$$

To directly get from matrix form

$$\frac{1}{n}\left(X^T_{p\times n} \times 1_{n\times 1}\right) = \bar{X} = \hat{\mu}$$

$\rightarrow$ no need to calculate individually

Covariance Matrix

$$V(X) = E(X-\mu)^2$$

$$= \sum (x-\mu)^2 \, p(x) \, d(x) \rightarrow \text{discrete}$$

$$= \int (x-\mu)^2 \, f(x) \, dx \rightarrow \text{continuous}$$

Univariate

$X_1, X_2 \quad j = 1, 2$

$$V(X_j) = E(X_j - \mu_j)^2$$

$$\sigma_{jj} = \sigma_j^2 = E(x_j - \mu_j)(x_j - \mu_j)$$

$$\text{cov}(X_j, X_k) = E[(X_j - \mu_j)(X_k - \mu_k)]$$

$$\sigma_{jk} = \sum_{\text{all } x_j} \sum_{\text{all } x_k} (x_j - \mu_j)(x_k - \mu_k) \, p(x_j, x_k) \rightarrow \text{discrete}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_j - \mu_j)(x_k - \mu_k) \, f(x_j, x_k) \cdot dx_j \cdot dx_k \rightarrow \text{continuous}$$

$$\Sigma = \begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_p \\ \hline X_1 & \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ X_2 & \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & & \\ X_p & \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{array}$$

same as σ_{p1} σ_{p2}

Sample covariance matrix

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & & & \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x - \bar{x})^2$$

$$s_{ij} = s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

$$x^*_{ij} = x_{ij} - \bar{x}_j$$

$$X^* = \begin{bmatrix} x^*_{11} & x^*_{12} & \cdots & x^*_{1p} \\ x^*_{21} & x^*_{22} & \cdots & x^*_{2p} \\ \vdots & \vdots & & \vdots \\ x^*_{n1} & x^*_{n2} & \cdots & x^*_{np} \end{bmatrix}_{n \times p}$$

directly in matrix form

$$\frac{1}{n-1} \left[X^{*T}_{p \times n} \cdot X^*_{n \times p} \right] =$$

Correlation Matrix

correlation

$$\rho_{xy} = \frac{cov(x, y)}{SD(x)\ SD(y)}$$

$$\rho_{jk} = \frac{cov(x_j, x_k)}{\sqrt{V(x_j)\ V(x_k)}}$$

$$= \frac{\sigma_{jk}}{\sigma_j \cdot \sigma_k}$$

$$\rho_{p \times p} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}$$

sample correlation matrix

$$r_{jk} = \frac{s_{jk}}{s_j \cdot s_k}$$

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix}$$

$$\tilde{x}_{jk} = \frac{x_{jk} - \bar{x}_k}{s_k} = \frac{x_{jk} - \bar{x}_k}{\sqrt{s_{kk}}}$$

$$\tilde{X} = \begin{bmatrix} \frac{x_{11} - \bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{12} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{1p} - \bar{x}_p}{\sqrt{s_{pp}}} \\ \frac{x_{21} - \bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{22} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{2p} - \bar{x}_p}{\sqrt{s_{pp}}} \\ \vdots & \vdots & & \vdots \\ \frac{x_{n1} - \bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{n2} - \bar{x}_2}{\sqrt{s_{22}}} & & \frac{x_{np} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix}$$

$$R = \frac{1}{n-1} \left[\tilde{X}^T_{p \times n} \cdot \tilde{X}_{n \times p} \right]$$

$$\underbrace{X^T X, \quad X^* X^{*T}, \quad \tilde{X}^T \tilde{X}}$$

Diagonal entries → sum squares

Non diagonal entries → cross products

S – Known

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & & & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix}$$

$$D_s = \begin{bmatrix} s_{11} & 0 & 0 & \cdots & 0 \\ 0 & s_{22} & 0 & 0 \cdots & 0 \\ 0 & & \ddots & & \vdots \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \cdots & s_{pp} \end{bmatrix}$$

$$\boxed{R = D_s^{-1/2} \, S \, D^{-1/2}}$$

correlation from covariance

$$\boxed{S = D_s^{1/2} R \, D^{1/2}}$$

covariance from correlation

$\tilde{X}^T X = (n-1)R$

$X^{*T} X^* = (n-1)S$

$$X = \begin{bmatrix} 4 & 5 & 10 \\ 5 & 4 & 12 \\ 6 & 6 & 15 \end{bmatrix}$$

find ① mean vector

② covariance matrix

③ correlation matrix

$$\frac{1}{n} \times \left(X^T_{p \times 1} \; 1_{n \times 1} \right) = \text{mean vector}$$

find X^*, $x^* = (x_{ij} - \bar{x}_j)$

$$\frac{1}{(n-1)} X^{*T} \times X^* = \text{covariance matrix}$$

$$R = D_s^{1/2} \, S \, D^{1/2} \rightarrow \text{correlation matrix}$$

Multivariate normal distribution

$$x \sim N(\underbrace{\mu, \sigma^2}_{\text{population parameters}})$$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \sim N_p(\mu, \Sigma)$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

Bivariate normal dist

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

$$X \sim N_2(\mu, \Sigma)$$

if $x_1, x_2 \rightarrow$ independent random variables

$$\rho_{12} = 0, \ \sigma_{12} = 0 \quad \Sigma = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}$$

joint pdf $\rightarrow$ univariate

$$f(x_1, x_2) = f(x_1) \times f(x_2)$$

$$= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \times \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2}$$

$$= \frac{1}{(2\pi)^{2/2}(\sigma_1^2\sigma_2^2)^{1/2}} e^{-\frac{1}{2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}$$

Multivariate normal distribution

Univariate — x

$$x \sim N(\mu, \sigma^2)$$

Multivariate

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \sim N_p(\mu, \Sigma)$$

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & & \vdots \\ \sigma_{1p} & \cdots & \sigma_{pp} \end{bmatrix}$$

Bivariate normal dist

x_1, x_2

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

$$x \sim N_2$$

$$f(x_1, x_2) = \frac{1}{(2\pi)^{2/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}[(x-\mu)^T \Sigma^{-1} (x-\mu)]}$$

Generalising for 3 independent features

x_1, x_2, x_3

$$f(x_1, x_2, x_3) = \frac{1}{(2\pi)^{3/2} |\Sigma|^{1/2}} e^{-1/2 [(x-\mu)^T \Sigma^{-1} (x-\mu)]}$$

