Multivariate Statistical Analysis

Hyoung-Moon Kim

Professor of Department of Applied Statistics
Konkuk University

Contents

List of Tables

List of Figures

Chapter 1

Matrix Algebra

1.1 Notation

- $m \times n$ (real) matrix
  $\boldsymbol{A} = (a_{ij})\ ,\ i = 1, \cdots, m,\ j = 1, \cdots, n$

- transpose of a matrix
  $\boldsymbol{A}^\top = (a_{ji})\ ,\ j = 1, \cdots, n,\ i = 1, \cdots, m$

- identity matrix

$$\boldsymbol{I} \text{ or } \boldsymbol{I}_n \qquad \text{ex) } \boldsymbol{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

- inverse of a non-singular matrix

  $\boldsymbol{A}^{-1}$ with the property that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$

  ex)

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix} \qquad \boldsymbol{A}^{-1} = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}$$

  $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}_2$ check!

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \qquad \boldsymbol{A}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}.$$

$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}_3$ check!

- product of two matrices
  $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}$, where $\boldsymbol{A}_{m\times p},\ \boldsymbol{B}_{p\times n}$ & $c_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}$

- $n \times n$ matrix of ones
  $\boldsymbol{J}_n = \mathbb{1}_n\mathbb{1}_n^\top$, where $\mathbb{1}_n^\top = (1\ 1\cdots 1)_{1\times n}$

  ex) $\boldsymbol{C}_n = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n,\ \boldsymbol{J}_n - \boldsymbol{I}_n = ?,\ a\boldsymbol{I}_n + b\boldsymbol{J}_n = ?$

$$\boldsymbol{C}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

- All vectors are column vectors

$$\boldsymbol{a} = (a_i) = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix},\ i = 1, \ldots, m$$

exception)

$$\mathbb{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

$\mathbb{0}$; vector or matrix of zeros. Use $\mathbb{0}_m$ or $\mathbb{0}_{m\times n}$ for emphasis.

Properties of vectors:

① $\boldsymbol{a}^\top\boldsymbol{b} = \sum_{i=1}^{m} a_ib_i,\ \boldsymbol{a}^\top\boldsymbol{a} = \sum_{i=1}^{m} a_i^2$

② length of $\boldsymbol{a} = \sqrt{\boldsymbol{a}^\top\boldsymbol{a}} = L_{\boldsymbol{a}}$

③ angle between $\boldsymbol{a}$ and $\boldsymbol{b}$

$$\cos\theta = \frac{\boldsymbol{a}^\top\boldsymbol{b}}{L_{\boldsymbol{a}}L_{\boldsymbol{b}}} = \frac{\sum_{i=1}^{m} a_ib_i}{\sqrt{\sum a_i^2}\sqrt{\sum b_i^2}}$$

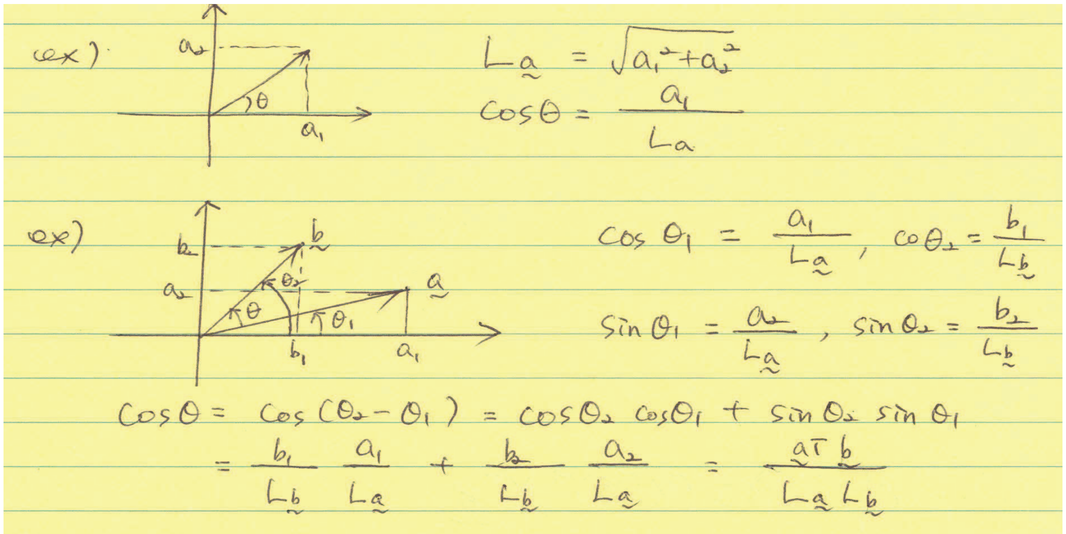

ex)
a_2
θ
a_1
$L_{\underset{\sim}{a}} = \sqrt{a_1^2+a_2^2}$
$\cos\theta = \frac{a_1}{L_a}$
ex)
b_2
$\underset{\sim}{b}$
a_2
θ_2
θ
θ_1
$\underset{\sim}{a}$
b_1
a_1
$\cos\theta_1 = \frac{a_1}{L_{\underset{\sim}{a}}}$, $\text{co}\,\theta_2 = \frac{b_1}{L_{\underset{\sim}{b}}}$
$\sin\theta_1 = \frac{a_2}{L_{\underset{\sim}{a}}}$, $\sin\theta_2 = \frac{b_2}{L_{\underset{\sim}{b}}}$
$\cos\theta = \cos(\theta_2 - \theta_1) = \cos\theta_2\cos\theta_1 + \sin\theta_2\sin\theta_1$
$= \frac{b_1}{L_{\underset{\sim}{b}}}\frac{a_1}{L_{\underset{\sim}{a}}} + \frac{b_2}{L_{\underset{\sim}{b}}}\frac{a_2}{L_{\underset{\sim}{a}}} = \frac{\underset{\sim}{a}^T\underset{\sim}{b}}{L_{\underset{\sim}{a}}L_{\underset{\sim}{b}}}$

Figure 1.1.1: Two examples

④ orthogonal vectors ($\perp$)

$$\boldsymbol{a}^\top\boldsymbol{b} = 0 \qquad i.e\ \ \theta = \frac{\pi}{2}$$

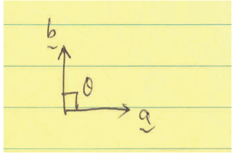


Figure 1.1.2: Orthogonal vectors

⑤ $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k$: linearly dependent if $\exists\, c_1, c_2, \cdots, c_k$, not all zero, such that

$$c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2 + \cdots + c_k\boldsymbol{x}_k = \mathbb{0}$$

i) ⇒ at least one vector can be written as a linear combination of the other vectors

ii) not linearly dep. ⇔ linearly independent (LIN) ex)

$$\boldsymbol{x}_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad \boldsymbol{x}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad \boldsymbol{x}_3 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \ ; \ \text{LIN}$$

⑥ Projection of a vector $\boldsymbol{y}$ on a vector $\boldsymbol{x}$

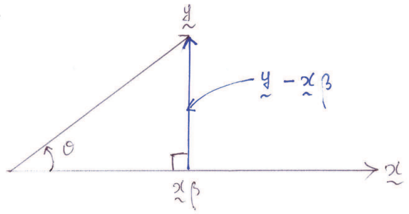

$\underset{\sim}{y}$
$\underset{\sim}{y} - \underset{\sim}{x}\beta$
θ
$\underset{\sim}{x}$
$\underset{\sim}{x}\beta$

Figure 1.1.3: Projection

Note that

$$\boldsymbol{x} \perp (\boldsymbol{y} - \boldsymbol{x}\beta) \Leftrightarrow \boldsymbol{x}^\top(\boldsymbol{y} - \boldsymbol{x}\beta) = 0$$
$$\boldsymbol{x}^\top\boldsymbol{y} = \boldsymbol{x}^\top\boldsymbol{x}\beta \Rightarrow \boldsymbol{(x^\top x)^{-1}x^\top y} = \beta$$
$$\therefore \boldsymbol{x}\beta = \boldsymbol{x(x^\top x)^{-1}x^\top y}$$

cf. Interpretation of LS estimator as a projection

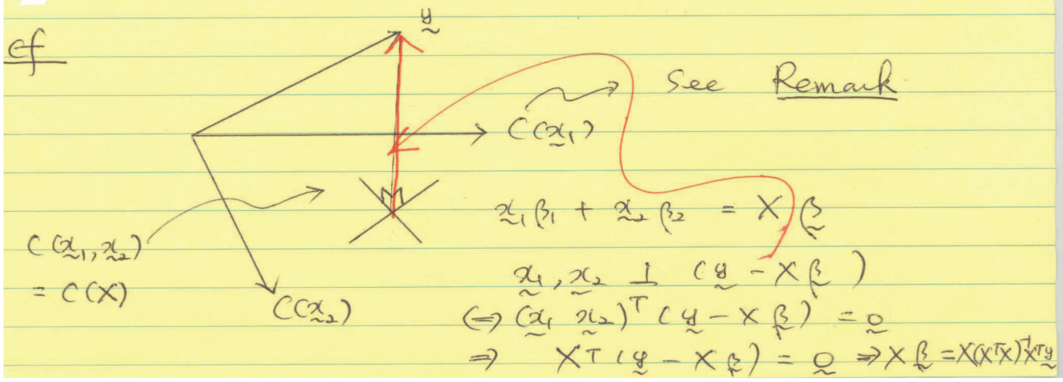

cf
$\underset{\sim}{y}$
See Remark
$C(\underset{\sim}{x}_1)$
$\underset{\sim}{x}_1\beta_1 + \underset{\sim}{x}_2\beta_2 = X\underset{\sim}{\beta}$
$C(\underset{\sim}{x}_1, \underset{\sim}{x}_2) = C(X)$
$\underset{\sim}{x}_1, \underset{\sim}{x}_2 \perp (\underset{\sim}{y} - X\underset{\sim}{\beta})$
$C(\underset{\sim}{x}_2)$
$\Leftrightarrow (\underset{\sim}{x}_1\ \underset{\sim}{x}_2)^T(\underset{\sim}{y} - X\underset{\sim}{\beta}) = \underset{\sim}{0}$
$\Rightarrow X^T(\underset{\sim}{y} - X\underset{\sim}{\beta}) = \underset{\sim}{0} \Rightarrow X\underset{\sim}{\beta} = X(X^TX)^{-1}X^T\underset{\sim}{y}$

Figure 1.1.4: LSE as a projection

Remark
$\boldsymbol{X}_{n\times p}(n \geq p),\ \boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$
Column space of $\boldsymbol{X}$: $C(\boldsymbol{X}) = \{\boldsymbol{x}_1 c_1 + \cdots + \boldsymbol{x}_p c_p;\ c_i^{\prime} s$ are real numbers $\} = \{\boldsymbol{X}\boldsymbol{c} : \boldsymbol{c} \in \mathbb{R}^p\}$

- Symmetric matrix if $a_{ij} = a_{ji}$

- Partitioned form of a matrix

  $\boldsymbol{A}_{m\times n} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n),$ where $\boldsymbol{a}_j,\ j = 1, \cdots, n;$ col. vector of length m

  $$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix},$$

  where $\boldsymbol{A}_{ij}$ is a rectangular matrix of appropriate dimension.

  ex)

  $$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & | & 5 \\ 3 & 4 & | & 6 \\ -- & -- & -- & -- \\ 7 & 8 & | & 9 \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}$$

1.2 Rank of a matrix

$r(\boldsymbol{A}) = \#$ of linearly indep. rows or columns of $\boldsymbol{A}$

- $r(\boldsymbol{AB}) \leq min(r(\boldsymbol{A}), r(\boldsymbol{B}))$

- $r(\boldsymbol{A} + \boldsymbol{B}) \leq r(\boldsymbol{A}) + r(\boldsymbol{B})$

- $r(\boldsymbol{A}) = r(\boldsymbol{A}^{\top}) = r(\boldsymbol{A}^{\top}\boldsymbol{A}) = r(\boldsymbol{A}\boldsymbol{A}^{\top})$

1.3 Trace of a matrix

$tr(\boldsymbol{A})$ = sum of diagonal elements = $\sum_{i=1}^{n} a_{ii}$

- Note that $\boldsymbol{A}$ is a square matrix, *i.e.* $\boldsymbol{A}_{n\times n}$
- $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$

Note : Determinant of $m \times m$ (square) matrix $\boldsymbol{A}, |\boldsymbol{A}|$ or $det(\boldsymbol{A})$

- $det(\boldsymbol{A}) = det(\boldsymbol{A}^\top),\ det(\boldsymbol{AB}) = det(\boldsymbol{A})det(\boldsymbol{B})$
- $det(\boldsymbol{A}) \neq 0 \ \Leftrightarrow\ \exists\ \boldsymbol{A}^{-1};\ det(\boldsymbol{A}^{-1}) = \frac{1}{det(\boldsymbol{A})}$
- $(\boldsymbol{A}^{-1})_{i,j} = (-1)^{i+j}\,\dfrac{det(\boldsymbol{M}_{ji})}{det(\boldsymbol{A})}$, where $\boldsymbol{M}_{ji}$ is obtained from $\boldsymbol{A}$ by deleting j^{th} row & i^{th} col

ex)

$$\boldsymbol{A} = \begin{pmatrix} 7 & 3 \\ 4 & 6 \end{pmatrix} \Rightarrow\ |A| = 7\cdot 6 - 3\cdot 4 = 30$$

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \Rightarrow\ |A| = a_{11}a_{22} - a_{12}a_{21}$$

ex)

$$\begin{aligned} |\boldsymbol{A}| &= \begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{vmatrix} = 1\cdot(+1)\begin{vmatrix} 5 & 6 \\ 8 & 10 \end{vmatrix} + 2\cdot(-1)\begin{vmatrix} 4 & 6 \\ 7 & 10 \end{vmatrix} + 3\cdot(+1)\begin{vmatrix} 4 & 5 \\ 7 & 10 \end{vmatrix} \\ &= \cdots = -3 \end{aligned}$$

$$\begin{aligned} |\boldsymbol{A}| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21} \\ &- (a_{31}a_{22}a_{13} + a_{11}a_{32}a_{23} + a_{12}a_{21}a_{33}) \end{aligned}$$

ex)

$$\begin{aligned} |\boldsymbol{A}| &= \sum_{j=1}^{n} a_{ij}(-1)^{i+j}|\boldsymbol{M}_{ij}|, \text{ for any } \ i \\ &= \sum_{i=1}^{n} a_{ij}(-1)^{i+j}|\boldsymbol{M}_{ij}|, \text{ for any } \ j, \end{aligned}$$

where $\boldsymbol{M}_{ij}$ is obtained from $\boldsymbol{A}$ by deleting i^{th} row and j^{th} column.

ex)

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix}$$

$$\begin{aligned} \left(\boldsymbol{A}^{-1}\right)_{1,1} &= (-1)^{1+1}\begin{vmatrix} 5 & 6 \\ 8 & 10 \end{vmatrix}/|\boldsymbol{A}| = 2/(-3) \\ \left(\boldsymbol{A}^{-1}\right)_{1,2} &= (-1)^{1+2}\begin{vmatrix} 2 & 3 \\ 8 & 10 \end{vmatrix}/|\boldsymbol{A}| = 4/(-3) \end{aligned}$$

Find $\boldsymbol{A}^{-1}$?

1.4 Eigenvalues and Eigenvectors

- $\boldsymbol{Ap} = \lambda \boldsymbol{p}$ for some non zero vector $\boldsymbol{p}$ & for someλ, where $\boldsymbol{A}$ is $m \times m$ matrix.

  $\Leftrightarrow \lambda$: eigenvalue of $\boldsymbol{A}$, $\boldsymbol{p}$: eigenvector of $\boldsymbol{A}$ associated with λ
  $\Leftrightarrow |\boldsymbol{A} - \lambda \boldsymbol{I}| = 0,\ \boldsymbol{Ap} = \lambda \boldsymbol{p}$ for $\boldsymbol{p} \neq \boldsymbol{0}$

  ex)

  $$\boldsymbol{A} = \begin{pmatrix} 1 & 4 \\ 9 & 1 \end{pmatrix}$$

  i) eigenvalue

  $$|\boldsymbol{A} - \lambda \boldsymbol{I}| = \left|\begin{pmatrix} 1 & 4 \\ 9 & 1 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\right| = \left|\begin{pmatrix} 1-\lambda & 4 \\ 9 & 1-\lambda \end{pmatrix}\right|$$

$\Rightarrow\ (1-\lambda)^2 - 36 = 0$ *i.e.* $\lambda = -5$ or 7

ii) eigenvector

$\boldsymbol{Ap} = \lambda \boldsymbol{p} \Leftrightarrow$

$$\begin{pmatrix} 1 & 4 \\ 9 & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} \lambda p_1 \\ \lambda p_2 \end{pmatrix}$$

$\Leftrightarrow p_1 + 4p_2 = \lambda p_1$
$\quad 9p_1 + p_2 = \lambda p_2$
$\Leftrightarrow (1-\lambda)p_1 + 4p_2 = 0$
$\quad 9p_1 + (1-\lambda)p_2 = 0$

$\underline{\lambda = -5}$

$$\begin{array}{l} 6p_1 + 4p_2 = 0 \\ 9p_1 + 6p_2 = 0 \end{array} \quad \Rightarrow \quad \boldsymbol{p} = \begin{pmatrix} 2 \\ -3 \end{pmatrix}$$

Similarly for $\underline{\lambda = 7}$ $\Rightarrow$

$$\boldsymbol{p} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

iii) Double check

$$\boldsymbol{Ap} = \lambda \boldsymbol{p} \Leftrightarrow \begin{pmatrix} 1 & 4 \\ 9 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ -3 \end{pmatrix} \overset{?}{=} -5 \begin{pmatrix} 2 \\ -3 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 4 \\ 9 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \end{pmatrix} \overset{?}{=} 7 \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

Since $\boldsymbol{Ap}_i = \lambda_i \boldsymbol{p}_i,$ define $\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_m)$ then

$\boldsymbol{AP} = \boldsymbol{A}(\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_m) = (\boldsymbol{Ap}_1\ \boldsymbol{Ap}_2\ \cdots\ \boldsymbol{Ap}_m) = (\lambda_1 \boldsymbol{p}_1\ \lambda_2 \boldsymbol{p}_2\ \cdots\ \lambda_m \boldsymbol{p}_m)$

$$= (\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_m) \begin{pmatrix} \lambda_1 & & & \phi \\ & \lambda_2 & & \\ & & \ddots & \\ \phi & & & \lambda_m \end{pmatrix}$$

$$= \boldsymbol{P}\boldsymbol{\Lambda}, \text{ where, } \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & & & \phi \\ & \lambda_2 & & \\ & & \ddots & \\ \phi & & & \lambda_m \end{pmatrix}$$

- $\mathrm{tr}(\boldsymbol{A}) = \sum_i \lambda_i$
  $\mathrm{r}(\boldsymbol{A}) = \#$ of nonzero eigenvalues
  $|\boldsymbol{A}| = \Pi_i \lambda_i$

- In R, use `eigen` function to compute eigenvalues and eigenvectors of numeric (double, integer, logical) or complex matrices.

- Eigenvalues & eigenvectors of symm. matrices
  $\boldsymbol{A}$: symm. $\Rightarrow$ eigenvalues & eigenvectors are real numbers

  Properties : Assume $\boldsymbol{p}^\top \boldsymbol{p} = 1$

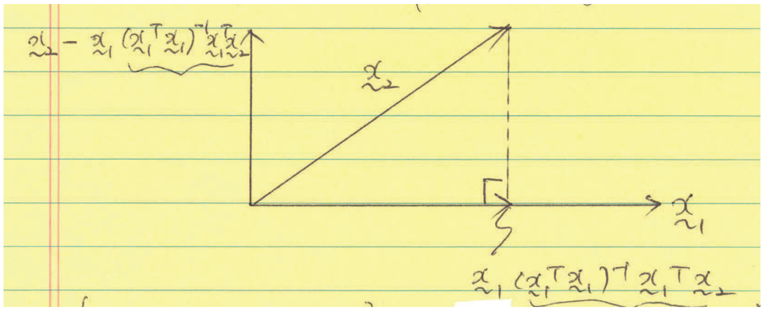
  i) $\lambda_1 \neq \lambda_2 \;\Rightarrow\; \boldsymbol{p}_1^\top \boldsymbol{p}_2 = 0$
  If all eigenvalues are distinct, then $\boldsymbol{P}$ is nonsingular and $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}$

  ii) If $\lambda_1 = \lambda_2 = \cdots = \lambda_r$, then the associated eigenvectors are **LIN but not always orthogonal**. However, by Gram-schmidt orthogonalization, we may always assume that $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}$.

  $\therefore\ \boldsymbol{P}^\top \boldsymbol{A} \boldsymbol{P} = \boldsymbol{\Lambda}$: **spectral decomposition**
  $(\because)\ \boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\Lambda} \;\Rightarrow\; \boldsymbol{P}^\top \boldsymbol{A}\boldsymbol{P} = \boldsymbol{\Lambda}$ since $\boldsymbol{P}^{-1} = \boldsymbol{P}^\top$

  Note $\boldsymbol{A}^{-1} = (\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top)^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^\top$.

cf. **Gram-schmidt process of orthogonalization**


$x_2 - \underbrace{x_1(x_1^\top x_1)^{-1}x_1^\top}x_2$
x_2
x_1
$x_1(x_1^\top x_1)^{-1}x_1^\top x_2$

$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p\}$; LIN vector in $\mathbb{R}^m$

$\boldsymbol{y}_1 = \boldsymbol{x}_1$
$\boldsymbol{y}_2 = \boldsymbol{x}_2 - \boldsymbol{y}_1(\boldsymbol{y}_1^\top \boldsymbol{y}_1)^{-1}\boldsymbol{y}_1^\top \boldsymbol{x}_2$
$\vdots$
$\boldsymbol{y}_k = \boldsymbol{x}_k - \sum_{j=1}^{k-1} \boldsymbol{y}_j(\boldsymbol{y}_j^\top \boldsymbol{y}_j)^{-1}\boldsymbol{y}_j^\top \boldsymbol{x}_k,\ k = 2, \cdots, p$

$\boldsymbol{b}_k = \frac{\boldsymbol{y}_k}{\sqrt{\boldsymbol{y}_k^\top \boldsymbol{y}_k}}$ (normalization), $k = 1, \cdots, p$
$\Rightarrow$

i) $\boldsymbol{y}_i^\top \boldsymbol{y}_j = 0 \quad {}^\forall i \neq j$ (orthogonal)

ii) $\boldsymbol{b}_i^\top \boldsymbol{b}_j = 0 \quad {}^\forall i \neq j$ & $\sqrt{\boldsymbol{b}_i^\top \boldsymbol{b}_i} = 1$ (orthonormal)

1.5 Quadratic forms & definite Matrices

$q(\boldsymbol{y}) = \boldsymbol{y}^\top \boldsymbol{A} \boldsymbol{y} = \sum_i \sum_j a_{ij} y_i y_j$: quadratic form

W.L.O.G we assume that A: symm

ex)

$$\begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$= a_{11}y_1^2 + a_{22}y_2^2 + a_{33}y_3^2 + (a_{12}+a_{21})y_1y_2 + (a_{13}+a_{31})y_1y_3 + (a_{23}+a_{32})y_2y_3$

ex)

$$\begin{pmatrix} y_1 & y_2 \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$= ay_1^2 + cy_2^2 + 2by_1y_2$

- $\boldsymbol{A}$: positive definite (p.d.) $\Leftrightarrow$ $\boldsymbol{y}^\top \boldsymbol{Ay} > 0 \ \forall \ \boldsymbol{y} \neq \mathbb{0}$
  $\Leftrightarrow$ All eigenvalues of $\boldsymbol{A}$ are > 0
  $\Leftrightarrow$ $\boldsymbol{A} = \boldsymbol{BB}^\top$ for some nonsingular $\boldsymbol{B}$
  ($\boldsymbol{B}$ can be taken as a symm. matrix)
  $\Leftrightarrow$ $\boldsymbol{A}$: n.n.d. & $\boldsymbol{A}$ is nonsingular

- $\boldsymbol{A}$: nonnegative definite (n.n.d.) or positive semi-definite (p.s.d.)
  $\Leftrightarrow$ $\boldsymbol{y}^\top \boldsymbol{Ay} \geq 0 \ \forall \ \boldsymbol{y} \neq \mathbb{0}$
  $\Leftrightarrow$ All eigenvalues of $\boldsymbol{A}$ are ≥ 0
  $\Leftrightarrow$ $\boldsymbol{A} = \boldsymbol{BB}^\top$ for some $\boldsymbol{B}$
  ($\boldsymbol{B}$ can be taken as a symmetric matrix)
  ($\boldsymbol{B}$ can be taken as a rank($\boldsymbol{B}$)=rank($\boldsymbol{A}$))
  ($\boldsymbol{B}$ can be taken as a full col. rank)
  $\Leftrightarrow$ $\boldsymbol{A}$ is a covariance matrix of a random vector

ex)

$$\boldsymbol{y}^\top \boldsymbol{Ay} = (y_1 \ y_2 \ y_3) \begin{pmatrix} 3 & 5 & 1 \\ 5 & 13 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$= 3y_1^2 + 13y_2^2 + y_3^2 + 10y_1y_2 + 2y_1y_3$
$= (y_1 + 2y_2)^2 + (y_1 + 3y_2)^2 + (y_1 + y_3)^2 > 0$ if $y_1 = y_2 = y_3 \neq 0$

$\therefore \boldsymbol{A}$: *p.d.*

ex)

$$\boldsymbol{y}^\top \boldsymbol{Ay} = (y_1 \ y_2 \ y_3) \begin{pmatrix} 37 & -2 & -24 \\ -2 & 13 & -3 \\ -24 & -3 & 17 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$= 37y_1^2 + 13y_2^2 + 17y_3^2 - 4y_1y_2 - 48y_1y_3 - 6y_2y_3$
$= (6y_1 - 4y_3)^2 + (y_1 - 2y_2)^2 + (3y_2 - y_3)^2 \geq 0$
Note that $\boldsymbol{y}^\top = (2, 1, 3) \;\Rightarrow\; \boldsymbol{y}^\top \boldsymbol{A} \boldsymbol{y} = 0$
$\therefore \boldsymbol{A}$: *p.s.d.*

- $\boldsymbol{A}$: *p.d.* $\Rightarrow q(\boldsymbol{y}) = c$ defines an ellipsoid.

  cf. An ellipsoid is a type of quadric surface that is a higher dimensional analogue of an ellipse.

  ex) $dim = 2 \;\Rightarrow\; ellipse \; i.e. \; \dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = 1$

- $\boldsymbol{P}$: matrix of orthogonal eigenvectors of $\boldsymbol{A}$
  Let $\boldsymbol{y} = \boldsymbol{P}\boldsymbol{z}$ $(\boldsymbol{z} = \boldsymbol{P}^\top \boldsymbol{y})$
  then $\boldsymbol{y}^\top \boldsymbol{A} \boldsymbol{y} = \boldsymbol{z}^\top \boldsymbol{P}^\top \boldsymbol{A} \boldsymbol{P} \boldsymbol{z} = \boldsymbol{z}^\top \boldsymbol{\Lambda} \boldsymbol{z} = \sum_i \lambda_i \; z_i^2$ (for $^\forall \; i$)

  Thus, in terms of the new coordinates, the axes of the ellipsoid $\boldsymbol{z}^\top \boldsymbol{\Lambda} \boldsymbol{z} = c$ are along the coordinate axes with radii $\sqrt{\frac{c}{\lambda_i}}$. The shape of the ellipsoid is not changed by this orthogonal transformation.

- $\boldsymbol{A}$: p.d. with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ & associated normalized eigenvectors $\boldsymbol{e}_1, \; \boldsymbol{e}_2, \; \cdots, \; \boldsymbol{e}_p$.
  $\Rightarrow$

  i) $\max_{\boldsymbol{y} \neq \mathbb{0}} \dfrac{\boldsymbol{y}^\top \boldsymbol{A} \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}} = \lambda_1$ (attained when $\boldsymbol{y} = \boldsymbol{e}_1$)

  ii) $\min_{\boldsymbol{y} \neq \mathbb{0}} \dfrac{\boldsymbol{y}^\top \boldsymbol{A} \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}} = \lambda_p$ (attained when $\boldsymbol{y} = \boldsymbol{e}_p$)

- $r(\boldsymbol{A}_{n \times p}) = p \;\Rightarrow\; \boldsymbol{A}^\top \boldsymbol{A}$: p.d.
  $r(\boldsymbol{A}_{n \times p}) = r < p \;\Rightarrow\; \boldsymbol{A}^\top \boldsymbol{A}$: p.s.d.

- Spectral decomposition since it is symm.
  $\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\Lambda} \Rightarrow \boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top = \sum \lambda_i \boldsymbol{p}_i \boldsymbol{p}_i^\top$
  $\boldsymbol{A}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^\top = \sum \frac{1}{\lambda_i} \boldsymbol{p}_i \boldsymbol{p}_i^\top$

  To emphasize $\boldsymbol{p}_i$'s are orthonormal, we use $\boldsymbol{e}_i$ instead of $\boldsymbol{p}_i$, *i.e.*,

$$\boldsymbol{A} = \sum \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^\top = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top, \text{ where } \boldsymbol{P} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_k]$$

1.6 Special matrices

- orthogonal matrices
  $\boldsymbol{P}_{m\times m}$: orthogonal if $\boldsymbol{P}^{-1} = \boldsymbol{P}^\top$

  i) $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$ (rows & cols are orthogonal & have length 1)
  ii) $|\boldsymbol{P}| = \pm 1$
  iii) $-1 \leq p_{ii} \leq 1$

  <u>Note</u> The matrix, $\boldsymbol{P}$, consisting of eigenvectors of a symm. matrix, $\boldsymbol{A}$, is orthogonal.

- Idempotent matrices
  $A_{n\times n}$: idempotent if $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$

  i) $r(\boldsymbol{A}) = n \Rightarrow \boldsymbol{A} = \boldsymbol{I}$
  ii) The non-zero eigenvalues of $\boldsymbol{A}$ are equal to one.
  ($\boldsymbol{A}$: symm & idemp. $\Leftrightarrow$ the nonzero eigenvalues are all equal to one.)
  iii) $tr(\boldsymbol{A}) = r(\boldsymbol{A})$
  iv) $\boldsymbol{A}$: symm & idemp $\Rightarrow$ A is at least p.s.d
  v) Suppose that $\boldsymbol{A}_i, i = 1, \cdots, m,$ are symm., $\boldsymbol{A} = \sum \boldsymbol{A}_i,\ r(\boldsymbol{A}_i) = r_i,$ and $r(\boldsymbol{A}) = r.$
  Then, if <u>$\boldsymbol{A}$ is idemp. & $r = \sum r_i$</u>, it can be shown that each $\boldsymbol{A}_i$ is idemp. & $\boldsymbol{A}_i\boldsymbol{A}_j = 0\ {}^\forall i \neq j$

- For a positive definite matrix $\boldsymbol{A}_{m\times m},$ a squre-root matrix of $\boldsymbol{A}$ is

$$\boldsymbol{A}^{\frac{1}{2}} = \sum_{i=1}^{m} \sqrt{\lambda_i}\boldsymbol{e}_i\boldsymbol{e}_i^\top = \boldsymbol{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}^\top,$$

$$\text{where } \boldsymbol{\Lambda}^{\frac{1}{2}} = \text{diag}\left(\sqrt{\lambda_1}\ \cdots\ \sqrt{\lambda_m}\right)$$

  <u>Note that</u>
  i) $\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{A}$
  ii) $(\boldsymbol{A}^{\frac{1}{2}})^\top = \boldsymbol{A}^{\frac{1}{2}}$

iii) $(\boldsymbol{A}^{\frac{1}{2}})^{-1} = \sum_{i=1}^{m} \frac{1}{\sqrt{\lambda_i}} \boldsymbol{e}_i \boldsymbol{e}_i^\top$
iv) $\boldsymbol{A}^{\frac{1}{2}} \boldsymbol{A}^{-\frac{1}{2}} = \boldsymbol{A}^{-\frac{1}{2}} \boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{I}$
$\boldsymbol{A}^{-\frac{1}{2}} = (\boldsymbol{A}^{\frac{1}{2}})^{-1}$

1.7 Decomposition of a matrix

- Diagonalization (spectral decomposition) of a real symm. matrix

$\boldsymbol{A} : n \times n$ real symm.
$\Rightarrow$
$\exists\, \boldsymbol{P}$ real matrix : $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$ such that $\boldsymbol{P}^\top \boldsymbol{A} \boldsymbol{P} = \mathrm{diag}(\lambda_i)$,
or equivalently $\boldsymbol{A} = \sum_{i=1}^{n} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^\top$, where $\lambda_1 \cdots \lambda_n$ & the col vectors $\boldsymbol{e}_1, \cdots \boldsymbol{e}_n$ of $\boldsymbol{P}$ are the eigenvalues &the associated eigenvectors of $\boldsymbol{A}$

- QR (or Householder) decomposition of a matrix

$\boldsymbol{A} :\ m \times n$ real matrix, $m \geq n,\ rank(\boldsymbol{A}) = n$
$\Rightarrow$
$\exists\, \boldsymbol{Q}_{m\times n}\ :\ \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_n$ &
$\exists\, \boldsymbol{R}_{n\times n}\ :$ (upper triangular with $r_{ii} > 0$) (= invertible upper trangular) s.t. $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$

Remark How to solve $\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^\top \boldsymbol{b}$!
Using QR decomposition, we have that $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$

$$\begin{aligned}
&\iff \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^\top \boldsymbol{b} \\
&\iff \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{R}\boldsymbol{x} = \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{b} \\
&\iff \boldsymbol{R}^\top \boldsymbol{R}\boldsymbol{x} = \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{b}\ (\because\ \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_n) \\
&\iff \boldsymbol{R}\boldsymbol{x} = \boldsymbol{Q}^\top \boldsymbol{b}
\end{aligned}$$

ex) $\boldsymbol{A} = \boldsymbol{X}$ (design matrix)

$$\begin{aligned}
& \boldsymbol{X} = \boldsymbol{QR} \\
& \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{y} \text{ ; Normal equations} \\
\iff \; & \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{QR}\hat{\boldsymbol{\beta}} = \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{y} \\
\iff \; & \boldsymbol{R}^\top \boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{y} \text{ (multiply } (\boldsymbol{R})^{-1}) \\
\iff \; & \boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{Q}^\top \boldsymbol{y}
\end{aligned}$$

ex)

$$\boldsymbol{X} = \begin{bmatrix} 3 & -6 \\ 4 & -8 \\ 0 & 1 \end{bmatrix} \text{ and } \boldsymbol{y} = \begin{bmatrix} -1 \\ 7 \\ 2 \end{bmatrix}$$

$$\begin{aligned}
\Longrightarrow \quad & \boldsymbol{X} = \boldsymbol{QR} \text{ with } \boldsymbol{Q} = \begin{bmatrix} 3/5 & 0 \\ 4/5 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \boldsymbol{R} = \begin{bmatrix} 5 & -10 \\ 0 & 1 \end{bmatrix} \\
\Longrightarrow \quad & \boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{Q}^\top \boldsymbol{y} \\
\iff \quad & \begin{bmatrix} 5 & -10 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 3/5 & 4/5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 7 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix} \\
\iff \quad & \hat{\beta}_0 = 5 \text{ and } \hat{\beta}_1 = 2.
\end{aligned}$$

- Singular-value decomposition

$\boldsymbol{A}_{m\times n}$: real matrix
$\Rightarrow$
$\exists\, \boldsymbol{U},\ \boldsymbol{V}$: orthogonal (real) matrices &
$\exists\, \gamma_1 \cdots \gamma_r$: positive #'s (singular value of $\boldsymbol{A}$) s.t.

$$\boldsymbol{A} = \boldsymbol{V} \begin{bmatrix} diag_r(\gamma_i) & 0 \\ 0 & 0 \end{bmatrix} \boldsymbol{U}^\top,$$

where $r = rank(\boldsymbol{A})$ & $\gamma_1^2 \cdots \gamma_r^2$ are the nonzero eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ (or $\boldsymbol{A}\boldsymbol{A}^\top$).

1.8 Kronecker(direct) products of Matrices

$\boldsymbol{A}_{p\times q} \otimes \boldsymbol{B}_{m\times n} = (a_{ij}\boldsymbol{B})_{pm\times qn}$

ex)

$$[1\ 2\ 3]_{1\times 3} \otimes \begin{pmatrix} 6 & 7 \\ 8 & 9 \end{pmatrix}_{2\times 2} = \left[1\begin{pmatrix} 6 & 7 \\ 8 & 9 \end{pmatrix} \ 2\begin{pmatrix} 6 & 7 \\ 8 & 9 \end{pmatrix} \ 3\begin{pmatrix} 6 & 7 \\ 8 & 9 \end{pmatrix} \right]$$

$$= \begin{bmatrix} 6 & 7 & 12 & 14 & 18 & 21 \\ 8 & 9 & 16 & 18 & 24 & 27 \end{bmatrix}_{(1\cdot 2)\times(3\cdot 2)=2\times 6}$$

- $(\boldsymbol{A}\otimes\boldsymbol{B})^\top = \boldsymbol{A}^\top \otimes \boldsymbol{B}^\top \ \Leftarrow\ (\boldsymbol{AB})^\top = \boldsymbol{B}^\top\boldsymbol{A}^\top$
  $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1} \ \Leftarrow\ (\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$
  $tr(\boldsymbol{A}\otimes\boldsymbol{B}) = tr(\boldsymbol{A})tr(\boldsymbol{B}) \ \Leftarrow\ tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$
  $\boldsymbol{A}\otimes(\boldsymbol{B}\otimes\boldsymbol{C}) = (\boldsymbol{A}\otimes\boldsymbol{B})\otimes\boldsymbol{C}$
  $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D}) = \boldsymbol{AC}\otimes\boldsymbol{BD}$
  $(\boldsymbol{A}+\boldsymbol{B})\otimes(\boldsymbol{C}+\boldsymbol{D}) = \boldsymbol{A}\otimes\boldsymbol{C}+\boldsymbol{A}\otimes\boldsymbol{D}+\boldsymbol{B}\otimes\boldsymbol{C}+\boldsymbol{B}\otimes\boldsymbol{D}$
  $\boldsymbol{ab}^\top = \boldsymbol{a}\otimes\boldsymbol{b}^\top = \boldsymbol{b}^\top\otimes\boldsymbol{a}$
- $r(\boldsymbol{A}\otimes\boldsymbol{B}) = r(\boldsymbol{A})r(\boldsymbol{B})$
  The eigenvalues of $\boldsymbol{A}\otimes\boldsymbol{B}$ are the products of the eigenvalues of $\boldsymbol{A}$ & $\boldsymbol{B}$
  That is,
  $\boldsymbol{A}_{m\times m}$ with eigenvalues $\lambda_1\cdots\lambda_m$
  $\boldsymbol{B}_{p\times p}$ with eigenvalues $\theta_1\cdots\theta_p$
  $\Rightarrow (\boldsymbol{A}\otimes\boldsymbol{B})_{mp\times mp}$ with eigenvalues $\lambda_i\theta_j\ :\ i=1,\cdots,m,\ j=1,\cdots,p$
- Let $\boldsymbol{A}$ be and $m\times m$ matrix and $\boldsymbol{B}$ be a $p\times p$ matrix. Then
  $|\boldsymbol{A}\otimes\boldsymbol{B}| = |\boldsymbol{A}|^p|\boldsymbol{B}|^m$.

1.9 Distance

- Euclidean distance

  $\boldsymbol{x}^\top = (x_1\ \cdots\ x_n) \ \Rightarrow\ \sqrt{\boldsymbol{x}^\top\boldsymbol{x}}$: length of a vector $\boldsymbol{x}$
  $\Leftrightarrow$ distance from the origin to the point $\boldsymbol{x}$

$\boldsymbol{y}^\top = (y_1 \ \cdots \ y_n) \ \Rightarrow \ \sqrt{(\boldsymbol{x}-\boldsymbol{y})^\top(\boldsymbol{x}-\boldsymbol{y})}$: distance from $\boldsymbol{x}$ to $\boldsymbol{y}$

Note that all point on a circle are equidistant from the center.

- Generalized distance

  $\boldsymbol{A}$: p.d. $\Rightarrow$ generalized(Mahalanobis) distance is $\sqrt{\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}}$

  This is called the length of $\boldsymbol{x}$ in the metric of $\boldsymbol{A}$

  Note that all points on an ellipsoid are equidistant from the origin. ($\because \ q(\boldsymbol{y}) = c$ defines an elipsoid)

  ex) of $\boldsymbol{A}$: $\hat{\boldsymbol{\Sigma}}^{-1}$ ($= \boldsymbol{S}^{-1}$) (inverse of sample covariance matrix)

1.10 Derivative

- $$\boldsymbol{c}^\top\boldsymbol{x} = (c_1 \ \cdots \ c_p)\begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = \sum_{i=1}^{p} c_i x_i = \boldsymbol{x}^\top\boldsymbol{c}$$

  $$\frac{\partial \boldsymbol{c}^\top\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^\top\boldsymbol{c}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial \boldsymbol{c}^\top\boldsymbol{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \boldsymbol{c}^\top\boldsymbol{x}}{\partial x_p} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix} = \boldsymbol{c}$$

  $$\frac{\partial \boldsymbol{c}^\top\boldsymbol{x}}{\partial \boldsymbol{x}^\top} = \frac{\partial \boldsymbol{x}^\top\boldsymbol{c}}{\partial \boldsymbol{x}^\top} = \boldsymbol{c}^\top$$

- $$\begin{aligned}\frac{\partial \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} &= \frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}^\top \boldsymbol{P}\right) + \frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{Q}\boldsymbol{x}\right) \text{ by the product rule of differentiation} \\ &= \begin{cases} 2\boldsymbol{A}\boldsymbol{x}, & \text{if } \boldsymbol{A}: \text{ symm.} \\ (\boldsymbol{A}+\boldsymbol{A}^\top)\boldsymbol{x}, & \text{if } \boldsymbol{A}: \text{ not symm.,} \end{cases}\end{aligned}$$

  where $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{Q} = \boldsymbol{x}^\top \boldsymbol{A}$.
  Similarly $\frac{\partial \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}^\top} = \begin{cases} 2\boldsymbol{x}^\top \boldsymbol{A}, & \text{if } \boldsymbol{A}: \text{ symm.} \\ \boldsymbol{x}^\top(\boldsymbol{A}^\top + \boldsymbol{A}), & \text{if } \boldsymbol{A}: \text{ not symm..} \end{cases}$

1.11 Matrix inequalities & Maximization

- Caucy-Schwarz inequality :

  Let $\mathbf{b}$ & $\mathbf{d}$ be any two $p \times 1$ vectors. Then

  $$(\mathbf{b}^\top \mathbf{d})^2 \leq (\mathbf{b}^\top \mathbf{b})(\mathbf{d}^\top \mathbf{d})$$

  with equality iff $\mathbf{b} = c\mathbf{d}$(or $\mathbf{d} = c\mathbf{b}$) for some constant c.

  ① C-S inequality $\Longleftrightarrow \left(\sum_{i=1}^p b_i d_i\right)^2 \leq \sum_{i=1}^p b_i^2 \sum_{i=1}^p d_i^2$

  ② Use the inner product of $\mathbf{d} - (\mathbf{b}^\top \mathbf{d}/\mathbf{b}^\top \mathbf{b})\mathbf{b}$ to prove C-S inequality.

- Extended Cauchy-Schwarz inequality:

  Let $\mathbf{b}$ & $\mathbf{d}$ be any two $p \times 1$ vectors, and let $\mathbf{B}_{p\times p}$ be a p.d matrix. Then

  $$(\mathbf{b}^\top \mathbf{d})^2 \leq (\mathbf{b}^\top \mathbf{B}\mathbf{b})(\mathbf{d}^\top \mathbf{B}^{-1}\mathbf{d})$$

  with equality iff $\mathbf{b} = c\mathbf{B}^{-1}\mathbf{d}$ (or $\mathbf{d} = c\mathbf{B}\mathbf{b}$) for some constant c.

- Maximization Lemma

$\mathbf{B}_{p\times p}$: p.d, $\mathbf{d}_{p\times 1}$: a given vector
$\Rightarrow$

$$max_{\boldsymbol{x}\neq \mathbb{0}} \frac{(\boldsymbol{x}^\top \mathbf{d})^2}{\boldsymbol{x}^\top \mathbf{B}\boldsymbol{x}} \ = \ \mathbf{d}^\top \mathbf{B}^{-1}\mathbf{d}$$

with the maximum attained when $\boldsymbol{x} = c\mathbf{B}^{-1}\mathbf{d}$ for any constant $c \neq 0$.

pf) By the extended Cauchy-schwarz inequality,

$$(\boldsymbol{x}^\top \mathbf{d})^2 \leq (\boldsymbol{x}^\top \mathbf{B}\boldsymbol{x})(\mathbf{d}^\top \mathbf{B}^{-1}\mathbf{d}).$$

Because $\boldsymbol{x} \neq \mathbb{0}$ & $\mathbf{B}$ is p.d, we have

$$\frac{(\boldsymbol{x}^\top \mathbf{d})^2}{\boldsymbol{x}^\top \mathbf{B}\boldsymbol{x}} \leq \mathbf{d}^\top \mathbf{B}^{-1}\mathbf{d}$$

$$\therefore max_{\boldsymbol{x}\neq \mathbb{0}} \ \frac{(\boldsymbol{x}^\top \mathbf{d})^2}{\boldsymbol{x}^\top \mathbf{B}\boldsymbol{x}}$$

The bound is attained for $\boldsymbol{x} = c\mathbf{B}^{-1}\mathbf{d}$ □

- Maximization of Quadratic forms

$\mathbf{B}_{p\times p}$: p.d. with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ & associated orthonormalized eigenvectors $\mathbf{e}_1 \cdots \mathbf{e}_p$.

Then

$$max_{\boldsymbol{x}\neq \mathbb{0}} \frac{\boldsymbol{x}^\top \mathbf{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \lambda_1 \text{ (attained when } \boldsymbol{x} = \boldsymbol{e}_1)$$

$$min_{\boldsymbol{x}\neq \mathbb{0}} \frac{\boldsymbol{x}^\top \mathbf{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \lambda_p \text{ (attanined when } \boldsymbol{x} = \boldsymbol{e}_p).$$

Moreover

$$max_{\boldsymbol{x} \perp \mathbf{e}_1 \cdots \mathbf{e}_k} \frac{\boldsymbol{x}^\top \mathbf{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \lambda_{k+1} \text{ (attained when } \boldsymbol{x} = \mathbf{e}_{k+1},\ k = 1, 2, \cdots, p-1).$$

1.12 Matrix Algebra in R

1.12.1 Notation

A vector: $\mathbf{a} = \begin{pmatrix} 1 & 2 & 3 & 4 \end{pmatrix}^\top$

```
> a <- c(1, 2, 3, 4)
> a
[1] 1 2 3 4
> a <- c(1:4)
> a
[1] 1 2 3 4
```

A matrix: $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$

```
> array(c(1:6), dim=c(2,3))
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> matrix(c(1:6), nrow=2, ncol=3)
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> cbind(c(1, 4), c(2, 5), c(3, 6))
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
> rbind(c(1, 2, 3), c(4, 5, 6))
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
```

A matrix: $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 \end{pmatrix} = \begin{pmatrix} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{pmatrix}$

```
> x1 <- c(1:4) ; x2 <- c(5:8) ; x3 <- c(9:12) ; x4 <- c(13:16)
> X <- matrix(c(x1,x2,x3,x4), nrow=4, ncol=4)
> X
     [,1] [,2] [,3] [,4]
[1,]    1    5    9   13
[2,]    2    6   10   14
[3,]    3    7   11   15
[4,]    4    8   12   16
```

Transpose of a vector: $\mathbf{a}^\top = \begin{pmatrix} 1 & 2 & 3 & 4 \end{pmatrix}$

```
> t(a)
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
```

Transpose of a matrix: $\mathbf{X}^\top = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 \end{pmatrix}^\top = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{pmatrix}$

```
> t(X)
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
[4,]   13   14   15   16
```

Identity matrix: $\mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

```
> diag(3)
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

Inverse matrix: $\mathbf{A}^{-1} = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}$ where $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}$

```
> A <- matrix(c(1, 2, 1, 3), 2, 2)
> solve(A)
     [,1] [,2]
[1,]    3   -1
[2,]   -2    1
```

Inverse of a diagonal matrix: $\mathbf{A}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}$ where $\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$

by using elementary row operation $\big[\mathbf{A}|\mathbf{I}\big]\sim\big[\mathbf{I}|\mathbf{A}^{-1}\big] \Longleftrightarrow \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 3 & 0 & 0 & 1 \end{pmatrix}$

$\sim \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 3 & 0 & 0 & 1 \end{pmatrix} : R_2 \longrightarrow \frac{1}{2}R_2 \sim \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{3} \end{pmatrix} : R_3 \longrightarrow \frac{1}{3}R_3$

```
> A <- matrix(c(1, 0, 0, 0, 2, 0, 0, 0, 3), 3, 3)
> solve(A)
     [,1] [,2]      [,3]
[1,]    1  0.0 0.0000000
[2,]    0  0.5 0.0000000
[3,]    0  0.0 0.3333333
```

Product of two matrices: Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$. Then $\mathbf{AB} = \begin{pmatrix} \sum\limits_{k=1}^{2} a_{1k}b_{k1} & \sum\limits_{k=1}^{2} a_{1k}b_{k2} \\ \sum\limits_{k=1}^{2} a_{2k}b_{k1} & \sum\limits_{k=1}^{2} a_{2k}b_{k2} \end{pmatrix} = \begin{pmatrix} 1+3 & 2+4 \\ 2+6 & 4+8 \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ 8 & 12 \end{pmatrix}$. But $\mathbf{BA} = \begin{pmatrix} 5 & 5 \\ 11 & 11 \end{pmatrix}$.

```
> A <- matrix(c(1, 2, 1, 2), 2, 2)
> B <- matrix(c(1, 3, 2, 4), 2, 2)
> A%*%B
     [,1] [,2]
[1,]    4    6
```

```
[2,]    8   12
> B%*%A
     [,1] [,2]
[1,]    5    5
[2,]   11   11
```

Vector and square matrix of ones: $\mathbf{1}_5 = \begin{pmatrix}1 & 1 & 1 & 1 & 1\end{pmatrix}^\top$, $\mathbf{J}_5 = \mathbf{1}_5\mathbf{1}_5^\top = \begin{pmatrix}1&1&1&1&1\\1&1&1&1&1\\1&1&1&1&1\\1&1&1&1&1\\1&1&1&1&1\end{pmatrix}$. Similarly, we can do vector and matrix of zeros.

```
> rep(1, 5)
[1] 1 1 1 1 1
> rep(1, 5)%*%t(rep(1, 5))
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    1    1    1    1
[2,]    1    1    1    1    1
[3,]    1    1    1    1    1
[4,]    1    1    1    1    1
[5,]    1    1    1    1    1
> matrix(c(1),5, 5)
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    1    1    1    1
[2,]    1    1    1    1    1
[3,]    1    1    1    1    1
[4,]    1    1    1    1    1
[5,]    1    1    1    1    1
```

Centering matrix: $\mathbf{C}_3 = \mathbf{I}_3 - \frac{1}{3}\mathbf{J}_3 = \begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} - \frac{1}{3}\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix} = \begin{pmatrix}\frac{2}{3}&-\frac{1}{3}&-\frac{1}{3}\\-\frac{1}{3}&\frac{2}{3}&-\frac{1}{3}\\-\frac{1}{3}&-\frac{1}{3}&\frac{2}{3}\end{pmatrix}$

```
> C3 <- diag(3)-(1/3)*matrix(c(1), 3, 3)
> C3
           [,1]       [,2]       [,3]
```

```
[1,]  0.6666667 -0.3333333 -0.3333333
[2,] -0.3333333  0.6666667 -0.3333333
[3,] -0.3333333 -0.3333333  0.6666667
```

An example of linear dependence: Let $\mathbf{V} = \begin{pmatrix}\mathbf{v}_1 & \mathbf{v}_2\end{pmatrix}$ where $\mathbf{v}_1 = \begin{pmatrix}1 & 2\end{pmatrix}^\top$, $\mathbf{v}_2 = \begin{pmatrix}2 & 4\end{pmatrix}^\top$. These column vectors are linearly dependent, since $span(\mathbf{v}_1, \mathbf{v}_2) = 0$, which is equation of argument r_1, r_2, holds even if the solutions are not zeros but the other values $\exists$ $(-2t, t)$, $t \in \mathbb{R}$. Thus, $\mathbf{V}$ is not invertible.

```
> v1 <- c(1, 2) ; v2 <- c(2, 4)
> V <- matrix(c(v1, v2), 2, 2)
> v.fun <- function(r1, r2){r1*v1 + r2*v2}
> v.fun(0, 0)
[1] 0 0
> v.fun(-2, 1)
[1] 0 0
> det(V)
[1] 0
```

Partitioned form of a matrix: Let $\mathbf{A} = \begin{pmatrix}\mathbf{A}_{11} \\ \mathbf{A}_{21}\end{pmatrix}$ where $\mathbf{A}_{11} = \begin{pmatrix}1 & 3 \\ 2 & 4\end{pmatrix}$, $\mathbf{A}_{21} = \begin{pmatrix}5 & 6\end{pmatrix}$. Then, $\mathbf{A}^\top\mathbf{A} = \mathbf{A}_{11}^\top\mathbf{A}_{11} + \mathbf{A}_{21}^\top\mathbf{A}_{21} = \begin{pmatrix}1 & 2 \\ 3 & 4\end{pmatrix}\begin{pmatrix}1 & 3 \\ 2 & 4\end{pmatrix} + \begin{pmatrix}5 \\ 6\end{pmatrix}\begin{pmatrix}5 & 6\end{pmatrix} = \begin{pmatrix}30 & 41 \\ 41 & 61\end{pmatrix}$.

```
> A11 <- matrix(c(1:4), 2, 2) ; A21 <- matrix(c(5:6), 1, 2)
> t(A11)%*%A11 + t(A21)%*%A21
     [,1] [,2]
[1,]   30   41
[2,]   41   61
> A <- matrix(c(1, 2, 5, 3, 4, 6), 3, 2)
> t(A)%*%A
     [,1] [,2]
[1,]   30   41
[2,]   41   61
```

1.12.2 Rank of a matrix

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. Then, $r(\mathbf{A}) = 1$, $r(\mathbf{B}) = 2$, $r(\mathbf{AB}) = 1$. We can easily see that the following properties hold.

- $r(\mathbf{AB}) \leq min(r(\mathbf{A}), r(\mathbf{B}))$
- $r(\mathbf{A} + \mathbf{B}) \leq r(\mathbf{A}) + r(\mathbf{B})$
- $r(\mathbf{A}) = r(\mathbf{A}^\top) = r(\mathbf{A}^\top \mathbf{A}) = r(\mathbf{A}\mathbf{A}^\top)$

```
> install.packages("matrixcalc")
> library(matrixcalc)
> A <- matrix(c(1, 2, 2, 4), 2, 2) ; B <- diag(1:2)
> A%*%B
     [,1] [,2]
[1,]    1    4
[2,]    2    8
> matrix.rank(A%*%B) <= min(matrix.rank(A), matrix.rank(B))
[1] TRUE
> matrix.rank(A+B) <= matrix.rank(A) + matrix.rank(B)
[1] TRUE
> all.equal(matrix.rank(A), matrix.rank(t(A)),
+ matrix.rank(t(A)%*%A), matrix.rank(A%*%t(A)))
[1] TRUE
```

1.12.3 Trace of a matrix

We can see that $tr(\mathbf{AB}) = tr(\mathbf{BA})$.
(The "matrixcalc" package also has a function for the trace of a matrix.)

```
> install.packages("matrixcalc")
> library(matrixcalc)
> A
     [,1] [,2]
[1,]    1    2
[2,]    2    4
> B
```

```
     [,1] [,2]
[1,]    1    0
[2,]    0    2
> A%*%B
     [,1] [,2]
[1,]    1    4
[2,]    2    8
> B%*%A
     [,1] [,2]
[1,]    1    2
[2,]    4    8
> matrix.trace(A%*%B) == matrix.trace(B%*%A)
[1] TRUE
```

1.12.4 Eigenvalues and Eigenvectors

Let $\mathbf{A} = \begin{pmatrix} 1 & 4 \\ 9 & 1 \end{pmatrix}$. Then its eigenvalues are $\lambda_1 = 7$, $\lambda_2 = -5$ and eigenvectors are $\mathbf{v}_1 = \left\{ \begin{pmatrix} s \\ -\frac{3}{2}s \end{pmatrix} | s \in \mathbb{R} \right\}$, $\mathbf{v}_2 = \left\{ \begin{pmatrix} t \\ \frac{3}{2}t \end{pmatrix} | t \in \mathbb{R} \right\}$. R calculates normalized eigenvectors whose lengths are one.

```
> A <- matrix(c(1, 9, 4, 1), 2, 2)
> eigen(A)
$values
[1]  7 -5

$vectors
          [,1]       [,2]
[1,] 0.5547002 -0.5547002
[2,] 0.8320503  0.8320503

> sum((eigen(A)$vectors[,1])^2)
[1] 1
> sum((eigen(A)$vectors[,2])^2)
[1] 1
```

Trace of **A** equals the sum of its eigenvalues for any square matrix **A**.

```
> matrix.trace(A) == sum(eigen(A)$values)
[1] TRUE
```

Determinant of $\mathbf{A}$ equals the product of its eigenvalues for any square matrix $\mathbf{A}$. Note that trace and determinant is only defined at square matrix.

```
> det(A)
[1] -35
> prod(eigen(A)$values)
[1] -35
> det(A) == prod(eigen(A)$values)
[1] FALSE
> det(A) - prod(eigen(A)$values) # Find the reason why?
[1] -7.105427e-15
```

Gram-Schmidt process of orthogonalization: Let $\mathbf{x}_1 = \begin{pmatrix}2 & 0 & 0\end{pmatrix}^\top$, $\mathbf{x}_2 = \begin{pmatrix}2 & 2 & 0\end{pmatrix}^\top$, $\mathbf{x}_3 = \begin{pmatrix}2 & 0 & 2\end{pmatrix}^\top$. $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ are LIN vectors, but they are not orthogonal. Using projection of a vector, however, we can make them orthogonal.

$$\mathbf{y}_1 = \mathbf{x}_1 = \begin{pmatrix}2 & 0 & 0\end{pmatrix}^\top$$
$$\mathbf{y}_2 = \mathbf{x}_2 - \mathbf{y}_1(\mathbf{y}_1^\top\mathbf{y}_1)^{-1}\mathbf{y}_1^\top\mathbf{x}_2 = \begin{pmatrix}0 & 2 & 0\end{pmatrix}^\top$$
$$\mathbf{y}_3 = \mathbf{x}_3 - \mathbf{y}_1(\mathbf{y}_1^\top\mathbf{y}_1)^{-1}\mathbf{y}_1^\top\mathbf{x}_3 - \mathbf{y}_2(\mathbf{y}_2^\top\mathbf{y}_2)^{-1}\mathbf{y}_2^\top\mathbf{x}_3 = \begin{pmatrix}0 & 0 & 2\end{pmatrix}^\top.$$

After normalization, $\frac{1}{2}\begin{pmatrix}\mathbf{y}_1 & \mathbf{y}_2 & \mathbf{y}_3\end{pmatrix} = \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}$ is an orthogonal matrix.

```
> install.packages("pracma")
> library(pracma)
> x1 <- c(2, 0, 0)
> x2 <- c(2, 2, 0)
> x3 <- c(2, 0, 2)
> X <- cbind(x1, x2, x3)
> gramSchmidt(X)
$Q
```

```
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1

$R
     [,1] [,2] [,3]
[1,]    2    2    2
[2,]    0    2    0
[3,]    0    0    2
```

1.12.5 Quadratic forms and definite matrices

All eigenvalues of a positive definite matrix are greater than zero. Let $\mathbf{A} = \begin{pmatrix} 3 & 5 & 1 \\ 5 & 13 & 0 \\ 1 & 0 & 1 \end{pmatrix}$. Its eigenvalues are approximately $\lambda_1 = 15.08$, $\lambda_2 = 1.88$ and $\lambda_3 = 0.03$. Since all the eigenvalues are positive, $\mathbf{A}$ is p.d. matrix.

```
> A <- matrix(c(3, 5, 1, 5, 13, 0, 1, 0, 1), 3, 3)
> all(eigen(A)$values > 0)
[1] TRUE
> eigen(A)$values
[1] 15.08151247  1.88327962  0.03520791
```

1.12.6 Special matrices

Orthogonal matrix: Let $\mathbf{P} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Then $\mathbf{P}$ is orthogonal matrix. We can verify that the following properties hold.

- $\mathbf{P}^\top = \mathbf{P}^{-1}$
- $det(\mathbf{P}) = \pm 1$
- $-1 \leq p_{ii} \leq 1$

```
> P <- matrix(c(-1, 0, 0, 0, -1, 0, 0, 0, 1), 3, 3)
> P
     [,1] [,2] [,3]
[1,]   -1    0    0
[2,]    0   -1    0
[3,]    0    0    1
> sum(t(P) - solve(P))
[1] 0
> det(P)
[1] 1
> all(abs(diag(P)) <= 1)
[1] TRUE
```

Idempotent matrix: Let $\mathbf{A} = \begin{pmatrix} 2 & -2 & -4 \\ -1 & 3 & 4 \\ 1 & -2 & -3 \end{pmatrix}$. Then, $\mathbf{AA} = \mathbf{A}$ and $tr(\mathbf{A}) = r(\mathbf{A}) = 2$.

```
> A <- matrix(c(2, -1, 1, -2, 3, -2, -4, 4, -3), 3, 3)
> sum(A%*%A - A)
[1] 0
> matrix.rank(A) == matrix.trace(A)
[1] TRUE
```

1.12.7 Decomposition of a matrix

Let $\mathbf{A} = \begin{pmatrix} 3 & 5 & 1 \\ 5 & 13 & 0 \\ 1 & 0 & 1 \end{pmatrix}$. It is a positive definite matrix. Note $\mathbf{\Lambda} = \mathbf{D} = diag(\lambda_i)$, $\mathbf{P} = \begin{pmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 \end{pmatrix}$ where $\mathbf{p}_i$ is a normalized eigenvector, $i = 1, 2, 3$. Then the spectral decomposition: $\mathbf{A} = \mathbf{P\Lambda P}^\top = \sum\limits_{i=1}^{3} \lambda_i \mathbf{p}_i \mathbf{p}_i^\top$ and $\mathbf{A}^{-1} = \mathbf{P\Lambda}^{-1}\mathbf{P}^\top = \sum\limits_{i=1}^{3} \frac{1}{\lambda_i} \mathbf{p}_i \mathbf{p}_i^\top$.

```
> A <- matrix(c(3, 5, 1, 5, 13, 0, 1, 0, 1), 3, 3)
> D <- diag(eigen(A)$values)
> D
         [,1]    [,2]       [,3]
[1,] 15.08151 0.00000 0.00000000
[2,]  0.00000 1.88328 0.00000000
[3,]  0.00000 0.00000 0.03520791
> p1 <- eigen(A)$vectors[,1]
> p2 <- eigen(A)$vectors[,2]
> p3 <- eigen(A)$vectors[,3]
> P <- cbind(p1, p2, p3)
> P
              p1         p2         p3
[1,] -0.38418583  0.6344832  0.6706954
[2,] -0.92285258 -0.2853734 -0.2586603
[3,] -0.02728299  0.7183266 -0.6951709

> P1 <- P%*%D%*%t(P)
> P11 <- D[1, 1]*p1%*%t(p1) + D[2, 2]*p2%*%t(p2) +
+ D[3, 3]*p3%*%t(p3)
> abs(sum(A-P1))
[1] 1.573915e-14
> abs(sum(A-P11))
[1] 1.307461e-14
> abs(sum(P1-P11))
[1] 2.664535e-15

> P2 <- P%*%solve(D)%*%t(P)
> P22 <- (1/D[1, 1])*p1%*%t(p1) + (1/D[2, 2])*p2%*%t(p2) +
+ (1/D[3, 3])*p3%*%t(p3)
> abs(sum(solve(A)-P2))
[1] 1.44329e-14
> abs(sum(solve(A)-P22))
[1] 1.798561e-14
> abs(sum(P2-P22))
[1] 3.552714e-15
```

1.12.8 Kronecker(direct) products of matrices

Let $\mathbf{A} = \begin{pmatrix}1 & 3\\ 2 & 4\end{pmatrix}$, $\mathbf{B} = \begin{pmatrix}2 & 4\\ 3 & 5\end{pmatrix}$, $\mathbf{C} = \begin{pmatrix}3 & 5\\ 4 & 6\end{pmatrix}$, $\mathbf{D} = \begin{pmatrix}4 & 6\\ 5 & 7\end{pmatrix}$. Then, $\mathbf{A}_{2\times 2} \otimes \mathbf{B}_{2\times 2} = (a_{ij}\mathbf{B})_{(2\times 2)\times(2\times 2)} = \begin{bmatrix}1 & 3\\ 2 & 4\end{bmatrix}_{2\times 2} \otimes \begin{pmatrix}2 & 4\\ 3 & 5\end{pmatrix}_{2\times 2} = \begin{bmatrix}1\begin{pmatrix}2 & 4\\ 3 & 5\end{pmatrix} & 3\begin{pmatrix}2 & 4\\ 3 & 5\end{pmatrix}\\ 2\begin{pmatrix}2 & 4\\ 3 & 5\end{pmatrix} & 4\begin{pmatrix}2 & 4\\ 3 & 5\end{pmatrix}\end{bmatrix} = \begin{bmatrix}2 & 4 & 6 & 12\\ 3 & 5 & 9 & 15\\ 4 & 8 & 8 & 16\\ 6 & 10 & 12 & 20\end{bmatrix}_{4\times 4}.$

```
> A <- matrix(c(1, 2, 3, 4), 2, 2)
> B <- matrix(c(2, 3, 4, 5), 2, 2)
> C <- matrix(c(3, 4, 5, 6), 2, 2)
> D <- matrix(c(4, 5, 6, 7), 2, 2)
> kronecker(A, B)
     [,1] [,2] [,3] [,4]
[1,]    2    4    6   12
[2,]    3    5    9   15
[3,]    4    8    8   16
[4,]    6   10   12   20
```

$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$

```
> sum(kronecker(A, B)%*%kronecker(C, D) -
+ kronecker(A%*%C, B%*%D))
[1] 0
```

Let $\mathbf{a} = \begin{pmatrix}1 & 2 & 3\end{pmatrix}^\top$, $\mathbf{b} = \begin{pmatrix}4 & 5 & 6\end{pmatrix}^\top$. Then, $\mathbf{a}\mathbf{b}^\top = \mathbf{a} \otimes \mathbf{b}^\top = \mathbf{b}^\top \otimes \mathbf{a} \iff \begin{pmatrix}1\\ 2\\ 3\end{pmatrix}\begin{pmatrix}4 & 5 & 6\end{pmatrix} = \begin{bmatrix}1\begin{pmatrix}4 & 5 & 6\end{pmatrix}\\ 2\begin{pmatrix}4 & 5 & 6\end{pmatrix}\\ 3\begin{pmatrix}4 & 5 & 6\end{pmatrix}\end{bmatrix} = \begin{bmatrix}4\begin{pmatrix}1\\ 2\\ 3\end{pmatrix} & 5\begin{pmatrix}1\\ 2\\ 3\end{pmatrix} & 6\begin{pmatrix}1\\ 2\\ 3\end{pmatrix}\end{bmatrix} = \begin{bmatrix}4 & 5 & 6\\ 8 & 10 & 12\\ 12 & 15 & 18\end{bmatrix}.$