$\hookrightarrow$ 3x3 matrix

Generalising for p independent features

$$f(x_1, --- x_p) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-1/2 [(x-\mu)^T \Sigma^{-1} (x-\mu)]}$$

$$-\infty < x_j < \infty, \; j = 1, 0, ----, p$$

If x_1 and x_2 are related

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \qquad \sigma_{12} \neq 0$$

$$|\Sigma| = \sigma_1^2 \sigma_2^2 - (\sigma_{12})^2$$

$$= \sigma_1^2 \sigma_2^2 - (\rho_{12} \sigma_1 \sigma_2)^2$$

$$|\Sigma| = \sigma_1^2 \sigma_2^2 (1 - \rho_{12}^2)$$

Properties of MND

1) If $X_{p\times 1} \sim N_p(\mu, \Sigma)$, then X_j is $N(\mu_j, \sigma_j^2)$

2) If $X_{p\times 1} \sim N_p(\mu, \Sigma)$, then the subset of $X_{p\times 1}$, is $X_{q\times 1}$ is $N_q(\mu_q, \Sigma_q)$

3) If $X_{p\times 1} \sim N_p(\mu, \Sigma)$, then the linear combination of X_j, $j = 1, 2, \ldots, p$ is univariate normal

$$a^T = [a_1, a_2, \ldots a_p]$$

$$E(a^T x) = a^T \mu \qquad a^T x \sim N(a^T \mu, a^T \Sigma a)$$

$$V(a^T x) = a^T \Sigma a \qquad a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}_{p\times 1} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}_{p\times 1}$$

4) If $X_{p\times 1} \sim N_p(\mu, \Sigma)$, then q linear combination X_j, $j = 1, 2, \ldots, p$ is multivariate (q-dimensional) normal

$$A = \begin{bmatrix} a_{11} & a_{21} & - & - & - & a_{q1} \\ a_{12} & a_{22} & - & - & - & a_{q2} \\ \vdots & & & & & \vdots \\ a_{1p} & - & - & - & - & - \; a_{qp} \end{bmatrix}$$

$$A^T X \sim N_q(A^T \mu, A^T \Sigma A)$$

Q. Consider the bivariate Normal distribution with

$$X = \begin{bmatrix} 10 \\ 12 \end{bmatrix} \quad \mu = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 16 \end{bmatrix}$$

Obtain its bivariate normal distribution function

$X \sim N_2(\mu, \Sigma)$

$\rightarrow$ p.d.f $f(x_1, x_2) = \dfrac{1}{2\pi |\Sigma|^{1/2}} e^{-\frac{1}{2}[(X-\mu)^T \Sigma^{-1} (X-\mu)]}$

$$\Sigma^{-1} = \begin{bmatrix} \frac{4}{15} & -1/30 \\ 1/30 & 1/15 \end{bmatrix} \qquad X - \mu = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$$

$$(x-\mu)^T \Sigma^{-1} (x-\mu) = \begin{bmatrix} 5 & 8 \end{bmatrix} \begin{bmatrix} 4/15 & -1/30 \\ -1/30 & 1/15 \end{bmatrix} \begin{bmatrix} 5 \\ 8 \end{bmatrix}$$

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{60}} e^{-1/2[(x-\mu)^T \Sigma^{-1} (x-\mu)]}$$

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{60}} e^{-4.42}$$

Q. $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ $\mu = \begin{bmatrix} -3 \\ 1 \\ 4 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

$x_1, x_2 \rightarrow$ independent or not?

$x^{(1)} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N_2(\mu^{(1)}, \Sigma^{(1)})$

$\mu^{(1)} = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$ $\Sigma^{(1)} = \begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix}$

The covariances are not 0 so x_1, x_2 are dependent

(x_1, x_2) and $x_3 \rightarrow$ independent?

$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

The covariances are 0, (x_1, x_2) and x_3 are independent.

$\rightarrow \left(\frac{X_1 + X_2}{2}, X_3\right)$ independent?

$$\frac{X_1 + X_2}{2} \sim N_2(\mu, \Sigma)$$

$$\frac{X_1 + X_2}{2}, X_3$$

coeff coeff

$$A = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$AX = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{X_1 + X_2}{2} \\ X_3 \end{bmatrix}$$

$$AX \sim N_2(A\mu, A\Sigma A^T)$$

given given

$$A\mu = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -3 \\ 1 \\ 4 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix}$$

$$A\Sigma A^T = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$AX \sim N_2\left(\begin{bmatrix} -1 \\ 4 \end{bmatrix}, \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}\right)$$

$$\begin{array}{c|cc} & \frac{X_1+X_2}{2} & X_3 \\ \hline \frac{X_1+X_2}{2} & 1/2 & 0 \\ X_3 & 0 & 2 \end{array}$$

since covariances are 0 $\frac{X_1 + X_2}{2}$ and X_3 are independent.

Find pdf $\frac{x_1+x_2}{2}, x_3$

$$f\left(\frac{x_1+x_2}{2}, x_3\right) = \frac{1}{2\pi|\Sigma^*|^{1/2}} e^{-1/2\left[(AX-\mu^*)^T(\Sigma^*)^{-1}(AX-\mu^*)\right]}$$

Q. Let $X_{3\times1} \sim N_3(\mu, \Sigma)$

with $\Sigma = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

Are x_1 and x_2 are independent?

→ No, covariance of submatrix not 0

Are (x_1, x_2) and x_3 are independent?