```
> a <- 1:3 ; b <- 4:6
> all.equal(a%*%t(b), kronecker(a, t(b)),
+ kronecker(t(b), a))
[1] TRUE
```

Chapter 2

Random vectors and matrices

2.1 Review of Univariate Results

2.1.1 Population parameter

We have a population of units that have a characteristic, X that can be measured.

The population may be

- Explicit - The individual in the US, X = age or
- Implicit - Particle boards to be produced by a factory, X = strength.

We are interested in the average value & a measure of variability in the population.

Define

- Average (expected) value = $EX = \mu$ &
- Variance = $Var(X) = E(X - \mu)^2 = \sigma^2$.

Properties

- $E(aX) = aEX = a\mu$
- $V(aX) = E(aX - a\mu)^2 = a^2VX = a^2\sigma^2$ where a is a constant.
- $V(aX + b) = E(aX + b - (a\mu + b))^2 = a^2VX = a^2\sigma^2$

2.1.2 Sample analogs

We have a random sample (r.s.) of size n from a population with observations, $x_1, x_2, \cdots, x_n$.

Define the sample mean, $\bar{x}$, as $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$, or if $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ then $\bar{x} = \dfrac{1}{n}\mathbb{1}^\top \boldsymbol{x} = \dfrac{1}{n}\boldsymbol{x}^\top \mathbb{1}$.

The sample variance, s^2, defined as

$$s_x^2 = \frac{1}{n-1}\textstyle\sum_{i=1}^{n}(x_i - \bar{x})^2 = \dfrac{1}{n-1}\boldsymbol{x}^\top \boldsymbol{C}_n \boldsymbol{x}, \quad \text{where } \boldsymbol{C}_n = \boldsymbol{I} - \dfrac{1}{n}\mathbb{1}\mathbb{1}^\top.$$

$$\begin{aligned}(\because)\ \textstyle\sum(x_i - \bar{x})^2 &= \textstyle\sum x_i^2 - n(\bar{x})^2 = \boldsymbol{x}^\top\boldsymbol{x} - \dfrac{1}{n}(\textstyle\sum x_i)^2 \\ &= \boldsymbol{x}^\top\boldsymbol{x} - \dfrac{1}{n}\boldsymbol{x}^\top\mathbb{1}\mathbb{1}^\top\boldsymbol{x} = \boldsymbol{x}^\top(\boldsymbol{I} - \dfrac{1}{n}\mathbb{1}\mathbb{1}^\top)\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{C}_n\boldsymbol{x}\end{aligned}$$

If a is a constant & $y_i = ax_i$, then

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} ax_1 \\ \vdots \\ ax_n \end{bmatrix} = a\boldsymbol{x} \ \&\ \bar{y} = \frac{1}{n}\mathbb{1}^\top\boldsymbol{y} = \frac{1}{n}\mathbb{1}^\top a\boldsymbol{x} = a\bar{x}$$

The sample variance of y_i is given by

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum(ax_i - a\bar{x})^2 = \frac{a^2}{n-1}\sum(x_i - \bar{x})^2 = a^2 s_x^2$$

Let $y_i^* = ax_i + b,\ \bar{y}^* = a\bar{x} + b$

$$s_{y^*}^2 = \frac{1}{n-1}\textstyle\sum(y_i^* - \bar{y}^*)^2 = \dfrac{1}{n-1}\textstyle\sum(ax_i + b - (a\bar{x} + b))^2$$

$$= \frac{a^2}{n-1}\textstyle\sum(x_i - \bar{x})^2 = a^2 s_x^2$$

Assumptions on the dist. in the pop.

Univariate Normal : $X \sim N(\mu,\ \sigma^2)$

Density ftn,

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$EX = \mu,\ VX = \sigma^2$$

Properties

- If $Y = aX,$ then $Y \sim N(a\mu,\ a^2\sigma^2)$.
  If $Y = aX + b,$ then $Y \sim N(a\mu + b,\ a^2\sigma^2)$.

- Sample statistics

  i) $\bar{X} = \frac{1}{n}\mathbb{1}^\top \boldsymbol{x} \sim N(\mu,\ \frac{\sigma^2}{n})$
  ii) $\frac{(n-1)s_x^2}{\sigma^2} = \frac{\boldsymbol{x}^\top \boldsymbol{C}_n \boldsymbol{x}}{\sigma^2} \sim \chi^2_{n-1}$
  iii) $\bar{X}$ is independent of s_x^2

2.2 Vectors

Let $X_1,\ X_2,\ \cdots,\ X_p$ be random variables such that $EX_i = \mu_i,\ VX_i = \sigma_{ii},$ & $cov(X_i, X_j) = \sigma_{ij}$.

Recall that $VX_i = E(X_i - \mu_i)^2$ & $cov(X_i, X_j) = E(X_i - \mu_i)(X_j - \mu_j)$.

Let $\boldsymbol{x}^\top = (X_1\ X_2\ \cdots\ X_p)$.

Define the mean vector $E\boldsymbol{x} = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$ &

the covariance matrix $V\boldsymbol{x} = \sum = (\sigma_{ij})$.

Note that $\sum = E(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top$

2.3 Matrices

Let $\boldsymbol{X} = (X_{ij})$ be a matrix of r.v.s. & define $E\boldsymbol{X} = (EX_{ij})$.

2.4 Data Matrix

The data matrix $\boldsymbol{X}$ consists of p observations on each of n experimental units, *i.e.*,

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [\boldsymbol{x}_1 \ \boldsymbol{x}_2 \ \cdots \ \boldsymbol{x}_p] = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix}$$

cf. Think about a design matrix in regression analysis, that is,

$$\boldsymbol{X} = [\mathbb{1} \ \boldsymbol{x}_1 \ \boldsymbol{x}_2 \ \cdots \ \boldsymbol{x}_{p-1}]$$

is $n \times p$ with $n =$ sample size and $p - 1 =$ number of independent variables.

Classic Blue Pullovers Data

This is a data set consisting of 10 measurements of 4 variables. The story: A textile shop manager is studying the sales of "classic blue" pullovers over 10 periods. He uses three different marketing methods and hopes to understand his sales as a fit of these variables using statistics. The variables measured are
X_1 : Numbers of sold pullovers,
X_2 : Price(in EURO),
X_3 : Advertisement costs in local newspapers(in EURO),
X_4 : Presense of a sales assistant(in hours per period).

	Sales	Price	Advert	Ass. Hours
1	230	125	200	109
2	181	99	55	107
3	165	97	105	98
4	150	115	85	71
5	97	120	0	82
6	192	100	150	103
7	181	80	85	111
8	189	90	120	93
9	172	95	110	86
10	170	125	130	78

The rows of $\boldsymbol{X}$ are denoted by

$$\boldsymbol{x}_i^\top = (x_{i1} \; x_{i2} \; \cdots \; x_{ip})$$

are assumed to be independent with mean vector $\boldsymbol{\mu}^\top$ and covariance matrix $\sum$.

The data matrix is a random sample of size n from a p-variate population.
(Think about p = 1 !)
Note that we may write

$$E\boldsymbol{X} = \begin{bmatrix} \boldsymbol{\mu}^\top \\ \vdots \\ \boldsymbol{\mu}^\top \end{bmatrix} = \mathbb{1}_n \boldsymbol{\mu}^\top = \mathbb{1}_n \otimes \boldsymbol{\mu}^\top.$$

2.5 Properties

If $\boldsymbol{a}$ & $\boldsymbol{b}$ are vectors of constants, then

$$1)\qquad E\boldsymbol{a}^\top\boldsymbol{x} = E\sum_{i=1}^p a_iX_i = \sum_{i=1}^p a_i\mu_i = \boldsymbol{a}^\top\boldsymbol{\mu}$$

$$2)\qquad V\boldsymbol{a}^\top\boldsymbol{x} = V\sum_{i=1}^p a_iX_i = \sum_{i=1}^p\sum_{j=1}^p a_ia_jcov(X_i,X_j) = \boldsymbol{a}^\top\Sigma\boldsymbol{a}$$

$$3)\qquad cov\left(\boldsymbol{a}^\top\boldsymbol{x},\boldsymbol{b}^\top\boldsymbol{x}\right) = cov\left(\sum_{i=1}^p a_iX_i, \sum_{i=1}^p b_iX_i\right)$$

$$= \sum_{i=1}^p\sum_{i=1}^p a_ib_jcov(X_i,X_j) = \boldsymbol{a}^\top\Sigma\boldsymbol{b}$$

$(\because)$ Let $X = \sum_{i=1}^p a_iX_i,\ Y = \sum_{i=1}^p b_iX_i,$ then

$$\begin{aligned} cov(X,Y) &= E(X-EX)(Y-EY) \\ &= E\sum_{i=1}^p(a_iX_i - a_i\mu_i)\sum_{j=1}^p(b_jX_j - b_j\mu_j) \\ &= E\sum_{i=1}^p\sum_{j=1}^p a_ib_j(X_i-\mu_i)(X_j-\mu_j) \\ &= \sum_{i=1}^p\sum_{j=1}^p a_ib_jcov(X_i,X_j) = \boldsymbol{a}^\top\Sigma\boldsymbol{b} \end{aligned}$$

ex)

$$\begin{aligned} V(a_1X_1+a_2X_2) &= E(a_1X_1+a_2X_2-E(a_1X_1+a_2X_2))^2 \\ &= E(a_1(X_1-EX_1)+a_2(X_2-EX_2))^2 \\ &= a_1^2E(X_1-EX_1)^2 + a_2^2E(X_2-EX_2)^2 \\ &\quad +2a_1a_2E(X_1-EX_1)(X_2-EX_2) \\ &= a_1^2VX_1 + a_2^2VX_2 + 2a_1a_2cov(X_1,X_2) \end{aligned}$$

$$\begin{aligned}
&cov(a_1X_1 + a_2X_2, b_1X_1 + b_2X_2)\\
&= E\{a_1X_1 + a_2X_2 - E(a_1X_1 + a_2X_2)\}\{b_1X_1 + b_2X_2 - E(b_1X_1 + b_2X_2)\}\\
&= E\{a_1(X_1 - EX_1) + a_2(X_2 - EX_2)\}\{b_1(X_1 - EX_1) + b_2(X_2 - EX_2)\}\\
&= a_1b_1E(X_1 - EX_1)^2 + a_2b_2E(X_2 - EX_2)^2\\
&\quad +a_1b_2E(X_1 - EX_1)(X_2 - EX_2) + a_2b_1E(X_2 - EX_2)(X_1 - EX_1)\\
&= a_1b_1VX_1 + a_2b_2VX_2 + a_1b_2cov(X_1, X_2) + a_2b_1cov(X_2, X_1)
\end{aligned}$$

In matrix form, let $\boldsymbol{C} = \begin{pmatrix} \boldsymbol{a}^\top \\ \boldsymbol{b}^\top \end{pmatrix}$ & $\boldsymbol{y} = \boldsymbol{Cx}$, then

$$E\boldsymbol{y} = \boldsymbol{C}\ E\boldsymbol{x}\ \&\ V\boldsymbol{y} = \boldsymbol{C}\ V\boldsymbol{x}\ \boldsymbol{C}^\top.$$

ex) Let the cols of $\boldsymbol{P}$ are the eigenvectors of $cov(\boldsymbol{x}) = \Sigma$, then the transformation $\boldsymbol{y} = \boldsymbol{P}^\top\boldsymbol{x}$ defines a new set of variables with variance

$$\begin{aligned}
V\boldsymbol{y} &= V\boldsymbol{P}^\top\boldsymbol{x} = \boldsymbol{P}^\top V\boldsymbol{x}\boldsymbol{P} = \boldsymbol{P}^\top\Sigma\boldsymbol{P}\\
&= \boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top\boldsymbol{P}\ (\because\ \Sigma\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\Lambda} \text{ by spectral decomposition})\\
&= \Lambda
\end{aligned}$$

That is, the new variables have zero covariances.

Correlation Matrix

$\boldsymbol{\rho} = [\rho_{ij}]$, where $\rho_{ij} = \dfrac{cov(X_i, X_j)}{\sqrt{VX_i}\sqrt{VX_j}} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$

Let $\boldsymbol{V}^{\frac{1}{2}} = diag(\sqrt{\sigma_{ii}}) = \begin{pmatrix} \sqrt{\sigma_{11}} & & & \phi \\ & \sqrt{\sigma_{22}} & & \\ & & \ddots & \\ \phi & & & \sqrt{\sigma_{pp}} \end{pmatrix}$, then

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

Note Σ: n.n.d.(p.s.d.)

$(\because)$ $VY = cov(Y,Y)$, where $Y = \boldsymbol{\alpha}^\top \boldsymbol{x}$

$$\begin{aligned} VY &= cov(\boldsymbol{\alpha}^\top \boldsymbol{x}, \boldsymbol{\alpha}^\top \boldsymbol{x}) = \boldsymbol{\alpha}^\top cov\boldsymbol{x}\ \boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}^\top \Sigma \boldsymbol{\alpha} \geq 0 \ \ {}^{\forall}\boldsymbol{\alpha} \\ &\Leftrightarrow \ \Sigma : n.n.d. \end{aligned}$$

2.6 Random sample

Let $\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \cdots\ ,\boldsymbol{x}_n$ denote a random sample from a p-variate distribution with $E\boldsymbol{x}_i = \boldsymbol{\mu}$ & $V\boldsymbol{x}_i = \Sigma$.

Let $\boldsymbol{X}_{n\times p}$ be the data matrix

$$\boldsymbol{X} = (\boldsymbol{x}_1 \boldsymbol{x}_2 \cdots \boldsymbol{x}_p) = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}$$

and define the sample mean vector as

$$\bar{\boldsymbol{x}}_{p\times 1} = \frac{1}{n}\boldsymbol{X}^\top \mathbb{1}_n = \frac{1}{n}\begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_p^\top \end{pmatrix} \mathbb{1}_n = \begin{pmatrix} \frac{1}{n}\boldsymbol{x}_1^\top \mathbb{1}_n \\ \vdots \\ \frac{1}{n}\boldsymbol{x}_p^\top \mathbb{1}_n \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

- Centered Data matrix

$$\begin{aligned} \boldsymbol{X}_c &= \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1p} - \bar{x}_p \\ \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{np} - \bar{x}_p \end{bmatrix} \\ &= \boldsymbol{X} - \begin{bmatrix} \bar{x}_1 & \cdots & \bar{x}_p \\ \vdots & & \vdots \\ \bar{x}_1 & \cdots & \bar{x}_p \end{bmatrix} = \boldsymbol{X} - \begin{bmatrix} \bar{\boldsymbol{x}}^\top \\ \vdots \\ \bar{\boldsymbol{x}}^\top \end{bmatrix} \\ &= \boldsymbol{X} - \mathbb{1}_n \bar{\boldsymbol{x}}^\top = \boldsymbol{X} - \frac{1}{n}\mathbb{1}_n \mathbb{1}_n^\top \boldsymbol{X} \\ &= (\boldsymbol{I}_n - \frac{1}{n}\mathbb{1}_n \mathbb{1}_n^\top)\boldsymbol{X} = (\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n)\boldsymbol{X} = \boldsymbol{C}_n \boldsymbol{X} \end{aligned}$$

- Sample variances and covariances

$$
\begin{aligned}
S_{ii} = \hat{\sigma}_{ii} &= \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2 \\
S_{ij} = \hat{\sigma}_{ij} &= \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)
\end{aligned}
$$

- Sample covariance matrix

$$
\begin{aligned}
\boldsymbol{S} = (S_{ij}) = (\hat{\sigma}_{ij}) &= \frac{1}{n-1}\boldsymbol{X}_c^\top \boldsymbol{X}_c \\
&(\because)\ \text{Note that } x_{ki}-\bar{x}_i \text{ is } i^{th} \text{ col of } X_c \\
&= \frac{1}{n-1}\boldsymbol{X}^\top \boldsymbol{C}_n^\top \boldsymbol{C}_n \boldsymbol{X} \\
&= \frac{1}{n-1}\boldsymbol{X}^\top \boldsymbol{C}_n \boldsymbol{X}
\end{aligned}
$$

$$
\begin{aligned}
(\because)\ \boldsymbol{C}_n^\top \boldsymbol{C}_n &= (\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}_n)^\top(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}_n) = (\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}_n)(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J}_n) \\
&\quad \text{Note that } \boldsymbol{I}-\frac{1}{n}\boldsymbol{J}_n \text{ is symm.} \\
&= \boldsymbol{I}-\frac{1}{n}\boldsymbol{J}_n-\frac{1}{n}\boldsymbol{J}_n+\frac{1}{n^2}n\boldsymbol{J}_n = \boldsymbol{I}-\frac{1}{n}\boldsymbol{J}_n = \boldsymbol{C}_n \\
&\quad \text{Note that } \boldsymbol{C}_n \text{ is idempotent.}
\end{aligned}
$$

- Sample correlation matrix

$$\boldsymbol{R} = \hat{\boldsymbol{\rho}} = \text{diag}\left(\frac{1}{\sqrt{S_{ii}}}\right)\ \boldsymbol{S}\ \text{diag}\left(\frac{1}{\sqrt{S_{ii}}}\right)$$

<u>Recall that</u> $\boldsymbol{\rho} = \text{diag}\left(\frac{1}{\sqrt{\sigma_{ii}}}\right)\ \Sigma\ \text{diag}\left(\frac{1}{\sqrt{\sigma_{ii}}}\right)$

- There are another notations useful for some proofs.
  $\bar{\boldsymbol{x}} = \frac{1}{n}\boldsymbol{X}^\top \mathbb{1}_n == \frac{1}{n}(\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n)\mathbb{1}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$ and

$\boldsymbol{S} = \dfrac{1}{n-1}\displaystyle\sum_{k=1}^{n}(\boldsymbol{x}_k - \bar{\boldsymbol{x}})(\boldsymbol{x}_k - \bar{\boldsymbol{x}})^\top$ since $(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$ is the $(i,j)^{th}$ element of $(\boldsymbol{x}_k - \bar{\boldsymbol{x}})(\boldsymbol{x}_k - \bar{\boldsymbol{x}})^\top$.

- Sample analogs of moment properties

Let $l = \boldsymbol{x}^\top \boldsymbol{a}$ with sample values, $\boldsymbol{l} = \boldsymbol{X}\boldsymbol{a}$.

Recall that $\boldsymbol{X}_{n\times p} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix} = (\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_p)$.

Note that we have defined new data $\boldsymbol{l}$, then,

$$\begin{aligned}
\bar{l} &= \frac{1}{n}\mathbf{1}^\top \boldsymbol{l} = \frac{1}{n}\mathbf{1}^\top \boldsymbol{X}\boldsymbol{a} = \bar{\boldsymbol{x}}^\top \boldsymbol{a} \\
\hat{V}l &= \frac{1}{n-1}\sum_{i=1}^{n}(l_i - \bar{l})^2 \\
&= \frac{1}{n-1}\boldsymbol{l}^\top(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{l} = \frac{1}{n-1}\boldsymbol{l}^\top \boldsymbol{C}_n \boldsymbol{l} \\
&= \frac{1}{n-1}\boldsymbol{a}^\top \boldsymbol{X}^\top \boldsymbol{C}_n \boldsymbol{X}\boldsymbol{a} = \boldsymbol{a}^\top \boldsymbol{S}\boldsymbol{a}
\end{aligned}$$

<u>Note</u>

$$\begin{aligned}
\sum (x_i - \bar{x})^2 &= \sum x_i^2 - n(\bar{x})^2 = \boldsymbol{x}^\top \boldsymbol{x} - n\left(\frac{1}{n}\mathbb{1}^\top \boldsymbol{x}\right)^2 \\
&= \boldsymbol{x}^\top \boldsymbol{x} - \frac{1}{n}\boldsymbol{x}^\top \mathbb{1}\mathbb{1}^\top \boldsymbol{x} \\
&= \boldsymbol{x}^\top \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{C}_n \boldsymbol{x}
\end{aligned}$$

If $m = \boldsymbol{x}^\top \boldsymbol{b}$ with sample values, $\boldsymbol{m} = \boldsymbol{X}\boldsymbol{b}$, then

$$\begin{aligned}
\widehat{cov}(l,m) &= \boldsymbol{a}^\top \boldsymbol{S}\boldsymbol{b} \\
(\because)\ \widehat{cov}(l,m) &= \frac{1}{n-1}\sum_{i=1}^{n}(l_i - \bar{l})(m_i - \bar{m}) \\
& \quad\ l_i - \bar{l} \ \Rightarrow\ \boldsymbol{l} - \frac{1}{n}\mathbb{1}\mathbb{1}^\top \boldsymbol{l} = (\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{l} \\
&= \frac{1}{n-1}\boldsymbol{l}^\top \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)^\top \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{m} \\
&= \frac{1}{n-1}\boldsymbol{l}^\top \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{m} \\
&= \frac{1}{n-1}\boldsymbol{a}^\top \boldsymbol{X}^\top \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{X}\boldsymbol{b} \\
&= \boldsymbol{a}^\top \boldsymbol{S}\boldsymbol{b},\ \text{where } \boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^\top \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{X} = \frac{1}{n-1}\boldsymbol{X}^\top \boldsymbol{C}_n \boldsymbol{X}
\end{aligned}$$

In general, if we define new variables by the relation

$$\boldsymbol{y} = C\boldsymbol{x}$$

Then, we define the new data matrix

$$Y = \boldsymbol{X}\boldsymbol{C}^\top$$

$$\begin{aligned}
(\because)\ \boldsymbol{y}_i &= \boldsymbol{C}\boldsymbol{x}_i \Leftrightarrow \boldsymbol{y}_i^\top = \boldsymbol{x}_i^\top \boldsymbol{C}^\top \\
\boldsymbol{Y} &= \begin{pmatrix} \boldsymbol{y}_1^\top \\ \vdots \\ \boldsymbol{y}_n^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \boldsymbol{C}^\top \\ \vdots \\ \boldsymbol{x}_n^\top \boldsymbol{C}^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix} \boldsymbol{C}^\top = \boldsymbol{X}\boldsymbol{C}^\top
\end{aligned}$$

It follows that the vector of sample means for the new data is given by

$$\begin{aligned}\bar{\boldsymbol{y}} &= \frac{1}{n}\boldsymbol{Y}^\top \mathbb{1} = \frac{1}{n}\boldsymbol{C}\boldsymbol{X}^\top \mathbb{1} = \boldsymbol{C}\bar{\boldsymbol{x}} \\ \boldsymbol{S}_y &= \frac{1}{n-1}\boldsymbol{Y}^\top \boldsymbol{C}_n \boldsymbol{Y} \text{ by definition} \\ &= \frac{1}{n-1}\boldsymbol{C}\boldsymbol{X}^\top \boldsymbol{C}_n \boldsymbol{X}\boldsymbol{C}^\top = \boldsymbol{C}\boldsymbol{S}_x\boldsymbol{C}^\top, \text{where} \\ & \boldsymbol{S}_x = \frac{1}{n-1}\boldsymbol{X}^\top \boldsymbol{C}_n \boldsymbol{X} \ \& \ \boldsymbol{C}_n = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n \\ & \boldsymbol{S}_x : \text{ sample cov. matrix of } \boldsymbol{X}\end{aligned}$$

ex) Let the cols of $\boldsymbol{P}$ are the eigenvectors of Σ then the transformation $\boldsymbol{y} = \boldsymbol{P}^\top \boldsymbol{x}$ had

$$V\boldsymbol{y} = \Lambda \text{ (see previous example)}$$

Similarly let the cols of $\widehat{\boldsymbol{P}}$ are the eigenvectors of $\boldsymbol{S}_x$, then

$$\boldsymbol{y}_i = \widehat{\boldsymbol{P}}^\top \boldsymbol{x}_i$$

defines a new data matrix

$$\boldsymbol{Y} = \boldsymbol{X}\widehat{\boldsymbol{P}}$$

with sample covariance matrix

$$\boldsymbol{S}_y = \widehat{\boldsymbol{P}}^\top \boldsymbol{S}_x \widehat{\boldsymbol{P}} = \widehat{\Lambda}, \text{where}$$

$\boldsymbol{S}_x = \widehat{\boldsymbol{P}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{P}}^\top$ by spectral decomposition.

That is, the transformed variables have zero (sample) covariances.

2.7 Addendum to Random sampling

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix}$$

$\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$: random sample from a joint distribution $f(\boldsymbol{x}) = f(x_1, \cdots, x_p)$ if $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ are independent observations from a common joint distribution $f(\boldsymbol{x})$.

<u>**Result 3.1**</u>

Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be a r.s from a joint distribution that has mean vector $\boldsymbol{\mu}$ and covariance matrix Σ

$\Longrightarrow$

① $E\bar{\boldsymbol{x}} = \boldsymbol{\mu} \quad \& \quad Cov(\bar{\boldsymbol{x}}) = \frac{1}{n}\Sigma$

② $E\boldsymbol{S}_n = \frac{n-1}{n}\Sigma \quad \& \quad E\boldsymbol{S} = E\frac{n}{n-1}\boldsymbol{S}_n = \Sigma$

pf) Note that

$$\begin{aligned} \bar{\boldsymbol{x}} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \ \& \\ \boldsymbol{S}_n &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \\ E\bar{\boldsymbol{x}} &= \frac{1}{n}\sum_{i=1}^{n} E\boldsymbol{x}_i = \frac{1}{n}n\boldsymbol{\mu} = \boldsymbol{\mu}_{p\times 1} \end{aligned}$$

$$\begin{aligned}
Cov(\bar{\boldsymbol{x}}) &= E(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \\
&= E\left\{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})\right\}\left\{\frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu})\right\}^\top \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_j - \boldsymbol{\mu})^\top \\
&= \frac{1}{n^2}\sum_{i=1}^{n}E(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \ (\because \text{independent}) \\
&= \frac{1}{n^2}n\Sigma \ = \frac{1}{n}\Sigma
\end{aligned}$$

$$\begin{aligned}
E\boldsymbol{S}_n &= \frac{1}{n}E\underbrace{\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top} \\
&= \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\boldsymbol{x}_i{}^\top - \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\bar{\boldsymbol{x}}^\top \\
&= \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i{}^\top - \bar{\boldsymbol{x}}\sum_{i=1}^{n}\boldsymbol{x}_i{}^\top \\
&= \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i{}^\top - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}E\boldsymbol{x}_i\boldsymbol{x}_i{}^\top - E\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top
\end{aligned}$$

Note

$\underline{E\mathbf{v} = \boldsymbol{\mu}_{\mathbf{v}},\ cov(\mathbf{v}) = \Sigma_{\mathbf{v}}}$

$\underline{\Rightarrow E\mathbf{v}\mathbf{v}^\top = \Sigma_{\mathbf{v}} + \boldsymbol{\mu}_{\mathbf{v}}\boldsymbol{\mu}_{\mathbf{v}}^\top}$

$$
\begin{aligned}
E\boldsymbol{S}_n &= \frac{1}{n}\sum_{i=1}^{n}\left(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right) - \left(\frac{1}{n}\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right) \\
&= \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top - \frac{1}{n}\Sigma - \boldsymbol{\mu}\boldsymbol{\mu}^\top \\
&= \frac{n-1}{n}\Sigma
\end{aligned}
$$

$$\therefore\ \boldsymbol{S} = \frac{n}{n-1}\boldsymbol{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \quad : \text{unbiased estimator of } \Sigma.$$

□

- **Generalized Variance**

$$\boldsymbol{S} = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix} = \left\{ S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j}) \right\},$$

where x_{ki} : i^{th} col of $\boldsymbol{X}$ and x_{kj} : j^{th} col of $\boldsymbol{X}$.

generalized sample variance = $|\boldsymbol{S}|$

generalized variance = $|\Sigma|$

<u>**Note**</u>

① $p = 1 \Rightarrow$ sample variance

② $n \le p$, i.e., (sample size) $\le$ (# of variables)
$\Rightarrow |\boldsymbol{S}| = 0 \ \leftarrow$ See Result 3.2

③ generalized sample variance of the standardized variables = $|\boldsymbol{R}|$, where

$\boldsymbol{R}$ is a sample correlation matrix and

$$\boldsymbol{R} = (r_{ij}),$$

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}}$$

④ total variance $= \operatorname{tr}(\Sigma) = \displaystyle\sum_{i=1}^{p} \sigma_{ii}$

total sample variance $= \operatorname{tr}(\boldsymbol{S}) = \displaystyle\sum_{i=1}^{p} S_{ii}$

Remark

Univariate	Multivariate
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\boldsymbol{x}^\top \mathbb{1}$	$\bar{\boldsymbol{x}} = \frac{1}{n}\boldsymbol{X}^\top \mathbb{1}$
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^\top \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{X}$
$= \frac{1}{n-1}\boldsymbol{x}^\top \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{x}$	

Chapter 3

The Multivariate Normal dist

3.1 Univariate normal distribution

$$X \sim N(\mu, \sigma^2) \quad \text{if} \quad f_x(x) = \frac{1}{(2\pi)^{1/2}|\sigma^2|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)\left(\sigma^2\right)^{-1}(x-\mu)\right\},\ x \in \mathbb{R}$$

3.2 Multivariate (p-dimensional) normal distribution

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma) \text{ if } f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},\ \boldsymbol{x} \in \mathbb{R}^p$$

Note

1) Assume that Σ is positive definite.

2) Σ is symm. $\Rightarrow$ $\Sigma = \boldsymbol{P\Lambda P}^\top$, where $P = (\boldsymbol{e}_1 \ \cdots \ \boldsymbol{e}_p)$, $\Lambda = \text{diag}(\lambda_i)$,
λ_i : eigenvalue of Σ, & $\boldsymbol{e}_i$: corresponding eigenvector of Σ.

Consider the special case $p = 2$ so that

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

1) $|\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1-\rho^2), \quad \text{where } \rho = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$

2) $$\Sigma^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \dfrac{1}{\sigma_{11}} & -\dfrac{\rho}{\sqrt{\sigma_{11}\sigma_{22}}} \\ -\dfrac{\rho}{\sqrt{\sigma_{11}\sigma_{22}}} & \dfrac{1}{\sigma_{22}} \end{pmatrix}$$

3) $$(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\ = \frac{1}{1-\rho^2}\left\{\frac{(x_1-\mu_1)^2}{\sigma_{11}} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sqrt{\sigma_{11}\sigma_{22}}} + \frac{(x_2-\mu_2)^2}{\sigma_{22}}\right\}$$

Note in the special case ρ=0, the bivariate normal density is the product of two univariate normal densities.

The quadratic form is an ellipse, centered at $(\mu_1,\ \mu_2)$. In the special case, $\sigma_{11} = \sigma_{22}$ it is a circle.

If $\rho \neq 0$ the quadratic form is also an ellipse, but the axes of the ellipse are not along the coordinate axes.

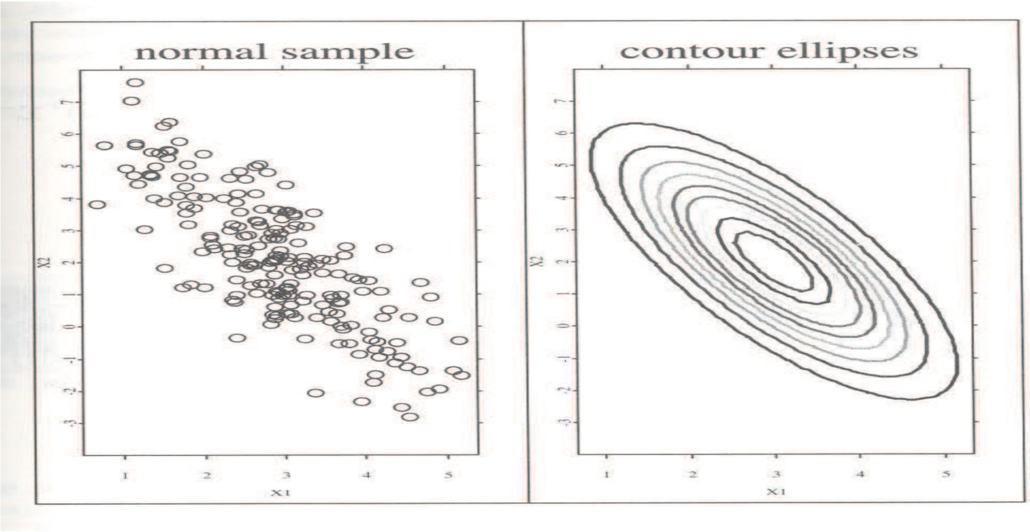


Figure 3.2.1: Scatterplot of a normal sample and contour ellipses

For Figure 3.2.1, $\mu = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 1 & -1.5 \\ -1.5 & 4 \end{bmatrix}$.

Examining the contours $(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ of the density helps us visualize it.

Since Σ is symm., $\Sigma = \boldsymbol{P\Lambda P}^\top$ so $\Sigma^{-1} = \boldsymbol{P\Lambda}^{-1}\boldsymbol{P}^\top$

We the make the transformation

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{P}^\top(\boldsymbol{x}-\boldsymbol{\mu}) \text{ followed by} \\ \boldsymbol{z} &= \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{y} \ \Leftrightarrow \ z_i = \frac{y_i}{\sqrt{\lambda_i}}. \end{aligned}$$

The quadratic form is then transformed as follows

$$\begin{aligned} &(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = (\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{P\Lambda}^{-1}\boldsymbol{P}^\top(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= \boldsymbol{y}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{y} = \boldsymbol{z}^\top\boldsymbol{z} = \sum_{i=1}^p \frac{y_i^2}{\lambda_i} = \sum_{i=1}^p z_i^2. \end{aligned}$$

Thus, in terms of the y coordinates, we have an ellipsoid lined up with the coordinate axes.
In terms of the z coordinates we have a sphere.
Note that

$$\begin{aligned} &\boldsymbol{y} \sim N_p(\mathbb{0}, \boldsymbol{\Lambda})\& \\ &\boldsymbol{z} \sim N_p(\mathbb{0}, \boldsymbol{I}). \text{ See next page.} \end{aligned}$$

Application to simulation.

$$\begin{aligned} &\boldsymbol{z} \sim N_p(\mathbb{0}, \boldsymbol{I}) \ \Leftrightarrow \ z_i \sim iid\ N(0,1),\ i = 1, \cdots, p \\ &\Rightarrow \boldsymbol{y} = \Lambda^{\frac{1}{2}}\boldsymbol{z} \\ &\Rightarrow \boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{Py} \sim N_p(\boldsymbol{\mu},\ \Sigma) \end{aligned}$$

Note: Points on the contours are equidistant from the center of the ellipse, and correspond to equal values of the density function, that is, are “equally likely”.

3.3 Properties of the multi. normal dist

1) Linear combinations of the components of $\boldsymbol{x}$ are normally distributed.

That is, $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \Sigma) \;\Rightarrow\; \boldsymbol{a}^\top\boldsymbol{x} \sim N(\boldsymbol{a}^\top\boldsymbol{\mu}, \boldsymbol{a}^\top\Sigma\boldsymbol{a})\ {}^\forall\boldsymbol{a}$.
Furthermore the inverse is also true, which is a working def. of the normal dist.

2) Several linear combinations of a multi. normal dist. are also normally distributed, i.e.

$\boldsymbol{x} \sim N_p(\mu,\ \Sigma)$ & $\underset{t\times1}{\boldsymbol{l}} = \underset{t\times p}{\boldsymbol{B}}\boldsymbol{x} + \underset{t\times1}{\boldsymbol{b}}$
$\Rightarrow\ \boldsymbol{l} \sim N_t(\boldsymbol{B}\boldsymbol{\mu} + \boldsymbol{b},\ \boldsymbol{B}\Sigma\boldsymbol{B}^\top)$.
ex)

$$\begin{array}{lll}
a. & & p=2,\ \boldsymbol{B}=(1\ 0),\ \boldsymbol{b}=0 \\
 & \Rightarrow & l=(1\ 0)\begin{pmatrix}x_1\\x_2\end{pmatrix}+0=x_1\sim N(\mu_1,\ \sigma_{11}) \\
b. & & p=2,\ \boldsymbol{B}=(1\ -1),\ \boldsymbol{b}=0 \\
 & \Rightarrow & l=(1-1)\begin{pmatrix}x_1\\x_2\end{pmatrix}=x_1-x_2\sim N(\mu_1-\mu_2,\ \sigma_{11}-2\sigma_{12}+\sigma_{22}) \\
c. & & \boldsymbol{x}=\begin{pmatrix}\boldsymbol{x}_1\\\boldsymbol{x}_2\end{pmatrix}\ (\boldsymbol{x}_1:\ t\times1,\ \boldsymbol{x}_2:\ (p-t)\times1), \boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu}_1\\\boldsymbol{\mu}_2\end{pmatrix}, \Sigma=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{pmatrix} \\
 & & \text{with } B=(I_t\ \mathbb{0}_{t\times(p-t)})\ \&\ \boldsymbol{b}=\mathbb{0}, \\
 & & \text{we have } \boldsymbol{l}=(I_t\ \mathbb{0})\begin{pmatrix}\boldsymbol{x}_1\\\boldsymbol{x}_2\end{pmatrix}=\boldsymbol{x}_1\sim N_t(\boldsymbol{\mu}_1,\ \Sigma_{11})
\end{array}$$

That is, all marginal distributions are normal.

3) The conditional dist. of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2$ is normal with density
$\boldsymbol{x}_1|\boldsymbol{x}_2 \sim N_t\left(\boldsymbol{\mu}_{1.2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \Sigma_{1.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$

Example

- Linear regression model was $y = \beta_0 + \beta_1 x + \epsilon, \epsilon \sim$ *iid* $(0, \sigma^2)$ and then normality assumption is added. In this case $x_1 = y$ is random, but $x_2 = x$ is fixed.

  This regression model can be thought as differently as follows (Bivariate normal model, Casella & Berger):

Let $x_1 = y$ be a dependent variable to the independent variable $x_2 = x$ having a bivariate normal distribution, then the population regression function is now a true conditional expectation and

$$X_1|X_2 = x_2 \overset{d}{=} \mu_1 + \beta(x_2 - \mu_2) + \varepsilon = \mu_1 - \beta\mu_2 + \beta x_2 + \varepsilon, \text{ where}$$
$$\beta = \frac{\sigma_{12}}{\sigma_{22}},\ \varepsilon \sim N(0, \sigma^2) \ \&\ \sigma^2 = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}} = \sigma_{11}(1 - \rho^2).$$

- Multiple correlation (when x_1 & x_2 are both random)

$$cov(x_1, \mu_1 + \beta(x_2 - \mu_2)) = \beta\sigma_{12}$$
$$corr(x_1, \mu_1 + \beta(x_2 - \mu_2)) = \frac{\beta\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\beta^2\sigma_{22}}}$$
$$corr^2(x_1, \mu_1 + \beta(x_2 - \mu_2)) = \left[\frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} \equiv \text{ pop. coefficient of determination.}\right]$$

Think about (sample) coefficient of determination R^2, which is defined as $R^2 = \frac{SSR}{SST}$

<u>Recall that</u>

① sample correlation coefficient

$$= r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

② sample coefficient of determunation.

$$= R^2 = \frac{SSR}{SST}, \text{ where } SSR = \sum(\widehat{y_i} - \bar{y})^2 \text{ and } SST = \sum(y_i - \bar{y})^2.$$

Recall that

$$\begin{aligned} SSR &= \sum (b_0 + b_1 x_i - \bar{y})^2, \text{ where } b_0 = \bar{y} - b_1\bar{x} \ \& \ b_1 = \frac{s_{xy}}{s_{xx}} \\ &= \sum (\bar{y} - b_1\bar{x} + b_1 x_i - \bar{y})^2 \\ &= b_1^2 \sum (x_i - \bar{x})^2 = b_1^2 s_{xx} \\ \therefore\ R^2 &= \frac{SSR}{SST} = \frac{b_1^2 s_{xx}}{s_{yy}} = \frac{\frac{s_{xy}^2}{s_{xx}^2} s_{xx}}{s_{yy}} = \frac{s_{xy}^2}{s_{xx}s_{yy}} \\ &= (r_{xy})^2 \end{aligned}$$

4) If $\boldsymbol{x}_1$ & $\boldsymbol{x}_2$ are indep., then $\text{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{0}$

$$\text{If } \boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix} \sim N_p\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right), \text{ then}$$
$$\boldsymbol{x}_1 \ \& \ \boldsymbol{x}_2 \text{ are indep. } \Leftrightarrow \Sigma_{12} = \mathbb{0}.$$

5) $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \Sigma) \underset{\nleftarrow}{\rightarrow} (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \sim \chi_p^2$

W.L.O.G. $\boldsymbol{z} \underset{\nleftarrow}{\rightarrow} \sum_{i=1}^p Z_i^2 = \boldsymbol{z}^\top \boldsymbol{z} \sim \chi_p^2$

$\underline{\text{Counter example}}$
$Y_i = |Z_i|, i = 1, 2 \Rightarrow Y_1^2 + Y_2^2 = Z_1^2 + Z_2^2 \sim \chi_2^2$. However $Y_i \sim TN(0, 1)$.

3.4 Distribution of sample statistics

Recall that

$$\begin{aligned} x_1 \cdots x_n &\sim \text{ iid N}(\mu, \sigma^2) \\ \Rightarrow \ & \textcircled{1}\ \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \\ & \textcircled{2}\ \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \\ & \textcircled{3}\ \bar{x} \text{ is independent of s}^2 \end{aligned}$$

Let $\boldsymbol{X}$ be the $n \times p$ data matrix for a random sample of size n from a p-variate normal dist, $N_p(\boldsymbol{\mu}, \Sigma)$.

1) $\bar{\boldsymbol{x}} = \dfrac{1}{n}\boldsymbol{X}^\top \mathbb{1} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\Sigma\right)$

2) $\boldsymbol{S} = \dfrac{1}{n-1}\boldsymbol{X}^\top \boldsymbol{C}_n \boldsymbol{X}$, where $\boldsymbol{C}_n = \boldsymbol{I} - \frac{1}{n}\boldsymbol{J}$

The joint dist. of $\boldsymbol{A} = (n-1)\boldsymbol{S} = \boldsymbol{X}^\top \boldsymbol{C}_n \boldsymbol{X}$ is known as the Wishart dist. with a df $n-1$, *i.e.* $\boldsymbol{A} \sim W_p(\Sigma, n-1) = W_{p,n-1}(\Sigma)$

Note that

$$f(\boldsymbol{A}) = \frac{|\boldsymbol{A}|^{\frac{n-p-2}{2}} \exp\left(-\frac{1}{2} tr\left(\boldsymbol{A}\Sigma^{-1}\right)\right)}{2^{\frac{p(n-1)}{2}} \pi^{\frac{p(p-1)}{4}} |\Sigma|^{\frac{n-1}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n-i)\right)}.$$

Fortunately, we will not have to deal directly with it.

3) $\bar{\boldsymbol{x}}$ is independent of $\boldsymbol{S}$.

3.5 Addendum to dist. of sample statistics

$\boldsymbol{x}_1 \cdots \boldsymbol{x}_n$: r.s. from $N_p(\boldsymbol{\mu}, \Sigma)$

$\Rightarrow \hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ & $\hat{\Sigma} = \boldsymbol{S}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ are the MLEs and sufficient statistics of $\boldsymbol{\mu}$ & Σ .
proof)

<u>Lemma</u>

① $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = tr(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top)$

② tr $\boldsymbol{A} = \displaystyle\sum_{i=1}^{k} \lambda_i$, λ_i : eigenvalues of symmetric matrix $\boldsymbol{A}$

$$f(\boldsymbol{x}_1 \cdots \boldsymbol{x}_n) = \prod_{i=1}^{n} \left\{ \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right)\right\}$$

$$= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left(-\frac{1}{2}\underbrace{\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})}_{\circledast}\right)$$

$$\circledast = \sum_i tr[(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})]$$

$$= \sum_i tr(\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top)$$

$$= tr\left[\Sigma^{-1}\left(\underbrace{\sum_i(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top}_{\circledast^1}\right)\right]$$

$$\circledast^1 = \sum_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} - \boldsymbol{\mu})(\boldsymbol{x}_i - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top$$

$$= \sum_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top$$

$$\therefore f(\boldsymbol{x}_1 \cdots \boldsymbol{x}_n) = (2\pi)^{-np/2}|\Sigma|^{-n/2}$$

$$\times \exp\left\{-\frac{1}{2}\underbrace{tr\left[\Sigma^{-1}\left(\sum_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top\right)\right]}_{\circledast^2}\right\}$$

$$\begin{aligned}
\circledast^2 &= tr\left[\Sigma^{-1}\left(\sum_i (\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top\right)\right] + n\cdot tr\left[\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top\right]\\
&= tr\left[\Sigma^{-1}\left(\sum_i (\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top\right)\right] + n\cdot tr(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\\
&= tr\left[\Sigma^{-1}\left(\sum_i (\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top\right)\right] + n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})
\end{aligned}$$

Note that Σ : p.d. $\Rightarrow$ Σ^{-1} : p.d. from Result 4.1.

The likelihood is maximized w.r.t. $\boldsymbol{\mu}$ at $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$
since i) $\boldsymbol{\mu}$ only appears on the 2^{nd} term &
ii) Σ^{-1} : p.d. $\Rightarrow$ $(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) > 0$.

It remains to maximize

$$L(\boldsymbol{\mu},\Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\left(-\frac{1}{2}\,tr\left[\Sigma^{-1}\left(\sum_i (\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top\right)\right]\right)$$

over Σ. Then the loglikelihood becomes

$$\begin{aligned}
l(\Sigma) &= -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\,tr\left(\Sigma^{-1}\boldsymbol{S}_n\right)\\
&= \frac{n}{2}\log\left|\Sigma^{-1}\boldsymbol{S}_n\right| - \frac{n}{2}\,tr\left(\Sigma^{-1}\boldsymbol{S}_n\right) - \frac{n}{2}\log|\boldsymbol{S}_n|\\
&= \frac{n}{2}\log\left|\boldsymbol{S}_n^{1/2}\Sigma^{-1}\boldsymbol{S}_n^{1/2}\right| - \frac{n}{2}\,tr\left(\boldsymbol{S}_n^{1/2}\Sigma^{-1}\boldsymbol{S}_n^{1/2}\right) - \frac{n}{2}\log|\boldsymbol{S}_n|\\
&= \frac{n}{2}\left\{\log(\lambda_1\cdots\lambda_p) - (\lambda_1+\cdots+\lambda_p)\right\} - \frac{n}{2}\log|\boldsymbol{S}_n|\\
&= \frac{n}{2}\left[\sum_i^p\left\{(\log\lambda_i) - (\lambda_i - 1)\right\} - p - \frac{n}{2}\log|\boldsymbol{S}_n|\right]\\
&\le \frac{n}{2}\left[-p - \frac{n}{2}\log|\boldsymbol{S}_n|\right]
\end{aligned}$$

with equality iff $\lambda_1 = \cdots = \lambda_p = 1$, where $\lambda_1,\cdots,\lambda_p$ are eigenvalues of positive definite $\boldsymbol{S}_n^{1/2}\Sigma^{-1}\boldsymbol{S}_n^{1/2}$. By spectral decomposition, $\boldsymbol{S}_n^{1/2}\Sigma^{-1}\boldsymbol{S}_n^{1/2}\boldsymbol{P} = \boldsymbol{P}\Lambda$, where $\Lambda = \boldsymbol{I}$.
Thus $l(\Sigma)$ is maximized when $\Sigma^{-1}\boldsymbol{S}_n = \boldsymbol{I}$, i.e., $\hat{\Sigma} = \boldsymbol{S}_n$.

(Another way is to use Result 4.10. See textbook.) By factorization theorem, sufficiency is clear from $\circledast^2$. □

Remark

MLE have an invariance property

That is, $\hat{\boldsymbol{\theta}}$: MLE of $\boldsymbol{\theta} \Rightarrow h(\hat{\boldsymbol{\theta}})$: MLE of $h(\boldsymbol{\theta})$ for any ftn h

ex) $\hat{\boldsymbol{\mu}}^\top \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}$: MLE of $\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{\mu}$, where $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ and $\hat{\Sigma} = \dfrac{n-1}{n}\mathbf{S}$ are MLE of $\boldsymbol{\mu}$ and Σ.

3.5.1 Sampling dist of sample mean vector and sample covariance matrix

Def Wishart distribution (generalization of scaled χ^2)

$\boldsymbol{z}_1 \cdots \boldsymbol{z}_m$: r.s. from $N_p\ (\mathbf{0}, \Sigma)$ and $\boldsymbol{W}_{p\times p} = \sum_{i=1}^{m} \boldsymbol{z}_i {\boldsymbol{z}_i}^\top$

$\Rightarrow$

The distribution of $\boldsymbol{W}$ is called a Wishart dist with m *df*, that is,

$\boldsymbol{W} \sim W_p(\Sigma, m) = W_{p,m}(\Sigma)$

Note $(n-1)S^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \sim \sigma^2 \chi^2_{n-1}$

So

$$\begin{aligned}(n-1)S^2 &= \sigma^2(z_1^2 + \cdots + z_{n-1}^2)\\ &= (\sigma z_1)^2 + \cdots + (\sigma z_{n-1})^2, \text{ where } z_i \sim iid\ N(0,1)\end{aligned}$$

$\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$: r.s. form $N_p(\boldsymbol{\mu}, \Sigma)$

$\Rightarrow$

$$\begin{cases} \text{i) } \bar{\boldsymbol{x}} \sim N_p\left(\boldsymbol{\mu}, \dfrac{1}{n}\Sigma\right) \\ \text{ii) } (n-1)\mathbf{S} \sim \text{Wishart distribution with } n-1\ df \\ \text{iii) } \bar{\mathbf{x}} \perp\!\!\!\perp \mathbf{S} \end{cases}$$

cf. See the H/O for a simple derivation of the Wishart distribution.

3.5.2 Large-Sample Behavior of sample mean vector and sample covariance matrix

$\boldsymbol{x}_1 \cdots \boldsymbol{x}_n$: r.s. from any population with mean $\boldsymbol{\mu}$ and finite covariance Σ

$\Rightarrow$

i) $\sqrt{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \dot{\sim} N_p(\mathbf{0}, \Sigma)$ as $n \to \infty$

$\Leftrightarrow \bar{\boldsymbol{x}} \dot{\sim} N_p\left(\boldsymbol{\mu}, \frac{1}{n}\Sigma\right)$; Central Limit Theorem

ii) $n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \dot{\sim} \chi^2_p$ for large $n - p$

<u>Note</u>

$$\left\{\begin{array}{l} ① \ \boldsymbol{x} \ \sim \ N_p(\boldsymbol{\mu}, \Sigma) \\ \Rightarrow (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \ \sim \ \chi^2_p \\ \\ ② \ \underline{\text{Univariate}} \\ \\ \dfrac{\bar{x} - \mu}{S/\sqrt{n}} \ \xrightarrow[n \to \infty]{} \ N(0,1) \\ \\ \Rightarrow \left(\dfrac{\bar{x} - \mu}{S/\sqrt{n}}\right)^2 \ \xrightarrow[n \to \infty]{} \ \chi^2_1 \end{array}\right.$$

3.6 Examining the data & assessing the Normality

3.6.1 Normality

Marginal distributions

① Histograms
② Q-Q plot(Quantile-Quantile plot)

③ Test; Shapiro-Wilk etc.

Remark Quantile function

$$\begin{aligned}\text{pop}: \quad Q(u) &= \inf\{x : F(x) \geq u\} \ (u : \text{prob.}) \\ &= F^{-1}(u) \text{ if F is conti \& strictly increasing} \\ \text{sample}: \quad \widehat{Q}_n(u) &= \inf\{x : \widehat{F}_n(x) \geq u\} \text{ for } \frac{1}{n} \leq u \leq \frac{n-1}{n}\end{aligned}$$

ex)

$$\widehat{Q}_n(u) = \begin{cases} x_{(1)} & if\ u \leq \frac{1}{n} \\ x_{(i)} & if\ \frac{i-1}{n} < u \leq \frac{i}{n} \\ x_{(n)} & if\ u > \frac{n-1}{n} \end{cases},$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ be an ordered sample.

Figure of $\widehat{Q}_n(u)$ in example:

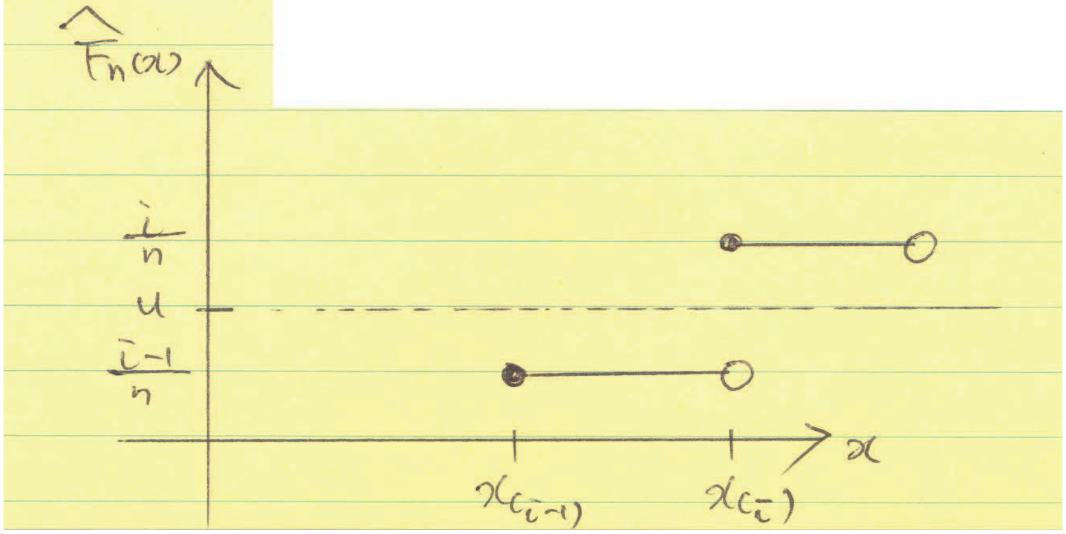


Figure 3.6.1: Empirical CDF

Plot sample quantile vs assumed pop. quantitle.

If they are similar, then assumption holds.
That is,

$$\begin{aligned} X \sim N(\mu, \sigma^2) \ &\Rightarrow \ \frac{X-\mu}{\sigma} \sim N(0,1) \\ &\Rightarrow \ \frac{X_{(i)} - \bar{X}}{S} \simeq \ q_N\left(\frac{i-\frac{1}{2}}{n}\right), \ i = 1, \cdots, n \\ q_n(\alpha) \ &: \ \alpha^{th} \text{ quantile of } N(0,1) \end{aligned}$$

Remark

① Plot of $\left(\frac{x_{(i)} - \bar{x}}{s} \ vs. \ q_N\left(\frac{i - \dfrac{1}{2}}{n}\right)\right)$ resemble a straight line through the origin $\Rightarrow$ Normality is O.K.

② Plot of $\left(x_{(i)} \ vs. \ q_N\left(\frac{i - \dfrac{1}{2}}{n}\right)\right)$ lie nearly along a straight line

$\Rightarrow$ Normality is O.K.
$\left(\because \ X_{(i)} \simeq \bar{x} + S \cdot q_N\left(\frac{i-\frac{1}{2}}{n}\right)\right)$

③ $\dfrac{1}{2}$: continuity correction

④ Improvement: Replace $\dfrac{i-\frac{1}{2}}{n}$ by $\dfrac{i-\frac{3}{8}}{n+\frac{1}{4}}$

⑤ See SAS program for proc univariate.

Multivariate normal dist.

$$\begin{aligned} \boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \Sigma) \ &\Rightarrow \ (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \sim \chi^2_p \\ &\Rightarrow \ d^2_{(i)} \simeq q_{c,p}\left(\frac{i-\frac{1}{2}}{n}\right), \ i = 1, \cdots, n, \\ \text{where} \quad d^2_{(1)} \ &\leq \ d^2_{(2)} \ \leq \ \cdots \ \leq \ d^2_{(n)} : \text{ordered } d^2_i = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \ \& \\ q_{c,p}(\alpha) \ &: \ \alpha^{th} \text{ quantile of } \chi^2_p \end{aligned}$$

Remark

① Plot of $\left(d^2_{(i)}\ vs.\ q_{c,p}\left(\frac{i-\frac{1}{2}}{n}\right)\right)$ resembles a straight line through the origin $\Rightarrow$ Normality is O.K.

② See SAS program.

3.6.2 Transformations to Normality

Variance stabilizing transformations

Suppose the variance of x depends on the mean. That is, $Ex = \mu,\ V(x) = \sigma^2 h(\mu)$.

The idea is to find a variable $z = g(x)$ such that the variance of z does not depend on Ez & $Ez \cong g(\mu)$.

ex) When sampling from the bivariate normal dist., sample correlation coefficient, r, has variance proportional to $(1-\mu^2)^2$.

In this case, the appropriate transformation is.

$$z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right), \quad \text{by Fisher.}$$

In has been shown that, z is approximately normal with

$$Ez = \frac{1}{2}\log_e\frac{1+\rho}{1-\rho} \quad \& \quad Vz = \frac{1}{n-3}.$$

Other useful transformations are

$$\begin{aligned} Vx \propto \mu \quad &\Rightarrow \quad z = \sqrt{x} \\ Vx \propto \mu^2 \quad &\Rightarrow \quad z = \log_e x \\ Vx \propto \mu^4 \quad &\Rightarrow \quad z = \frac{1}{x}. \end{aligned}$$

Box-Cox transformation

The last three transformations above have z as a power of x. The Box-Cox transformation uses the data to determine the power. That is,

$$z_i^{(\lambda)} = \begin{cases} \dfrac{x_i^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log_e x_i, & \lambda = 0 \end{cases} .$$

The transformation hold for $x_i > 0$. The parameter λ is estimated using the profile likelihood function. `boxCox` or `powerTransform` will do the job in R.

The Yeo-Johnson transformation (2000) allows also for zero and negative values of x_i. The allowed range of the parameter is $0 \leq \lambda \leq 2$, where $\lambda = 1$ produces the identity transformation. The transformation law reads:

$$z_i^{(\lambda)} = \begin{cases} ((x_i+1)^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, x \geq 0 \\ \log_e(x_i+1) & \text{if } \lambda = 0, x \geq 0 \\ -[(-x_i+1)^{(2-\lambda)} - 1]/(2-\lambda) & \text{if } \lambda \neq 2, x < 0 \\ -\log_e(-x_i+1) & \text{if } \lambda = 2, x < 0 \end{cases}$$

In R, `yeo.johnson` computes the Yeo-Johnson transformation.

3.7 Miscellanea

3.7.1 Distribution of quadratic forms in normal variates

Note

1. **A**: symmetric and idempotent ⇔ the nonzero eigenvalues are all equal to one

2. $Z_i \sim N(0,1), i = 1, \ldots, r$ independently $\Rightarrow \sum\limits_{i=1}^r Z_i^2 \sim \chi_r^2$

3. $Y_i \sim N(\mu_i, 1), i = 1, \ldots, r$ independently $\Rightarrow \sum\limits_{i=1}^r Y_i^2 \sim \chi_r^2(\lambda)$, where $\lambda = \frac{1}{2}\sum\limits_{i=1}^r \mu_i^2$. λ is the noncentrality parameter.

Theorem 1 *Let $\mathbf{x} \sim N_m(\mathbf{0}, \mathbf{I}_m)$, and suppose that the $m \times m$ matrix $\mathbf{A}$ is symmetric, idempotent, and has rank r. Then $\mathbf{x}'\mathbf{A}\mathbf{x} \sim \chi^2_r$.*

pf) Since $\mathbf{A}$ is symmetric and idempotent, there exists an orthogonal matrix $\mathbf{P}$ such that

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}',$$

where $\mathbf{\Lambda} = diag(\mathbf{I}_r, (0))$. Let $\mathbf{z} = \mathbf{P}'\mathbf{x}$ and note that since $\mathbf{x} \sim N_m(\mathbf{0}, \mathbf{I}_m)$,

$$\begin{aligned} E(\mathbf{z}) &= E(\mathbf{P}'\mathbf{x}) = \mathbf{P}'E(\mathbf{x}) = \mathbf{P}'\mathbf{0} = \mathbf{0} \\ Var(\mathbf{z}) &= Var(\mathbf{P}'\mathbf{x}) = \mathbf{P}'\{Var(\mathbf{x})\}\mathbf{P} = \mathbf{P}'\mathbf{I}_m\mathbf{P} = \mathbf{P}'\mathbf{P} = \mathbf{I}_m, \end{aligned}$$

and so $\mathbf{z} \sim N_m(\mathbf{0}, \mathbf{I}_m)$; that is, the components of $\mathbf{z}$ are, like the components of $\mathbf{x}$, independent standard normal random variables. Now due to the form of $\mathbf{\Lambda}$, we find that

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{P}\mathbf{\Lambda}\mathbf{P}'\mathbf{x} = \mathbf{z}'\mathbf{\Lambda}\mathbf{z} = \sum_{i=1}^{r} Z_i^2,$$

and the result then follows. □

The next theorem is for a positive definite matrix $\mathbf{\Sigma}$.

Theorem 2 *Let $\mathbf{x} \sim N_m(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a positive definite matrix, and let $\mathbf{A}$ be an $m \times m$ symmetric matrix. If $\mathbf{A}\mathbf{\Sigma}$ is idempotent and rank($\mathbf{A}\mathbf{\Sigma}$) $= r$, then $\mathbf{x}'\mathbf{A}\mathbf{x} \sim \chi^2_r$.*

pf) Since $\mathbf{\Sigma}$ is positive definite, there exists a nonsingular matrix $\mathbf{T}$ such that $\mathbf{\Sigma} = \mathbf{T}\mathbf{T}'$. Let $\mathbf{z} = \mathbf{T}^{-1}\mathbf{x}$, then $E\mathbf{z} = \mathbf{T}^{-1}E\mathbf{x} = \mathbf{0}$, and

$$Var(\mathbf{z}) = Var(\mathbf{T}^{-1}\mathbf{x}) = \mathbf{T}^{-1}Var(\mathbf{x})\mathbf{T}^{'-1} = \mathbf{T}^{-1}\mathbf{T}\mathbf{T}'\mathbf{T}^{'-1} = \mathbf{I}_m,$$

so that $\mathbf{z} \sim N_m(\mathbf{0}, \mathbf{I}_m)$. Note that

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{T}^{'-1}\mathbf{T}'\mathbf{A}\mathbf{T}\mathbf{T}^{-1}\mathbf{x} = \mathbf{z}'\mathbf{T}'\mathbf{A}\mathbf{T}\mathbf{z}.$$

All that remains is to show that $\mathbf{T}'\mathbf{A}\mathbf{T}$ satisfies the conditions of Theorem 1. First, $\mathbf{T}'\mathbf{A}\mathbf{T}$ is symmetric and idempotent since

$$(\mathbf{T}'\mathbf{A}\mathbf{T})^2 = \mathbf{T}'\mathbf{A}\mathbf{T}\mathbf{T}'\mathbf{A}\mathbf{T} = \mathbf{T}'\mathbf{A}\mathbf{\Sigma}\mathbf{A}\mathbf{T} = \mathbf{T}'\mathbf{A}\mathbf{T},$$

where the last equality follows from the identity $\mathbf{A\Sigma A} = \mathbf{A}$. ($\because$ $\mathbf{A\Sigma}$ is idempotent, $\mathbf{A\Sigma A\Sigma} = \mathbf{A\Sigma}$. Multiplying $\mathbf{\Sigma}^{-1}$ to both sides, the result follows.) Finally

$$rank(\mathbf{T}'\mathbf{AT}) = tr(\mathbf{T}'\mathbf{AT}) = tr(\mathbf{ATT}') = tr(\mathbf{A\Sigma}) = rank(\mathbf{A\Sigma}) = r,$$

and so the proof is complete. □

Finally previous two theorems are extended for general $\boldsymbol{\mu} \neq \mathbf{0}$.

Theorem 3 *Let $\mathbf{x} \sim N_m(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a positive definite matrix, and let $\mathbf{A}$ be an $m \times m$ symmetric matrix. If $\mathbf{A\Sigma}$ is idempotent and $rank(\mathbf{A\Sigma}) = r$, then $\mathbf{x}'\mathbf{Ax} \sim \chi^2_r(\lambda)$, where $\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$.*

pf) Since $\mathbf{\Sigma}$ is positive definite, there exists a nonsingular matrix $\mathbf{T}$ such that $\mathbf{\Sigma} = \mathbf{TT}'$. Let $\mathbf{y} = \mathbf{T}^{-1}\mathbf{x}$, then $E\mathbf{y} = \mathbf{T}^{-1}E\mathbf{x} = \mathbf{T}^{-1}\boldsymbol{\mu}$, and

$$Var(\mathbf{y}) = Var(\mathbf{T}^{-1}\mathbf{x}) = \mathbf{T}^{-1}Var(\mathbf{x})\mathbf{T}^{'-1} = \mathbf{T}^{-1}\mathbf{TT}'\mathbf{T}^{'-1} = \mathbf{I}_m,$$

so that $\mathbf{y} \sim N_m(\mathbf{T}^{-1}\boldsymbol{\mu}, \mathbf{I}_m)$. The quadratic form $\mathbf{x}'\mathbf{Ax}$ can be written as

$$\mathbf{x}'\mathbf{Ax} = \mathbf{x}'\mathbf{T}^{'-1}\mathbf{T}'\mathbf{ATT}^{-1}\mathbf{x} = \mathbf{y}'\mathbf{T}'\mathbf{ATy}.$$

Note that $\mathbf{T}'\mathbf{AT}$ is symmetric and idempotent since

$$(\mathbf{T}'\mathbf{AT})^2 = \mathbf{T}'\mathbf{ATT}'\mathbf{AT} = \mathbf{T}'\mathbf{A\Sigma AT} = \mathbf{T}'\mathbf{AT}.$$

($\because$ $\mathbf{A\Sigma}$ is idempotent, $\mathbf{A\Sigma A\Sigma} = \mathbf{A\Sigma}$. Multiplying $\mathbf{\Sigma}^{-1}$ to both sides, the result follows.) Since $\mathbf{T}'\mathbf{AT}$ is symmetric and idempotent, there exists an orthogonal matrix $\mathbf{P}$ such that

$$\mathbf{T}'\mathbf{AT} = \mathbf{P\Lambda P}',$$

where $\mathbf{\Lambda} = diag(\mathbf{I}_r, (0))$. Because

$$rank(\mathbf{T}'\mathbf{AT}) = tr(\mathbf{T}'\mathbf{AT}) = tr(\mathbf{ATT}') = tr(\mathbf{A\Sigma}) = rank(\mathbf{A\Sigma}) = r.$$

Let $\mathbf{z} = \mathbf{P}'\mathbf{y}$, then $\mathbf{z} \sim N_m(\mathbf{P}'\mathbf{T}^{-1}\boldsymbol{\mu}, \mathbf{P}'\mathbf{I}_m\mathbf{P} = \mathbf{I}_m)$. The quadratic form $\mathbf{x}'\mathbf{Ax}$ can be written as

$$\mathbf{x}'\mathbf{Ax} = \mathbf{y}'\mathbf{T}'\mathbf{ATy} = \mathbf{y}'\mathbf{P\Lambda P}'\mathbf{y} = \mathbf{z}'\mathbf{\Lambda z} = \sum_{i=1}^{r} Z_i^2,$$

where $Z_i \sim N(E(Z_i), 1), i = 1, \ldots, m$. Therefore, by Note 3,

$$\mathbf{x}'\mathbf{A}\mathbf{x} \sim \chi^2_r(\lambda),$$

where

$$\begin{aligned}\lambda &= \frac{1}{2}\sum_{i=1}^{r}(EZ_i)^2 = \frac{1}{2}\boldsymbol{\mu}'\mathbf{T}'^{-1}\mathbf{P}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\mathbf{P}'\mathbf{T}^{-1}\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}'\mathbf{T}'^{-1}\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'\mathbf{T}^{-1}\boldsymbol{\mu} \\ &= \frac{1}{2}\boldsymbol{\mu}'\mathbf{T}'^{-1}\mathbf{T}'\mathbf{A}\mathbf{T}\mathbf{T}^{-1}\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}.\end{aligned}$$

Note that $(EZ_1\ \ldots\ EZ_r\ 0\ \ldots\ 0)' = \boldsymbol{\Lambda}\mathbf{P}'\mathbf{T}^{-1}\boldsymbol{\mu}$. □

Remark There are similar theorems for positive semidefinite matrix $\boldsymbol{\Sigma}$. Here we did for positive definite matrix $\boldsymbol{\Sigma}$.

We give a lemma useful for proving independence of two quadratic forms.

Lemma 1 *Suppose that* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $m \times m$ *symmetric matrices. Then an orthogonal matrix* $\mathbf{P}$ *exists, such that* $\mathbf{P}'\mathbf{A}\mathbf{P}$ *and* $\mathbf{P}'\mathbf{B}\mathbf{P}$ *are both diagonal if and only if* $\mathbf{A}$ *and* $\mathbf{B}$ *commute; that is, if and only if* $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$.

pf) See Theorem 4.17 of Schott (2005) for proof. □

We begin with the following basic result regarding the statistical independence of two quadratic forms in the same normal vector.

Theorem 4 *Let* $\mathbf{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Omega})$, *where* $\boldsymbol{\Omega}$ *is positive definite, and suppose that* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $m \times m$ *symmetric matrices. If* $\mathbf{A}\boldsymbol{\Omega}\mathbf{B} = \mathbf{0}$, *then* $\mathbf{x}'\mathbf{A}\mathbf{x}$ *and* $\mathbf{x}'\mathbf{B}\mathbf{x}$ *are independently distributed.*

pf) Since $\boldsymbol{\Omega}$ is positive definite, a nonsingular matrix $\mathbf{T}$ exists, such that $\boldsymbol{\Omega} = \mathbf{T}\mathbf{T}'$. Define $\mathbf{G} = \mathbf{T}'\mathbf{A}\mathbf{T}$ and $\mathbf{H} = \mathbf{T}'\mathbf{B}\mathbf{T}$, and note that if $\mathbf{A}\boldsymbol{\Omega}\mathbf{B} = \mathbf{0}$, then

$$\mathbf{G}\mathbf{H} = (\mathbf{T}'\mathbf{A}\mathbf{T})(\mathbf{T}'\mathbf{B}\mathbf{T}) = \mathbf{T}'\mathbf{A}\boldsymbol{\Omega}\mathbf{B}\mathbf{T} = \mathbf{T}'\mathbf{0}\mathbf{T} = \mathbf{0}. \qquad (3.7.1)$$

Consequently, because of the symmetry of $\mathbf{G}$ and $\mathbf{H}$, we also have

$$\mathbf{0} = \mathbf{0}' = (\mathbf{G}\mathbf{H})' = \mathbf{H}'\mathbf{G}' = \mathbf{H}\mathbf{G},$$

and so we have established that $\mathbf{GH} = \mathbf{HG}$ (commute). By Lemma 1, we know that an orthogonal matrix $\mathbf{P}$ exists that simultaneously diagonalizes $\mathbf{G}$ and $\mathbf{H}$; that is, for some diagonal matrices $\mathbf{C}$ and $\mathbf{D}$,

$$\mathbf{P}'\mathbf{G}\mathbf{P} = \mathbf{P}'\mathbf{T}'\mathbf{A}\mathbf{T}\mathbf{P}, \quad \mathbf{P}'\mathbf{T}'\mathbf{B}\mathbf{T}\mathbf{P} = \mathbf{D}. \tag{3.7.2}$$

However, using (3.7.1) and (3.7.2), we find that

$$\mathbf{0} = \mathbf{GH} = \mathbf{PCP}'\mathbf{PDP}' = \mathbf{PCDP}'.$$

Furthermore

$$\mathbf{CD} = \mathbf{0}$$

since $\mathbf{P}$ is an orthogonal matrix. Since $\mathbf{C}$ and $\mathbf{D}$ are diagonal matrices, this means that if the ith diagonal element of one of these matrices is nonzero, the ith diagonal element of the other must be zero. As a result, by choosing $\mathbf{P}$ appropriately, we may obtain $\mathbf{C}$ and $\mathbf{D}$ in the form

$$\mathbf{C} = diag(c_1, \cdots, c_{m_1}, 0, \cdots, 0) \text{ and } \mathbf{D} = diag(0, \cdots, 0, d_{m_1+1}, \cdots, d_m)$$

for some integer m_1. If we let $\mathbf{y} = \mathbf{P}'\mathbf{T}^{-1}\mathbf{x}$, then

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{T}^{-1'}\mathbf{PP}'\mathbf{T}'\mathbf{ATPP}'\mathbf{T}^{-1}\mathbf{x} = \mathbf{y}'\mathbf{C}\mathbf{y} = \sum_{i=1}^{m_1} c_i y_i^2$$

and

$$\mathbf{x}'\mathbf{B}\mathbf{x} = \mathbf{x}'\mathbf{T}^{-1'}\mathbf{PP}'\mathbf{T}'\mathbf{BTPP}'\mathbf{T}^{-1}\mathbf{x} = \mathbf{y}'\mathbf{D}\mathbf{y} = \sum_{i=m_1+1}^{m} d_i y_i^2;$$

that is, the first quadratic form is a function only of $y_1, \cdots, y_{m_1}$, whereas the second quadratic form is a function of $y_{m_1+1}, \cdots, y_m$. Note that

$$\mathbf{y} \sim N_m(\mathbf{P}'\mathbf{T}^{-1}\boldsymbol{\mu}, \mathbf{I}_m)$$

since

$$var(\mathbf{y}) = var(\mathbf{P}'\mathbf{T}^{-1}\mathbf{x}) = \mathbf{P}'\mathbf{T}^{-1}\boldsymbol{\Omega}\mathbf{T}^{-1'}\mathbf{P} = \mathbf{I}_m.$$

The result follows from the independence of $y_1, \cdots, y_m$. □

The proof Theorem 5, which is similar to the proof of Theorem 4, is left to the reader as an exercise.

Theorem 5 *Let $\mathbf{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is positive definite, and suppose that $\mathbf{A}$ is an $m \times m$ symmetric matrix, whereas $\mathbf{B}$ is an $n \times m$ matrix. If $\mathbf{B\Omega A} = \mathbf{0}$, then $\mathbf{x}'\mathbf{Ax}$ and $\mathbf{Bx}$ are independently distributed.*

Our final result can be helpful in establishing that several quadratic forms in the same normal random vector are independently distributed, with each having a chi-squared distribution. We assume that Cochran's Theorem and related stuffs will be appeared.

Theorem 6 *Let $\mathbf{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is positive definite. Suppose that $\mathbf{A}_i$ is an $m \times m$ symmetric matrix of rank r_i, for $i = 1, \cdots k$, and $\mathbf{A} = \mathbf{A}_1 + \cdots \mathbf{A}_k$ is of rank r. Consider the conditions*

(a) $\mathbf{A}_i\boldsymbol{\Omega}$ *is idempotent for each* i,

(b) $\mathbf{A}\boldsymbol{\Omega}$ *is idempotent,*

(c) $\mathbf{A}_i\boldsymbol{\Omega}\mathbf{A}_j = \mathbf{0}$, *for all* $i \neq j$,

(d) $r = \sum_{i=1}^{k} r_i$.

If any two of (a), (b), and (c) hold, or if (b) and (d) hold, then

(i) $\mathbf{x}'\mathbf{A}_i\mathbf{x} \sim \chi^2_{r_i}\left(\frac{1}{2}\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu}\right)$,

(ii) $\mathbf{x}'\mathbf{Ax} \sim \chi^2_{r}\left(\frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\right)$,

(iii) $\mathbf{x}'\mathbf{A}_1\mathbf{x}, \cdots, \mathbf{x}'\mathbf{A}_k\mathbf{x}$ *are independently distributed.*

pf) Since $\boldsymbol{\Omega}$ is positive definite, a nonsingular matrix $\mathbf{T}$ satisfying $\boldsymbol{\Omega} = \mathbf{TT}'$ exists, and the conditions (a)-(d) can be equivalently expressed as

(a) $\mathbf{T}'\mathbf{A}_i\mathbf{T}$ is idempotent for each i,

(b) $\mathbf{T}'\mathbf{AT}$ is idempotent,

(c) $(\mathbf{T}'\mathbf{A}_i\mathbf{T})(\mathbf{T}'\mathbf{A}_j\mathbf{T}) = \mathbf{0}$, for all $i \neq j$,

(d) $rank(\mathbf{T}'\mathbf{AT}) = \sum_{i=1}^{k} rank(\mathbf{T}'\mathbf{A}_i\mathbf{T})$.

Since$\mathbf{T}'\mathbf{A}_1\mathbf{T}, \cdots, \mathbf{T}'\mathbf{A}_k\mathbf{T}$ and $\mathbf{T}'\mathbf{AT}$ satisfy the conditions of Corollary 10.7.1, we are ensured that if any two of (a), (b), and (c) hold or if (b) and (d) hold, then all four of the conditions (a)-(d) hold. Now using Theorem 3, (a) implies (i) and (b) implies (ii), whereas Theorem 4, along with (c), guarantees that (iii) holds. □

3.7.2 Cochran's Theorem

3.8 The multivariate normal distribution in R

3.8.1 Density and visualization

A bivariate normal distribution with

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1/\sqrt{2} \\ 1/\sqrt{2} & 1 \end{pmatrix}$$

will be used.