→ Yes, x_3 independent of x_1 and x_2

Q. Let $x_1\ x_2\ x_3$ be trivariate normal random vector with $-3, 1, 4$ and variance covariance matrix

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & -3 \\ 0 & -3 & 4 \end{bmatrix}$$

Find pdf of $\frac{x_1+x_2}{2}, x_3$

Multiple Correlation

2 random variables $\rightarrow x, y$

ρ_{xy}

$-1 < \rho < 1$

$$r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{V(x) \cdot V(y)}}$$

variance

* To study multiple impact of independent variables on dependent variables
* To study the direction b/w them
  → Positive correlation etc. - -

$$0 \leq R_{1.23} \leq 1$$ always positive

impact of x_2, x_3 on x_1

2 independent random variables

dependent random variable

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

$r_{12}, r_{13}, r_{23} \rightarrow$ simple or zero order correlation coefficient

Partial Correlation

* To study the partial of one independent variable on the dependent variable keeping other independent variables constant

* To study the direction

* $-1 \leq r_{12\cdot3} \leq 1$

DV IV fixed IV

fix age, on basis of height what could be the weight

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\cdot r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

Q. $r_{12} = 0.7$, $r_{13} = 0.61$, $r_{23} = 0.4$

$$r_{12\cdot3} = \frac{0.7 - 0.61 \times 0.4}{\sqrt{1-(0.61)^2}\sqrt{1-(0.4)^2}} = 0.628$$

$$R_{1\cdot23} = \sqrt{\frac{(0.7)^2 + (0.61)^2 - 2\times0.7\times0.61\times0.4}{1-(0.4)^2}} = 0.726$$

Q. Find the partial correlation b/w x_1 and x_3 given x_2

$$\Sigma = \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1 & 0.80 & -0.40 \\ 0.80 & 1 & -0.56 \\ -0.40 & -0.56 & 1 \end{bmatrix} \end{array}$$

σ_{12} σ_{13} σ_{23}

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$\rho_{13 \cdot 2}$

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_{11} \cdot \sigma_{22}}$$

apply formula

Hotelling's T^2

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \sqrt{n}(\bar{x} - \mu)s^{-1}$$

$$t^2 = n(\bar{x} - \mu)(s^2)^{-1}(\bar{x} - \mu)$$

univariate

$x_1, x_2, x_3 - - - x_p$

$n \rightarrow n$

$$\bar{x} \longrightarrow \bar{X} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \qquad \mu \rightarrow \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

$$T^2 = n(\bar{x} - \mu)^T s^{-1}(\bar{x} - \mu)$$

Hotelling's T^2

$$\chi^2 = \left[\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right]^2$$

$$\chi^2 = n(\bar{x}-\mu)(\sigma^2)^{-1}(\bar{x}-\mu)$$

univariate

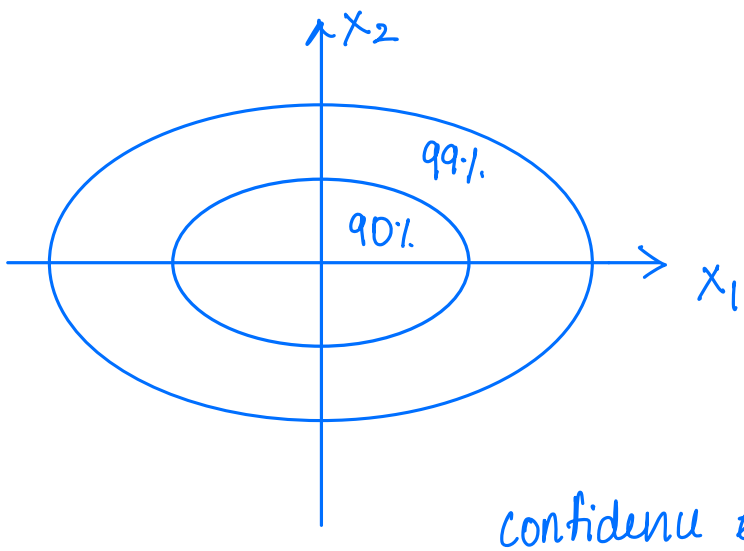
chi-square dist.

$$n(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) \sim \chi^2_p$$

$$\sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

Confidence region

$$P\left(n(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) \le \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)\right) = 1-\alpha$$



confidence ellipse

Case (i): Sampling from multivariate normal population with known Σ

→ population: $X \sim N_p(\mu, \Sigma)$, Σ Known

→ Sample statistic: $n(\bar{x}-\mu)^T \Sigma^{-1} (\bar{x}-\mu)$

→ Sampling distribution: χ^2_p

→ $100(1-\alpha)\%$ CR: $n(\bar{x}-\mu)^T \Sigma^{-1} (\bar{x}-\mu) \leq \chi^2_p(\alpha)$

ie, $P\left[n(\bar{x}-\mu)^T \Sigma^{-1} (\bar{x}-\mu) \leq \chi^2_p(\alpha)\right]$

$= 1-\alpha$

Case (ii): Sampling from multivariate normal population with large sample size and unknown Σ

→ Population: $X \sim N_p(\mu, \Sigma)$, Σ unknown but $n-p \geq 40$

→ Sample statistic: $n(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu)$

→ Sampling distribution: χ^2_p

→ $100(1-\alpha)\%$ CR: $n(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) \leq \chi^2_p(\alpha)$

ie, $P[n(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) \leq \chi^2_p(\alpha)] = 1-\alpha$

Case (iii): Sampling from multivariate normal population with small size and unknown Σ

→ Population: $X \sim N_p(\mu, \Sigma)$, Σ unknown but $n-p < 40$

→ Sample statistic: $n(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu)$

→ Sampling distribution: $\frac{(n-1)p}{n-p} F_{p,\, n-p}(\alpha)$

→ $100(1-\alpha)\%$ CR: $n(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) \leq \chi^2_p(\alpha)$

ie, $P[n(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) \leq \chi^2_p(\alpha)] = 1-\alpha$

Problem: A random sample with $n = 20$ were collected from a bivariate normal process. The sample mean vector and covariance matrix are given below.

(a) Obtain Hotelling's T^2 statistic

(b) What will be the distribution of it?