```
> mu.vec <- c(0,2)
> sig.mat <- matrix(c(2,1/sqrt(2),1/sqrt(2),1),ncol=2)
> variable.num <- 2
> x.vec <- c(2,1) # given x
```

We can use the package 'mvtnorm' for multivariate normal and t distributions. This package computes multivariate normal and t probabilities, quantiles, random deviates and densities.

Density calculation at a given $\mathbf{x} = (2, 1)^\top$ with the mean and the covariance matrix is as follows:

```
> install.packages("mvtnorm")
> library(mvtnorm)
> density1 <- dmvnorm(x=x.vec,mean=mu.vec, sigma=sig.mat)
> density1
[1] 0.006850603
```

To visualize above bivariate normal distribution, first generate samples and do kernel smoothing using package 'MASS' (n is the number of grid points).

```
> install.packages("MASS")
> library(MASS)
> sample.vec <- mvrnorm(1000, mu=mu.vec, Sigma=sig.mat)
> # random mvnorm given in MASS package
> sample.kde <- kde2d(sample.vec[,1], sample.vec[,2], n = 50)
> # kernel density estimation for smoothing the observations
```

Next, plotting code for contour, image, and perspective plots (phi, theta are for the angle).

```
> layout(matrix(1:4,2,2))
> plot1 <- contour(sample.kde)
> plot2 <- persp(sample.kde, phi = 45, theta = 30)
> plot3 <- image(sample.kde) ; contour(sample.kde, add = T)
> plot4 <- persp(sample.kde, phi = 30, theta = 30)
> layout(matrix(1))
```

Another way to plot a bivariate normal probability density is as follows:

```
> # bivariate normal probability density
> mu1 <- 0   # mean of x1
> mu2 <- 0   # mean of x2
> s11 <- 10  # variance of x1
> s12 <- 15  # covariance between x1 and x2
> s22 <- 10  # variance of x2
> rho <- 0.5 # correlation between x1 and x2
> x1 <- seq(-10, 10, length=41)
> x2 <- x1
>
> f <- function(x1, x2) {
+  term1 <- 1/(2*pi*sqrt(s11*s22*(1-rho^2)))
+  term2 <- -1/(2*(1-rho^2))
+  term3 <- (x1-mu1)^2/s11
+  term4 <- (x2-mu2)^2/s22
+  term5 <- 2*rho*((x1-mu1)*(x2-mu2))/(sqrt(s11)*sqrt(s22))
+  term1*exp(term2*(term3+term4-term5))
+ }
>
> z <- outer(x1,x2,f) # calculating the density values
>
> # perspective plot
```

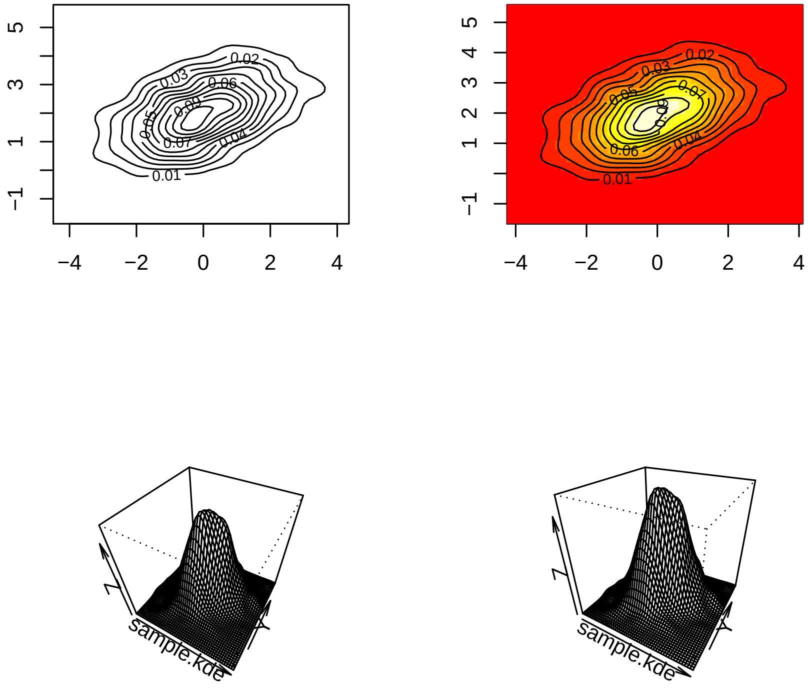


Figure 3.8.1: contour, image and perspective plots.

```
> layout(matrix(1:2,2,1))
> persp(x1,x2,z,main="Two dimensional Normal Distribution",
+  theta=30)
> # Contour plot
> contour(x1,x2,z,xlab="x1",ylab="x2",main="contour plot")
> layout(matrix(1))
```

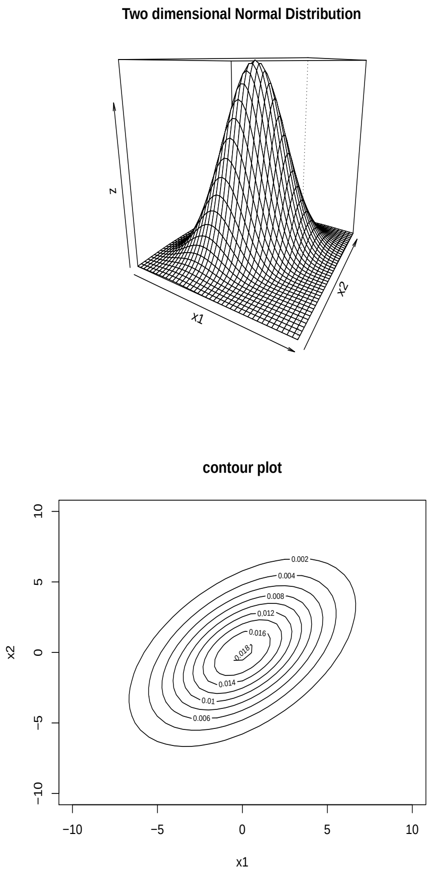


Figure 3.8.2: Density and contour plots.

3.8.2 Conditional distribution

We can use the package 'condMVNnorm'.

```
> install.packages("condMVNorm")
> library(condMVNorm)
> result.con <- condMVN(mean=mu.vec, sigma=sig.mat,
+  dependent.ind=1, given.ind=2, X.given=1)
> result.con
$condMean
[1] -0.7071068
```

```

$condVar
     [,1]
[1,]  1.5
```

- ‘condMVN’ returns conditional mean and conditional covariance matrix.
- ‘dependent.ind’ is a vector of integers denoting the indices of dependent variable Y.
- ‘given.ind’ is a vector of integers denoting the indices of conditioning variable X.
- ’X.given’ is a vector of reals denoting the conditioning value of X.

Hence we know that

$$(X_1|X_2=1) \sim N_1\left(\mu_{1|2}=-0.71, \sigma^2_{1|2}=1.5\right).$$

ex) $\mathbf{x}=(x_1,x_2,x_3)^\top$ follows a multivariate normal distribution with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\mu}=\begin{pmatrix} -3 \\ 1 \\ 4 \end{pmatrix}, \boldsymbol{\Sigma}=\begin{pmatrix} -1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

We want to know the conditional distribution of $(X_2|X_1=x_1, X_3=x_3)$. Need to partition $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_{22}$, where

$$\boldsymbol{\Sigma}_{11}=\begin{pmatrix} 5 \end{pmatrix}, \boldsymbol{\Sigma}_{12}=\begin{pmatrix} -2 & 0 \end{pmatrix}, \boldsymbol{\Sigma}_{22}=\begin{pmatrix} -1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Similarly partition $\boldsymbol{\mu}$ as $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, where

$$\boldsymbol{\mu}_1=\begin{pmatrix} 1 \end{pmatrix}, \boldsymbol{\mu}_2=\begin{pmatrix} -3 \\ 4 \end{pmatrix}.$$

Above partitions are useful for directly calculating the conditional distributions. R program for finding a conditional distribution is as follows:

```
> ## conditional dist with X1 = 2, X3 = 3)
> mu.vec <- c(-3,1,4)
> sig.mat <- matrix(c(1,-2,0,-2,5,0,0,0,2),3,3)
> given.x.vec <- c(2,3) # let x_1 = 2, x_3 = 3 (constant)
> given.x.num <- c(1,3) # given X index
> result.con <- condMVN(mean=mu.vec, sigma=sig.mat,
+ dependent.ind=2, given.ind=given.x.num, X.given=given.x.vec)
> result.con
$condMean
[1] -9

$condVar
     [,1]
[1,]    1
```

So we find that

$$(X_2|X_1 = 2, X_3 = 3) \sim N_1\left(\mu_{2|1,3} = -9, \sigma^2_{2|1,3} = 1\right).$$

3.8.3 Normality tests

Data from Table 1.2, p. 15 in Applied Multivariate Statistical Analysis (Johnson and Wichern, 2007) will be used. All data can be downloaded at class website.

Specimen	Density	Machine direction	Cross direction
1	0.801	121.41	70.42
2	0.824	127.70	72.47
3	0.841	129.20	78.20
4	0.816	131.80	74.89
5	0.840	135.10	71.21

Table 3.8.1: Paper quality measurement data; 1 to 5 rows

Paper quality measurement data (sample size 41)

- ‘Specimen’ is the observation number.
- ‘Density’ stands for density (grams/cubic centimeter).

- ‘Machine direction’ means strength (pounds) in the machine direction.
- ‘Cross direction’ denotes strength (pounds) in the cross direction.

```
> exercise <- read.table("T1-2.dat")
> library(stats)
> ??stats::qqnorm # Q-Q plot function
> ??stats::shapiro.test # Shapiro-wilk test function
> exercise <- as.data.frame(exercise) # to give the name
> names(exercise) <- c("Density","Machine","Cross")
> boxplot(exercise)
```

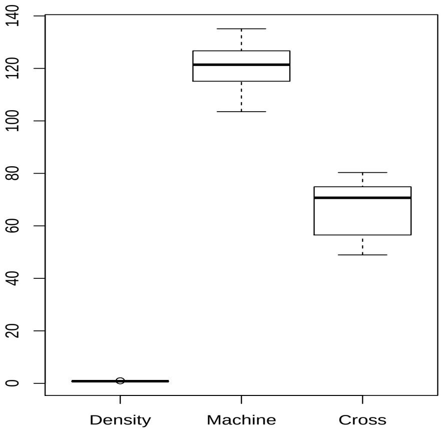


Figure 3.8.3: Box plots of raw data.

By the scale of each variables, it needs to be standardized to see properly. Note that, in R, `scale(x)=(x-mean(x))/sd(x)`.

```
> exercise1 <- as.data.frame(cbind(scale(exercise[,1]),
+  scale(exercise[,2]),scale(exercise[,3])))
> names(exercise1) <- c("Density","Machine","Cross")
> boxplot(exercise1)
```

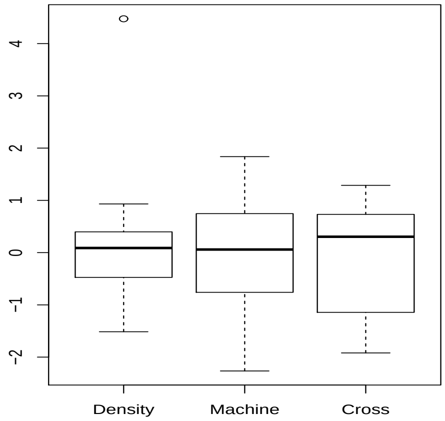


Figure 3.8.4: Box plots of standardized data.

Now it looks better, but there is one outlier in the Density so it will be removed.

```
> exercise[exercise$Density==max(exercise$Density),]
   Density Machine Cross
25   0.971   126.1  72.1
> exercise2 <- exercise1[-25,] # removing
> names(exercise2) <- c("Density","Machine","Cross")
> boxplot(exercise2)
```

After removing one outlier in the Density, it looks nice. Three data sets (raw, standardized, standardized without outlier) will be used to make histograms, Q-Q plots and Shapiro-Wilk tests.

```
> # Histograms of raw data
> p <- dim(exercise)[2]
> layout(matrix(1:p,ncol=p))
> hist(exercise$Density,main="")
> hist(exercise$Machine,main="")
```

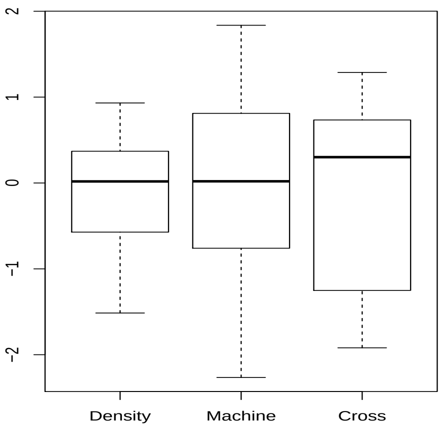


Figure 3.8.5: Box plots of standardized data with removing an ourlier.

```
> hist(exercise$Cross,main="")
> layout(matrix(1))
```

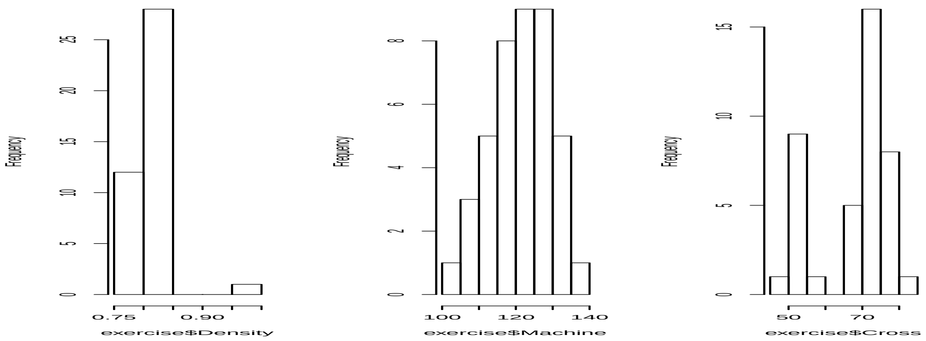


Figure 3.8.6: Histograms of raw data

See Figure 3.8.6. Histogram of Machine looks bit like a normal distribution, but not others.

```
> # Histograms of standardized data without outlier
> p <- dim(exercise2)[2]
> layout(matrix(1:p,ncol=p))
> hist(exercise2$Density,main="")
> hist(exercise2$Machine,main="")
> hist(exercise2$Cross,main="")
> layout(matrix(1))
```

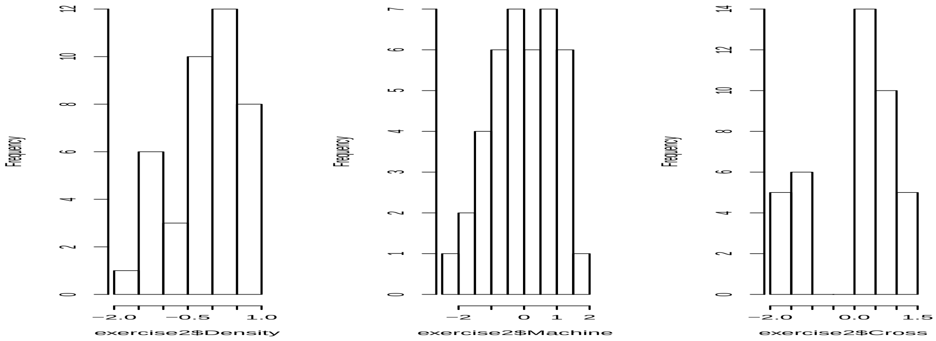


Figure 3.8.7: Histograms of standardized data with removing an ourlier.

See Figure 3.8.7. After transforming the data (standardizing and removing an outlier), the plots look better than before.

```
> # Q-Q plots of raw data
> p <- dim(exercise)[2]
> layout(matrix(1:p,ncol=p))
> qqnorm(exercise$Density)
> qqnorm(exercise$Machine)
> qqnorm(exercise$Cross)
> layout(matrix(1))
```

See Figure 3.8.8. Only Machine follows a line so it may follow a normal distribution. Similarly Q-Q plots of standardized data without outlier are drawn. See Figure 3.8.9.

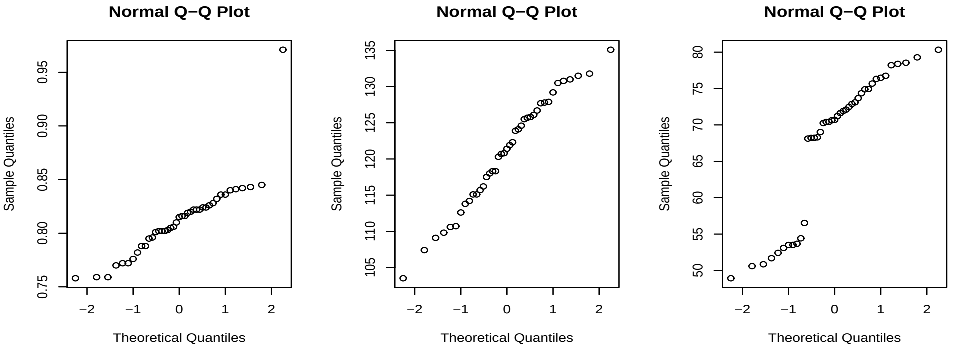


Figure 3.8.8: Normal Q-Q plots of raw data.

```
> # Q-Q plots of standardized data without outlier
> p <- dim(exercise2)[2]
> layout(matrix(1:p,ncol=p))
> qqnorm(exercise2$Density)
> qqnorm(exercise2$Machine)
> qqnorm(exercise2$Cross)
> layout(matrix(1))
```

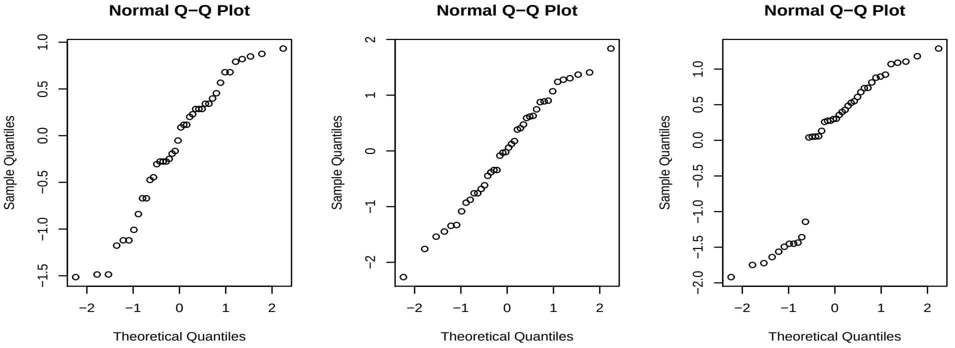


Figure 3.8.9: Normal Q-Q plots of standardized data without outlier.

So far we examined some figures. There are so many tests for normality.

Shapiro-Wilk test is a test of normality in frequentist statistics. It was published in 1965 by Samuel Sanford Shapiro and Martin Wilk. The Shapiro-Wilk test tests the null hypothesis that a sample $x_1, \cdots, x_n$ came from a normally distributed population.

```
> # Shapiro-Wilk test of raw data
> shapiro.test(exercise$Density)

        Shapiro-Wilk normality test

data:  exercise$Density
W = 0.8219, p-value = 1.642e-05

> shapiro.test(exercise$Machine)

        Shapiro-Wilk normality test

data:  exercise$Machine
W = 0.9773, p-value = 0.5741

> shapiro.test(exercise$Cross)

        Shapiro-Wilk normality test

data:  exercise$Cross
W = 0.8435, p-value = 5.212e-05
```

Only Machine follows a normal distribution with p-value of 0.5741.

```
> # Shapiro-Wilk test of standardized data without outlier
> shapiro.test(exercise2$Density)

        Shapiro-Wilk normality test

data:  exercise2$Density
W = 0.9462, p-value = 0.05599

> shapiro.test(exercise2$Machine)
```

```

        Shapiro-Wilk normality test

data:  exercise2$Machine
W = 0.9786, p-value = 0.6376

> shapiro.test(exercise2$Cross)

        Shapiro-Wilk normality test

data:  exercise2$Cross
W = 0.8463, p-value = 7.332e-05
```

Using $\alpha = 0.05$, Density and Machine follow normal distributions.

Box-Cox transformation

```
> # Box-Cox transformation (Univariate)
> # getting lambda (for each variable, univariate)
> exercise3 <- exercise[-25,]
> library(car)
> lambda.den <- powerTransform(exercise3$Density)
> lambda.machine <- powerTransform(exercise3$Machine)
> lambda.cross <- powerTransform(exercise3$Cross)
> lambda.vec <- c(as.vector(lambda.den$lambda),
+  as.vector(lambda.machine$lambda),
+  as.vector(lambda.cross$lambda))
>
> # powertransforming using the lambdas
> trans.Density <- (exercise3[,1]^lambda.vec[1] - 1)/lambda.vec[1]
> trans.Machine <- (exercise3[,2]^lambda.vec[2] - 1)/lambda.vec[2]
> trans.Cross   <- (exercise3[,3]^lambda.vec[3] - 1)/lambda.vec[3]
> trans.x.mat.uni <- cbind(trans.Density,trans.Machine,trans.Cross)
> # Q-Q plots
> p <- dim(exercise3)[2]
> layout(matrix(1:p,ncol=p))
> qqnorm(trans.x.mat.uni[,1])
> qqnorm(trans.x.mat.uni[,2])
> qqnorm(trans.x.mat.uni[,3])
> layout(matrix(1))
> lambda.vec
[1] 8.630528 2.539020 4.123345
```

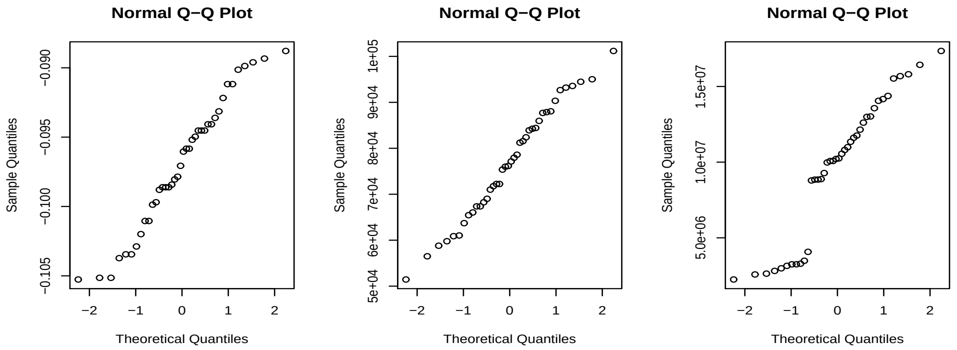


Figure 3.8.10: Normal Q-Q plots of box-cox transformed data without outlier (univariate).

Note that the scales of Machine and Cross in Sample Quantiles in Figure 3.8.10.

```
> # Box-cox transformation (multivariate)
> library(car)
> ??car::powerTransform
> # getting lambda (for all variables)
> exercise3 <- exercise[-25,]
> lambda.fit <- powerTransform(exercise3)
> lambda.vec <- as.vector(lambda.fit$lambda)
>
> # powertransforming using the lambdas
> trans.Density <- (exercise3[,1]^lambda.vec[1] - 1)/lambda.vec[1]
> trans.Machine <- (exercise3[,2]^lambda.vec[2] - 1)/lambda.vec[2]
> trans.Cross   <- (exercise3[,3]^lambda.vec[3] - 1)/lambda.vec[3]
> trans.x.mat.multi <- cbind(trans.Density,trans.Machine,trans.Cross)
>
```

```
> # Q-Q plots
> p <- dim(exercise3)[2]
> layout(matrix(1:p,ncol=p))
> qqnorm(trans.x.mat.multi[,1])
> qqnorm(trans.x.mat.multi[,2])
> qqnorm(trans.x.mat.multi[,3])
> layout(matrix(1))
> lambda.vec
[1] 3.6927590 0.7919834 3.0973121
```

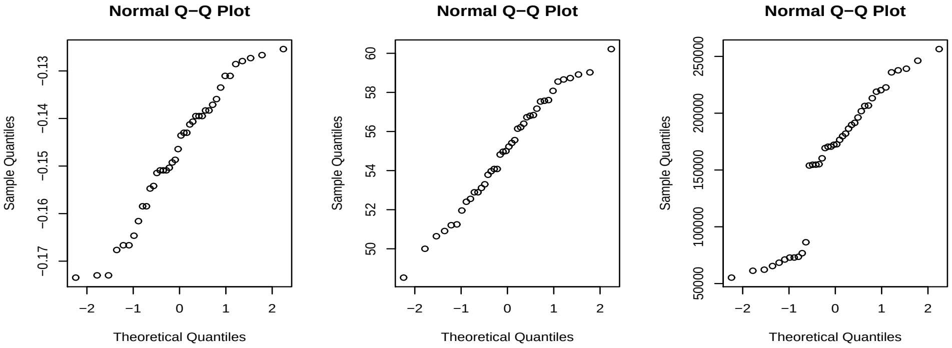


Figure 3.8.11: Normal Q-Q plots of box-cox transformed data without outlier (Multivariate).

The scales of Machine and Cross in Sample Quantiles in Figure 3.8.11 are a lot better than those of Figure 3.8.10.

Shapiro-Wilk multivariate normality test using R package 'mvnormtest' and the function 'mshapiro.test' in the package will be done.

```
> exercise <- as.matrix(exercise)   # raw data
> exercise1 <- as.matrix(exercise1) # standardized
> exercise2 <- as.matrix(exercise2) # standardized & removed 1 outlier
> exercise3 <- as.matrix(exercise3) # removed 1 outlier
> trans.x.mat.uni # powertransformed (univariate)
> trans.x.mat.multi # powertransformed (multivariate)
> # multivariate shapiro-wilk test
> library(mvnormtest)
> ??mvnormtest::mshapiro.test
> mshapiro.test(t(exercise))

        Shapiro-Wilk normality test

data:  Z
W = 0.5691, p-value = 8.969e-10

> mshapiro.test(t(exercise1))

        Shapiro-Wilk normality test

data:  Z
W = 0.5691, p-value = 8.969e-10

> mshapiro.test(t(exercise2))

        Shapiro-Wilk normality test

data:  Z
W = 0.978, p-value = 0.6154

> mshapiro.test(t(exercise3))

        Shapiro-Wilk normality test

data:  Z
W = 0.978, p-value = 0.6154

> mshapiro.test(t(trans.x.mat.uni))
  solve.default(R %*% t(R), tol = 1e-18) :
  system is computationally singular: reciprocal condition number = 3.0282e-1
> mshapiro.test(t(trans.x.mat.multi))