(c) Obtain 95% confidence region

$$\bar{x} = \begin{bmatrix} 10 \\ 20 \end{bmatrix} \qquad S = \begin{bmatrix} 40 & -50 \\ -50 & 100 \end{bmatrix}$$

$n = 20 \quad p = 2$ (Bivariate) $\quad n - p < 40$

$$\bar{x} = \begin{bmatrix} 10 \\ 20 \end{bmatrix} \qquad S = \begin{bmatrix} 40 & -50 \\ -50 & 100 \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$T^2 = n(\bar{x} - \mu)^T S^{-1} (\bar{x} - \mu)$$ To find

$$S^{-1} = \frac{1}{1500} \begin{bmatrix} 100 & 50 \\ 50 & 40 \end{bmatrix} \qquad \bar{x} - \mu = \begin{bmatrix} 10 - \mu_1 \\ 20 - \mu_2 \end{bmatrix}$$

$$(\bar{x} - \mu)^T = \begin{bmatrix} 10 - \mu_1 & 20 - \mu_2 \end{bmatrix}$$

$$S^{-1}(\bar{x} - \mu) = \frac{1}{1500} \begin{bmatrix} 100 & 50 \\ 50 & 40 \end{bmatrix} \begin{bmatrix} 10 - \mu_1 \\ 20 - \mu_2 \end{bmatrix}$$

$$S^{-1}(\bar{x} - \mu) = \begin{bmatrix} 100(10 - \mu_1) + 50(20 - \mu_2) \\ 50(10 - \mu_1) + 40(20 - \mu_2) \end{bmatrix} \frac{1}{1500}$$

$$T^2 = \frac{20}{1500} \begin{bmatrix} 10 - \mu_1 & 20 - \mu_2 \end{bmatrix} \begin{bmatrix} 100(10 - \mu_1) + 50(20 - \mu_2) \\ 50(10 - \mu_1) + 40(20 - \mu_2) \end{bmatrix}$$

b) Follows F distribution

$$T^2 \sim \frac{(n-1)p}{(n-p)} F_{p,n-p}$$

$$\boxed{T^2 \sim \frac{38}{18} F_{2,18}}$$

c) 95% $\Longrightarrow$ $100(1-\alpha)\%$

$\alpha = 0.05$

$$P\left[n(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)\right] = 1-\alpha$$

i.e $P\left[T^2 \leq \frac{38}{18} F_{2,18}(0.05)\right] = 0.95$

i.e $\boxed{P\left[T^2 \leq 7.5042\right] = 0.95}$

Expectation, Covariance Matrix, Properties

If x is a random vector with probability density function $f(x)$, then

$$E(x) = \begin{bmatrix} E(x_1) \\ \vdots \\ E(x_p) \end{bmatrix}$$

→ $E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$

α, β are some row vectors with constants

→

Cumulative distribution function

Let $X = (X_1\ X_2\ \cdots X_p)^T$ be a random vector with pdf $f(x)$

The cdf of x is defined by

$$F(x) = P[X \leq x]$$
$$= P[X_1 \leq x_1, X_2 \leq x_2, \cdots]$$

Marginal pdf

$$f_{x_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2$$

Marginal pdf of x_1 in case of bi-variate

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

$$f_{x_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2, x_3)\, dx_2\, dx_3$$

for trivariate

Note: All marginal distributions of $X \sim N_p(\mu, \Sigma)$ are normal

→ when X, Y are independent

$$E(XY^T) = E(X) \cdot E(Y^T)$$

$$\mathrm{cov}(X,Y) = E(XY^T) - E(X) \cdot E(Y^T)$$

$$\mathrm{var}(X,X) = E(XX^T) - E(X) \cdot E(X^T)$$

$$\mathrm{var}(aX) = a\,\mathrm{Var}(X)\,a^T$$

$$\mathrm{var}(AX+b) = A\,\mathrm{var}(X)\,A^T$$

$$\mathrm{cov}(X+Y,Z) = \mathrm{cov}(X,Z) + \mathrm{cov}(Y,Z)$$

$$\mathrm{cov}(AX,BY) = A(\mathrm{cov}(X,Y))B^T$$

Note: Let $X \sim N_p(\mu, \Sigma)$

Let us partition into k and $p-k$

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}$$

$$\Sigma = \left[\begin{array}{c|c} \underset{(k \times k)}{\Sigma_{11}} & \underset{k \times (p-k)}{\Sigma_{12}} \\ \hline \underset{(p-k) \times k}{\Sigma_{21}} & \underset{(p-k) \times (p-k)}{\Sigma_{22}} \end{array}\right]$$

$$\boxed{\begin{aligned} \text{Then} \quad X^{(1)} &\sim N_k(\mu^{(1)}, \Sigma_{11}) \\ X^{(2)} &\sim N_{p-k}(\mu^{(2)}, \Sigma_{22}) \end{aligned}}$$

Conditional expectations

$$\rightarrow \quad E(X_2 | X_1) = \int x_2 f(x_2 | x_1)\, dx_2$$

$$E(X_1 | X_2) = \int x_1 f(x_1 | x_2)\, dx_1$$

$$\mathrm{var}(X_2 | X_1 = x_1) = E(X_2 X^T | X_1 = x_1) - E(X_2 | X_1 = x_1)\, E(X_2^T | X_1 = x_1)$$

$$\mathrm{cov}(X_2, X_3 | X_1 = x_1) = E(X_2 X_3^T | X_1 = x_1) - E(X_2 | X_1 = x_1)\, E(X_2^T | X_1 = x_1)$$

Properties of conditional Expectation

$$E(X_2) = E\{E(X_2 | X_1)\}$$

$$\mathrm{var}(X_2) = E\{\mathrm{var}(X_2 | X_1)\} + \mathrm{var}\{E(X_2 | X_1)\}$$

The conditional distribution of $X^{(1)}$ given $X^{(2)} = x^{(2)}$ is normally distributed

$$N_k\left[\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)}),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right]$$

$\beta = \Sigma_{12}\Sigma_{22}^{-1}$ → matrix regression coeffs of $X^{(1)}$ or $X^{(2)}$

The vector $\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(X^{(2)} - \mu^{(2)})$ is called regression function

$$\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

partial covariance

Correlation coefficient

$$\rho_{12\cdot 3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}$$

Generalisation

$$\rho_{ij\cdot q+1,\ldots,p} = \frac{\sigma_{ij\cdot q+1,\ldots,p}}{\sqrt{\sigma_{ii\cdot q+1,\ldots,p}}\sqrt{\sigma_{jj\cdot q+1,\ldots,p}}}$$

$$i,j = 1,\ldots,p$$

Linear and quadratic forms

Let $X \sim N_p(\mu, \Sigma)$

Let A be $m \times p$ & 'a' be $m \times 1$ arrays

Then $AX + a \sim N_m(A\mu + a, A\Sigma A^T)$

In case of a symmetric matrix, if the eigen values are positive, then it is called positive definite matrix

-ve eigen values → negative indefinite matrix

Quadratic forms

$$Q(x) = x^T A x = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} x_i x_j$$

where $A_{p \times p}$ - symmetric matrix

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

Note: 1) If $Q(x) > 0$ for all $x \neq 0$ Positive definite.

2) If $Q(x) \geq 0$ for all $x \neq 0$ Positive semi-definite.

Q. $Q(x) = x_1^2 + x_2^2$

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

$$X^T A X = Q(x)$$

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + x_2^2$$

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + x_2^2$$

$$|A - \lambda I| = 0 \quad \boxed{\lambda_1 = \lambda_2 = 1}$$

positive definite.

Q. $Q(x) = (x_1 - x_2)^2$

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

$$X^T A X = Q(x)$$

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (x_1 - x_2)^2$$

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (x_1 - x_2)^2$$

$\lambda_1 = 2 \quad \lambda_2 = 0$

positive semi-definite

Q. $Q(x) = x_1^2 - x_2^2$

$$x^T A x = Q(x)$$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

Q. Consider the pdf

$$f(x_1, x_2) = \begin{cases} \frac{1}{2}x_1 + \frac{3}{2}x_2 & 0 < x_1 x_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

① Find the marginal densities $f_{X_1}(x_1)$ & $f_{X_2}(x_2)$

② Find the conditional densities $f(x_2/x_1)$ & $f(x_1/x_2)$

③ Find the mean vector

④ Find the covariance matrix

A1. $$f_{X_1}(x_1) = \int_0^1 \left(\tfrac{1}{2}x_1 + \tfrac{3}{2}x_2\right) dx_2$$

$$f_{X_2}(x_2) = \int_0^1 \left(\tfrac{1}{2}x_1 + \tfrac{3}{2}x_2\right) dx_1$$

A2. conditional densities

$$f(x_2/x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)}$$

$$f(x_1/x_2) = \frac{f(x_1, x_2)}{f_{X_2}(x_2)}$$

A3. $\mathrm{Cov}(x_1, x_2) = E(x_1, x_2) - E(x_1)E(x_2)$

Wishart's distribution (Chi sq. dist. in multivariate)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$(n-1)s^2 = \sum (x_i - \bar{x})^2$$

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

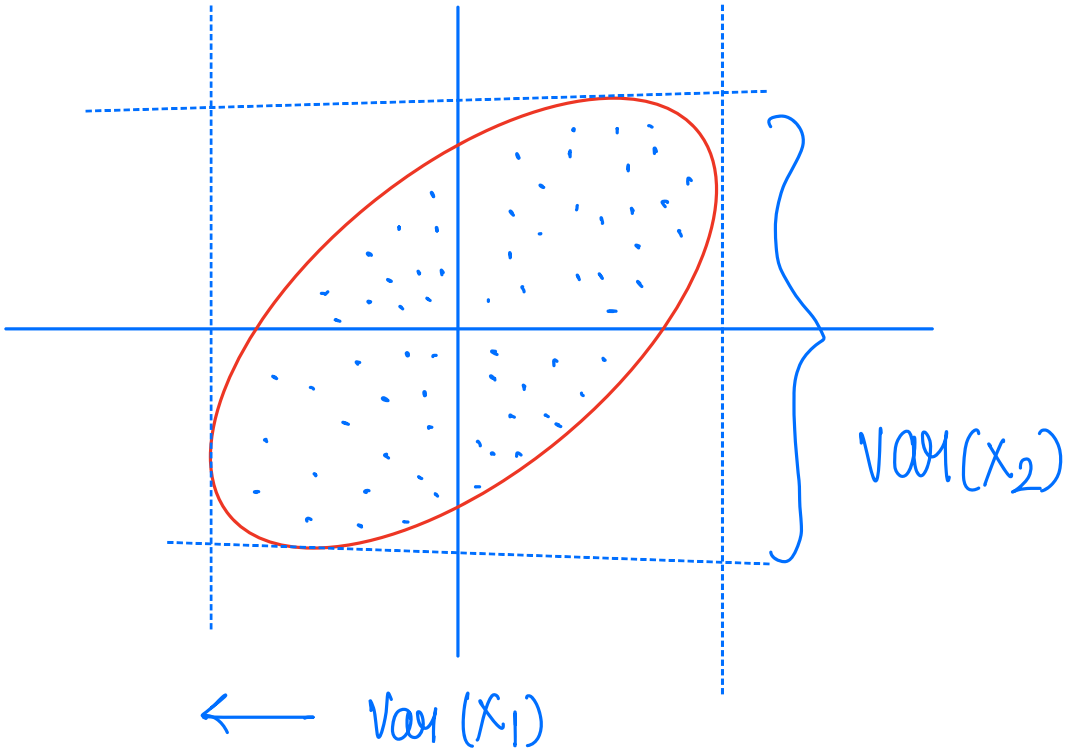
univariate

→ When $\mu = 0$

$$\sum (X_i - \bar{X})(X_i - \bar{X})^T \rightarrow \text{Wishart dist.}$$

$$\sum_{i=1}^{p} X_i X_i^T \sim W$$

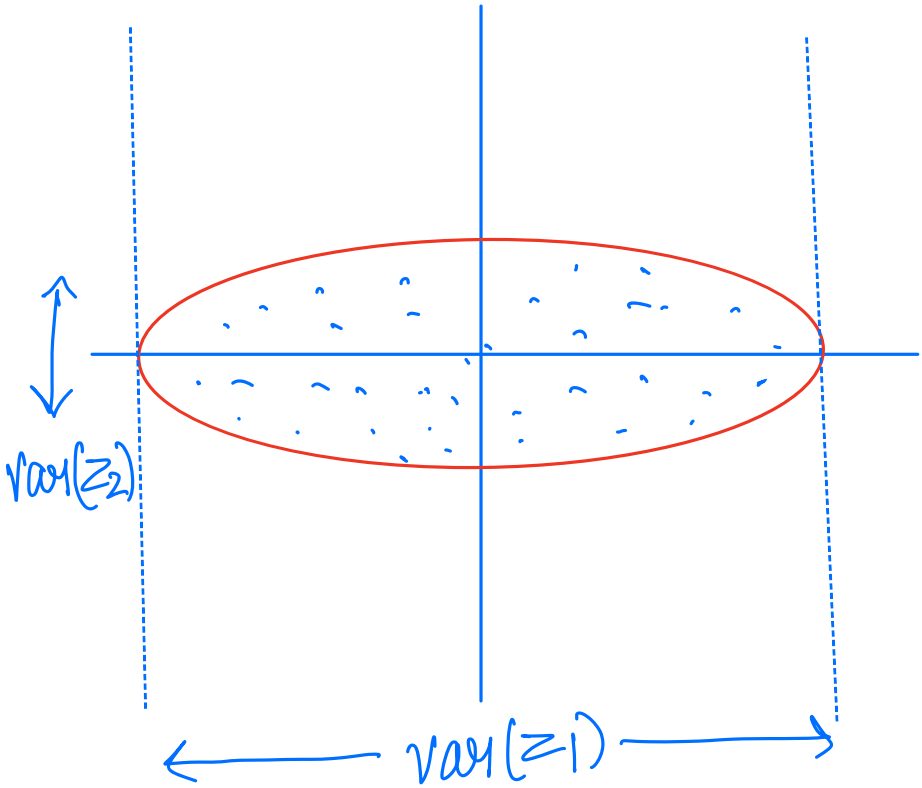
Principal Component Analysis



$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

→ we try to visualise the data from different angles and try to make the random variables independent