        Shapiro-Wilk normality test
```

```
data:  Z
W = 0.9708, p-value = 0.3811
```

Each `exercise2`, `exercise3` and `trans.x.mat.multi` follows a multivariate normal distribution, but not `exercise` and `exercise1`. And an error occurred at `trans.x.mat.uni`. The p-values of `exercise2` and `exercise3` are the same. The p-value of `trans.x.mat.multi` is little bit lower as 0.3811.

Chapter 4

Inferences about a mean vector

4.1 Hotelling's T^2

Recall that $X_1 \cdots X_n \sim\ iid\ N(\mu, \sigma^2),\ \sigma^2$: unknown

$\Rightarrow$

$$\begin{aligned}
t &= \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1),\ S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \ \ \& \\
t^2 &= n(\bar{X}-\mu)(S^2)^{-1}(\bar{X}-\mu) \\
&= \sqrt{n}(\bar{X}-\mu)(S^2)^{-1}\sqrt{n}(\bar{X}-\mu) \\
&= N(0,\sigma^2)\left(\frac{\sigma^2}{n-1}\chi^2_{n-1}\right)^{-1} N(0,\sigma^2) \sim F_1, n-1
\end{aligned}$$

$\boldsymbol{x}_1 \cdots \boldsymbol{x}_n$ r.s. from $N_p(\boldsymbol{\mu}, \Sigma),\ \Sigma$: unknown

$\Rightarrow$

$$\begin{aligned}
T^2 &= n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\ ;\ \text{Hotelling's } T^2 \\
&\sim \frac{(n-1)p}{n-p}\ F_{p,n-p},\ \text{where} \\
\bar{\boldsymbol{x}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i,\quad \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top,
\end{aligned}$$

and $\boldsymbol{x}_i$: transpose of i^{th} row of $\boldsymbol{X}$, data matrix, i.e.

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1{}^\top \\ \boldsymbol{x}_2{}^\top \\ \vdots \\ \boldsymbol{x}_n{}^\top \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ X_{21} & \cdots & X_{2p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}$$

<u>Remark</u>

$$\begin{aligned} T^2 &= \sqrt{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1} \sqrt{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \\ &= N_p(\mathbb{0}, \Sigma)^\top \left(\frac{1}{n-1} W_{p,n-1}(\Sigma) \right)^{-1} N_p(\mathbb{0}, \Sigma) \end{aligned}$$

$H_0 \ : \ \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1 \ : \ \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$

Reject $H_0 \ : \ \boldsymbol{\mu} = \boldsymbol{\mu}_0$ if $T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) > \dfrac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$, where $F_{p,n-p}(\alpha)$ is the upper $(100\alpha)^{th}$ percentile of the $F_{p,n-p}$ dist.

ex) See examaple 5.2, pg.214

<u>Remark</u> Invariance property of T^2 - statistic (invariant (unchanged) under changes in the units of measurements for $\boldsymbol{X}$).

$\boldsymbol{x} \sim N_p \ (\boldsymbol{\mu}_0, \, \Sigma)$

Let $\underset{p\times 1}{\boldsymbol{y}} = \underset{p\times p}{\boldsymbol{C}}\ \underset{p\times 1}{\boldsymbol{x}} + \underset{p\times 1}{\boldsymbol{d}}$, $\boldsymbol{C}$: nonsingular, then

$$\begin{aligned}
\text{i)}\quad & \bar{\boldsymbol{y}} = \boldsymbol{C}\bar{\boldsymbol{x}} + \boldsymbol{d},\\
\text{ii)}\quad & S_{\boldsymbol{y}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^\top\\
& = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{C}\boldsymbol{x}_i + \boldsymbol{d} - \boldsymbol{C}\bar{\boldsymbol{x}} - \boldsymbol{d})(\boldsymbol{C}\boldsymbol{x}_i + \boldsymbol{d} - \boldsymbol{C}\bar{\boldsymbol{x}} - \boldsymbol{d})^\top\\
& = \boldsymbol{C}\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top\boldsymbol{C}^\top\\
& = \boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top,\\
\text{iii)}\quad & E\boldsymbol{y} = \boldsymbol{C}\boldsymbol{\mu}_0 + \boldsymbol{d},\ V\boldsymbol{y} = \boldsymbol{C}\Sigma\boldsymbol{C}^\top,\ \text{and}
\end{aligned}$$

$$\begin{aligned}
T^2_{\boldsymbol{y}} \quad & = n(\bar{\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{y}})^\top\boldsymbol{S}^{-1}_{\boldsymbol{y}}(\bar{\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{y}})\\
& = n(\boldsymbol{C}\bar{\boldsymbol{x}} + \boldsymbol{d} - \boldsymbol{C}\boldsymbol{\mu}_0 - \boldsymbol{d})^\top(\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\bar{\boldsymbol{x}} - \boldsymbol{C}\boldsymbol{\mu}_0)\\
& = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top)^{-1}\boldsymbol{C}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\\
& = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) = T^2_{\boldsymbol{x}}\\
& \text{since } (\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top)^{-1} = (\boldsymbol{C}^\top)^{-1}\boldsymbol{S}^{-1}\boldsymbol{C}^{-1}.
\end{aligned}$$

Note

① For example, $Y_i = a_i(X_i - b_i)$ maybe the process of converting temperature from a Fahrenheit to a Celsius reading.

② Hotelling's T^2 test is equivalent to the likelihood test of

$H_0\ :\ \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1\ :\ \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ because

$$\Lambda^{2/n} = \left(1 + \frac{T^2}{n-1}\right)^{-1},\ \Lambda^{2/n} \text{ is called Wilks' lambda.}$$

See Section 5.3 of textbook to find it out.

③ H_0 is rejected for large values of T^2
$\Leftrightarrow$
H_0 is rejected for small values of $\Lambda^{2/n}$

4.2 Confidence regions

Def The region $R(\boldsymbol{x})$ is a $100(1-\alpha)\%$ confidence region if

$$P(R(\boldsymbol{x}) \text{ will cover the true } \theta) = 1-\alpha,$$

where $R(\boldsymbol{x})$ is determined by the data $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)^\top$.

So a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}$ is the ellipsoid determined by all $\boldsymbol{\mu}$ such that

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \le \frac{(n-1)p}{n-p}\, F_{p,n-p}(\alpha)$$

ex) See example 5.3 pg. 211 (See Addendum too!)

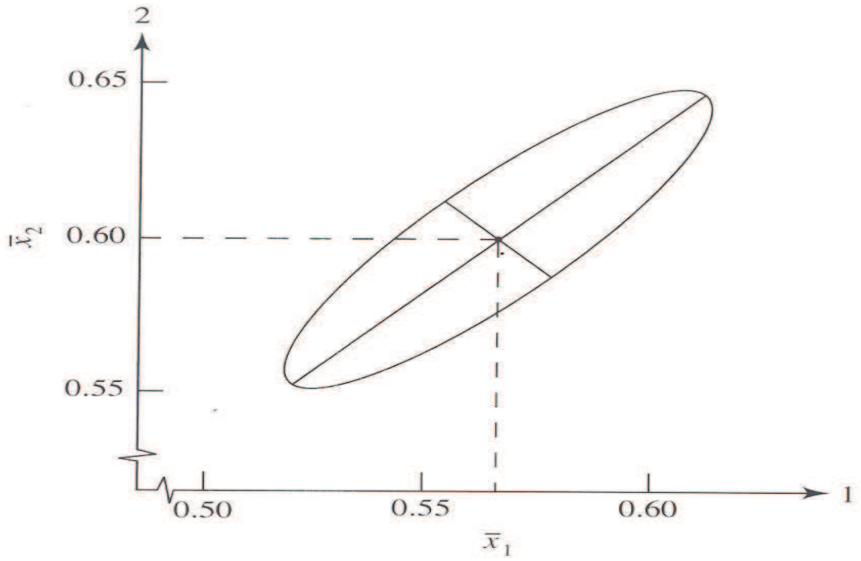


Figure 4.2.1: A 95% confidence ellipse for μ based on microwave radiation data.

Sometimes, we need the simultaneous intervals of $\boldsymbol{a}^\top \boldsymbol{\mu}$ **<u>for any</u>** $\underset{p\times 1}{\boldsymbol{a}}$ with probability $1-\alpha$. This is called as T^2 - intervals as follows (See Result 5.3.):

$$\boldsymbol{a}^\top \bar{\boldsymbol{x}} \pm \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha)\ \boldsymbol{a}^\top \boldsymbol{S} \boldsymbol{a}}$$

ex) $\boldsymbol{a}^\top = (0, 0, \cdots, 0, 1, 0, \cdots, 0)$, where 1 is i^{th} position

$$\Rightarrow \bar{x}_i \pm \sqrt{\frac{p(n-1)}{n-p} F_{p,n-p}(\alpha)} \sqrt{\frac{S_{ii}}{n}}, \quad i = 1, \cdots, p$$

$\boldsymbol{a}^\top = (0, \cdots, 0, a_i, 0, \cdots, 0, a_j, 0, \cdots, 0)$, where $a_i = 1, a_j = -1$

$$\Rightarrow \bar{x}_i - \bar{x}_j \pm \sqrt{\frac{p(n-1)}{n-p} F_{p,n-p}(\alpha)} \sqrt{\frac{S_{ii} - 2S_{ij} + S_{jj}}{n}}$$

ex) See example 5.4 pg.226 & the Figure 4.2.2.

<u>Remark</u>

- Simultaneous confidence intervals (T^2 - intervals) are shadows (projections) of the confidence ellipsoid.
- For **<u>a fixed</u>** $\boldsymbol{a}$, $100(1-\alpha)\%$ confidence interval for $\boldsymbol{a}^\top \boldsymbol{\mu}$ is

$$\boldsymbol{a}^\top \bar{\boldsymbol{x}} \pm t_{n-1}(\alpha/2) \frac{\sqrt{\boldsymbol{a}^\top \boldsymbol{S} \boldsymbol{a}}}{\sqrt{n}}.$$

  This interval is called as one-at-a-time interval which ignores the covariance structure of the p variables.

4.3 Bonferroni method

Suppose that we are interested in m linear combinations $\boldsymbol{a}_1^\top \boldsymbol{\mu}, \cdots, \boldsymbol{a}_m^\top \boldsymbol{\mu}$ prior to sampling.

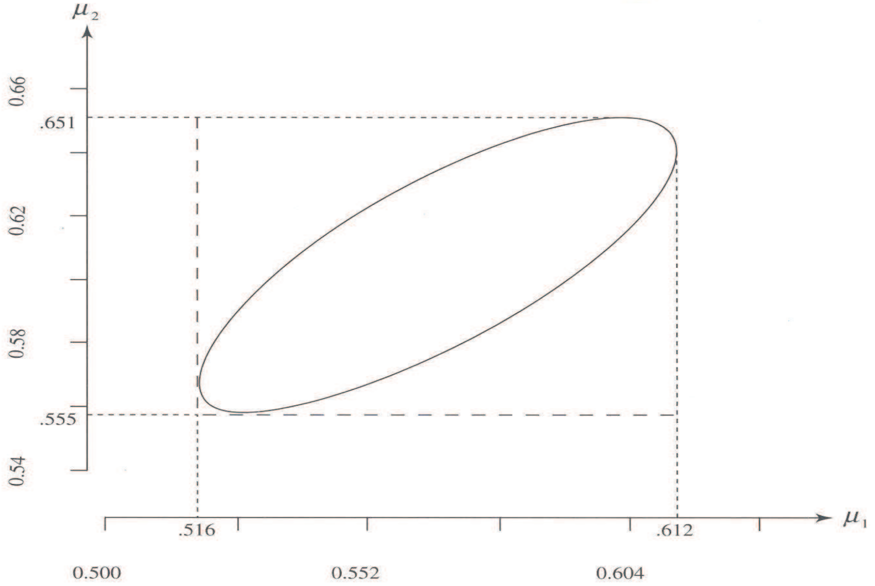


Figure 4.2.2: Simultaneous T^2- intervals for the component means as shadows of the confidence ellipse on the axes-microwave radiation data.

Let c_i denote a confidence statements about $\boldsymbol{a}_i^\top \boldsymbol{\mu}$ with $P(c_i \text{ true}) = 1-\alpha_i,\ i = 1, \cdots, m$, then

$$\begin{aligned} P(\text{all } c_i \text{ true}) &= 1 - P(\text{at least one } c_i \text{ false}) \\ &\geq 1 - \sum_{i=1}^{m} P(c_i \text{ false}) \\ &= 1 - \sum_{i} (1 - P(c_i \text{ true})) \\ &= 1 - (\alpha_1 + \cdots + \alpha_m). \end{aligned} \tag{4.3.1}$$

Try to develop simultaneous intervals for $\mu_i, i = 1, \cdots, m \leq p$. Since $P\left(\bar{x}_i \pm t_{n-1}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{S_{ii}}{n}} \text{ contains } \mu_i\right) = 1 - \alpha/m, i = 1, \cdots, m$, where $\alpha_i =$

α/m, then we have, from (4.3.1),

$$P\left(\bar{x}_i \pm t_{n-1}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{S_{ii}}{n}} \text{ contains } \mu_i \ \forall i\right)$$

$$\geq 1 - \underbrace{\left(\frac{\alpha}{m} + \cdots + \frac{\alpha}{m}\right)}_{\text{m terms}} \quad = 1 - \alpha$$

$$\therefore \ \bar{x}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}} \text{ are simultaneous interval estimates for}$$

$\mu_i, \ i = 1, \cdots, p$

ex) See example 5.6, pg. 233

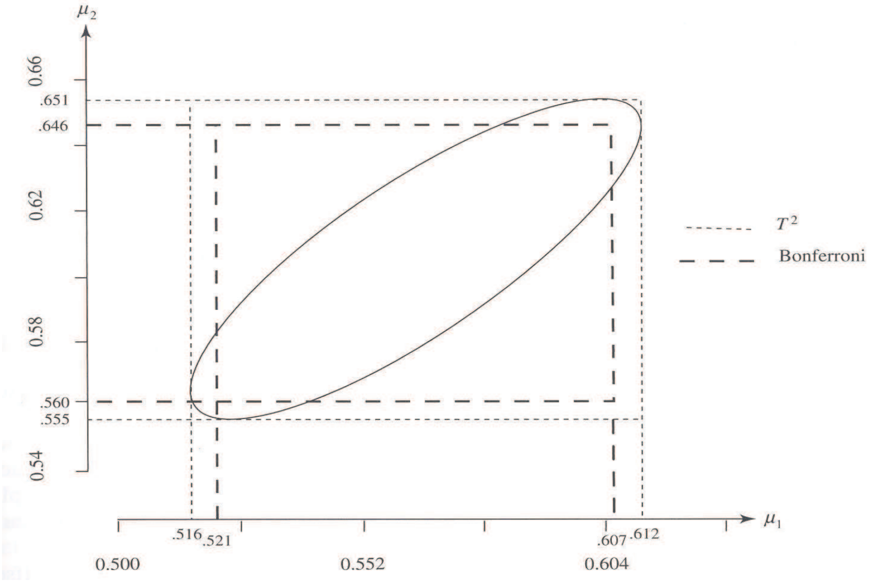


Figure 4.3.1: The 95% T^2 and 95% Bonferroni simultaneous confidence intervals for the component means-microwave radiation data.

Remark

- Note that the difference between T^2 & 95% Bonferroni simultaneous confidence intervals.

- T^2 intervals are wider than those of Bonferroni's since T^2 intervals are for any $\boldsymbol{a}$, whereas Bonferroni's are for specific $\boldsymbol{a}$.

4.4 Large sample inferences

$\boldsymbol{x}_1 \cdots \boldsymbol{x}_n$: random sample from a population with mean $\boldsymbol{\mu}$ & p.d. covariance matrix Σ

$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$

Reject H_0 if

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) > \chi_p^2(\alpha), \text{ where } P(\chi_p^2 > \chi_p^2(\alpha)) = \alpha.$$

See (4.28) of textbook for χ^2 approximation.

So a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}$ is the ellipsoid such that

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) \leq \chi_p^2(\alpha)$$

A $100(1-\alpha)\%$ simultaneous confidence intervals are

$$\boldsymbol{a}^\top \bar{\boldsymbol{x}} \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{\boldsymbol{a}^\top \boldsymbol{S} \boldsymbol{a}}{n}} \text{ for any } \boldsymbol{a} \text{ if } n-p \text{ is large}$$

ex) See example 5.7 pg.236

Chapter 5

Comparisons of several multivariate means

5.1 Paired comparisons and a repeated measures design

5.1.1 Paired Comparisons

Univariate

$$
\begin{aligned}
&X_{j1} : \text{response to treatment1 (or response before treatment)}\\
&X_{j2} : \text{response to treatment2 (or response after treatment)}
\end{aligned}
$$

Let $D_j = X_{j1} - X_{j2}$, $j = 1, \cdots, n$.
Assume $D_j \sim$ iid $N(\delta, \sigma_d^2)$

$$
\begin{aligned}
&t = \frac{\bar{D} - \delta}{s_d/\sqrt{n}} \sim t(n-1), \text{ where } \bar{D} = \frac{1}{n}\sum_{j=1}^{n} D_j,\ s_d^2 = \frac{1}{n-1}\sum_{j}(D_j - \bar{D})^2\\
&\text{Reject } H_0:\ \delta = 0 \text{ if } |t| \ > \ t_{n-1}\left(\frac{\alpha}{2}\right)\\
&\bar{d} \ \pm \ t_{n-1}\left(\frac{\alpha}{2}\right)\frac{s_d}{\sqrt{n}}\ ;\ 100(1-\alpha)\%\ C.I. \text{ for } \delta = E(D_j) = E(X_{j1} - X_{j2})
\end{aligned}
$$

Multivariate

$$
\begin{aligned}
X_{1j1} &= \text{variable 1 under treatment 1} \\
X_{1j2} &= \text{variable 2 under treatment 1} \\
&\vdots \\
X_{1jp} &= \text{variable } p \text{ under treatment 1} \\
\\
X_{2j1} &= \text{variable 1 under treatment 2} \\
X_{2j2} &= \text{variable 2 under treatment 2} \\
&\vdots \\
X_{2jp} &= \text{variable } p \text{ under treatment 2} \\
\Longrightarrow \\
D_{j1} &= X_{1j1} - X_{2j1} \\
D_{j2} &= X_{1j2} - X_{2j2} \\
&\vdots \\
D_{jp} &= X_{1jp} - X_{2jp}
\end{aligned}
$$

Let $\boldsymbol{D}_j^\top \;=\; [D_{j1}\ D_{j2}\ \cdots\ D_{jp}]$, $j = 1, \cdots, n$.
Assume $\boldsymbol{D}_j^\top \;\sim\;$ indep $N_p(\boldsymbol{\delta}, \Sigma_d)$, where

$$
E\boldsymbol{D}_j \;=\; \boldsymbol{\delta} \;=\; \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_p \end{bmatrix} \;\&\; cov(\boldsymbol{D}_j) \;=\; \Sigma_d
$$

Reject $H_0 : \boldsymbol{\delta} = \boldsymbol{0}$ if $T^2 = n\bar{\boldsymbol{d}}^\top \boldsymbol{S}_d^{-1} \bar{\boldsymbol{d}} \;>\; \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$, where
$\bar{\boldsymbol{D}} \;=\; \frac{1}{n} \sum_{j=1}^n \boldsymbol{D}_j$ & $\boldsymbol{S}_d \;=\; \frac{1}{n-1} \sum_{j=1}^n (\boldsymbol{D}_j - \bar{\boldsymbol{D}})(\boldsymbol{D}_j - \bar{\boldsymbol{D}})^\top$

$100(1-\alpha)\%$ confidence region for $\boldsymbol{\delta}$;
$(\bar{\boldsymbol{d}} - \boldsymbol{\delta})^\top \boldsymbol{S}_d^{-1} (\bar{\boldsymbol{d}} - \boldsymbol{\delta}) \;\leq\; \frac{(n-1)p}{n(n-p)} F_{p,n-p}(\alpha)$

$100(1-\alpha)\%$ simultaneous confidence intervals ;
$\delta_i \;:\; \bar{d}_i \;\pm\; \sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{S_{d_i}^2}{n}}$,

where $\bar{d}_i$: i^{th} element of $\bar{\boldsymbol{d}}$, $S^2_{d_i}$: i^{th} diagonal element of $\boldsymbol{S}_d$

Bonferroni $100(1-\alpha)\%$ simultaneous confidence interval
$\delta_i \ : \ \bar{d}_i \ \pm \ t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{S^2_{d_i}}{n}}$

ex) See example 6.1 pg. 276

5.1.2 A Repeated Measures design

Each subject receives each treatment once over successive periods of time

The j^{th} observation is $\boldsymbol{X}_j^\top = (X_{j1}, X_{j2}, \cdots, X_{jq})$, $j = 1, \cdots, n$, where X_{ji} : response to the i^{th} treatment on the j^{th} unit .

Note Repeated measures stem from the fact that all treatments are administered to each unit.

For comparative purposes,

$$\begin{pmatrix} \mu_1-\mu_2 \\ \mu_1-\mu_3 \\ \vdots \\ \mu_1-\mu_q \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix} = \boldsymbol{C}_1\boldsymbol{\mu}$$

or

$$\begin{pmatrix} \mu_2-\mu_1 \\ \mu_3-\mu_2 \\ \vdots \\ \mu_q-\mu_{q-1} \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix} = \boldsymbol{C}_2\boldsymbol{\mu}.$$

When the treatment means are equal, $\boldsymbol{C_1\mu = C_2\mu} = \mathbb{0}$

Reject H_0: $\boldsymbol{C\mu} = \mathbb{0}$ (equal treatment means) if

$$T^2 = n(\boldsymbol{C\bar{x}})^\top(\boldsymbol{CSC}^\top)^{-1}\boldsymbol{C\bar{x}} > \frac{(n-1)(q-1)}{(n-q+1)}F_{q-1,n-q+1}(\alpha),$$

where

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_j \ \& \ \boldsymbol{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^\top.$$

ex) See example 6.2. (pg. 281)

5.2 Comparing Mean Vectors from two Populations

Under the assumptions that
i) same covariance matrix &
ii) different covariance matrix

See Text. We will skip this because next section contains this one as a special case.

5.3 Comparing Several Multivariate Population Means (One-Way MANOVA)

5.3.1 Univariate

$$indep \begin{cases} Pop\ 1 : X_{11},\ , X_{12},\ ,\cdots,\ , X_{1n_1} & \sim iid\ N(\mu_1,\ \sigma^2) \\ Pop\ 2 : X_{21},\ , X_{22},\ ,\cdots,\ , X_{2n_2} & \sim iid\ N(\mu_2,\ \sigma^2) \\ \vdots & \\ Pop\ g : X_{g1},\ , X_{g2},\ ,\cdots,\ , X_{gn_g} & \sim iid\ N(\mu_g,\ \sigma^2) \end{cases}$$

Model

$$X_{lj} = \mu + \tau_l + e_{lj}\ ,\ l = 1, \cdots, g\ ,\ j = 1, \cdots, n_l,$$

where $e_{lj} \ \sim \ iid\ N(0,\ \sigma^2)$,
μ : overall mean
τ_l : trt effect
$\sum_{l=1}^{g} n_l\tau_l \ = 0.$

$$x_{lj} \;=\; \bar{x} \;+\; (\bar{x}_l \;-\; \bar{x}) + (\bar{x}_{lj} \;-\; \bar{x}_l)$$

$$\Rightarrow \sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj}-\bar{x})^2 = \sum_{l=1}^{g} n_l(\bar{x}_l-\bar{x})^2 \;+\; \sum_{l}\sum_{j}(x_{lj}-\bar{x}_l)^2$$

$$SS_{cor} \qquad = \qquad SS_{tr} \qquad + \qquad SS_{res}$$

ANOVA Table

Source of Variation (SV)	Sum of Squares (SS)	df
Trt	SS_{tr}	$g-1$
$Error$	SS_{res}	$\sum_{l=1}^{g} n_l - g$
$Total$	SS_{cor}	$\sum_{l=1}^{g} n_l - 1$

Reject $H_0 : \tau_1 \;=\; \tau_2 \;=\; \cdots \;=\; 0$ if

$$F = \frac{SS_{tr}/(g-1)}{SS_{res}/(\sum_{\rho=1}^{q} n_l - g)} \quad > \quad F_{g-1,\sum n_l - q}.$$

$$\Leftrightarrow reject\ H_0\ for\ large\ values\ of\ \frac{SS_{tr}}{SS_{res}}$$

$$\Leftrightarrow reject\ H_0\ for\ large\ values\ of\ 1+\frac{SS_{tr}}{SS_{res}}$$

$$\Leftrightarrow reject\ H_0\ for\ small\ values\ of\ \frac{1}{1+\dfrac{SS_{tr}}{SS_{res}}} \;=\; \frac{SS_{res}}{SS_{res}+SS_{tr}}$$

5.3.2 Multivariate

$$indep \begin{cases} Pop\ 1 : \boldsymbol{X}_{11},\ ,\boldsymbol{X}_{12},\ ,\cdots,\ ,\boldsymbol{X}_{1n_1} & \sim iid\ N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \\ Pop\ 2 : \boldsymbol{X}_{21},\ ,\boldsymbol{X}_{22},\ ,\cdots,\ ,\boldsymbol{X}_{2n_2} & \sim iid\ N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \\ \vdots & \\ Pop\ g : \boldsymbol{X}_{g1},\ ,\boldsymbol{X}_{g2},\ ,\cdots,\ ,\boldsymbol{X}_{gn_g} & \sim iid\ N_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}) \end{cases}$$

Model

$$\boldsymbol{X}_{lj} = \boldsymbol{\mu} + \boldsymbol{\tau}_l + \boldsymbol{e}_{lj}\ ,\ l = 1,\cdots,g\ ,\ j = 1,\cdots,n_l,$$

where $\boldsymbol{e}_{lj} \sim iid\ N(\mathbf{0}, \boldsymbol{\Sigma})$,

$\boldsymbol{\mu}$: overall mean
$\boldsymbol{\tau}_l$: treatment effect
$\sum_{l=1}^{g} n_l \boldsymbol{\tau}_l = \mathbf{0}$

$$\boldsymbol{x}_{lj} = \bar{\boldsymbol{x}} + (\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}}) + (\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)$$

$\Longrightarrow$

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}})^\top = \sum_l n_l(\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}})^\top + \sum_l\sum_j(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)^\top$$

$$Total\ S.S. = trt(Between)\ S.S. + Residual(Within)\ S.S.$$

Remark

$\boldsymbol{W} = \sum_l\sum_j(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)^\top = \sum_l(n_l - 1)\boldsymbol{S}_l$,

$\boldsymbol{S}_l = \frac{1}{n_l-1}\sum_j(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)^\top$

MANOVA Table

SV	Matrix S.S. & Cross products (SSP)	df
Trt	$\boldsymbol{B} = \sum_l n_l(\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}})^\top$	$g-1$
$Error$	$\boldsymbol{W} = \sum_l\sum_j(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)^\top$	$\sum_l n_l - g$
$Total$	$\boldsymbol{B} + \boldsymbol{W} = \sum_l\sum_j(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}})^\top$	$\sum_l n_l - 1$

Reject $H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \mathbf{0}$ if Λ^* is too small, where

$$\Lambda^* = \frac{|\boldsymbol{W}|}{|\boldsymbol{B} + \boldsymbol{W}|}$$

is Wilks' lambda which is a ratio of generalized variances.

Remark

- See Table 6.3 (distribution of Wilks' lambda)

- For other cases and large sample sizes,

$$-\left(n-1-\frac{p+q}{2}\right)\ln\Lambda^* \;\sim\; \chi^2_{p(g-1)}$$

under H_0 is true and $\sum n_l = n$ is large. So reject H_0 if

$$-\left(n-1-\frac{p+q}{2}\right)\ln\Lambda^* \;>\; \chi^2_{p(g-1)}(\alpha).$$

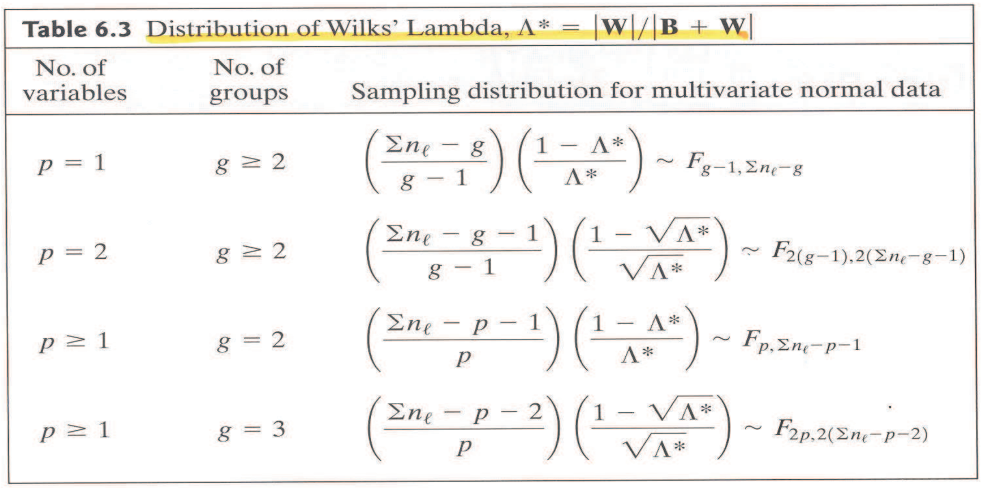


Table 6.3 Distribution of Wilks' Lambda, $\Lambda^* = |\mathbf{W}|/|\mathbf{B}+\mathbf{W}|$

No. of variables	No. of groups	Sampling distribution for multivariate normal data
$p = 1$	$g \geq 2$	$\left(\frac{\Sigma n_\ell - g}{g-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{g-1,\Sigma n_\ell - g}$
$p = 2$	$g \geq 2$	$\left(\frac{\Sigma n_\ell - g - 1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1),2(\Sigma n_\ell - g - 1)}$
$p \geq 1$	$g = 2$	$\left(\frac{\Sigma n_\ell - p - 1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p,\Sigma n_\ell - p - 1}$
$p \geq 1$	$g = 3$	$\left(\frac{\Sigma n_\ell - p - 2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p,2(\Sigma n_\ell - p - 2)}$

ex) example 6.9 & 6.10 (p.g 304 - 308)

5.4 Simultaneous Confidence intervals for treatment effects

Model $\boldsymbol{X}_{lj} = \boldsymbol{\mu} + \boldsymbol{\tau}_l + \boldsymbol{e}_{lj}$, $l = 1, \cdots, g$, $j = 1, \cdots, n_l$,
μ : overall mean, τ_l : trt effect, $\boldsymbol{e}_{lj} \sim iid\ N(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{\Sigma} = (\sigma_{ij})$

If $H_0 : \boldsymbol{\tau}_1 = \cdots = \boldsymbol{\tau}_g = \mathbf{0}$ is rejected, those effects that led to the rejection of H_0 are of interest.

How to find them!
$\Longrightarrow$
Use Bonferroni approach to construct simultaneous CIs for the components of the differences $\boldsymbol{\tau}_k - \boldsymbol{\tau}_l$.

1. $\boldsymbol{\tau}_k$ is estimated by $\hat{\boldsymbol{\tau}}_k = \bar{\boldsymbol{x}}_k$ - $\bar{\boldsymbol{x}}$.
τ_{ki} be the i^{th} component of $\boldsymbol{\tau}_k$.
$\hat{\tau}_{ki} - \hat{\tau}_{li} = (\bar{x}_{ki} - \bar{x}_i) - (\bar{x}_{li} - \bar{x}_i) = \bar{x}_{ki} - \bar{x}_{li}$
$Var(\hat{\tau}_{ki} - \hat{\tau}_{li}) = Var\bar{x}_{ki} + Var\bar{x}_{li} = \frac{1}{n_k}\sigma_{ii} + \frac{1}{n_l}\sigma_{ii} = \left(\frac{1}{n_k} + \frac{1}{n_l}\right)\sigma_{ii}$

2. $\boldsymbol{\Sigma}$ is estimated by

$$\boldsymbol{\Sigma} = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2 + \cdots + (n_g - 1)\boldsymbol{S}_g}{\sum_{l=1}^{g}(n_l - 1)} = \frac{\boldsymbol{W}}{n - g},$$

where

$$\boldsymbol{S}_l = \frac{1}{n_l - 1}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)^\top$$

is the sample covariance matrix for trt l.
Therefore, $Var(\hat{\tau}_{ki} - \hat{\tau}_{li})$ is estimated by $\left(\frac{1}{n_k} + \frac{1}{n_l}\right)\frac{w_{ii}}{n-g}$, where $n = \sum_{l=1}^{g} n_l$.

3. How may CIs involved?

p components (variables)

g groups $\Rightarrow$ $\frac{g(g-1)}{2}$ pairwise differences

$\Rightarrow m = p\frac{g(g-1)}{2}$ CIs involved

Combining 1, 2 & 3, simultaneous CIs of $\hat{\tau}_{ki} - \hat{\tau}_{li}$ belongs to

$$\bar{x}_{ki} - \bar{x}_{li} \pm t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\frac{w_{ii}}{n-g}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)}$$

with confidence at least $1 - \alpha$.

ex 6.11) pg.309 Wisconsin nursing home data

Type of nursing home : private ($l = 1$), non-profit ($l = 2$), government $(1 = 3)$

$$\begin{aligned} \text{Variables ;} \quad X_1 &= \text{Cost of nursing labor} \\ X_2 &= \text{Cost of dietary labor} \\ X_3 &= \text{Cost of plant operation} \\ X_4 &= \text{Cost of housekeeping \& laundry labor} \\ X^\top &= (X_1,\ X_2,\ X_3,\ X_4) \\ & p = 4,\ g = 3 \\ & n_1 = 271,\ n_2 = 138,\ n_3 = 107,\ n_4 = 516 \\ & \vdots \end{aligned}$$

$$\Lambda^* = \frac{|\boldsymbol{W}|}{|\boldsymbol{B} + \boldsymbol{W}|} = 0.7714$$

Test statistic (Use Table 3, $g = 3$)

$$\left(\frac{\sum n_l - p - 2}{p}\right)\left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) = 17.67 \ \sim \ F_{2(4),2(510)}.$$

For $\alpha = 0.01$, $F_{8,1020}(0.01) \ \doteq \ \chi^2_8(0.01)/8 \ \doteq \ 2.5$.
Since $17.67 > 2.51$, reject $H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \boldsymbol{\tau}_3$ (no ownership effects).

Where are these differences?
Construct simultaneous CIs

Focus on X_3

$$\exists \text{ difference} \begin{cases} \tau_{13} - \tau_{33} & (-0.061, -0, 025) \\ \tau_{13} - \tau_{23} & (-0.058, -0, 026) \end{cases}$$

$$\tau_{23} - \tau_{33} \qquad (-0.021, 0, 019) \qquad \text{No diff}$$

Total $\quad 12 = \ 4\cdot 3\cdot 2/2$ intervals.

5.5 Testing for equality of covariance Matrices

Box's M-test (See Text)

<u>**Remark**</u>

- M-test is sensitive to some forms of non-normality
- With reasonably large samples, the MANOVA tests of trt effects are rather robust to nonnormality
- Thus the M-test may reject $H_0 : \ \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_g = \mathbf{\Sigma}$ in some non-normal cases where it is not damaging to the MANOVA tests

$\therefore$ We may decide to continue with the usual MANOVA tests even though the M-test leads to rejection of $H_0 : \ \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_g = \mathbf{\Sigma}$.

5.6 Two-way Multivariate ANOVA

5.6.1 Univariate

$$\begin{aligned} X_{lkr} &= \mu \quad + \underset{factor\ 1}{\tau_l} + \underset{factor\ 2}{\beta_k} + \underset{interaction}{\gamma_{lk}} + e_{lkr}, \\ & \quad l = 1, \cdots, g, \quad k = 1, \cdots, b, \quad r = 1, \cdots, n \\ & \quad \sum_l \tau_l = \sum_k \beta_k = \sum_l \gamma_{lk} = \sum_k \gamma_{lk} = 0, \quad e_{lkr} \sim iid\ N(0, \sigma^2) \end{aligned}$$

<u>**sample**</u> $x_{lkr} = \bar{x} + (\bar{x}_{l\cdot} - \bar{x}) + (\bar{x}_{\cdot k} - \bar{x}) + (\bar{x}_{lk} - \bar{x}_{l\cdot} - \bar{x}_{\cdot k} + \bar{x}) + (x_{lkr} - \bar{x}_{lk}),$
where

$$\begin{aligned} \bar{x}_{lk} &= \frac{1}{n} \sum_r x_{lkr} \\ \bar{x}_{l\cdot} &= \frac{1}{nb} \sum_k \sum_r x_{lkr} \\ \bar{x}_{\cdot k} &= \frac{1}{ng} \sum_l \sum_r x_{lkr} \\ \bar{x} &= \frac{1}{gbn} \sum_l \sum_k \sum_r x_{lkr}. \end{aligned}$$

$$\begin{aligned}\sum_l\sum_k\sum_r(x_{lkr}-\bar{x})^2 &= \sum_l bn(\bar{x}_{l\cdot}-\bar{x})^2+\sum_k gn(\bar{x}_{\cdot k}-\bar{x})^2\\ &+ \sum_l\sum_k n(\bar{x}_{lk}-\bar{x}_{l\cdot}-\bar{x}_{\cdot k}+\bar{x})+\sum_l\sum_k\sum_r(x_{lkr}-\bar{x}_{lk})^2\end{aligned}$$

$$\Leftrightarrow SS_{cor} \quad = SS_{fac1} \;+ SS_{fac2} \;+ SS_{int} \;+ SS_{res}$$

ANOVA Table

S.V	S.S	d.f	
$Factor\ 1$	SS_{fac1}	$g-1$	$F_1=MS_{fac1}/MS_{res}$
$Factor\ 2$	SS_{fac2}	$b-1$	$F_2=MS_{fac2}/MS_{res}$
$Interaction$	SS_{int}	$(g-1)(b-1)$	$F_3=MS_{int}/MS_{res}$
$Error$	SS_{res}	$gb(n-1)$	
$Total$	SS_{cor}	$gbn-1$	

$$MS=\frac{SS}{d.f.}$$

5.6.2 Multivariate

$$\boldsymbol{X}_{lkr} = \boldsymbol{\mu} + \underset{factor\ 1}{\boldsymbol{\tau}_l} + \underset{factor\ 2}{\boldsymbol{\beta}_k} + \underset{interaction}{\boldsymbol{\gamma}_{lk}} + \boldsymbol{e}_{lkr},$$

$$\begin{aligned}&l=1,\cdots,g,\quad k=1,\cdots,b,\quad r=1,\cdots,n\\ &\sum_l\boldsymbol{\tau}_l = \sum_k\boldsymbol{\beta}_k = \sum_l\boldsymbol{\gamma}_{lk} = \sum_k\boldsymbol{\gamma}_{lk} = \mathbb{0}\\ &\boldsymbol{e}_{lkr} \sim iid\ N_p(\mathbb{0},\boldsymbol{\Sigma})\end{aligned}$$

sample $\boldsymbol{x}_{lkr}=\bar{\boldsymbol{x}}+(\bar{\boldsymbol{x}}_{l\cdot}-\bar{\boldsymbol{x}})+(\bar{\boldsymbol{x}}_{\cdot k}-\bar{\boldsymbol{x}})+(\bar{\boldsymbol{x}}_{lk}-\bar{\boldsymbol{x}}_{l\cdot}-\bar{\boldsymbol{x}}_{\cdot k}+\bar{\boldsymbol{x}})+(\boldsymbol{x}_{lkr}-\bar{\boldsymbol{x}}_{lk})$

MANOVA Table
Tests can be performed as the univariate case.

$i)$ H_0 : $\boldsymbol{\gamma}_{11} = \boldsymbol{\gamma}_{12} = \cdots = \boldsymbol{\gamma}_{gb} = \mathbb{0}$ (no interaction effects)
$vs.$ H_1 : At least one $\boldsymbol{\gamma}_{lk} \neq \mathbb{0}$

$$\Lambda^*=\frac{|SSP_{res}|}{|SSP_{int}+SSP_{res}|}\ ;\ \text{Wilks' lambda}$$

SV	Matrix S.S. & Cross products (SSP)	df
$Factor\ 1$	$SSP_{fac1} = \sum_l bn(\bar{\boldsymbol{x}}_{l\cdot} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{l\cdot} - \bar{\boldsymbol{x}})^\top$	$g-1$
$Factor\ 2$	$SSP_{fac2} = \sum_k gn(\bar{\boldsymbol{x}}_{\cdot k} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{\cdot k} - \bar{\boldsymbol{x}})^\top$	$b-1$
$Interaction$	$SSP_{int} = \sum_l \sum_k n(\bar{\boldsymbol{x}}_{lk} - \bar{\boldsymbol{x}}_{l\cdot} - \bar{\boldsymbol{x}}_{\cdot k} + \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{lk} - \bar{\boldsymbol{x}}_{l\cdot} - \bar{\boldsymbol{x}}_{\cdot k} + \bar{\boldsymbol{x}})^\top$	$(g-1)(b-1)$
$Error$	$SSP_{res} = \sum_l \sum_k \sum_r (\boldsymbol{x}_{lkr} - \bar{\boldsymbol{x}}_{lk})(x_{lkr} - \bar{\boldsymbol{x}}_{lk})^\top$	$gb(n-1)$
$Total$	$SSP_{cor} = \sum_l \sum_k \sum_r (\boldsymbol{x}_{lkr} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{lkr} - \bar{\boldsymbol{x}})^\top$	$gbn-1$

For large samples (Bartlett's approx),

$$-\left[gb(n-1) - \frac{p+1-(g-1)(b-1)}{2}\right] \ln \Lambda^* \;\sim\; \chi^2_{(g-1)(b-1)p}$$

Reject H_0 if $$-\left[gb(n-1) - \frac{p+1-(g-1)(b-1)}{2}\right] \ln \Lambda^* \;>\; \chi^2_{(g-1)(b-1)p}(\alpha).$$

$$\begin{array}{lll} ii) & H_0 & : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\tau}_g = \mathbb{0} \text{ (no factor 1 effects)} \\ vs. & H_1 & : \text{At least one } \boldsymbol{\tau}_l \neq \mathbb{0} \end{array}$$

$$\Lambda^* = \frac{|SSP_{res}|}{\left|SSP_{fac1} + SSP_{res}\right|} \text{ ; Wilks' lambda}$$

Reject H_0 if,

$$-\left[gb(n-1) - \frac{p+1-(g-1)}{2}\right] \ln \Lambda^* > \chi^2_{(g-1)p}(\alpha) \text{ ; Bartlett's approx}$$

$$\begin{array}{lll} ii) & H_0 & : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_b = \mathbb{0} \text{ (no factor 2 effects)} \\ vs. & H_1 & : \text{At least one } \boldsymbol{\beta}_k \neq \mathbb{0} \end{array}$$

$$\Lambda^* = \frac{|SSP_{res}|}{\left|SSP_{fac2} + SSP_{res}\right|} \text{ ; Wilks' lambda}$$

Rejct H_0 if

$$-\left[gb(n-1) - \frac{p+1-(b-1)}{2}\right] \ln \Lambda^* > \chi^2_{(b-1)p}(\alpha) \text{ ; Bartlett's approx}$$

Remark If a null hypothesis is rejected, use simultaneous CIs to study the differences.

Bonferroni approach

i) $100(1-\alpha)\%$ simultaneous CIs for $\tau_{li} - \tau_{mi}$

$$(\bar{x}_{l\cdot i} - \bar{x}_{m\cdot i}) \pm t_\nu \left(\frac{\alpha}{pg(g-1)}\right) \sqrt{\frac{E_{ii}}{\nu} \frac{2}{bn}},$$

where $\nu = gb(n-1)$, E_{ii} is the i^{th} diagonal element of $E = SSP_{res}$ & $(\bar{x}_{l\cdot i} - \bar{x}_{m\cdot i})$ is the $i^t h$ component of $\bar{\boldsymbol{x}}_{l\cdot} - \bar{\boldsymbol{x}}_{m\cdot}$.

ii) $100(1-\alpha)\%$ simultaneous CIs for $\beta_{ki} - \beta_{qi}$

$$(\bar{x}_{\cdot ki} - \bar{x}_{\cdot qi}) \pm t_\nu \left(\frac{\alpha}{pb(b-1)}\right) \sqrt{\frac{E_{ii}}{\nu} \frac{2}{gn}},$$

where $(\bar{x}_{\cdot ki} - \bar{x}_{\cdot qi})$ is the i^{th} component of $\bar{\boldsymbol{x}}_{\cdot k} - \bar{\boldsymbol{x}}_{\cdot q}$.

ex) See Example 6.13 (pg. 318 - 323)

$i)$ Do not reject, $H_0 : \boldsymbol{\gamma}_{11} = \boldsymbol{\gamma}_{12} = \boldsymbol{\gamma}_{21} = \boldsymbol{\gamma}_{22} = \mathbb{0}$ (no interaction effects).
$ii)$ Rejected $H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \mathbb{0}$ (no factor 1 effects)
& $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \mathbb{0}$ (no factor 2 effects).

5.7 Profile analysis

P treatments (tests, questions, etc) are administered to two or more groups of subjects.

Assume that

i) the responses for the different groups are independent of one another and

ii) all responses must be expressed in similar units.

Are the population mean vectors the same?

Consider the case of two groups.

$$\begin{aligned} \boldsymbol{\mu}_1 &= [\mu_{11}\ \mu_{12}\ \cdots\ \mu_{1p}]^\top \quad \& \\ \boldsymbol{\mu}_2 &= [\mu_{21}\ \mu_{22}\ \cdots\ \mu_{2p}]^\top \end{aligned}$$

Profile of group 1 (See Fig 5.7.1).

$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ implies that the treatments have the same effect on the tow populations.

In terms of the population profiles, we formulate the question of equality in a stepwise fashion.

① Are the profiles parallel?
$H_{01} : \mu_{1i} - \mu_{1i-1} = \mu_{2i} - \mu_{2i-1}, i = 2, 3, \cdots, p$

② Under H_{01}, are the profiles coincident?
$H_{02} : \mu_{1i} = \mu_{2i}, i = 1, 2, \cdots, p$

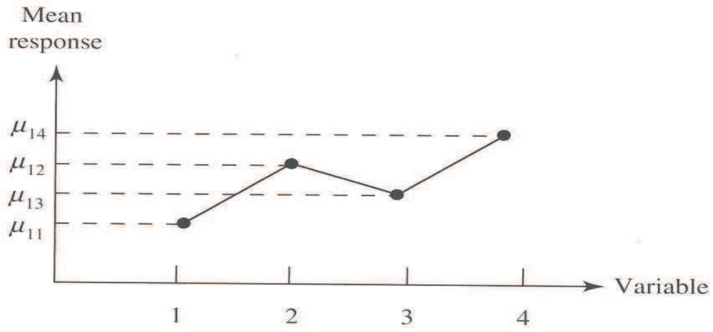


Figure 5.7.1: The population profile $p = 4$.

③ Under H_{02}, are the profiles level?
$H_{03} : \mu_{11} = \mu_{12} = \cdots = \mu_{1p} = \mu_{21} = \mu_{22} = \cdots = \mu_{2p}$

For ①, H_{01} can be written $H_{01}; C\boldsymbol{\mu}_1 = C\boldsymbol{\mu}_2$, where

$$C_{(p-1)\times p} = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}$$

$$S_{pooled} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

from $\boldsymbol{x}_{1j}, j = 1, \dots, n_1$ independent of $\boldsymbol{x}_{2j}, j = 1, \dots, n_2$.

Note that $\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i = 1, 2$ then $C\boldsymbol{x}_i \sim N_{p-1}(C\boldsymbol{\mu}, C\boldsymbol{\Sigma}C^\top)$.

Reject $H_{01} : C\boldsymbol{\mu}_1 = C\boldsymbol{\mu}_2$ if

$$T^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top C^\top \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) C S_{pooled} C^\top \right]^{-1} C(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) > c^2, \quad (5.7.1)$$

where $c^2 = \frac{(n_1+n_2-2)(p-1)}{n_1+n_2-p} F_{p-1,n_1+n_2-p}(\alpha)$.

For ②, $H_{02} : \mathbf{1}^\top \boldsymbol{\mu}_1 = \mathbf{1}^\top \boldsymbol{\mu}_2$ since $\mu_{1i} > \mu_{2i}, \forall i$ or vice versa under H_{01}.

Reject $H_{02} : \mathbf{1}^\top \boldsymbol{\mu}_1 = \mathbf{1}^\top \boldsymbol{\mu}_2$, if

$$\begin{aligned} T^2 &= \mathbf{1}^\top(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathbf{1}^\top S_{pooled}\mathbf{1}\right]^{-1}\mathbf{1}^\top(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \\ &= \left(\frac{\mathbf{1}^\top(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathbf{1}^\top S_{pooled}\mathbf{1}}}\right)^2 > t^2_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) = F_{1,n_1+n_2-2}(\alpha). \end{aligned}$$

For ③, $H_{03} : C\boldsymbol{\mu} = \mathbf{0}$, where C is given in ① since they have the same mean vector.

Need to pool sample mean vectors,

$$\bar{\boldsymbol{x}} = \frac{1}{n_1+n_2}\left(\sum_j \boldsymbol{x}_{1j} + \sum_j \boldsymbol{x}_{2j}\right) = \frac{n_1}{n_1+n_2}\bar{\boldsymbol{x}}_1 + \frac{n_2}{n_1+n_2}\bar{\boldsymbol{x}}_2.$$

Reject $H_{03} : C\boldsymbol{\mu} = \mathbf{0}$, where $\boldsymbol{\mu} = (\mu_1 \, \ldots \, \mu_p)^\top$ if

$$(n_1+n_2)\bar{\boldsymbol{x}}^\top C^\top [CSC^\top]^{-1} C\bar{\boldsymbol{x}} > c^2,$$

where

$$\begin{aligned} S &= \frac{1}{n_1+n_2-1}\left[\sum_j(\boldsymbol{x}_{1j}-\bar{\boldsymbol{x}})(\boldsymbol{x}_{1j}-\bar{\boldsymbol{x}})^\top + \sum_j(\boldsymbol{x}_{2j}-\bar{\boldsymbol{x}})(\boldsymbol{x}_{2j}-\bar{\boldsymbol{x}})^\top\right] \text{ and} \\ c^2 &= \frac{(n_1+n_2-1)(p-1)}{(n_1+n_2-p+1)}F_{p-1,n_1+n_2-p+1}(\alpha). \end{aligned}$$

ex) See Example 6.14 pg. 325-326

Fail to reject $H_{01} : C\boldsymbol{\mu}_1 = C\boldsymbol{\mu}_2$ (parallel profiles)
Fail to reject $H_{02} : \mathbf{1}^\top \boldsymbol{\mu}_1 = \mathbf{1}^\top \boldsymbol{\mu}_2$ (profiles coincident)
No need to test for level profiles since the different scales.

5.8 Repeated Measure Designs and Growth Curves

- Repeated Measures (diff. treatments); observations taken from the same subject at different times or locations
  ex) See Example 6.2 pg. 281.

- Growth curves (single treatment); observations are taken from the same subject over a period of time.

ex) of growth curve; Calcium measurements of the dominant ulna bone in older women.

- Control group Table 6.5
  15 subjects and measured initial, after 1 year, 2 years and 3 years

- Treatment group Table 6.6 (received special help)
  16 subjects and measured initial, after 1 year, 2 years and 3years

Note that a profile, constructed from $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4)$, summarizes the growth which here is a loss of calcium over time.

How to model it!

For example, quadratic growth model

$$E\boldsymbol{x} = E\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 t_1 + \beta_2 t_1^2 \\ \beta_0 + \beta_1 t_2 + \beta_2 t_2^2 \\ \vdots \\ \beta_0 + \beta_1 t_p + \beta_2 t_p^2 \end{bmatrix}$$

In our case, $t_1 = 0, t_2 = 1, t_3 = 2, t_4 = 3$ $(p = 4)$.

In general, groups need to be compared.

Let $\boldsymbol{x}_{l1} \ldots \boldsymbol{x}_{ln_l}$ be the n_l vectors of measurements on the n_l subjects in group l, $l = 1, \ldots, g$.

In our case, $g = 2, n_1 = 15, n_2 = 16$

Assume that $\boldsymbol{x}_{lj} \sim$ indep the mean $B\boldsymbol{\beta}_l$ and the same covariance $\boldsymbol{\Sigma}$, where

$$E\boldsymbol{x}_{lj} = \begin{bmatrix} \beta_{l0} + \beta_{l1}t_1 + \beta_{l2}t_1^2 \\ \beta_{l0} + \beta_{l1}t_2 + \beta_{l2}t_2^2 \\ \vdots \\ \beta_{l0} + \beta_{l1}t_p + \beta_{l2}t_p^2 \end{bmatrix} = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ & \vdots & \\ 1 & t_p & t_p^2 \end{bmatrix} \begin{bmatrix} \beta_{l0} \\ \beta_{l1} \\ \beta_{l2} \end{bmatrix}. \tag{5.8.1}$$

Under normality assumption,

$$\hat{\boldsymbol{\beta}}_l = (B'S_{pooled}^{-1}B)^{-1}B'S_{pooled}^{-1}\bar{\boldsymbol{x}}_l \text{ for } l = 1, \cdots, g,$$

where

$$S_{pooled} = \tfrac{1}{N-g}\{(n_1-1)S_1 + \cdots + (n_g-1)S_g\} = \tfrac{1}{N-g}W,\ N = \textstyle\sum_{l=1}^{g} n_l.$$

$$\widehat{Cov(\widehat{\boldsymbol{\beta}}_l)} = \frac{k}{n_l}(B'S_{pooled}^{-1}B)^{-1}$$

for $l = 1, \ldots, g$, where

$$k = (N-g)(N-g-1)/(N-g-p+q)(N-g-p+q+1)$$

and $q =$ degree of the polynomial.

To test H_0 : qth-order polynomial is adequate,
Reject H_0 if

$$-\left(N - \frac{1}{2}(p-q+g)\right)\ln\Lambda^* > \chi^2_{(p-q-1)g}(\alpha),$$

where

$$\Lambda^* = \frac{|W|}{W_q}$$

and

$$W_q = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj} - B\hat{\boldsymbol{\beta}}_l)(\boldsymbol{x}_{lj} - B\hat{\boldsymbol{\beta}}_l)^\top.$$

ex)See Example 6.15 pg.331

Estimated growth curves are

- Control group: $73.07 + 3.64t - 2.03t^2$
- Trt group: $70.14 + 4.09t - 1.85t^2$

5.9 Perspectives and a Strategy for Analyzing Multivariate models

A single multivariate test is recommended over, say, p univariate tests.
See example 6.16 and 6.17.

Remark: Some useful test statistics for MANOVA
We used Wilk's lambda $\Lambda^* = \dfrac{|W|}{|B+W|}$

1. Lawley-Hotelling trace $= tr(BW^{-1})$
2. Pillai trace $= tr(B(B+W)^{-1})$
3. Roy's largest root $=$ maximum eigenvalue of $W(B+W)^{-1}$

Chapter 6

Multivariate Linear Regression Models

6.1 The Classical Linear Regression Model

Y : response (dependent) variable
$Z_1, \ldots, Z_r$: predictor (independent) variables

Model $Y_j = \beta_0 + \beta_1 z_{j1} + \cdots + \beta_r z_{jr} + \varepsilon_j, j = 1, \ldots, n$
Assume that

① $E\varepsilon_j = 0$

② $V\varepsilon_j = \sigma^2$ (constant)

③ $Cov(\varepsilon_j, \varepsilon_k) = 0, j \neq k$

Matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1r} \\ 1 & z_{21} & z_{22} & \cdots & z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

or

$$\boldsymbol{y}_{n\times 1} = \boldsymbol{Z}_{n\times(r+1)}\boldsymbol{\beta}_{(r+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

Assume that

① $E\boldsymbol{\varepsilon} = \mathbf{0}$

② $Cov(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2\boldsymbol{I}$

For convenience, let $z_{j0} = 1, j = 1, \ldots, n$,

$$\boldsymbol{Z} = \begin{bmatrix} z_{10} & z_{11} & z_{12} & \cdots & z_{1r} \\ z_{20} & z_{21} & z_{22} & \cdots & z_{2r} \\ \vdots & \vdots & \vdots & & \vdots \\ z_{n0} & z_{n1} & z_{n2} & \cdots & z_{nr} \end{bmatrix}$$

is a design matrix.

ex) See Examples 7.1 and 7.2 pg.362-363.

6.2 Least Squares Estimation

From the model,

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}.$$

LSE $\hat{\boldsymbol{\beta}}$ minimizes error sum of squares which is

$$\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) = \sum_{j=1}^{n}(y_j - \beta_0 - \beta_1 z_{j1} - \cdots - \beta_r z_{jr})^2.$$

<u>Result</u>
If $\boldsymbol{Z}$ is full column rank $r + 1 \le n$, then

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}.$$

$\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}}$; fitted values of $\boldsymbol{y}$ and

$\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}} = \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{y}$, where $\boldsymbol{H}$; hat matrix

$\Rightarrow$

① $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$; residuals

② $\boldsymbol{Z}^\top\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$

③ $\hat{\boldsymbol{y}}^\top \hat{\boldsymbol{\varepsilon}} = 0$

④

$$\begin{aligned}
\text{residual s.s.} &= \sum_j (y_j - \hat{\beta}_0 - \hat{\beta}_1 z_{j1} - \cdots - \hat{\beta}_r z_{jr})^2 \\
&= \hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)\boldsymbol{y} \\
&= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{Z}\hat{\boldsymbol{\beta}}
\end{aligned}$$

Note Properties of $\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$
(a) Symmetric (b) Idempotent (c) $\boldsymbol{Z}^\top(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top) = \boldsymbol{0}$

pf) Let $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{y}$ as asserted.

② $\boldsymbol{Z}^\top \hat{\boldsymbol{\varepsilon}} = \boldsymbol{Z}^\top(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{y} = \boldsymbol{0}$ by (c)

③ $\hat{\boldsymbol{y}}^\top \hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}^\top \hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$ by ②

④ $\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{H})^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ by (a), (b)

To verify $\hat{\boldsymbol{\beta}}$, we write

$$\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

so

$$\begin{aligned}
S(\boldsymbol{\beta}) &= \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) \\
&= [\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^\top[\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\
&= (\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{Z}^\top \boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&\quad + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{Z}^\top(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}) + (\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})^\top \boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).
\end{aligned}$$

By ②, $\boldsymbol{Z}^\top(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ and $(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})^\top \boldsymbol{Z} = \boldsymbol{0}$.

$\therefore\ S(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{Z}^\top \boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

Because $\boldsymbol{Z}$ has full column rank, $\boldsymbol{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \neq \boldsymbol{0}$ if $\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$.

Hence, if $\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$, the 2^{nd} term is positive.

Consequently $\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} \geq (\boldsymbol{y}-\boldsymbol{Z}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\boldsymbol{Z}\hat{\boldsymbol{\beta}})$ with equality hold when $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

$\therefore\ \hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}$ is the LSE of $\boldsymbol{\beta}$. □

ex) See Example 7.3 for calculating LSE.

Tip

$\boldsymbol{X}\boldsymbol{a}$; linear combinations of the columns of $\boldsymbol{X}$
$\boldsymbol{a}^\top\boldsymbol{X}$; linear combinations of the rows of $\boldsymbol{X}$

6.2.1 Sum-of-Squares Decomposition

$$\begin{aligned}\boldsymbol{y}^\top\boldsymbol{y} &= (\hat{\boldsymbol{y}}+\boldsymbol{y}-\hat{\boldsymbol{y}})^\top(\hat{\boldsymbol{y}}+\boldsymbol{y}-\hat{\boldsymbol{y}}) = (\hat{\boldsymbol{y}}+\hat{\varepsilon})^\top(\hat{\boldsymbol{y}}+\hat{\varepsilon}) \\ &= \hat{\boldsymbol{y}}^\top\hat{\boldsymbol{y}}+\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} \qquad (6.2.1)\end{aligned}$$

since $\hat{\boldsymbol{y}}^\top\hat{\boldsymbol{\varepsilon}} = 0$ by ②.

Because the first column of $\boldsymbol{Z}$ is $\boldsymbol{1}$, $\boldsymbol{Z}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$ includes $\boldsymbol{1}^\top\hat{\boldsymbol{\varepsilon}} = 0 = \sum_{j=1}^n \hat{\varepsilon}_j = \sum_j y_j - \sum_j \hat{y}_j \Leftrightarrow \bar{y} = \bar{\hat{y}}$.

$\Rightarrow n(\bar{y})^2 = n(\bar{\hat{y}})^2$
Subtracting $n(\bar{y})^2 = n(\bar{\hat{y}})^2$ from (6.2.1),

$$\begin{aligned}\boldsymbol{y}^\top\boldsymbol{y} - n(\bar{y})^2 &= \hat{\boldsymbol{y}}^\top\hat{\boldsymbol{y}} - n(\bar{\hat{y}})^2 + \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} \\ \Leftrightarrow \sum_{j=1}^n (y_j-\bar{y})^2 &= \sum_{j=1}^n (\hat{y}_j-\bar{y})^2 + \sum_{j=1}^n \hat{\varepsilon}_j^2\end{aligned}$$

SST=SSR+SSE

ANOVA table

SV	S.S.	df	M.S	F
Regression	SSR	r	M.S.= $\frac{S.S.}{df}$	$\frac{MSR}{MSE}$
Error	SSE	$n-r-1$		
Total	SST	$n-1$		

Reject $H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0$ if $F > F_{r,n-r-1}(\alpha)$.

$R^2 = \dfrac{SSR}{SST} = 1 - \dfrac{SSE}{SST}$; coefficient of determination.

6.2.2 Sampling properties of LSE

Result

① $E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, where $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{y}$
$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}$

② $E\hat{\boldsymbol{\varepsilon}} = \mathbf{0}, Cov(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$, where $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$

③ $S^2 \equiv MSE = \dfrac{SSE}{n-r-1} = \dfrac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n-r-1} = \dfrac{\boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}}{n-r-1}$
$\Rightarrow ES^2 = \sigma^2$

④ $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are uncorrelated

pf)

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{y} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{\varepsilon} \\
\hat{\boldsymbol{\varepsilon}} &= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)(\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= (\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)\boldsymbol{\varepsilon} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}
\end{aligned}$$

①

$$\begin{aligned}
E\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top E\boldsymbol{\varepsilon} = \boldsymbol{\beta} \\
Cov(\hat{\boldsymbol{\beta}}) &= Cov((\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{\varepsilon}) = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \left[Cov(\boldsymbol{\varepsilon})\right] \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \\
&= \sigma^2(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \text{ by } \ Cov(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}
\end{aligned}$$

②

$$\begin{aligned}
E\hat{\boldsymbol{\varepsilon}} &= (\boldsymbol{I} - \boldsymbol{H})E\boldsymbol{\varepsilon} = \mathbf{0} \\
Cov\hat{\boldsymbol{\varepsilon}} &= Cov((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}) = (\boldsymbol{I} - \boldsymbol{H})(Cov(\boldsymbol{\varepsilon}))(\boldsymbol{I} - \boldsymbol{H})^\top \\
&= \sigma^2(\boldsymbol{I} - \boldsymbol{H}) \text{ by } \boldsymbol{I} - \boldsymbol{H} : \text{ symmetric and idempotent.}
\end{aligned}$$

③

$$\begin{aligned}
ESSE &= E\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = E\boldsymbol{\varepsilon}^\top(\boldsymbol{I}-\boldsymbol{H})^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon} \\
&= E\boldsymbol{\varepsilon}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon} = E\ tr[\boldsymbol{\varepsilon}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}] \\
&= E\ tr[(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] \\
&= tr[(\boldsymbol{I}-\boldsymbol{H})\ E\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2 tr(\boldsymbol{I}-\boldsymbol{H}) \\
&= \sigma^2[tr\boldsymbol{I}_n - tr(\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)] = \sigma^2(n-(r+1)) \\
\text{since } tr(\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top) &= tr(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{Z} = tr\boldsymbol{I}_{r+1} \\
\therefore ES^2 &= E\frac{SSE}{n-r-1} = \sigma^2
\end{aligned}$$

④

$$\begin{aligned}
Cov(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\varepsilon}}) &= Cov(\boldsymbol{\beta}+(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{\varepsilon},(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}) \\
&= Cov((\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{\varepsilon},(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}) \\
&= (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top Cov(\boldsymbol{\varepsilon})(\boldsymbol{I}-\boldsymbol{H})^\top \\
&= \sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top(\boldsymbol{I}-\boldsymbol{H}) \\
&= \sigma^2\left\{(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top-(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\right\} = \boldsymbol{0}
\end{aligned}$$

□

6.2.3 Gauss' least squares theorem

Result
$\boldsymbol{y} = \boldsymbol{Z\beta}+\boldsymbol{\varepsilon}, E\boldsymbol{\varepsilon} = \boldsymbol{0}$ and $Cov\boldsymbol{\varepsilon} = \sigma^2\boldsymbol{I}$
$\boldsymbol{Z}$ has full col rank $r+1$

For any $\boldsymbol{c}$, the estimator

$$\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} = c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \cdots + c_r\hat{\beta}_r \text{ of } \boldsymbol{c}^\top\boldsymbol{\beta}$$

has the smallest possible variance among all linear estimators of the form

$$\boldsymbol{a}^\top\boldsymbol{y} = a_1y_1 + a_2y_2 + \cdots + a_ny_n$$

that are unbiased for $\boldsymbol{c}^\top\boldsymbol{\beta}$.

pf) $\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}$: BLUE (Best Linear Unbiased Estimator) of $\boldsymbol{c}^\top\boldsymbol{\beta}$

For any fixed $\boldsymbol{c}$, $E(\boldsymbol{a}^\top\boldsymbol{y}) = \boldsymbol{c}^\top\boldsymbol{\beta}\ \forall\,\boldsymbol{\beta}$.

By assumption, $E(\boldsymbol{a}^\top\boldsymbol{y}) = E(\boldsymbol{a}^\top(\boldsymbol{Z}\boldsymbol{\beta}+\boldsymbol{\varepsilon})) = \boldsymbol{a}^\top\boldsymbol{Z}\boldsymbol{\beta}$.
$\therefore\ \boldsymbol{c}^\top\boldsymbol{\beta} = \boldsymbol{a}^\top\boldsymbol{Z}\boldsymbol{\beta} \Rightarrow (\boldsymbol{c}^\top - \boldsymbol{a}^\top\boldsymbol{Z})\boldsymbol{\beta} = 0,\ \forall\,\boldsymbol{\beta}$
Since the above result holds for all $\boldsymbol{\beta}$, we can choose $\boldsymbol{\beta} = (\boldsymbol{c}^\top - \boldsymbol{a}^\top\boldsymbol{Z})^\top$, resulting in

$$\boldsymbol{c}^\top = \boldsymbol{a}^\top\boldsymbol{Z}.$$

Because $(\boldsymbol{c}^\top - \boldsymbol{a}^\top\boldsymbol{Z})(\boldsymbol{c}^\top - \boldsymbol{a}^\top\boldsymbol{Z})^\top = 0$ results in $\boldsymbol{c}^\top = \boldsymbol{a}^\top\boldsymbol{Z}$ for any linear unbiased estimator.

Now, consider $\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} = \boldsymbol{c}^\top(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y} = \boldsymbol{a}^{*\top}\boldsymbol{y}$ with $\boldsymbol{a}^* = \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{c}$.
$E\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} = \boldsymbol{c}E\hat{\boldsymbol{\beta}} = \boldsymbol{c}^\top\boldsymbol{\beta}$ by the above result.
$\therefore\ \boldsymbol{c}^\top\hat{\boldsymbol{\beta}} = \boldsymbol{a}^{*\top}\boldsymbol{y}$: UE of $\boldsymbol{c}^\top\boldsymbol{\beta}$
Hence

$$\boldsymbol{c}^\top = \boldsymbol{a}^{*\top}\boldsymbol{Z}$$

since $E\boldsymbol{a}^{*\top}\boldsymbol{y} = E\boldsymbol{a}^{*\top}(\boldsymbol{Z}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = \boldsymbol{a}^{*\top}\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{c}^\top\boldsymbol{\beta}$.

$$\begin{aligned}
V\boldsymbol{a}^\top\boldsymbol{y} &= V(\boldsymbol{a}^\top\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{a}^\top\boldsymbol{\varepsilon}) = V\boldsymbol{a}^\top\boldsymbol{\varepsilon} = \boldsymbol{a}^\top\sigma^2\boldsymbol{I}\boldsymbol{a}\\
&= \sigma^2(\boldsymbol{a} - \boldsymbol{a}^* + \boldsymbol{a}^*)^\top(\boldsymbol{a} - \boldsymbol{a}^* + \boldsymbol{a}^*)\\
&= \sigma^2[(\boldsymbol{a} - \boldsymbol{a}^*)^\top(\boldsymbol{a} - \boldsymbol{a}^*) + \boldsymbol{a}^*\boldsymbol{a}^*]
\end{aligned}$$

Since

$$\begin{aligned}
(\boldsymbol{a} - \boldsymbol{a}^*)^\top\boldsymbol{a}^* &= (\boldsymbol{a} - \boldsymbol{a}^*)^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{c}\\
&= (\boldsymbol{a}^\top\boldsymbol{Z} - \boldsymbol{a}^{*\top}\boldsymbol{Z})(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{c}\\
&= (\boldsymbol{c}^\top - \boldsymbol{c}^\top)(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{c} = 0.
\end{aligned}$$

Note that $(\boldsymbol{a} - \boldsymbol{a}^*)^\top(\boldsymbol{a} - \boldsymbol{a}^*) \geq 0$ and $\boldsymbol{a}^{*\top}\boldsymbol{a}^* \geq 0$

$\therefore\ V\boldsymbol{a}^\top\boldsymbol{a} \geq \sigma^2\boldsymbol{a}^{*\top}\boldsymbol{a}^*$ with equality holds when $\boldsymbol{a} = \boldsymbol{a}^*$.

$\boldsymbol{a}^{*\top}\boldsymbol{y} = \boldsymbol{c}^\top\hat{\boldsymbol{\beta}}$ is the BLUE (Best: minimum variance). □

6.3 Inferences about the Regression Model

Result
$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \ \boldsymbol{\varepsilon}$, $\boldsymbol{Z}_{n\times(r+1)}$; full column rank
$\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$
$\Rightarrow$

① $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}; LSE = MLE$

② $\hat{\boldsymbol{\beta}} \sim N_{r+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1})$

③ $\hat{\boldsymbol{\beta}}$ independent of $\boldsymbol{vare\hat{p}silon} = \boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}$

④ $n\hat{\sigma^2} = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} \sim \sigma^2\chi^2_{n-r-1}$, where $\hat{\sigma^2} = SSE/n$ is the MLE of σ^2

pf) See class homepage. □

Result
$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \ \boldsymbol{\varepsilon}$, $\boldsymbol{Z}_{n\times(r+1)}$; full column rank
$\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, $S^2 = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}/(n-r-1)$
$\Rightarrow$

① $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq (r+1)S^2F_{r+1,n-r-1}(\alpha)$;
100$(1-\alpha)$% confidence region for $\boldsymbol{\beta}$

② $\hat{\beta}_i \pm \sqrt{\widehat{V\hat{\beta}_i}}\sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}, i = 0, 1, \ldots, r$;
simultaneous 100$(1-\alpha)$% CIs for β_i, where $\widehat{V\hat{\beta}_i}$ is the diagonal elements of $S^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}$ corresponding to $\hat{\beta}_i$

pf) Let $\boldsymbol{v} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, then $E\boldsymbol{v} = \boldsymbol{0}$ since $E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. And $Cov(\boldsymbol{v}) = (\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2}Cov(\hat{\boldsymbol{\beta}})(\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2} = \sigma^2\boldsymbol{I}$ since $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}$.

$$\therefore \boldsymbol{v} \sim N_{r+1}(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

Hence

$$\begin{aligned} \boldsymbol{v}^\top\boldsymbol{v} &= (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top(\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2}(\boldsymbol{Z}^\top\boldsymbol{Z})^{1/2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ &= (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \sim \sigma^2\chi^2_{r+1}. \end{aligned}$$

We know that $(n-r-1)S^2 = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} \sim \sigma^2\chi^2_{n-r-1}$ and $\hat{\boldsymbol{\beta}}$ independent of $\hat{\boldsymbol{\varepsilon}}$ by the above result. Therefore

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\boldsymbol{Z}^\top\boldsymbol{Z})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/(r+1)}{(n-r-1)S^2/(n-r-1)} = \frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\boldsymbol{Z}^\top\boldsymbol{Z})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(r+1)S^2} \sim F_{r+1,n-r-1}$$

□

ex) See example 7.4 for SAS analysis.

6.3.1 Likelihood Ratio Test for the Regression Parameters

Full model; $Y_j = \beta_0 z_{j0} + \beta_1 z_{j1} + \cdots + \beta_r z_{jr} + \varepsilon_j, j = 1, \cdots, n$
Reduced model; $Y_j = \beta_0 z_{0j} + \beta_1 z_{j1} + \cdots + \beta_q z_{jq} + \varepsilon_j, j = 1, \cdots, n$ and $q < r$

Want to test

$$\begin{aligned} H_0 &: \beta_{q+1} = \cdots = \beta_r = 0 \quad \text{or} \\ H_0 &: \boldsymbol{\beta}_{(2)} = \mathbf{0},\ \text{where} \quad \boldsymbol{\beta}_{(2)} = (\beta_{q+1}\beta_{q+2}\cdots\beta_r)^\top \end{aligned}$$