correlated data → uncorrelated dataset



$x_1, x_2 \rightarrow z_1, z_2$

$\text{var}\, z_1 > \text{var}\, z_2$

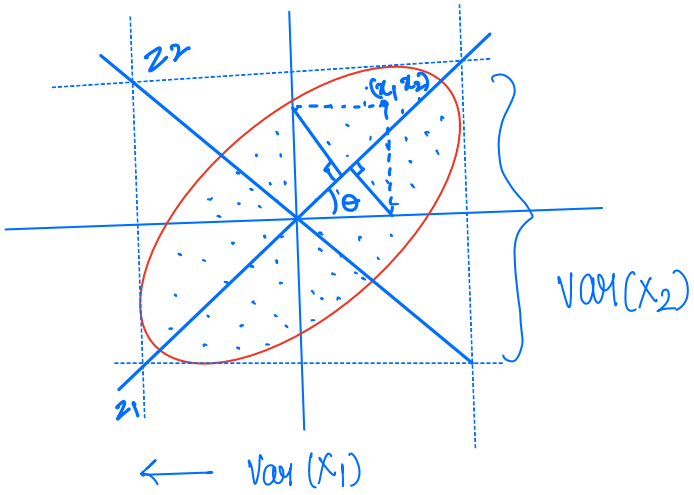
first PCA

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \qquad A^T = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$z = A^T x$$

$z_1 = \cos\theta\, x_1 + \sin\theta\, x_2 \rightarrow$ First PCA

$z_2 = \sin\theta\, x_1 + \cos\theta\, x_2 \rightarrow$ second PCA



For p variables

$$Z = \begin{bmatrix} z_1 \\ \vdots \\ z_p \end{bmatrix} \quad A^T = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & & & \vdots \\ a_{p1} & & & a_{pp} \end{bmatrix} \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

$$z_1 = a_{11}x_1 + a_{12}x_2 + \cdots a_{1p}x_p$$

$\hookrightarrow$ first PC

$$z_2 = a_{21}x_1 + a_{22}x_2 + \cdots a_{2p}x_p$$

$\hookrightarrow$ second PC

$$\boxed{[\ \ ]_{p\times p} \longrightarrow [\ \ ]_{m\times m} \quad m<p}$$

$V(z_1) \geq V(z_2) \geq V(z_3)$

The PC with very low variance can be ignored and thats how dimentionality decreases

$\rightarrow \quad z_j = a_j^T x \qquad E(z_j) = a_j^T \bar{x}$

$v(z_j)$ – maximize $\quad v(z_j) = a_j^T S a_j$

& $a_j^T a_j = 1 \quad$ is $\quad a_j^T a_j - 1 = 0$

$$L = a_j^T S a_j - \lambda (a_j^T a_j - 1)$$

$$\frac{\partial L}{\partial a_j} = 0 \Longrightarrow 2 a_j S - 2 \lambda a_j = 0$$

$$(S - \lambda I) a_j = 0 \qquad (S - \lambda I) = 0$$

Q. $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ sample size $n = 20$

$$S = \begin{bmatrix} 9 & 3 \\ 3 & 4 \end{bmatrix}$$

1. Find the first principle component Y_1
2. Find the variance explained by the first principal component

$$|S - \lambda I| = 0$$

$$\begin{vmatrix} 9-\lambda & 3 \\ 3 & 4-\lambda \end{vmatrix} = 0 \qquad (9-\lambda)(4-\lambda) - 9 = 0$$

$$\lambda^2 - 13\lambda + 27 = 0$$

$$\lambda_1 = 10.41 \qquad \lambda_2 = 2.59$$

$$\begin{pmatrix} 9-10.41 & 3 \\ 3 & 4-10.41 \end{pmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$a_{11} = 2.13 a_{12}$ | $a_{11} = 2.14 a_{12}$

infinitely many solutions

use orthogonality condition

$a_j^T a_j = 1$ $\quad \begin{bmatrix} a_{11} & a_{12} \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = 1$

$$a_{11}^2 + a_{12}^2 = 1$$

$$\begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = \begin{bmatrix} 0.89 \\ 0.42 \end{bmatrix}$$

First PC $Y_1 = 0.89 X_1 + 0.42 X_2$

Second PC $Y_2 = -0.42 X_1 + 0.89 X_2$

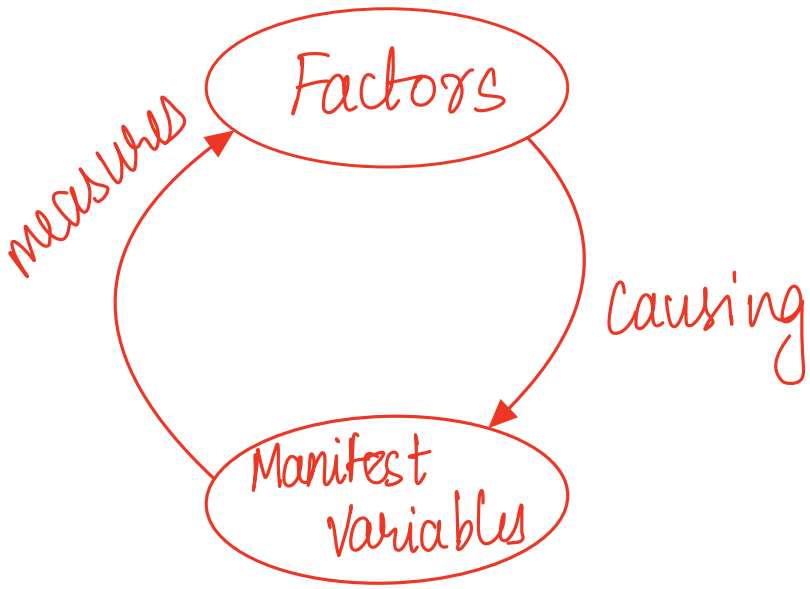
A2. $\left(\dfrac{\text{Eigen value of } 1^{st} \text{ PC}}{\text{sum of eigen values}} \right) \times 100$

$$= \left(\frac{10.41}{10.41 + 2.59} \right) \times 100 = \underline{\underline{80.01\%}}$$

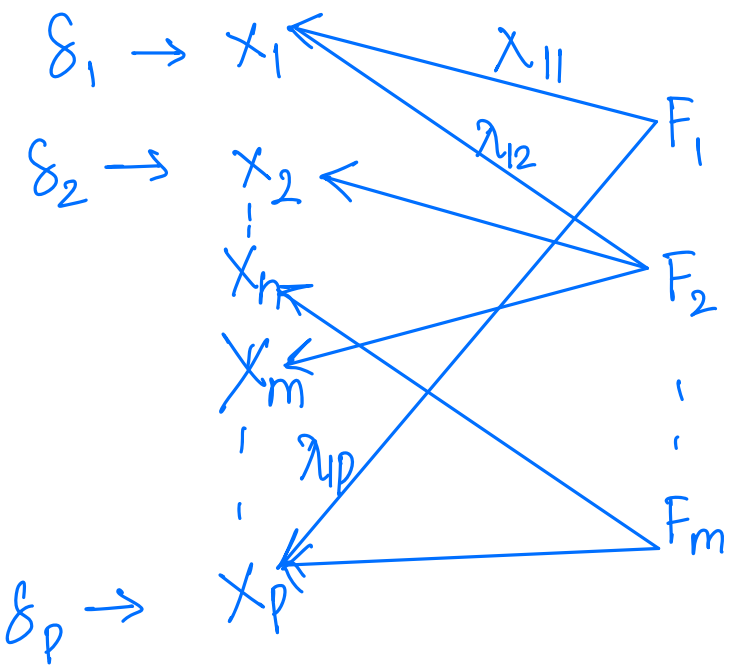
$$X = \begin{bmatrix} 4 & 11 \\ 8 & 4 \\ 13 & 5 \\ 7 & 14 \end{bmatrix}$$

Factor Analysis

→ measuring unobservables, concepts, constructs



→ By observing manifest variables quantifying Factors



$$X_1 = \mu_1 + \lambda_{11}F_1 + \lambda_{12}F_2 \; - - - \; \lambda_{1m}F_m$$

$$X_2 = \mu_2 + \lambda_{21}F_1 + \; - \; - \; - \; \lambda_{2m}F_m$$

$$\vdots$$

$$X_p = \mu_m + \lambda_{p1}F_1 + \; - - - \; \lambda_{pm}F_m$$

In matrix form:

$$X = \mu + \lambda F + \delta \quad \longleftarrow \text{factor model}$$

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \qquad F = \begin{bmatrix} F_1 \\ \vdots \\ F_n \end{bmatrix}$$

$$\delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_p \end{bmatrix} \qquad \lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & - & - & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & - & - & \lambda_{2m} \\ \vdots & \vdots & & & \vdots \\ \lambda_{p1} & \lambda_{p2} & & & \lambda_{pm} \end{bmatrix}$$

$$\boxed{\begin{aligned} \because\ X - \mu &= \lambda F + \delta \\ E(X-\mu) &= 0 \end{aligned}}$$

① $E(F) = 0$

② $E(\delta) = 0$

③ $\mathrm{cov}(F) = E(FF^T) = I$

④ $\mathrm{cov}(\delta) = \chi = \begin{bmatrix} \chi_{11} & - & - & 0 \\ 0 & \chi_{22} & - & 0 \\ \vdots & & \ddots & \vdots \\ 0 & - & - & \chi_{pp} \end{bmatrix}$

2 types of factor model

→ exploratory factor model

→ confirmatory factor model

$$\mathrm{cov}(x) = \Sigma = E[(x-\mu)(x-\mu)^T]$$

use

$$x - \mu = \lambda F + \delta$$

After simplification

$$\Sigma = \Lambda\Lambda^T + \chi$$

$$\Downarrow$$

$$\sigma_{jj} = \sum_{k=1}^{m} \lambda_{jk}^2 + \chi_{jj}$$

communality (h_j^2)

specific variance of x_j

m common factors

Estimation of λ by PCA

$$\Sigma = \lambda\lambda^T + \chi$$

↳ estimator

$$S = \hat{\lambda}\hat{\lambda}^T + \hat{\chi}$$

$Q_j \rightarrow$ eigen values

$e_j \rightarrow$ eigen vector

after manipulation

$$\Lambda = \left[\sqrt{Q_1}e_1 : \sqrt{Q_2}e_2 : \; - - - - \sqrt{Q_m}e_m\right.$$

→ The first factor, $F_1 = \sqrt{Q_1}e_1$

$F_2 = \sqrt{Q_2}\,e_2$

$\vdots$

where $e_1 = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1p} \end{bmatrix}$

When samples are given $S = \hat{\Lambda}\hat{\Lambda}^T + \hat{\chi}$

$$\therefore \hat{\chi} = S - \hat{\Lambda}\hat{\Lambda}^T$$

<u>Steps included</u>

→ Find S

→ Find out Eigen values and Eigen vectors of S

→ Select m (no. of factors)

→ Obtain loading matrix $\hat{\Lambda}$

Problem:

In an orthogonal exploratory factor model (assume m = 1 factor model), suppose the manifest variables X_1, X_2, and X_3 have the sample covariance matrix (S) as

$$\begin{bmatrix} 100 & 40 & 0 \\ 40 & 60 & 0 \\ 0 & 0 & 30 \end{bmatrix}$$

The estimated Eigen values of the covariance matrix (S) are

$\hat{Q}_1 = 124.72$, $\hat{Q}_2 = 35.28$, and $\hat{Q}_3 = 30$

a) Find the estimated Eigen vector corresponding to the largest Eigen value $\hat{Q}_1$

b) Find the estimated loading matrix $\hat{\Lambda}$

c) Find the communality for the variable X_2

d) Find the estimated specific variance matrix $\hat{\psi}$

e) Find the proportion of the variability explained by the common factor

$\hat{Q}_1 = 124.72 \quad \hat{Q}_2 = 35.28 \quad \hat{Q}_3 = 30$

a) largest Eigen vector $\hat{Q}_1 = 124.72$

$$(S - \hat{Q}_1 I)\begin{bmatrix} e_{11} \\ e_{21} \\ e_{31} \end{bmatrix} = 0$$

$$\begin{bmatrix} 100 - 124.72 & 40 & 0 \\ 40 & 60 - 124.72 & 0 \\ 0 & 0 & 30 - 124.72 \end{bmatrix}\begin{bmatrix} \hat{e}_{11} \\ \hat{e}_{21} \\ \hat{e}_{31} \end{bmatrix} = 0$$

subjected to $\hat{e}_{11}^2 + \hat{e}_{21}^2 + \hat{e}_{31}^2 = 0$

b) $\hat{\lambda} = \sqrt{Q_1}\, e_1$ $\quad (m=1)$

↳ consider only one factor

$$= \sqrt{124.72} \times \begin{bmatrix} 0.86 \\ 0.53 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 9.60 \\ 5.92 \\ 0 \end{bmatrix}$$

c) communality for the variable X_2 is

$$\hat{\Lambda}\Lambda^T = \begin{bmatrix} 9.60 \\ 5.92 \\ 0 \end{bmatrix} \begin{bmatrix} 9.60 & 5.92 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 92.16 & 56.83 & 0 \\ 56.83 & 35.05 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

d) The estimated specific variance matrix, $\hat{\Psi}$

$$\hat{\Psi} = S - \hat{\Lambda}\hat{\Lambda}^T$$

$$=$$

e) variability explained by the common factor

$$= \frac{\text{Sum of communality for variable } X_1, X_2 X_3}{\text{sum of eigen values}} \times 100$$

$$= \left[\frac{92.16 + 35.05 + 0}{190}\right] \times 100$$

$$= 66.95\%$$

Cluster Analysis

→ Grouping → informal means for assessing dimentionality, identifying the outliers and suggesting interesting hypothesis concerning relationships

Clustering

- Hierarchical: Usually applied when <100 obs
- Non-Hierarchical (K-means): huge no. of observations

Similarity measures

1. Euclidean measures

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \quad \text{and} \quad x_j = \begin{matrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jp} \end{matrix}$$

$$(x_{i1} - x_{j1})^2$$

2. Statistical distance

$$d(i,j) = \sqrt{(x_i - \mu_i)^T S^{-1} (x_j - \mu_j)}$$

$S \rightarrow$ sample variance covariance matrix

3. Minkowski distance

$$d(i,j) = \{ |x_{i1} - x_{j1}|^m + \; - \; . \}^{1/m}$$

Hierarchial joining algorithms

1. Single linkage → nearest neighbour
2. Complete linkage → further neighbour
3. Centroid linkage

Agglomerative Hierarchial Clustering Algo:

Step1: Identify variables(p) and objects(n)

Step2: Collect data $X_{n \times p}$

Step3: Select similarity or disimilarity measures

Step4: obtain distance matrix

Step5: Start with n clusters where each cluster contains a single entity

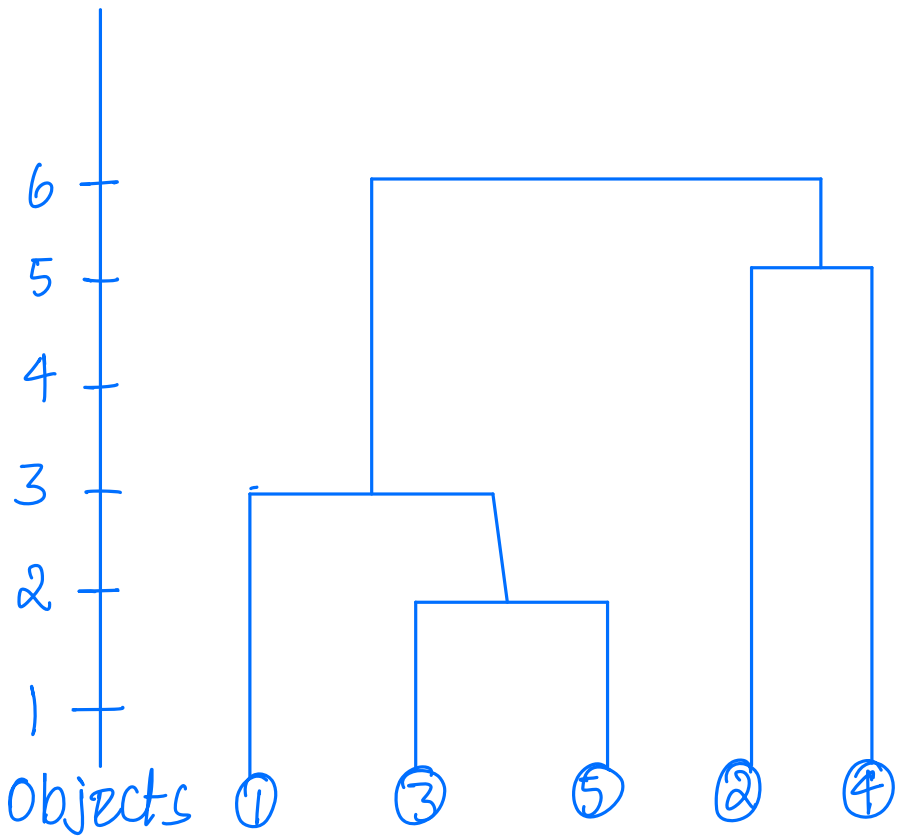
Step6: Find out the nearest pairs of clusters from $D_{n \times n}$
Let distance b/w most similar clusters P & Q be d_{PQ}

Problem: Using single linkage

consider the distances b/w the pair of five objects as follows

Object	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

	(3,5)	1	2	4
(3,5)	0			
1	3	0		
2	7	9	0	
4	8	6	5	0

	(1,3,5)	2	4
(1,3,5)	0		
2	7	0	
4	6	5	0

	(1,3,5)	(2,4)
(1,3,5)	0	
(2,4)	6	0