Setting

$$\boldsymbol{Z} = [\underset{n\times(q+1)}{\boldsymbol{Z}_1} \ \vdots\ \underset{n\times(r-q)}{\boldsymbol{Z}_2}],\ \boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_{(1)} \\ \cdots \\ \boldsymbol{\beta}_{(2)}\end{bmatrix}\begin{matrix} {\scriptstyle (q+1)\times 1} \\ \\ {\scriptstyle (r-q)\times 1}\end{matrix}$$

Model;

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = (\boldsymbol{Z}_1 \vdots \boldsymbol{Z}_2)\begin{pmatrix}\boldsymbol{\beta}_{(1)} \\ \boldsymbol{\beta}_{(2)}\end{pmatrix} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}; \quad \text{Full model} \\ \boldsymbol{y} &= \boldsymbol{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}; \quad \text{Reduced model} \end{aligned}$$

$$\begin{aligned} SS_{res}(\boldsymbol{Z}) &= SS_{res}(\text{Full}) = (\boldsymbol{y}-\boldsymbol{Z}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\boldsymbol{Z}\hat{\boldsymbol{\beta}}), \quad \widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y} \\ SS_{res}(\boldsymbol{Z}_1) &= SS_{res}(\text{Reduced}) = (\boldsymbol{y}-\boldsymbol{Z}_1\hat{\boldsymbol{\beta}}_{(1)})'(\boldsymbol{y}-\boldsymbol{Z}_1\hat{\boldsymbol{\beta}}_{(1)}), \quad \hat{\boldsymbol{\beta}}_{(1)} = (\boldsymbol{Z}_1^\top\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1^\top\boldsymbol{y} \end{aligned}$$

$$\begin{aligned}\text{Extra } SS &= SS_{res}(\boldsymbol{Z}_1) - SS_{res}(\boldsymbol{Z}) \\ &= SS_{res}(\text{Reduced}) - SS_{res}(\text{Full})\end{aligned}$$

Result $\boldsymbol{Z}$; full column rank, $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$
LRT of $H_0 : \boldsymbol{\beta}_{(2)} = \boldsymbol{0}$ is to reject the null hypothesis if

$$\frac{[SS_{res}(\boldsymbol{Z}_1) - SS_{res}(\boldsymbol{Z})]/(r-q)}{S^2} > F_{r-q,n-r-1}(\alpha),$$

where $S^2 = \dfrac{SS_{res}(\boldsymbol{Z})}{n-r-1}$; MSE under Full model.

pf)

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[(-(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})/(2\sigma^2)\right] \le \frac{1}{(2\pi)^{n/2}\widehat{\sigma}^n} \exp\left(-\frac{n}{2}\right)$$

with the maximum occurring at $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}$ and

$$\widehat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})}{n}.$$

Similarly, under $H_0 \ : \ \boldsymbol{\beta}_{(2)} = \mathbb{0}$,

$$\max_{\boldsymbol{\beta}_{(1)},\ \sigma^2} L(\boldsymbol{\beta}_{(1)}, \sigma^2) = \frac{1}{(2\pi)^{n/2}\widehat{\sigma}_1^n} \exp(-n/2),$$

where the maximum occurs at $\widehat{\boldsymbol{\beta}}_{(1)} = (\boldsymbol{Z}_1^\top\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1^\top\boldsymbol{y}$

$$\widehat{\sigma}_1^2 = (\boldsymbol{y} - \boldsymbol{Z}_1\widehat{\boldsymbol{\beta}}_{(1)})^\top(\boldsymbol{y} - \boldsymbol{Z}_1\widehat{\boldsymbol{\beta}}_{(1)}/n$$

Rejecting $H_0 \ : \ \boldsymbol{\beta}_{(2)} = \mathbb{0}$ for small values of the likelihood ratio

$$\begin{aligned}\frac{\max\limits_{\boldsymbol{\beta}_{(1)},\sigma^2} L(\boldsymbol{\beta}_{(1)}, \sigma^2)}{\max\limits_{\boldsymbol{\beta},\sigma^2} L(\widehat{\boldsymbol{\beta}}, \sigma^2)} &= \left(\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}^2}\right)^{-n/2} = \left(\frac{\widehat{\sigma}^2 + \widehat{\sigma}_1^2 - \widehat{\sigma}^2}{\widehat{\sigma}^2}\right)^{-n/2} \\ &= \left(1 + \frac{\widehat{\sigma}_1^2 - \widehat{\sigma}^2}{\widehat{\sigma}^2}\right)^{-n/2}\end{aligned}$$

is equivalent to rejecting H_0 for large values of $(\widehat{\sigma}_1^2 - \widehat{\sigma}^2)/\widehat{\sigma}^2$

$$\begin{aligned}
\Leftrightarrow &\quad \frac{n(\widehat{\sigma}_1^2 - \widehat{\sigma}^2)}{n\widehat{\sigma}^2} \quad \text{is large} \\
\Leftrightarrow &\quad \frac{SS_{res}(\boldsymbol{Z}_1) - SS_{res}(\boldsymbol{Z})}{SS_{res}(\boldsymbol{Z})} \quad \text{is large} \\
\Leftrightarrow &\quad \text{Reject} \quad H_0 : \ \boldsymbol{\beta}_{(2)} = \mathbb{0} \\
&\quad \text{if} \quad \frac{[SS_{res}(\boldsymbol{Z}_1) - SS_{res}(\boldsymbol{Z})]/(r-q)}{SS_{res}(\boldsymbol{Z})/(n-r-1)} > F_{r-q,n-r-1}(\alpha)
\end{aligned}$$

□

<u>Remark</u>
Under H_0, both numerator and denominator are independent χ^2 with $r-q$ & $n-r-1$ dfs. Hence F distribution is obtained.

ex) See Example 7.5 pg. 376-378.

6.4 Inferences from the Estimated Regression Function

6.4.1 Estimating the Regression function at a point

$$\begin{aligned}
E(y_0|\boldsymbol{z}_0) &= \beta_0 + \beta_1 z_{01} + \cdots + \beta_r z_{0r} = \boldsymbol{z}_0^\top \boldsymbol{\beta} \\
&= (1 \ z_{01} \ \cdots \ z_{0r})^\top \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{pmatrix} ; \ \text{a linear combination of } \boldsymbol{\beta}
\end{aligned}$$

<u>Result</u>

① $\boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}$: unbiased estimator of $E(y_0|\boldsymbol{z}_0) = \boldsymbol{z}_0^\top \boldsymbol{\beta}$

② $V(\boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}) = \boldsymbol{z}_0^\top (V\widehat{\boldsymbol{\beta}})\boldsymbol{z}_0 = \sigma^2 \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0$;
BLUE for $E(y_0|\boldsymbol{z}_0) = \boldsymbol{z}_0^\top \boldsymbol{\beta}$ by Gauss' least squares theorem

③ $\boldsymbol{\varepsilon} \sim N_n(\mathbb{0}, \sigma^2 \boldsymbol{I})$
$\Rightarrow \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}} \ \pm \ t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{S^2 \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{z}_0}$;
a $100(1-\alpha)\%$ CI for $E(y_0|\boldsymbol{z}_0) = \boldsymbol{z}_0^\top \boldsymbol{\beta}$

pf)

$$\begin{aligned} &\widehat{\boldsymbol{\beta}} \sim N_{r+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}) \text{ and} \\ &\frac{(n-r-1)S^2}{\sigma^2} \ \sim \ \chi^2_{n-r-1}. \end{aligned}$$

They are independent since $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are. Furthermore $S^2 = \dfrac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n-r-1}$, which is a function of $\hat{\boldsymbol{\varepsilon}}$.
$\Rightarrow$

$$\begin{aligned} &\boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}} \ \sim \ N(\boldsymbol{z}_0^\top \boldsymbol{\beta}, \ \sigma^2 \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{z}_0) \\ &\frac{\dfrac{\boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{z}_0^\top \boldsymbol{\beta}}{\sqrt{\sigma^2 \boldsymbol{z}_0^\top \left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} \boldsymbol{z}_0}}}{\sqrt{\dfrac{(n-r-1)S^2}{\sigma^2}/(n-r-1)}} = \frac{\boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{z}_0^\top \boldsymbol{\beta}}{\sqrt{S^2 \boldsymbol{z}_0^\top \left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} \boldsymbol{z}_0}} \sim t(n-r+1) \end{aligned}$$

□

6.4.2 Forcasting a New observation at a point

<u>Result</u> Note that, from the model, $y_0 = \boldsymbol{z}_0^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon}_0$

① $\boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}$: unbiased predictor of y_0

② $V(y_0 - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}) = \sigma^2(1 + \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{z}_0)$: forecast error

③ $\boldsymbol{\epsilon} \sim N_n(\mathbb{0}, \sigma^2 \boldsymbol{I})$
$\Longrightarrow$
$\boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}} \ \pm \ t_{\left(n-r-1, \frac{\alpha}{2}\right)}\sqrt{S^2(1 + \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{z}_0)}$;
a $100(1-\alpha)\%$ prediction interval.

pf) ②

$$\begin{aligned} & y_0 - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}} = \boldsymbol{z}_0^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon}_0 - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}} = \boldsymbol{\varepsilon}_0 + \boldsymbol{z}_0^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \\ & V(y_0 - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}) \ = \ V\boldsymbol{\varepsilon}_0 + V(\boldsymbol{z}_0^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})) \quad (\because \widehat{\boldsymbol{\beta}} \text{ independent of } \ \varepsilon_0) \\ & = \sigma^2 + \sigma^2 \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0 \quad = \quad \sigma^2 \left[1 + \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0\right] \end{aligned}$$

□

ex) See Example 7.6 pg. 380-381.

6.5 Model checking and other Aspects of Regression

Tools

① Studentized (standardized) residual

② residual plot

③ normal probability plot

④ high leverage measure

⑤ Cook's distance

⑥ Residual ACF (autocorrelation function)

6.5.1 Standardized (studentized) residual

$$\begin{aligned} \widehat{\boldsymbol{\epsilon}} &= \boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}} \ = \ \boldsymbol{y} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{y} \ = \ (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \\ E\widehat{\boldsymbol{\varepsilon}} &= \mathbb{0} \\ V\widehat{\boldsymbol{\varepsilon}} &= (\boldsymbol{I} - \boldsymbol{H})\left\{Var\left(\boldsymbol{y}\right)\right\}(\boldsymbol{I} - \boldsymbol{H})^\top \ = \ \sigma^2(\boldsymbol{I} - \boldsymbol{H}) \end{aligned}$$

Note that

$$E\boldsymbol{\varepsilon} = \mathbb{0} \ , \ V\boldsymbol{y} \ = \ V\boldsymbol{\varepsilon} \ = \ \sigma^2 \boldsymbol{I}.$$

$$\therefore\ V\widehat{\varepsilon}_j \ = \ \sigma^2(1-h_{jj})\ ,\ \text{where } \boldsymbol{H} = (h_{ij})_{n\times n}$$

<u>standardized residual</u>;

$$\widehat{\varepsilon}_j^* = \frac{\widehat{\varepsilon}_j}{\sqrt{S^2(1-h_{jj})}},\ j = 1,\cdots,n$$

If the model is correct,

$$\widehat{\varepsilon}_j^* \ \dot{\sim}\ \text{iid } N(0,1).$$

Standardized residual is sometimes called as internally studentized residual.

<u>studentized residual</u>;

$$\widehat{r}_j^* = \frac{\widehat{\varepsilon}_j}{\sqrt{S^2(j)(1-h_{jj})}},\ j = 1,\cdots,n,$$

where $S(j)^2$ is the calculated S^2 after removing j^{th} measurement y_j. This measure is sometimes called as externally studentized residual to distinguish it from the standardized residual. There is a relationship between standardized residual and studentized residual, that is,

$$S^2(j) = \left[(n-r-1)S^2 - \frac{\widehat{\varepsilon}_j^2}{1-h_{jj}}\right]/(n-r-2).$$

So no need to refit n regression models.

6.5.2 Leverage measure

Consider simple linear regression ($r = 1$), then

$$\boldsymbol{Z} = \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{bmatrix}_{n\times 2} \qquad \boldsymbol{Z}^\top\boldsymbol{Z} = \begin{pmatrix} n & \sum z_j \\ \sum z_j & \sum z_j^2 \end{pmatrix} |\boldsymbol{Z}^\top\boldsymbol{Z}| = n\sum_j (z_j - \bar{z})$$

$$\therefore\ \ (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} = \begin{bmatrix} \frac{\sum z_j^2}{n} & -\bar{z} \\ -\bar{z} & 1 \end{bmatrix} \frac{1}{\sum_j (z_j - \bar{z})^2}$$

$$h_{ii} \ = \ (1 \quad z_i)(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\begin{pmatrix} 1 \\ z_i \end{pmatrix} \ = \ \frac{1}{n} + \frac{(z_i - \bar{z})^2}{\sum_j (z_j - \bar{z})^2}$$

$\therefore h_{ii}$ measure the relative distance of z_i to the center $\bar{z}$.

Remark

i)

$$\begin{aligned} tr(\boldsymbol{H}) &= tr(\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top) = tr((\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{Z}) \\ &= tr\boldsymbol{I}_{r+1} = r+1 \end{aligned}$$

$$\therefore \text{ average leverage } = \frac{r+1}{n}$$

ii) High leverage points play a more dominating role in determining the regression function.

6.5.3 Cook's distance

A data point is called an influential observation if deleting it results in "big" change in $\widehat{\boldsymbol{\beta}}$.

Cook's distance; distance measure between $\widehat{\boldsymbol{\beta}}$ & $\widehat{\boldsymbol{\beta}}(i)$, where $\widehat{\boldsymbol{\beta}}(i)$ is LSE without i^{th} data point

$$D(i) = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}(i))^\top \boldsymbol{Z}^\top \boldsymbol{Z}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}(i))}{(r+1)S^2}$$

6.5.4 Residual ACF

$$\text{lag-}k \text{ ACF} \quad ; \quad \widehat{\rho}_k = \frac{\sum_{j=k+1}^n \widehat{\varepsilon}_j \widehat{\varepsilon}_{j-k}}{\sum_{j=1}^n \widehat{\varepsilon}_j^2}$$

Residual ACF includes DW-statistic because

$$\begin{aligned} &DW \approx 2(1 - \widehat{\rho}_1), \\ &DW = \sum_{j=2}^n (\widehat{\varepsilon}_j - \widehat{\varepsilon}_{j-1})^2 / \sum_{j=1}^n \widehat{\varepsilon}_j^2. \end{aligned}$$

6.5.5 Final Remark

- Residual plot; outlier? nonlinearity? homogeneity?
- Normal probability plot; normality
- h_{ii} or Cook's distance; influential points
- residual ACF; independence
- variable selection (model selection) by stepwise regression, etc
- multi collinearity; $\boldsymbol{Z}^\top \boldsymbol{Z}$ matrix close to singularity because of high correlation between the independent (explanatory) variables.

6.6 Multivariate Multiple regression

6.6.1 Introduction

Multiple regression;

$$\underset{n\times 1}{\boldsymbol{y}} = \underset{n\times(r+1)}{\boldsymbol{Z}} \underset{(r+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

Goal: Find a relationship between

$$z_1, z_2, \cdots, z_r$$

(independent variables) and

$$y_1, y_2, \cdots, y_m$$

(dependent variables) ($y_2, \cdots, y_m$; extra compared to multiple regression).

Put m multiple regression together. This leads to multivariate multiple regression.

1^{st} **regression;**

$$\begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n1} \end{bmatrix} = \begin{bmatrix} z_{10} & z_{11} & \cdots & z_{1r} \\ z_{20} & z_{21} & \cdots & z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n0} & z_{n1} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_{01} \\ \beta_{11} \\ \vdots \\ \beta_{r1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{n1} \end{bmatrix}$$

$$\Leftrightarrow \quad \boldsymbol{y}_{(1)} = \boldsymbol{Z}\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}_{(1)}$$

2^{nd} **regression;**

$$\begin{aligned}
\begin{bmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{n2} \end{bmatrix} &= \begin{bmatrix} z_{10} & z_{11} & \cdots & z_{1r} \\ z_{20} & z_{21} & \cdots & z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n0} & z_{n1} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_{02} \\ \beta_{12} \\ \vdots \\ \beta_{r2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{12} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{n2} \end{bmatrix} \\
\Leftrightarrow \quad \boldsymbol{y}_{(2)} &= \boldsymbol{Z}\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}_{(2)} \\
&\vdots
\end{aligned}$$

m^{th} **regression;**

$$\begin{aligned}
\begin{bmatrix} y_{1m} \\ y_{2m} \\ \vdots \\ y_{nm} \end{bmatrix} &= \begin{bmatrix} z_{10} & z_{11} & \cdots & z_{1r} \\ z_{20} & z_{21} & \cdots & z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n0} & z_{n1} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_{0m} \\ \beta_{1m} \\ \vdots \\ \beta_{rm} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1m} \\ \varepsilon_{2m} \\ \vdots \\ \varepsilon_{nm} \end{bmatrix} \\
\Leftrightarrow \quad \boldsymbol{y}_{(m)} &= \boldsymbol{Z}\boldsymbol{\beta}_{(m)} + \boldsymbol{\varepsilon}_{(m)}
\end{aligned}$$

Putting the m multiple regression together, we have

$$\begin{aligned}
\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \cdots & \cdots & \ddots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{bmatrix} &= \begin{bmatrix} z_{10} & z_{11} & \cdots & z_{1r} \\ z_{20} & z_{21} & \cdots & z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n0} & z_{n1} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rm} \end{bmatrix} \\
&+ \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{bmatrix} \\
\Leftrightarrow \begin{bmatrix} \boldsymbol{y}_{(1)} & \boldsymbol{y}_{(2)} & \cdots & \boldsymbol{y}_{(m)} \end{bmatrix} &= \boldsymbol{Z} \begin{bmatrix} \boldsymbol{\beta}_{(1)} & \boldsymbol{\beta}_{(2)} & \cdots & \boldsymbol{\beta}_{(m)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{(1)} & \boldsymbol{\varepsilon}_{(2)} & \cdots & \boldsymbol{\varepsilon}_{(m)} \end{bmatrix}
\end{aligned}$$

But allow for contemporaneous correlation in $\boldsymbol{\varepsilon}$.

$$\begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}_1^\top \\ \boldsymbol{\varepsilon}_2^\top \\ \vdots \\ \boldsymbol{\varepsilon}_n^\top \end{bmatrix}, \text{ where } \boldsymbol{\varepsilon}_j^\top = [\varepsilon_{j1}\ \varepsilon_{j2}\ \cdots\ \varepsilon_{jm}]$$

with $Cov(\boldsymbol{\varepsilon}_j) = \boldsymbol{\Sigma}_{m\times m}$ ($\boldsymbol{\Sigma}:$ positive definite).

In summary, the multivariate multiple regression is

$$\begin{aligned}
&\underset{n\times m}{\boldsymbol{Y}} = \underset{n\times(r+1)}{\boldsymbol{Z}}\underset{(r+1)\times m}{\boldsymbol{\beta}} + \underset{n\times m}{\boldsymbol{\varepsilon}}, \quad \text{where} \\
&E\boldsymbol{\varepsilon}_{(i)} = \mathbb{0} \text{ and } Cov(\boldsymbol{\varepsilon}_{(i)}, \boldsymbol{\varepsilon}_{(k)}) = \sigma_{ik}\boldsymbol{I}_n, \quad i,k = 1,2,\cdots,m \\
&Cov(\boldsymbol{\varepsilon}_j) = \boldsymbol{\Sigma} = (\sigma_{ik})_{m\times m}.
\end{aligned}$$

Remark

① $\boldsymbol{y}_{(i)} = \boldsymbol{Z}\boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}, \ i = 1,\cdots,m$

$$Cov(\boldsymbol{\varepsilon}_{(i)}) = \sigma_{ii}\boldsymbol{I}_n$$

② $Cov(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_k) = \boldsymbol{0}_{m\times m}$, if $i \neq k$. That is, the errors for different trials are uncorrelated.

However, $Cov(\boldsymbol{\varepsilon}_j) = \boldsymbol{\Sigma} = (\sigma_{ik})_{m\times m}$ means that the errors for different responses on the same trial can be correlated. σ_{ik} may not be zero.

For each equation, $\boldsymbol{y}_{(i)} = \boldsymbol{Z}\boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}, \ i = 1,\cdots,m$, the LSE of $\boldsymbol{\beta}_{(i)}$ is

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}_{(i)}.$$

Collecting these estimates, we obtain

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\beta}}_{(1)}\ \widehat{\boldsymbol{\beta}}_{(2)}\ \cdots\ \widehat{\boldsymbol{\beta}}_{(m)}\right) = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\left[\boldsymbol{y}_{(1)}\ \boldsymbol{y}_{(2)}\ \cdots\ \boldsymbol{y}_{(m)}\right]$$

or

$$\underset{(r+1)\times m}{\widehat{\boldsymbol{\beta}}} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}$$

Note

i) LSE $\widehat{\boldsymbol{\beta}}$ minimizes $tr\underset{m\times m}{(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})}$

ii) LSE $\widehat{\boldsymbol{\beta}}$ minimizes $|(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})|$

pf of i) For any choice of parameters $\boldsymbol{B} = \left[\boldsymbol{b}_{(1)}\ \boldsymbol{b}_{(2)}\ \cdots\ \boldsymbol{b}_{(m)}\right]$, the matrix of errors is $\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{B}$. The error sum of squares and cross products matrix is

$$\begin{aligned}&(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})\\ &= \begin{bmatrix} (\boldsymbol{Y}_{(1)} - \boldsymbol{Z}\boldsymbol{b}_{(1)})^\top(\boldsymbol{Y}_{(1)} - \boldsymbol{Z}\boldsymbol{b}_{(1)}) & \cdots & (\boldsymbol{Y}_{(1)} - \boldsymbol{Z}\boldsymbol{b}_{(1)})^\top(\boldsymbol{Y}_{(m)} - \boldsymbol{Z}\boldsymbol{b}_{(m)}) \\ \vdots & & \vdots \\ (\boldsymbol{Y}_{(m)} - \boldsymbol{Z}\boldsymbol{b}_{(m)})^\top(\boldsymbol{Y}_{(1)} - \boldsymbol{Z}\boldsymbol{b}_{(1)}) & \cdots & (\boldsymbol{Y}_{(m)} - \boldsymbol{Z}\boldsymbol{b}_{(m)})^\top(\boldsymbol{Y}_{(m)} - \boldsymbol{Z}\boldsymbol{b}_{(m)}) \end{bmatrix}.\end{aligned}$$

The selection $\boldsymbol{b}_{(i)} = \widehat{\boldsymbol{\beta}}_{(i)}$ minimizes the ith diagonal sum of squares $(\boldsymbol{Y}_{(i)} - \boldsymbol{Z}\boldsymbol{b}_{(i)})^\top(\boldsymbol{Y}_{(i)} - \boldsymbol{Z}\boldsymbol{b}_{(i)})$. Consequently, $tr(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})$ is minimized by the choice $\boldsymbol{B} = \widehat{\boldsymbol{\beta}}$. For a proof of ii), see Exercise 7.11. □

Using the LSE $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}$,

$$\begin{aligned}\text{predicted values: } \widehat{\boldsymbol{Y}} &= \boldsymbol{Z}\widehat{\boldsymbol{\beta}} = \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y},\\ \text{residuals: } \widehat{\boldsymbol{\varepsilon}} &= \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)\boldsymbol{Y}.\end{aligned}$$

Remark

- $\boldsymbol{Z}^\top\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Z}^\top(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)\boldsymbol{Y} = \boldsymbol{0}$
- $\widehat{\boldsymbol{Y}}^\top\widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\beta}}^\top\boldsymbol{Z}^\top(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)\boldsymbol{Y} = \boldsymbol{0}$

Because

$$\begin{aligned}&\boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{\varepsilon}} \iff \boldsymbol{Y} = \widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{\varepsilon}}\\ &\boldsymbol{Y}^\top\boldsymbol{Y} = (\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{\varepsilon}})^\top(\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{\varepsilon}}) = \widehat{\boldsymbol{Y}}^\top\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}\end{aligned}$$

(total s.s. & cross products) = (predicted s.s. & cross products) + (residual (error) s.s. & cross product)

Remark
$\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y}^\top\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^\top\widehat{\boldsymbol{Y}} = \boldsymbol{Y}^\top\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}^\top\boldsymbol{Z}^\top\boldsymbol{Z}\widehat{\beta}$

ex) See example 7.8 for fitting a multivariate straight-line regression model.

$$\widehat{y}_1 = 1 + 2z_1 \quad \& \quad \widehat{y}_2 = -1 + z_1$$

Result For multivariate multiple regression, the LSE

$$\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{(1)}\ \widehat{\boldsymbol{\beta}}_{(2)}\ \cdots\ \widehat{\boldsymbol{\beta}}_{(m)})\ \&\ \boldsymbol{Z}$$

with full col rank $r+1 < n$,

① $E\widehat{\boldsymbol{\beta}}_{(i)} = \boldsymbol{\beta}_{(i)}$ or $E\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$

② $Cov(\widehat{\boldsymbol{\beta}}_{(i)}, \widehat{\boldsymbol{\beta}}_{(k)}) = \sigma_{ik}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1},\ i,k = 1,\cdots,m$

③ $\widehat{\boldsymbol{\varepsilon}} = (\widehat{\boldsymbol{\varepsilon}}_{(1)}\ \widehat{\boldsymbol{\varepsilon}}_{(2)}\cdots\widehat{\boldsymbol{\varepsilon}}_{(m)}) = \boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$

$$E\widehat{\boldsymbol{\varepsilon}}_{(i)} = \mathbb{0},\ E\widehat{\boldsymbol{\varepsilon}}_{(i)}^\top\widehat{\boldsymbol{\varepsilon}}_{(k)} = (n-r-1)\sigma_{ik}$$

so

$$E\widehat{\boldsymbol{\varepsilon}} = \mathbf{0}, E\left(\frac{1}{n-r-1}\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}\right) = \boldsymbol{\Sigma}$$

④ $\widehat{\boldsymbol{\varepsilon}}$ & $\widehat{\boldsymbol{\beta}}$ are uncorrelated.

pf) $\boldsymbol{y}_{(i)} = \boldsymbol{Z}\boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}$, $E\boldsymbol{\varepsilon}_{(i)} = \mathbb{0}$ & $E\boldsymbol{\varepsilon}_{(i)}\boldsymbol{\varepsilon}_{(i)}^\top = \sigma_{ii}\boldsymbol{I}$

① $E\widehat{\boldsymbol{\beta}}_{(i)} = E(\boldsymbol{Z}^\top\boldsymbol{Y})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}_{(i)} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top E\boldsymbol{y}_{(i)} = \boldsymbol{\beta}_{(i)}$

② $Cov(\widehat{\boldsymbol{\beta}}_{(i)}, \widehat{\boldsymbol{\beta}}_{(k)}) = E(\widehat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})(\widehat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})^\top$.
Note that

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)} &= (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}_{(i)} - \boldsymbol{\beta}_{(i)} \\
&= (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}) - \boldsymbol{\beta}_{(i)} \\
&= \boldsymbol{\beta}_{(i)} + (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{\varepsilon}_{(i)}) - \boldsymbol{\beta}_{(i)} \\
&= (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{\varepsilon}_{(i)}
\end{aligned}$$

Hence

$$\begin{aligned}
Cov(\widehat{\boldsymbol{\beta}}_{(i)}, \widehat{\boldsymbol{\beta}}_{(k)}) &= E(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{\varepsilon}_{(i)}\boldsymbol{\varepsilon}_{(i)}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} \\
&= (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top[E\boldsymbol{\varepsilon}_{(i)}\boldsymbol{\varepsilon}_{(k)}^\top]\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} = \sigma_{ik}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}
\end{aligned}$$

since $E\boldsymbol{\varepsilon}_{(i)}\boldsymbol{\varepsilon}_{(k)}^\top = \sigma_{ik}\boldsymbol{I}_n$.

③

$$\begin{aligned} E\widehat{\boldsymbol{\varepsilon}}_{(i)} &= E(\boldsymbol{y}_{(i)} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{(i)}) = \boldsymbol{Z}\boldsymbol{\beta}_{(i)} - \boldsymbol{Z}\boldsymbol{\beta}_{(i)} = \mathbb{0} \\ E\widehat{\boldsymbol{\varepsilon}}_{(i)}^\top \widehat{\boldsymbol{\varepsilon}}_{(k)} &= E\boldsymbol{\varepsilon}_{(i)}^\top (\boldsymbol{I}-\boldsymbol{H})^\top (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}_{(k)} \end{aligned}$$

since

$$\begin{aligned} \widehat{\boldsymbol{\varepsilon}}_{(i)} &= \boldsymbol{y}_{(i)} - \widehat{\boldsymbol{y}}_{(i)} = \boldsymbol{y}_{(i)} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{(i)} = \boldsymbol{y}_{(i)} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}_{(i)} \\ &= (\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)\boldsymbol{y}_{(i)} \\ &= (\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)(\boldsymbol{Z}\boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}) \\ &= (\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)\boldsymbol{\varepsilon}_{(i)} \\ &= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}_{(i)}. \end{aligned}$$

Therefore,

$$\begin{aligned} E\widehat{\boldsymbol{\varepsilon}}_{(i)}^\top \widehat{\boldsymbol{\varepsilon}}_{(k)} &= E\boldsymbol{\varepsilon}_{(i)}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}_{(k)} = tr[E\boldsymbol{\varepsilon}_{(i)}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}_{(k)}] \\ &= E\ tr[(\boldsymbol{\varepsilon}_{(i)}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}_{(k)})] \\ &= E\ tr[(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\varepsilon}_{(k)}\boldsymbol{\varepsilon}_{(i)}^\top] = tr[(\boldsymbol{I}-\boldsymbol{H})E\boldsymbol{\varepsilon}_{(k)}\boldsymbol{\varepsilon}'_{(i)}] \\ &= tr[(\boldsymbol{I}-\boldsymbol{H})\sigma_{ki}\boldsymbol{I}] \\ &= \sigma_{ki}\ tr(\boldsymbol{I}-\boldsymbol{H}) = \sigma_{ki}\ tr(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top) \\ &= \sigma_{ki}(n-r-1) \text{ as in Multiple linear regression.} \end{aligned}$$

④

$$\begin{aligned} Cov(\widehat{\boldsymbol{\beta}}_{(i)}, \widehat{\boldsymbol{\varepsilon}}_{(k)}) &= E(\widehat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})(\widehat{\boldsymbol{\varepsilon}}_{(k)} - \mathbb{0})^\top \\ &= E(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{\varepsilon}_{(i)}\boldsymbol{\varepsilon}_{(k)}^\top(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top) \\ &= (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top[E\boldsymbol{\varepsilon}_{(i)}\boldsymbol{\varepsilon}'_{(k)}](\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top) \\ &= \sigma_{ik}[(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}(\boldsymbol{Z}^\top - \boldsymbol{Z}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)] = \boldsymbol{0} \end{aligned}$$

Above result enables us to obtain the sampling properties of the least squares predictors.

i) Estimating the mean vectors for $\boldsymbol{z}_0 = (1\ z_{01}\ \cdots\ z_{0r})$; a point of interest

$\boldsymbol{z}_0^\top\widehat{\boldsymbol{\beta}} = (\boldsymbol{z}_0^\top\widehat{\boldsymbol{\beta}}_{(1)}\ \boldsymbol{z}_0^\top\widehat{\boldsymbol{\beta}}_{(2)} \cdots \boldsymbol{z}_0^\top\widehat{\boldsymbol{\beta}}_{(m)})$

: Unbiased estimator of $\boldsymbol{z}_0^\top\boldsymbol{\beta}$ since $E\boldsymbol{z}_0^\top\widehat{\boldsymbol{\beta}}_{(i)} = \boldsymbol{z}_0^\top E\widehat{\boldsymbol{\beta}}_{(i)} = \boldsymbol{z}_0^\top\boldsymbol{\beta}_{(i)},\quad i = 1, \cdots, m$

Estimation error; $\boldsymbol{z}_0^\top \boldsymbol{\beta}_{(i)} - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(i)} = \boldsymbol{z}_0^\top (\boldsymbol{\beta}_{(i)} - \widehat{\boldsymbol{\beta}}_{(i)})$

$$\begin{aligned} & Cov(\boldsymbol{z}_0^\top (\boldsymbol{\beta}_{(i)} - \widehat{\boldsymbol{\beta}}_{(i)}),\ \boldsymbol{z}_0^\top (\boldsymbol{\beta}_{(k)} - \widehat{\boldsymbol{\beta}}_{(k)})) \\ = & E\boldsymbol{z}_0^\top (\boldsymbol{\beta}_{(i)} - \widehat{\boldsymbol{\beta}}_{(i)})[\boldsymbol{z}_0^\top (\boldsymbol{\beta}_{(k)} - \widehat{\boldsymbol{\beta}}_{(k)})]^\top \\ = & \boldsymbol{z}_0^\top [E(\widehat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})(\widehat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})^\top]\boldsymbol{z}_0 \\ = & \sigma_{ik}\boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{z}_0 \end{aligned}$$

ii) Forecasting a new obs. $\boldsymbol{y}_0^\top$=$(y_{01}\ y_{02}\ \cdots\ y_{0m})$ at $\boldsymbol{z}_0$

$$\begin{aligned} y_{0i} &= \boldsymbol{z}_0^\top \boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{0i} \\ \boldsymbol{\varepsilon}_0^\top &= (\varepsilon_{01}\ \varepsilon_{02}\ \cdots\ \varepsilon_{0m}) \text{ independent of } \boldsymbol{\varepsilon}\ \& \\ E\varepsilon_{0i} &= 0,\ E\varepsilon_{0i}\ \varepsilon_{0k} = \sigma_{ik} \end{aligned}$$

$$\begin{aligned} \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(i)} \quad &: \text{ Unbiased estimator of } y_{0i} \\ (\because) \quad & y_{0i} - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(i)} = y_{0i} - \boldsymbol{z}_0^\top \boldsymbol{\beta}_{(i)} + \boldsymbol{z}_0^\top \boldsymbol{\beta}_{(i)} - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(i)} \\ & = \varepsilon_{0i} - \boldsymbol{z}_0^\top (\widehat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)}) \\ & E(y_{0i} - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(i)}) = 0 \ \ (\because)\ E\varepsilon_{0i} = 0\ \&\ E\widehat{\boldsymbol{\beta}}_{(i)} = \boldsymbol{\beta}_{(i)} \end{aligned}$$

Forecast error; $y_{0i} - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(i)}$

$$\begin{aligned} & Cov(y_{0i} - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(i)},\ y_{0k} - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(k)}) \\ = & E[(y_{0i} - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(i)})(y_{0k} - \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(k)})] \\ = & E(\varepsilon_{0i} - \boldsymbol{z}_0^\top (\widehat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)}))(\varepsilon_{0k} - \boldsymbol{z}_0^\top (\widehat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})) \\ = & E\varepsilon_{0i}\varepsilon_{0k} + \boldsymbol{z}_0^\top E(\widehat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})(\widehat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_k)^\top \boldsymbol{z}_0 \\ & -\boldsymbol{z}_0^\top [E(\widehat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})\varepsilon_{0k}] - [E\varepsilon_{0i}(\widehat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})^\top]\boldsymbol{z}_0 \\ = & \sigma_{ik}(1 + \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{z}_0) \end{aligned}$$

since $\widehat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)}$ is a function of $\boldsymbol{\varepsilon}_{(i)}$ and independent of $\boldsymbol{\varepsilon}_0$ by assumption.

Result

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \vdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \vdots & \varepsilon_{2m} \\ \cdots & \cdots & \vdots & \cdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{bmatrix} = \begin{pmatrix} \boldsymbol{\varepsilon}_1' \\ \boldsymbol{\varepsilon}_2' \\ \vdots \\ \boldsymbol{\varepsilon}_n' \end{pmatrix}$$

Assume that

$$\begin{aligned}&\boldsymbol{\varepsilon}_i \sim \text{ independent } N_m(\mathbb{0},\ \boldsymbol{\Sigma})\ \&\\ &r(\boldsymbol{Z}) = r+1 \text{ (full column rank)},\\ &n \geqq (r+1)+m\end{aligned}$$

$\Rightarrow$

i) $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{Y}$;MLE of $\boldsymbol{\beta}$
$\sim$ a Normal distribuion with $E\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ & $Cov(\widehat{\boldsymbol{\beta}}_{(i)},\ \widehat{\boldsymbol{\beta}}_{(k)}) = \sigma_{ik}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}$

ii) $\widehat{\boldsymbol{\Sigma}} = \dfrac{1}{n}\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}} = \dfrac{1}{n}(\boldsymbol{Y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{Y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}})$
$n\widehat{\boldsymbol{\Sigma}}\ \sim\ W_{m,n-r-1}(\boldsymbol{\Sigma})$

iii) $\widehat{\boldsymbol{\beta}}\ \perp\!\!\!\perp\ \widehat{\boldsymbol{\Sigma}}$

iv) $L(\widehat{\boldsymbol{\beta}},\ \widehat{\boldsymbol{\Sigma}}) = (2\pi)^{-\dfrac{mn}{2}}\left|\widehat{\boldsymbol{\Sigma}}\right|^{-\dfrac{n}{2}} e^{-\dfrac{m}{2}}$; maximized likelihood

Note that the dimension of $\widehat{\boldsymbol{\beta}}$ is $(r+1)\times m$ so its distribution is a matrix variate distribution. Need to do more, but we will keep it this way.

6.6.2 Likelihood ratio tests for regression parameters

$$H_0\ :\ \boldsymbol{\beta}_{(2)} = \mathbf{0}\ v.s.\ H_1\ :\ \boldsymbol{\beta}_{(2)} \neq \mathbf{0},$$

where

$$\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_{(1)}\\ \boldsymbol{\beta}_{(2)}\end{bmatrix}\begin{matrix}(q+1)\times m\\ (r-q)\times m\end{matrix}$$

Partition

$$\boldsymbol{Z} = \begin{bmatrix}\underset{n\times(q+1)}{\boldsymbol{Z}_1} & \underset{n\times(r-q)}{\boldsymbol{Z}_2}\end{bmatrix}$$

$$E\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} = (\boldsymbol{Z}_1\ \boldsymbol{Z}_2)\begin{pmatrix}\boldsymbol{\beta}_{(1)}\\ \boldsymbol{\beta}_{(2)}\end{pmatrix} = \boldsymbol{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_2\boldsymbol{\beta}_{(2)}$$

Under $H_0\ :\ \boldsymbol{\beta}_{(2)} = \mathbf{0},\ \boldsymbol{Y} = \boldsymbol{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}$

Extra s.s. and cross products

$$\begin{aligned} &= (\boldsymbol{Y} - \boldsymbol{Z}_1\widehat{\boldsymbol{\beta}}_{(1)})^\top(\boldsymbol{Y} - \boldsymbol{Z}_1\widehat{\boldsymbol{\beta}}_{(1)}) - (\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}) \\ &= n(\widehat{\boldsymbol{\Sigma}}_1 - \widehat{\boldsymbol{\Sigma}}), \end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_{(1)} = (\boldsymbol{Z}_1^\top \boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1\boldsymbol{Y}$ and $\widehat{\boldsymbol{\Sigma}}_1 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{Z}_1\widehat{\boldsymbol{\beta}}_{(1)})^\top(\boldsymbol{Y} - \boldsymbol{Z}_1\widehat{\boldsymbol{\beta}}_{(1)})$.
$\Rightarrow$

$$\Lambda = \frac{\max\limits_{\boldsymbol{\beta}_{(1)},\boldsymbol{\Sigma}} L(\boldsymbol{\beta}_{(1)}, \boldsymbol{\Sigma})}{\max\limits_{\boldsymbol{\beta},\boldsymbol{\Sigma}} L(\boldsymbol{\beta}, \boldsymbol{\Sigma})} = \frac{L(\widehat{\boldsymbol{\beta}}_{(1)}, \widehat{\boldsymbol{\Sigma}}_1)}{L(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})} = \left(\frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_1|}\right)^{\frac{n}{2}}$$

Result $\boldsymbol{Z}$; full column rank
LRT of $H_0 \ : \ \boldsymbol{\beta}_{(2)} = \boldsymbol{0}$ is to reject H_0 for large values of

$$-2\ln\Lambda = -n\ln\left(\frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_1|}\right) = -n\ln\frac{|n\widehat{\boldsymbol{\Sigma}}|}{|n\widehat{\boldsymbol{\Sigma}} + n(\widehat{\boldsymbol{\Sigma}}_1 - \widehat{\boldsymbol{\Sigma}})|}.$$

For n large,

$$-\left[n - r - 1 - \frac{1}{2}(m - r + q + 1)\right]\ln\left(\frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_1|}\right) \sim \chi^2_{m(r-q)}.$$

ex) See Example 7.9.

6.6.3 Predictions from Multivariate Multiple regressions

$$\begin{aligned} \boldsymbol{z}_0 &= (1 \ z_{01} \ z_{02} \ \cdots \ z_{0r})^\top \ ; \ \text{new fixed values} \\ \underset{n\times m}{\boldsymbol{Y}} &= \underset{n\times(r+1)}{\boldsymbol{Z}}\underset{(r+1)\times m}{\boldsymbol{\beta}} + \underset{n\times m}{\boldsymbol{\varepsilon}}, \ \text{with normal errors } \boldsymbol{\varepsilon} \end{aligned}$$

$\Rightarrow$
We know that

$$\begin{aligned} i) &\quad \underset{(r+1)\times m}{\widehat{\boldsymbol{\beta}}}{}^\top \underset{(r+1)\times 1}{\boldsymbol{z}_0} \sim N_m(\boldsymbol{\beta}^\top\boldsymbol{z}, \ \boldsymbol{z}_0^\top \ (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{z}_0\boldsymbol{\Sigma}) \\ ii) &\quad n\widehat{\boldsymbol{\Sigma}} \ \sim \ W_{m,n-r-1}(\boldsymbol{\Sigma}) \text{ and} \\ iii) &\quad \text{They are independent.} \end{aligned}$$

i) Predict the mean responses at $\boldsymbol{z}_0$

$$\widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}_0 \ : \ \text{Unbiased Estimator of } \boldsymbol{\beta}^\top \boldsymbol{z}_0$$

$$T^2 = \left(\frac{\widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}_0 - \boldsymbol{\beta}^\top \boldsymbol{z}_0}{\sqrt{\boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0}} \right)^\top \left(\frac{n}{n-r-1} \widehat{\boldsymbol{\Sigma}} \right)^{-1} \left(\frac{\widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}_0 - \boldsymbol{\beta}^\top \boldsymbol{z}_0}{\sqrt{\boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0}} \right)$$

& $100(1-\alpha)\%$ confidence ellipsoid for $\boldsymbol{\beta}^\top \boldsymbol{z}_0$ is

$$\begin{aligned} & \left(\boldsymbol{\beta}^\top \boldsymbol{z}_0 - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}_0 \right)^\top \left(\frac{n}{n-r-1} \widehat{\boldsymbol{\Sigma}} \right)^{-1} \left(\boldsymbol{\beta}^\top \boldsymbol{z}_0 - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}_0 \right) \\ \leq \quad & \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0 \left(\frac{m(n-r-1)}{n-r-m} \right) F_{m,n-r-m}(\alpha) \end{aligned}$$

$\therefore$ $100(1-\alpha)\%$ simultaneous CIs for $Ey_i = \boldsymbol{z}_0^\top \boldsymbol{\beta}_{(i)}$

$$\boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(i)} \pm \sqrt{\frac{m(n-r-1)}{n-r-m} F_{m,n-r-m}(\alpha)} \sqrt{\boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0 \left(\frac{n}{n-r-1} \widehat{\sigma}_{ii} \right)}$$

ii) Forecasting new responses $\boldsymbol{y_0} = \boldsymbol{\beta}^\top \boldsymbol{z_0} + \boldsymbol{\varepsilon_0}$ at $\boldsymbol{z}_0$
Note that $\boldsymbol{\varepsilon}_0 \perp\!\!\!\perp \boldsymbol{\varepsilon}$

$$\boldsymbol{y}_0 - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}_0 = (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{z}_0 + \boldsymbol{\varepsilon}_0 \sim N_m \big(\mathbb{0}, \ (1 + \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0 \boldsymbol{\Sigma} \big)$$

$100(1-\alpha)\%$ prediction ellipsoid for $\boldsymbol{y}_0$;

$$\begin{aligned} & (\boldsymbol{y}_0 - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}_0)^\top \left(\frac{n}{n-r-1} \boldsymbol{\Sigma} \right)^{-1} (\boldsymbol{y}_0 - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}_0) \\ \leq \ & (1 + \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0) \frac{m(n-r-1)}{n-r-m} F_{m,n-r-m}(\alpha) \end{aligned}$$

$100(1-\alpha)\%$ simultaneous prediction intervals for $\boldsymbol{y}_{0i}$

$$\boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_{(i)} \pm \sqrt{\frac{m(n-r-1)}{n-r-m} F_{m,n-r-m}(\alpha)} \sqrt{(1 + \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0) \left(\frac{n}{n-r-1} \right) \widehat{\sigma}_{ii}}$$

ex) See example 7.10 pg.400

Prediction ellipse is larger than confidence ellipse

Chapter 7

Principal Components Analysis

(*) Recall that maximization of Quadratic forms.

$$\begin{aligned} \boldsymbol{B}_{p\times p} \quad &: \quad \text{p.d. matrix with eigenvalues } \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0 \\ &\quad\ \ \&\text{ associated orthonormalized eigenvectors } \boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_p. \end{aligned}$$

$\Rightarrow$

$$\max_{\boldsymbol{x}\neq \mathbb{0}} \frac{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \lambda_1 \quad (\text{attained when } \boldsymbol{x} = \boldsymbol{e}_1)$$

$$\min_{\boldsymbol{x}\neq \mathbb{0}} \frac{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \lambda_p \quad (\text{attained when } \boldsymbol{x} = \boldsymbol{e}_p)$$

Moreover,

$$\max_{\boldsymbol{x}\perp\boldsymbol{e}_1,\cdots,\boldsymbol{e}_k} \frac{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \lambda_{k+1} \quad (\text{attained when } \boldsymbol{x} = \boldsymbol{e}_{k+1},\ k = 1, \cdots, p-1).$$

Goal:

1) From the data matrix, we want to determine a new set of variables that are uncorrelated.

2) These new variables give us another look at the data that may reveal unusual observations (outliers) or patterns in the data.

3) We may choose to use only a small number of these new variables in our analysis.

4) In regression analysis, this may reveal collinearities in the predictors. We may reduce the number of predictors needed.

7.1 Population principal components

Suppose you have a pop. on which you can observe the vector $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ & covariance matrix Σ.

Let $y_1 = \boldsymbol{a}_1^\top \boldsymbol{x} = \sum_{i=1}^p a_{i1} x_i = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p$,
then

$$Ey_1 = \boldsymbol{a}_1^\top \boldsymbol{\mu} \ \& \ Vy_1 = \boldsymbol{a}_1^\top \Sigma \boldsymbol{a}_1$$

Problem: Determine $\boldsymbol{a}_1$ to maximize the variance of y_1 subject to $\boldsymbol{a}_1^\top \boldsymbol{a}_1 = 1$

Mathematical Statement: $\max \boldsymbol{a}_1^\top \Sigma \boldsymbol{a}_1$ subject to $\boldsymbol{a}_1^\top \boldsymbol{a}_1 = 1$.

From (*) with $\boldsymbol{B} = \Sigma$, $\max \boldsymbol{a}_1^\top \sum \boldsymbol{a}_1 = \lambda_1$ & the maximum is attained when $\boldsymbol{a}_1 = \boldsymbol{e}_1$.

Note that y_1 is called the “first principal component”.

Our next problem is to determine the linear combination, y_2, defined by the vector $\boldsymbol{a}_2$, such that it has maximum variance among all variables that are uncorrelated with y_1 & is normalized to have length 1.

Mathematical statement: max $\boldsymbol{a}_2^\top \Sigma \boldsymbol{a}_2$ subject to $\boldsymbol{a}_2^\top \boldsymbol{a}_2 = 1$ & $\boldsymbol{a}_1^\top \Sigma \boldsymbol{a}_2 = 0$.

From (*) again with $\boldsymbol{B} = \Sigma$,

$$\max \boldsymbol{a}_2^\top \Sigma \boldsymbol{a}_2 = \lambda_2 \ \&$$

the maximum is attained when $\boldsymbol{a}_2 = \boldsymbol{e}_2$.

Note

$$i) \ Vy_2 = \boldsymbol{a}_2^\top \Sigma \boldsymbol{a}_2$$

$$ii)\ Cov(y_1, y_2) \quad = \quad Cov(\boldsymbol{a}_1^\top \boldsymbol{x}, \boldsymbol{a}_2^\top \boldsymbol{x}) = \boldsymbol{a}_1^\top \Sigma \boldsymbol{a}_2.$$

Similarly, we define $y_3, y_4, \cdots, y_p$ so that y_j is uncorrelated with the first $(j-1)$ variables & has maximum variance among such variables. These new variables are defined by the remaining eigenvectors with variances given by the eigenvalues.

ex)

$$\boldsymbol{x}^\top = (x_1\ x_2\ x_3) \text{ with } \Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

It may be verified that

$$\begin{aligned} \lambda_1 = 5.83 & \qquad \boldsymbol{e}_1^\top = [0.383\ \ -0.924\ 0] \\ \lambda_2 = 2 & \qquad \boldsymbol{e}_2^\top = [0\ 0\ 1] \\ \lambda_3 = 0.17 & \qquad \boldsymbol{e}_3^\top = [0.924\ 0.383\ 0] \end{aligned}$$

$\therefore$ the principal components are

$$\begin{aligned} y_1 &= \boldsymbol{e}_1^\top \boldsymbol{x} = 0.383x_1 - 0.924x_2 \\ y_2 &= \boldsymbol{e}_2^\top \boldsymbol{x} = x_3 \\ y_3 &= \boldsymbol{e}_3^\top \boldsymbol{x} = 0.924x_1 + 0.393x_2 \end{aligned}$$

For example,

$$\begin{aligned} Vy_1 &= V(0.383x_1 - 0.924x_2) \\ &= (0.383)^2 Vx_1 + (-0.924)^2 Vx_2 \\ &\quad +2(0.383)(-0.924)Cov(x_1, x_2) \\ &= 0.147(1) + 0.854(5) - 0.708(-2) \\ &= 5.83 = \lambda_1 \end{aligned}$$

$$\begin{aligned} Cov(y_1, y_2) &= cov(0.383x_1 - 0.924x_2,\ x_3) \\ &= 0.383Cov(x_1, x_3) - 0.924Cov(x_2, x_3) \\ &= 0.383(0) - 0.924(0) = 0 \end{aligned}$$

Note that we may write

$$\Sigma = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^\top$$

by spectral decomposition.

Thus, for example, using only λ_1, the largest eigenvalue, we get an approximation for Σ. Including successive terms improves the approximation and we can assess the value of including additional terms.

In summary, let $\boldsymbol{x}^\top = (x_1\ x_2\ \cdots\ x_p)$ have covariance matrix Σ with eigenvalue-eigenvector pairs $(\lambda_1, \boldsymbol{e}_1),\ (\lambda_2, \boldsymbol{e}_2),\ \cdots,\ (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Let $y_1 = \boldsymbol{e}_1^\top \boldsymbol{x},\ y_2 = \boldsymbol{e}_2^\top \boldsymbol{x}, \cdots,\ y_p = \boldsymbol{e}_p^\top \boldsymbol{x}$ be the principal components. Then

the total variance is given by $\text{tr}(\Sigma) = \sum_{i=1}^{p} \sigma_{ii}$ & $\text{tr}(\sum) = \sum_{i=1}^{p} \lambda_i$.

$(\because)$ By spectral decomposition,

$$\Sigma \boldsymbol{P} = \boldsymbol{P}\Lambda, \qquad \text{where } \boldsymbol{P} = (\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_p) \ \&\ \Lambda = diag(\lambda_i).$$

$$\begin{aligned} \text{tr}(\Sigma) &= \text{tr}(\boldsymbol{P}\Lambda\boldsymbol{P}^\top) = \text{tr}(\Lambda\boldsymbol{P}^\top\boldsymbol{P}) \\ &= \text{tr}(\Lambda) = \sum_{i=1}^{p} \lambda_i \end{aligned}$$

Hence the ratio $\dfrac{\lambda_i}{\text{tr}(\Sigma)} = \dfrac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$ gives the proportion of the total variance accounted for by y_i, i^{th} principal component.

Furthermore, let $\boldsymbol{y}^\top = (y_1\ y_2\ \cdots\ y_p)$ then

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} \boldsymbol{e}_1^\top \boldsymbol{x} \\ \boldsymbol{e}_2^\top \boldsymbol{x} \\ \vdots \\ \boldsymbol{e}_p^\top \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{e}_1^\top \\ \boldsymbol{e}_2^\top \\ \vdots \\ \boldsymbol{e}_p^\top \end{pmatrix} \boldsymbol{x} = \boldsymbol{P}^\top \boldsymbol{x},$$

where $\boldsymbol{P} = (\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_p)$

Note that

$$\begin{aligned}
V\boldsymbol{y} &= V\boldsymbol{P}^\top\boldsymbol{x} = \boldsymbol{P}^\top(V\boldsymbol{x})\boldsymbol{P} = \boldsymbol{P}^\top\Sigma\boldsymbol{P} \\
&= \boldsymbol{P}^\top\boldsymbol{P}\Lambda\boldsymbol{P}^\top\boldsymbol{P} = \Lambda \ \& \\
Cov(\boldsymbol{x},\boldsymbol{y}) &= Cov(\boldsymbol{x},\boldsymbol{P}^\top\boldsymbol{x}) = (V\boldsymbol{x})\boldsymbol{P} = \Sigma\boldsymbol{P} \\
&= \boldsymbol{P}\Lambda
\end{aligned}$$

Recall that

$$\boldsymbol{P} = (\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_p) = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ \vdots & \vdots & & \vdots \\ e_{p1} & e_{p2} & \cdots & e_{pp} \end{pmatrix}.$$

So

$$\begin{aligned}
Cov(x_i,y_j) &= (P\Lambda)_{ij} \\
&= e_{ij}\lambda_j \ \& \\
corr(x_i,y_j) &= \frac{cov(x_i,y_j)}{\sqrt{Vx_i}\sqrt{Vy_j}} = \frac{e_{ij}\lambda_j}{\sqrt{\sigma_{ii}}\sqrt{\lambda_j}} \\
&= \frac{e_{ij}\sqrt{\lambda_j}}{\sqrt{\sigma_{ii}}}.
\end{aligned}$$

ex) continued
Note that

$$\begin{aligned}
\sigma_{11}+\sigma_{22}+\sigma_{33} = 1+5+2 &= \lambda_1+\lambda_2+\lambda_3 \\
&= 5.83+2+0.17.
\end{aligned}$$

For example, the proportion of total variance explained by the first principal component is

$$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = \frac{5.83}{8}.$$

$$\begin{aligned}
corr(x_1,y_1) &= \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{0.383\sqrt{5.83}}{\sqrt{1}} = 0.925 \\
corr(x_2,y_1) &= \frac{e_{21}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-0.924\sqrt{5.83}}{\sqrt{5}} = -0.998.
\end{aligned}$$

Similarly

$$\begin{aligned} corr(x_1, y_2) &= corr(x_2, y_2) = 0 \ \& \\ corr(x_3, y_2) &= \frac{e_{32}\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}} = \frac{\sqrt{2}}{\sqrt{2}} = 1 \end{aligned}$$

<u>Procedure of PCA</u> (Principal Components Analysis)

Suppose λ_1 is very large compared to $\lambda_2 + \cdots + \lambda_p$. Then $\mathrm{tr}(\Sigma) \approx \lambda_1$. So, one can explain Σ by use of λ_1 approximately.

That is, most of information in $\boldsymbol{x}$ can be represented by $y_1 =$ the first principal component $= PC_1$.

When $\mathrm{tr}(\Sigma) \gg \lambda_1$, consider $\lambda_1 + \lambda_2$ and do similarly.

Keep doing this until m such that

$$\mathrm{tr}(\Sigma) \approx \lambda_1 + \cdots + \lambda_m.$$

Then one can think that $(PC_1, \cdots, PC_m)$ has most of information in $\boldsymbol{x}$.

If $\dfrac{\lambda_1 + \cdots + \lambda_m}{tr(\Sigma)} = 0.9$, one can think $(PC_1, \cdots, PC_m)$ has 90% of the information in $\boldsymbol{x}$.

<u>PCA with the correlation matrix</u>

(**Warning**!!!) The result of PCA with the covariance matrix is different from that with the correlation matrix.

ex)

$$\begin{aligned} \text{Suppose} \quad \Sigma &= \begin{pmatrix} 1 & 4 \\ 4 & 400 \end{pmatrix}, \\ \text{then} \quad \lambda_1 &= 400.04, \ \boldsymbol{e}_1^\top = (0.01, \ 0.999) \ \& \\ \lambda_2 &= 0.96, \ \boldsymbol{e}_2^\top = (-0.999, \ 0.01). \end{aligned}$$

Hence, the first principal component dominates significantly the second one.

The corresponding correlation matrix is

$$\rho = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$$

the corresponding eigenvalue & eigenvectors are

$$\begin{aligned} \lambda_1 &= 1.2, \ \boldsymbol{e}_1^\top = (1/\sqrt{2}, \ 1/\sqrt{2}) \\ \lambda_2 &= 0.8, \ \boldsymbol{e}_2^\top = (-1/\sqrt{2}, \ 1/\sqrt{2}). \end{aligned}$$

The first principal component does not dominate the second one significantly. Also, the eigenvectors are different.

Which one (covariance matrix or correlation matrix) should one use in practice?

- The fact that PCA based on the covariance matrix is different from that based on the correlation matrix implies that PCA is "not scale invariant". So, when using PCA, one should be careful for the scale of the data.

- One of the easiest ways of avoiding the scale problem is to standardize the data, i.e., use the correlation matrix.

- However, using the corr. matrix makes inference hard.

- But, inference is not of the main interest of PCA !!!

Remark

PCA with the correlation matrix is equivalent to PCA obtained from standardized variables.

$(\because)$
Let $z_i = \dfrac{x_i - \mu_i}{\sqrt{\sigma_{ii}}}, i = 1, \cdots, p$, then $\boldsymbol{z} = (\boldsymbol{V}^{1/2})^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$, where $\boldsymbol{V}^{1/2} = diag(\sqrt{\sigma_{ii}})$. Hence $cov(\boldsymbol{z}) = (\boldsymbol{V}^{1/2})^{-1}\Sigma(\boldsymbol{V}^{1/2})^{-1} = \boldsymbol{\rho}$, which is the correlation matrix. □

<u>The number of principal components</u>

There is no definite answer.
Standard approaches are

- Make the cumulative percentage of variance $\sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{p} \lambda_i$ larger than some numbers (eg. 80 %).

- When using the correlation matrix, choose principal components whose eigenvalues are greater than 1 (or 0.7).

- When using the covariance matrix, choose principal components whose eigenvalues are greater than 1 (or 0.7) times the mean of the eigenvalues.

A graphical approach : Scree plot which is a plot of λ_i vs i.

7.2 Sample principal components analysis

In practice, we do not know the parameters $\boldsymbol{\mu}$ & Σ but have estimates, $\bar{\boldsymbol{x}}$ & $\boldsymbol{S}$ based on the data matrix $\boldsymbol{X}$. These quantities are used to provide estimates of the principal components.

$$\begin{aligned} \widehat{y}_i &= \widehat{\boldsymbol{a}}_i^\top \boldsymbol{x} \quad , \quad i = 1, \cdots, p \\ &= \widehat{\boldsymbol{e}}_i^\top \boldsymbol{x} \end{aligned}$$

<u>Note</u>

① $\boldsymbol{S}_{p\times p}$ is the sample covariance matrix with eigenvalue-eigenvector pairs $(\widehat{\lambda}_1, \widehat{\boldsymbol{e}}_1), \cdots, (\widehat{\lambda}_p, \widehat{\boldsymbol{e}}_p)$, where $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_p > 0$.

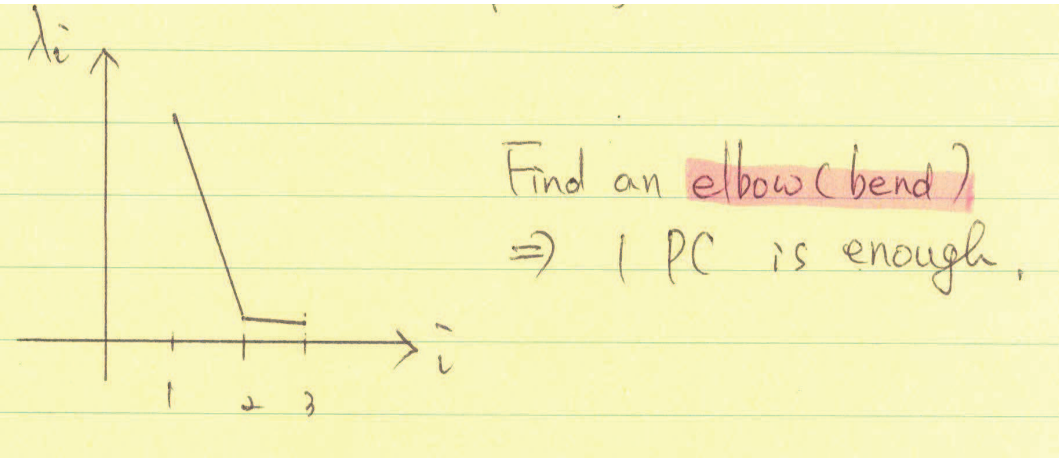


Figure 7.1.1: Scree plot

② Sample variance $(\widehat{y}_i) = \widehat{\lambda}_i, \; i = 1, \cdots, p$
Sample covariance $(\widehat{y}_i, \widehat{y}_j) = 0, \; i \neq j$
Total sample variance $= \sum_{i=1}^{p} S_{ii} = \widehat{\lambda}_1 + \widehat{\lambda}_2 + \cdots + \widehat{\lambda}_p$
Sample correlation $(x_i, \widehat{y}_j) = \dfrac{\widehat{e}_{ij}\sqrt{\widehat{\lambda}_j}}{\sqrt{S_{ii}}}, \; i, j = 1, \cdots, p$

③ Basically everything is the same as the population case except population quantities are replaced by corresponding sample quantities.

④ Because we have samples of size n,

$$\widehat{y}_{ji} = \widehat{\boldsymbol{e}}_i^\top \boldsymbol{x}_j, \; i = 1, \cdots, p, \; j = 1, \cdots, n.$$

These are called "principal component scores". Will be used later.

Procedures of PCA

The same as population case expcept population quantities are replaced by corresponding sample quantities.

Principal components regression analysis

① Standardize the independent & dependent variables.

② Get principal components scores from sample correlation matrix.

③ Regress the standardized dependent variable on some principal components scores.

④ Back transform to original variables.

Remark

Data matrix $\boldsymbol{X} = (x_{ij}), i = 1, \cdots, n, j = 1, \cdots, p$

Sample covariance matrix $\boldsymbol{S} = (s_{ij}), i, j = 1, \cdots, p$, where

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j}).$$

Let

$$z_{ki} = \frac{x_{ki} - \bar{x_i}}{s_{ii}},$$

i.e. standardized variable, then

$$\bar{z_i} = \frac{1}{n} \sum_{k=1}^{n} z_{ki} = 0$$

and the sample covariance of standardized variable is

$$\begin{aligned}
s_{z_{ij}} &= \frac{1}{n-1}\sum_{k=1}^{n}(z_{ki}-\bar{z}_i)(z_{kj}-\bar{z}_j) \\
&= \frac{1}{n-1}\sum_{k=1}^{n}\left(\frac{x_{ki}-\bar{x}_i}{\sqrt{s_{ii}}}\right)\left(\frac{x_{kj}-\bar{x}_j}{\sqrt{s_{jj}}}\right) \\
&= \frac{\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{s_{ii}}\sqrt{s_{jj}}} \\
&= \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} = r_{ij}
\end{aligned}$$

which is the sample correlation coefficient of i^{th} column of $\boldsymbol{X}$ and j^{th} column of $\boldsymbol{X}$.

Summary The sample covariance of standardized variable is the sample correlation coefficient of original variable.

7.3 Principal Component Analysis in R

7.3.1 Sample Principal Component Analysis in R

First, read the data and then obtain covariance and correlation matrices of sweat data.

```
> # data read
> sweat <- read.table("T5-1.dat")
> sweat <- as.data.frame(sweat) # to give the name
> names(sweat) <- c("Sweat","Sodium","Potassium")
>
> # covariance and correlation
> s.cov <- cov(sweat)
> s.cor <- cor(sweat)
> s.cov
               Sweat   Sodium Potassium
Sweat       2.879368  10.0100 -1.809053
Sodium     10.010000 199.7884 -5.640000
Potassium -1.809053  -5.6400  3.627658
```

Individual	sweat rate	sodium	potassium
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14.0
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5	58.8	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Table 7.3.1: Sweat Data

```
> s.cor
               Sweat     Sodium  Potassium
Sweat      1.0000000  0.4173499 -0.5597440
Sodium     0.4173499  1.0000000 -0.2094984
Potassium -0.5597440 -0.2094984  1.0000000
```

Second, obtain eigenvalues and eigenvectors.

```
> # eigenvalues and eigenvectors
> eigen.cov <- eigen(s.cov)
> eigen.cor <- eigen(s.cor)
> eigen.cov
$values
[1] 200.462464   4.531591   1.301392
```

```

$vectors
            [,1]        [,2]        [,3]
[1,] -0.05084144 -0.57370364  0.81748351
[2,] -0.99828352  0.05302042 -0.02487655
[3,]  0.02907156  0.81734508  0.57541452

> eigen.cor
$values
[1] 1.8078076 0.8009595 0.3912329

$vectors
           [,1]       [,2]       [,3]
[1,] -0.6532613 -0.1017681  0.7502619
[2,] -0.4876988  0.8145266 -0.3141596
[3,]  0.5791369  0.5711301  0.5817309
```

Third, try pca for covariance matrix. `prcomp` returns

- `sdev` the standard deviations of the principal components (i.e., the square roots of the eigenvalues of the covariance/correlation matrix)
- `rotation` a matrix whose columns contain the eigenvectors (Note that the difference between this and eigenvectors appeared before. There are sign differences.)
- `x` if `retx` is true the value of the rotated data (the centred (and scaled if requested) data multiplied by the rotation matrix) is returned.

```
> # pca (covariance)
> # sdev ; the sd's of the pc (the square roots of the eigenvalues)
> # rotation ; a matrix whose columns contain the eigenvectors
> pca.cov <- prcomp(sweat) # default is covariance matrix
> pca.cov
Standard deviations:
[1] 14.158477  2.128753  1.140786

Rotation:
                  PC1         PC2         PC3
```

```
Sweat      0.05084144 -0.57370364 -0.81748351
Sodium     0.99828352  0.05302042  0.02487655
Potassium -0.02907156  0.81734508 -0.57541452
> summary(pca.cov)
Importance of components:
                           PC1     PC2     PC3
Standard deviation     14.1585 2.12875 1.14079
Proportion of Variance  0.9717 0.02197 0.00631
Cumulative Proportion   0.9717 0.99369 1.00000
```

Similarly we obtain the pca results for correlation matrix.

```
> # pca (correlation)
> pca.cor <- prcomp(sweat, center = T ,scale = T)
> pca.cor
Standard deviations:
[1] 1.3445474 0.8949634 0.6254861

Rotation:
                 PC1        PC2        PC3
Sweat      0.6532613 -0.1017681 -0.7502619
Sodium     0.4876988  0.8145266  0.3141596
Potassium -0.5791369  0.5711301 -0.5817309
> summary(pca.cor)
Importance of components:
                          PC1    PC2    PC3
Standard deviation     1.3445 0.8950 0.6255
Proportion of Variance 0.6026 0.2670 0.1304
Cumulative Proportion  0.6026 0.8696 1.0000
> screeplot(pca.cor,type="lines")
```

Figure 7.3.1: A scree plot for the sweat data.

Finally the principal scores are obtained using two different methods. Only the first line among $n = 20$ lines is displayed to save space. The third line is used for double checking.

```
> # pc scores (2 methods)
> pca.cor$x[1,]
        PC1         PC2         PC3
-0.05271564  0.03560838  0.68762676
> predict(pca.cor)[1,]
        PC1         PC2         PC3
-0.05271564  0.03560838  0.68762676
> (pca.cor$x == predict(pca.cor))[1,]
 PC1  PC2  PC3
TRUE TRUE TRUE
```

7.3.2 Principal Component Regression analysis in R

Principal components regression analysis

① Standardize the independent & dependent variables.

② Get principal components scores from sample correlation matrix.

③ Regress the standardized dependent variable on some principal components scores.

④ Back transform to original variables.

There are 5 functions to do principal component analysis in R. We will use `prcomp()` mainly.

- prcomp() (stats)
- princomp() (stats)
- PCA() (FactoMineR)
- dudi.pca() (ade4)
- acp() (amap)

First, we read a data file from a website

https://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test.

The data set includes 103 data points. There are 7 input variables, and 3 output variables in the data set. However we will use only one output variable, “Y = ” Compressive Strength.

```
> rm(list = ls())
> par(mfrow=c(2,2))
> #https://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test
> #After saving the data file as "Cement.csv", read it.
> cement <- read.csv("Cement.csv", header=T)[,-1]
> cement <- data.frame(X1=cement[,1],X2=cement[,2],X3=cement[,3],
+                      X4=cement[,4],X5=cement[,5],X6=cement[,6],
+                      X7=cement[,7],Y=cement[,8])
>
> head(cement) # y is dependent var. & X1~X7 indep vars
   X1  X2  X3  X4 X5  X6  X7     Y
1 273  82 105 210  9 904 680 34.99
2 163 149 191 180 12 843 746 41.14
3 162 148 191 179 16 840 743 41.81
4 162 148 190 179 19 838 741 42.08
5 154 112 144 220 10 923 658 26.82
6 147  89 115 202  9 860 829 25.21
```

Step ① : Standardize the independent & dependent variables

```
> y.mean <- apply(cement[8], 2, mean)
> y.std <- apply(cement[8], 2, sd)
> scale.data <- as.data.frame(scale(cement))
> head(round(scale.data, 3))
      X1    X2     X3     X4    X5     X6     X7      Y
1  0.546 0.067 -0.515  0.635 0.164  0.227 -0.941 -0.134
2 -0.848 1.175  0.492 -0.850 1.232 -0.464  0.101  0.651
3 -0.861 1.158  0.492 -0.899 2.657 -0.498  0.054  0.736
4 -0.861 1.158  0.480 -0.899 3.726 -0.520  0.022  0.771
5 -0.962 0.563 -0.059  1.130 0.520  0.441 -1.288 -1.176
6 -1.051 0.182 -0.398  0.239 0.164 -0.271  1.411 -1.382
```

Step ② : Get principal components scores from sample correlation matrix.

```
> par(mfrow=c(1,1))
> data.pca <- prcomp(scale.data[1:7])
> plot(data.pca, type="l")  # pc1~4
> head(round(data.pca$x, 3)) # pca scores
        PC1   PC2    PC3    PC4    PC5    PC6    PC7
[1,] -0.266 0.049 -0.940 -0.743  0.470 -0.454 -0.026
[2,]  0.487 1.836  0.828  0.287  0.269  0.598 -0.029
[3,]  0.451 2.686  0.898  0.780  1.243  0.258 -0.014
[4,]  0.397 3.325  0.946  1.130  1.971 -0.017  0.015
[5,]  0.488 0.845  0.053 -1.462  0.352 -1.194 -0.061
[6,] -0.566 0.486  1.139  0.686 -0.894 -0.568 -0.005
> eigen$values
[1] 2.224568561 1.513425597 1.108659467 1.006115259 0.681279524
[6] 0.462898789 0.003052803
```

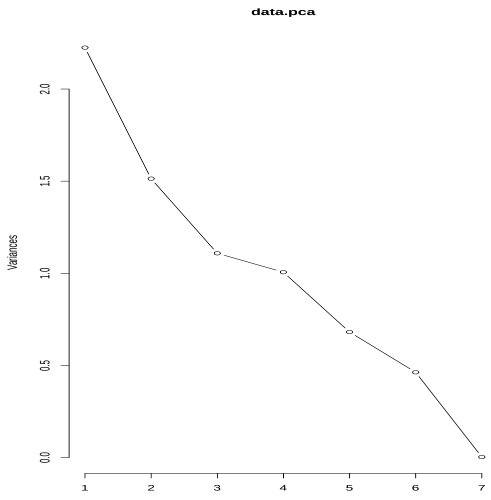


Figure 7.3.2: A scree plot for the cement data.

Four PCs will be used as independent variables based on a scree plot Figure 7.3.2, where 4 eigenvalues are greater than 1.

Step ③ : Regress the standardized dependent variable on some principal components scores.

```
> #make new data use pca score & scale y
> y <- scale(cement[,8])
> pca.score <- data.pca$x
> data <- cbind(as.data.frame(pca.score), y)
>
> #regression (pc1~pc4)
> data.pcr <- glm(y~ PC1 + PC2 + PC3 + PC4,  data = data)
> summary(data.pcr)

Call:
glm(formula = y ~ PC1 + PC2 + PC3 + PC4, data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.83044  -0.72372  -0.06759   0.70428   2.19425

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.708e-16  9.354e-02   0.000 1.000000
PC1          5.970e-02  6.302e-02   0.947 0.345811
PC2         -2.741e-01  7.641e-02  -3.588 0.000523 ***
PC3         -5.438e-02  8.927e-02  -0.609 0.543840
PC4          9.583e-02  9.371e-02   1.023 0.308978
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for gaussian family taken to be 0.9011689)

    Null deviance: 102.000  on 102  degrees of freedom
Residual deviance:  88.315  on  98  degrees of freedom
AIC: 288.46

Number of Fisher Scoring iterations: 2
```

To calculate R^2 and adj-R^2.

```
> # coefficient of determination in pcr
> one.vec <- rep(1,dim(pca.score)[1])
> x.mat <- cbind(one.vec,pca.score[,(1:4)])
> y.hat <- x.mat%*%as.vector(data.pcr$coefficients)
> head(y.hat, 3)
            [,1]
[1,] -0.04939851
[2,] -0.49166604
[3,] -0.68347490
> original.y.hat <- (y.hat * y.std) + y.mean
> head(original.y.hat, 3)
         [,1]
[1,] 35.65222
[2,] 32.18563
[3,] 30.68218
>
> r <-sum((original.y.hat - y.mean)ˆ2) / sum((cement[,8]
+ - y.mean)ˆ2)
> ad.r <- 1-((dim(x.mat)[1]-1)*(1-r) / (dim(x.mat)[1]
+ -dim(x.mat)[2]))
```

The other way around is to use a package called “pls” instead of calculating all minors.

```
> # use pcr code using correlation matrix
> install.packages("pls")
> library(pls)
> #scale.data <- as.data.frame(scale(cement))
> pcr <-pcr(Y˜ . ,data = scale.data, ncomp= 4)
> head(pcr$fitted.values, 3) # pcr all y hat
[1] -0.01589309  0.02904829  0.02690038
> ftn <- c(1:(103*3))
> pcr.yhat <- as.vector(pcr$fitted.values[-ftn]) # 4 comps
> head(pcr.yhat, 3)
[1] -0.04939851 -0.49166604 -0.68347490
```

Comparison can be done as follows:

```
> # coefficient of determination
> x.mat1 <- cbind(one.vec,pcr$scores[,(1:4)])
> original.pcr.y <- (pcr.yhat * y.std) + y.mean
> r2<-sum((original.pcr.y - y.mean)^2) / sum((cement[,8]
+ - y.mean)^2)
> ad.r2 <- 1-((dim(x.mat1)[1]-1)*(1-r2)/ (dim(x.mat1)[1]
+ -dim(x.mat1)[2]))
>
> # pca + glm y.hat & pcr y.hat
> head(round(y.hat, 5) == round(pcr.yhat, 5) , 5)
     [,1]
[1,] TRUE
[2,] TRUE
[3,] TRUE
[4,] TRUE
[5,] TRUE
> round(r, 5) == round(r2, 5)
[1] TRUE
> round(ad.r, 5) == round(ad.r2, 5)
[1] TRUE
> ad.r2
[1] 0.09883108
```

Adjusted-$R^2 = 0.0988$ is too low using 4 principal components, so we tried stepwise regression as follows:

```
> #general liner model + stepwise
> scale.data <- as.data.frame(scale(cement))
> lm.c <- glm(Y~ ., data = scale.data)
> #stepwise
> step.lm.c <-step(lm.c, direction="both")
> #Y ~ X1 + X2 + X3 + X4 + X6 + X7 (Chosen model)
> #summary(step.lm.c)
> # coefficient of determination in glm
> one.vec2 <- rep(1,dim(cement)[1])
> x.mat.glm <- as.matrix(cbind(one.vec2, scale.data[,-c(5,8)]))
> y.hat.glm <- x.mat.glm%*% step.lm.c$coefficients
> original.y.glm <- (y.hat.glm * y.std) + y.mean
>
> r3 <- sum((original.y.glm - y.mean)^2) / sum((cement[,8]
+ - y.mean)^2)
> ad.r3 <- 1-((dim(x.mat.glm)[1]-1)*(1-r3) / (dim(x.mat.glm)[1]
+ -dim(x.mat.glm)[2]))
> ad.r3
[1] 0.8897122
```

Since 6 independent variables were chosen, we decided to use 6 principal components.

```
> # use pcr code using correlation matrix (PC1 ~ PC6)
> scale.data <- as.data.frame(scale(cement))
> pcr2 <-pcr(Y~ . ,data = scale.data, ncomp= 6)
> head(pcr2$fitted.values,2) # pcr all y hat
[1] -0.01589309  0.02904829
> ftn2 <- c(1:(103*5))
> pcr.yhat2 <- as.vector(pcr2$fitted.values[-ftn2]) # 4 comps
> head(pcr.yhat2, 3)
[1] -0.1876975  0.2839852  0.4023519
>
> # coefficient of determination
> x.mat2 <- cbind(one.vec,pcr2$scores[,(1:6)])
> original.pcr.y2 <- (pcr.yhat2 * y.std) + y.mean
> pcr.r<-sum((original.pcr.y2 - y.mean)^2) / sum((cement[,8] -
+ y.mean)^2)
> pcr.ad.r <- 1-((dim(x.mat2)[1]-1)*(1-pcr.r)/ (dim(x.mat2)[1]
+ -dim(x.mat2)[2]))
> pcr.ad.r
[1] 0.8901821
```

Adjusted-R^2 for pcr ($= 0.8902$) is slightly higher than that of stepwise result ($= 0.8897$).

Be aware of that the small number of principal components is not always enough. For this example, we used 6 principal components from 7 independent variables.

However it will be always better to use the same number of principal components instead of original independent variables since principal components are uncorrelated.

Chapter 8

Factor Analysis

8.1 Introduction

Motivation example : Spearman(1904)

Data were taken on test scores of 33 students in various topics with the following correlation matrix.

Subjects	1	2	3	4	5	6
Classics	1.00					
French	0.83	1.00				
English	0.78	0.67	1.00			
Mathematrics	0.70	0.67	0.64	1.00		
Discrimination	0.66	0.65	0.54	0.45	1.00	
Music	0.63	0.57	0.51	0.51	0.40	1.00

You can see that correlation decrease systematically. Spearman conjectured this feature by arguing that there exists a certain hidden variable which affects all the observed variables (subjects).

That is, he conjectured that each student could be given a score on each topic consisting of a measure of general intelligence, F, &

a measure of ability in the specific topic, ε_i.

Thus, for each topic, the score would look like

$$Topic_i = l_i F + \varepsilon_i,\ i = 1, \cdots, 6.$$

That is,

$$\begin{aligned} \text{Classics} &= l_1 F + \varepsilon_1 \\ \text{French} &= l_2 F + \varepsilon_2 \\ &\vdots \\ \text{Music} &= l_6 F + \varepsilon_6. \end{aligned}$$

Here F is a random component common (common factor) to all topics, and ε_i is a random component specific to the topic (errors or specific factor). The l_i are parameters reflecting the importance of F in describing the response and is called the loading of the i^{th} response on the common factor (factor loading).

Further, he assumed that
1) F is an unobserved random variable with mean 0 and variance 1.
2) ε_i is an independent random variable with mean 0 and variance ψ_i.
3) F and ε_i are independent for all i.

The assumption that the means are zero is consistent with the fact that, in the correlation matrix, the data have been standardized.

Similarly, the assumption that the variance of F is 1 is not restrictive since any different assumption could be absorbed in the $l's$. The $\varepsilon's$ are allowed to have different variances. **The strong assumption is that the random variables are independent**.

Denoting the topic scores by X_i, we have

$$\begin{aligned} X_i &= l_i F + \varepsilon_i,\ i = 1, \cdots, 6,\ \text{where} \\ F &\sim (0, 1) \\ \varepsilon_i &\sim \text{independent}(0, \psi_i),\ \textit{and} \\ F &\ \&\ \varepsilon_i \text{ are independent.} \end{aligned}$$

Under this model,

$$\begin{aligned}
cov(X_i, X_j) &= E(X_i - EX_i)(X_j - EX_j) = EX_iX_j \\
&= E(l_iF + \varepsilon_i)(l_jF + \varepsilon_j) \\
&= l_il_jEF^2 + l_iEF\varepsilon_j + l_jE\varepsilon_iF + E\varepsilon_i\varepsilon_j \\
&= l_il_j, \\
VX_i = V(l_iF + \varepsilon_i) &= l_i^2VF + V\varepsilon_i = l_i^2 + \psi_i
\end{aligned}$$

Note: The factor F could be interpreted as general intelligence, for example IQ.

The model can be written in matrix form,

$$\begin{aligned}
\boldsymbol{X} &= \boldsymbol{l}F + \boldsymbol{\varepsilon}, \text{ where} \\
\boldsymbol{X} &= (x_1, \cdots, x_6)^\top, \ \boldsymbol{l} = (l_1, \cdots, l_6)^\top, \ and \\
\boldsymbol{\varepsilon} &= (\varepsilon_1, \cdots, \varepsilon_6)^\top.
\end{aligned}$$

The assumptions are
1) $F \sim (0, 1)$
2) $\boldsymbol{\varepsilon} \sim (0, \Psi)$, where Ψ =diag(ψ_i)
3) F & $\boldsymbol{\varepsilon}$ are independent.

8.2 Orthogonal Factor Model

We generalize Spearman's idea.

The observable random vector $\boldsymbol{X}_{p\times 1}$ has mean $\boldsymbol{\mu}$ and covariance Σ. Assume that $\boldsymbol{X}$ is linearly dependent upon a few unobservable random variables $F_1, \cdots, F_m$ (common factors) and p additional sources of variation $\varepsilon_1, \cdots, \varepsilon_p$ (errors or specific factors). That is,

$$\begin{aligned}
X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\
&\vdots \\
X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p
\end{aligned}$$

or, in matrix notation

$$\underset{p\times 1}{\boldsymbol{X}} - \underset{p\times 1}{\boldsymbol{\mu}} = \underset{p\times m}{\boldsymbol{L}}\ \underset{m\times 1}{\boldsymbol{F}} + \underset{p\times 1}{\boldsymbol{\varepsilon}},\ \text{where}$$

<u>**Note**</u> $m \le p$

$$\begin{aligned}
\boldsymbol{X} &= (X_1\ X_2\ \cdots\ X_p)^\top,\ \boldsymbol{\mu} = (\mu_1\ \mu_2\ \cdots\ \mu_p)^\top, \\
\boldsymbol{L} &= (l_{ij}),\ \boldsymbol{F} = (F_1\ F_2\ \cdots\ F_m)^\top,\ \& \\
\boldsymbol{\varepsilon} &= (\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_p)^\top.
\end{aligned}$$

l_{ij} are called factor loadings & $\boldsymbol{L}$ is the matrix of factor loadings.

We assume that

$$\begin{array}{ll}
1) & E\boldsymbol{F} = \mathbb{0},\ cov(\boldsymbol{F}) = I \text{ (orthogonal factors)} \\
2) & E\boldsymbol{\varepsilon} = \mathbb{0},\ cov(\boldsymbol{\varepsilon}) = \Psi,\ \text{where } \Psi = diag(\psi_i) \\
3) & \boldsymbol{F}\ \amalg\ \boldsymbol{\varepsilon}
\end{array}$$

<u>Remark</u> Allowing the factors $\boldsymbol{F}$ to be correlated so that cov($\boldsymbol{F}$) is not diagonal gives the "oblique factor" model.

<u>Covariance structure</u> for the orthogonal factor model.

$$\begin{aligned}
\Sigma = cov(\boldsymbol{X}) &= E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top \\
&= E(\boldsymbol{LF}+\boldsymbol{\varepsilon})(\boldsymbol{LF}+\boldsymbol{\varepsilon})^\top \\
&= \boldsymbol{L}(E\boldsymbol{FF}^\top)\boldsymbol{L}^\top + \boldsymbol{L}(E\boldsymbol{F\varepsilon}^\top) \\
&\quad +(E\boldsymbol{\varepsilon F}^\top)\boldsymbol{L}^\top + E\boldsymbol{\varepsilon\varepsilon}^\top \\
&= \boldsymbol{LL}^\top + \Psi \quad \text{by assumption.}
\end{aligned}$$

Hence,

$$\begin{aligned}
&VX_i = l_{i1}^2 + \cdots + l_{im}^2 + \psi_i - \circledast \text{ and} \\
&cov(X_i, X_j) = l_{i1}l_{j1} + \cdots + l_{im}l_{jm}.
\end{aligned}$$

$$\begin{aligned} cov(\boldsymbol{X}, \boldsymbol{F}) &= E(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{F}^\top = E(\boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon})\boldsymbol{F}^\top \\ &= \boldsymbol{L}(E\boldsymbol{F}\boldsymbol{F}^\top) + E\boldsymbol{\varepsilon}\boldsymbol{F}^\top \\ &= \boldsymbol{L}. \end{aligned}$$

That is,

$$cov(X_i, F_j) = l_{ij}.$$

From $\textcircled{*}$

$$\begin{aligned} VX_i = \sigma_{ii} &= \underbrace{l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2}_{\text{communality}} + \psi_i \qquad (\psi_i : \text{ specific variance or uniquness}) \\ &= h_i^2 + \psi_i. \end{aligned}$$

Remark : The i^{th} communality, h_i^2, is the sum of square of the loadings of the i^th variable on the m common factors.

ex) Verifying the relation $\Sigma = \boldsymbol{L}\boldsymbol{L}^\top + \Psi$ for two factors

$$\Sigma = \begin{pmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{pmatrix}$$

Then we can write

$$\Sigma = \boldsymbol{L}\boldsymbol{L}^\top + \Psi, \text{ where}$$
$$\boldsymbol{L} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix} \text{ and } \Psi = diag(2, 4, 1, 3).$$

For example,

$$\begin{aligned} & \sigma_{11} = l_{11}^2 + l_{12}^2 + \psi_1 \\ \Leftrightarrow \quad & \underbrace{19}_{\text{variance}} = \underbrace{4^2 + 1^2}_{\text{communality}} + \underbrace{2}_{\text{specific variance}} . \end{aligned}$$

Remark

1. Σ has $\dfrac{p(p+1)}{2}$ parameters whereas the m OFM(orthogonal factor model) has $p \times m + p + p$ parameters since $\Sigma_{p\times p} = L_{p\times m}L^\top + \Psi_{p\times p}$, where the last p parameters from $\boldsymbol{\mu}$.

2. When $m = p$, $\Sigma = \Sigma^{\frac{1}{2}}(\Sigma^{\frac{1}{2}})^\top + \mathbb{0}$, i.e. $\boldsymbol{L} = \Sigma^{\frac{1}{2}}$ & $\Psi = \mathbb{0}$.

3. When m is much smaller than p, we can reduce the number of parameters significantly using the OFM.

4. However, in some cases, the factor model does not exist. That is nonexistence of solution for $\Sigma = \boldsymbol{L}\boldsymbol{L}^\top + \Psi$.

5. When $m > 1$, some inherent ambiguity exists.
For any $m \times m$ orthogonal matrix $\boldsymbol{T}$ such that $\boldsymbol{T}^\top\boldsymbol{T} = \boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}$, we have

$$\begin{aligned}
\boldsymbol{X} - \boldsymbol{\mu} &= \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}^\top\boldsymbol{F} + \boldsymbol{\varepsilon} \\
&= \boldsymbol{L}^*\boldsymbol{F}^* + \boldsymbol{\varepsilon};\ \text{a new factor model} \\
\text{since } E\boldsymbol{F}^* &= \boldsymbol{T}^\top E\boldsymbol{F} = \mathbb{0}\ \& \\
cov(\boldsymbol{F}^*) &= \boldsymbol{T}^\top(cov\boldsymbol{F})\boldsymbol{T} = \boldsymbol{T}^\top\boldsymbol{T} = \boldsymbol{I}
\end{aligned}$$

$\therefore$ Factor loading $\boldsymbol{L}$ are determined only up to an orthogonal matrix $\boldsymbol{T}$.

Need “additional conditions” to make $\boldsymbol{L}$ be unique (identifiable).

Goal of factor analysis

- To describe the covariance relationships among many variables in terms of a few underlying but unobservable, random quantities called factors.

- That is, Factor analysis decomposes the structure of multivariate observations, which are not directly interpretable due to the correlation, into few independent variables, which have some practical meanings.

- PCA ; Find $y_i = \boldsymbol{a}_i^\top \boldsymbol{x},\ \boldsymbol{x}$: obsevable random vector & dimension reduction is primary goal.

- OFM ; $\boldsymbol{x} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon},\ \boldsymbol{F}$: unobservable random vector & $\boldsymbol{x}$: observable

- Regression Analysis ; $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{y}$ & $\boldsymbol{X}$ can be observed. Goal is to obtain a relationship between independent variables and dependent variables, which are both observables.

- That is the primary question in factor analysis is whether the data are consistent with a prescribed structure.

8.3 Method of Estimation

Three methods, principal component method, principal factor method, and maximum likelihood method, will be introduced.

8.3.1 Principal component method

Recall that

OFM

$$\underset{p\times1}{\boldsymbol{X}} - \underset{p\times1}{\boldsymbol{\mu}} = \underset{p\times m}{\boldsymbol{L}}\,\underset{m\times1}{\boldsymbol{F}} + \underset{p\times1}{\boldsymbol{\varepsilon}},\ \text{where}$$

$$\boldsymbol{F} \sim (\mathbb{0}, \boldsymbol{I}),\ \boldsymbol{\varepsilon} \sim (\mathbb{0}, \Psi),\ \Psi = \text{diag}(\psi_i),\ \&$$

$$\boldsymbol{F} \amalg \boldsymbol{\varepsilon}$$

Recall that

$$cov(\boldsymbol{X}) = \underset{p\times p}{\Sigma} = \underset{p\times m}{\boldsymbol{L}}\boldsymbol{L}^\top + \underset{p\times p}{\Psi}.$$

By spectral decomposition of Σ,

$$\begin{aligned}
\Sigma &= \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^\top + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^\top \\
&= \left[\sqrt{\lambda_1}\boldsymbol{e}_1 \ \sqrt{\lambda_2}\boldsymbol{e}_2 \ \cdots \ \sqrt{\lambda_p}\boldsymbol{e}_p\right] \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1^\top \\ \sqrt{\lambda_2}\boldsymbol{e}_2^\top \\ \vdots \\ \sqrt{\lambda_p}\boldsymbol{e}_p^\top \end{bmatrix} \\
&= \boldsymbol{L}\boldsymbol{L}^\top + \mathbb{0}, \text{ where} \\
\boldsymbol{L} &= \left[\sqrt{\lambda_1}\boldsymbol{e}_1 \ \sqrt{\lambda_2}\boldsymbol{e}_2 \ \cdots \ \sqrt{\lambda_p}\boldsymbol{e}_p\right].
\end{aligned}$$

By defining $\boldsymbol{L}$ as above, we obtain the solution, that is,

$$\begin{aligned}
l_{ij} &= \sqrt{\lambda_j} e_{ij}, \ i,j = 1,\cdots,p \quad \& \\
\psi_i &= 0, \quad \forall\, i = 1,\cdots,p.
\end{aligned}$$

Although this solution is exact, it is not particularly useful since we want just a few common factors.

One approach, when the last $p-m$ eigenvalues are small, is to neglect this contribution. So

$$\begin{aligned}
\Sigma &\cong \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^\top + \cdots + \lambda_m \boldsymbol{e}_m \boldsymbol{e}_m^\top \\
&= \left[\sqrt{\lambda_1}\boldsymbol{e}_1 \cdots \sqrt{\lambda_m}\boldsymbol{e}_m\right] \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1^\top \\ \vdots \\ \sqrt{\lambda_m}\boldsymbol{e}_m^\top \end{bmatrix} \\
&= \boldsymbol{L}\boldsymbol{L}^\top.
\end{aligned}$$

Further allowing for specific factors,

$$\begin{aligned}
\Sigma &\cong \boldsymbol{L}\boldsymbol{L}^\top + \Psi \\
&= \left[\sqrt{\lambda_1}\boldsymbol{e}_1 \ \cdots \ \sqrt{\lambda_m}\boldsymbol{e}_m\right]_{p\times m} \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1^\top \\ \vdots \\ \sqrt{\lambda_m}\boldsymbol{e}_m^\top \end{bmatrix}_{m\times p} + \begin{bmatrix} \psi_1 & \cdots & \phi \\ \vdots & \ddots & \vdots \\ \phi & \cdots & \psi_p \end{bmatrix}_{p\times p}, \\
&\quad \text{where } \psi_i = \sigma_{ii} - \sum_{j=1}^m l_{ij}^2 \text{ for } i = 1, \cdots, p.
\end{aligned}$$

Note that $l_{ij} = \sqrt{\lambda_j}e_{ij}, \ i = 1, \cdots, p, \ j = 1, \cdots, m$ and $\psi_i = \sigma_{ii} - \sum_{j=1}^m l_{ij}^2$.

Question: For given data $\boldsymbol{x}_1 \cdots \boldsymbol{x}_n$, how do we identify a (good) factor model?

Basic idea is (i) to use the sample covariance matrix $\boldsymbol{S}$ instead of Σ and (ii) to find the factor model which gives a good approximation of $\boldsymbol{S}$.

Often, work with standardized variables, *i.e.*, with the sample correlation matrix $\boldsymbol{R}$ instead of $\boldsymbol{S}$.

Let $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p > 0$ be the eigenvalues of $\boldsymbol{S}$ (or $\boldsymbol{R}$) & let $\widehat{\boldsymbol{e}}_1 \cdots \widehat{\boldsymbol{e}}_p$ be the corresponding eigenvectors.

Using the spectral decomposition,

$$\boldsymbol{S} = \widehat{\lambda}_1\widehat{\boldsymbol{e}}_1\widehat{\boldsymbol{e}}_1^\top + \cdots + \widehat{\lambda}_p\widehat{\boldsymbol{e}}_p\widehat{\boldsymbol{e}}_p^\top = \widehat{\boldsymbol{L}}_p\widehat{\boldsymbol{L}}_p^\top,$$

where

$$\widehat{L}_p = \left[\sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{e}}_1 \cdots \sqrt{\widehat{\lambda}_p}\boldsymbol{e}_p\right].$$

For a given m ($m \leq p$), let

$$\widehat{L}_m = \left[\sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{e}}_1 \cdots \sqrt{\widehat{\lambda}_m}\boldsymbol{e}_m\right] \equiv \left[\widehat{l}_{ij}\right]$$

and

$$\begin{aligned}
\widehat{\Psi} &= \text{diag}\left[\boldsymbol{S} - \widehat{\boldsymbol{L}}_m \widehat{\boldsymbol{L}}_m^\top\right], \textit{i.e.}, \\
\widehat{\psi}_i &= s_{ii} - \widehat{h}_i^2 \qquad \text{(specific variance)} \\
\widehat{h}_i^2 &= \widehat{l}_{i1}^2 + \cdots + \widehat{l}_{im}^2 \qquad \text{(communality)} \\
&= \sum_{j=1}^{m} \widehat{l}_{ij}^2 \qquad \text{for } i = 1, \cdots, p \;\&\; j = 1, \cdots, m.
\end{aligned}$$

<u>How to choose m?</u>

1. Subject similar to PCA.

2. Small values of off-diagonal element of $\boldsymbol{S} - (\widehat{\boldsymbol{L}}_m \widehat{\boldsymbol{L}}_m^\top + \widehat{\Psi})$ since $\Sigma = \boldsymbol{L}\boldsymbol{L}^\top + \Psi$.

3. Total sample variance $= tr(\boldsymbol{S}) = s_{11} + s_{22} + \cdots + s_{pp}$, where $s_{ii} = \widehat{h}_i^2 + \widehat{\psi}_i$, and $\widehat{h}_i^2 = \widehat{l}_{i1}^2 + \cdots + \widehat{l}_{im}^2$, $i = 1, \cdots, p$.

   $\Rightarrow \widehat{l}_{i1}^2$: contribution to s_{ii} from the first common factor.

   $\Rightarrow$ Contribution to $tr(\boldsymbol{S})$ from the first common factor is $\widehat{l}_{11}^2 + \widehat{l}_{21}^2 + \cdots + \widehat{l}_{p1}^2$.

   Note that $\widehat{l}_{11}^2 + \widehat{l}_{21}^2 + \cdots + \widehat{l}_{p1}^2 = \left(\sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{e}}_1\right)^\top \left(\sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{e}}_1\right) = \widehat{\lambda}_1 =$ sum of squares of the first column of $\widehat{\boldsymbol{L}}_m$.

   In general, the contribution of the j^{th} common factor to the total sample variance is

   $$\widehat{l}_{1j}^2 + \widehat{l}_{2j}^2 + \cdots + \widehat{l}_{pj}^2 = \left(\sqrt{\widehat{\lambda}_j}\widehat{\boldsymbol{e}}_j\right)^\top \left(\sqrt{\widehat{\lambda}_j}\widehat{\boldsymbol{e}}_j\right) = \widehat{\lambda}_j.$$

4. Use $m =$ number of $\widehat{\lambda}_j > 1$ if $\boldsymbol{R}$ is used since $tr(\boldsymbol{R}) = p$.

5. Proportion of total sample variance due to m common factors is

$$\begin{array}{ll} \dfrac{\widehat{\lambda}_1 + \cdots + \widehat{\lambda}_m}{tr(\boldsymbol{S})} & \text{with } \boldsymbol{S} \text{ or} \\ \dfrac{\widehat{\lambda}_1 + \cdots + \widehat{\lambda}_m}{tr(\boldsymbol{R}) = p} & \text{with } \boldsymbol{R}. \end{array}$$

6. Use scree plot, *i.e.* plot of $(j, \widehat{\lambda}_j),\ j = 1, \cdots, p$.

Example : Stock price data

- Data description
  - Weekly rates of return for five stocks - JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell, ExxonMobil.
  - January 2004 - December 2005
- Objective of Data analysis
  - Investigate common forces which affects the prices of the five.
- Two-Factor solution (with the correlation matrix) by PC method.

	One-factor solution		Two-factor solution		
Variable	Factor 1	Specific variances	Factor 1	Factor 2	Specific variances
1.JP Morgan	.732	.46	.732	-.437	.27
2.Citibank	.831	.31	.831	-.280	.23
3.Wells Fargo	.726	.47	.726	-.374	.33
4.Royal Dutch Shell	.605	.63	.605	.694	.15
5.ExxonMobil	.563	.68	.563	.719	.17
Cumulative proportion of total sample variance explained	.487		.487	.769	

$$\boldsymbol{R} - \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^\top - \widehat{\Psi} = \begin{pmatrix} 0 & -.099 & -.185 & -.025 & .056 \\ -.099 & 0 & -.134 & .014 & -.054 \\ -.185 & -.134 & 0 & .003 & .006 \\ -.025 & .014 & .003 & 0 & -.156 \\ .056 & -.054 & .006 & -.156 & 0 \end{pmatrix}$$

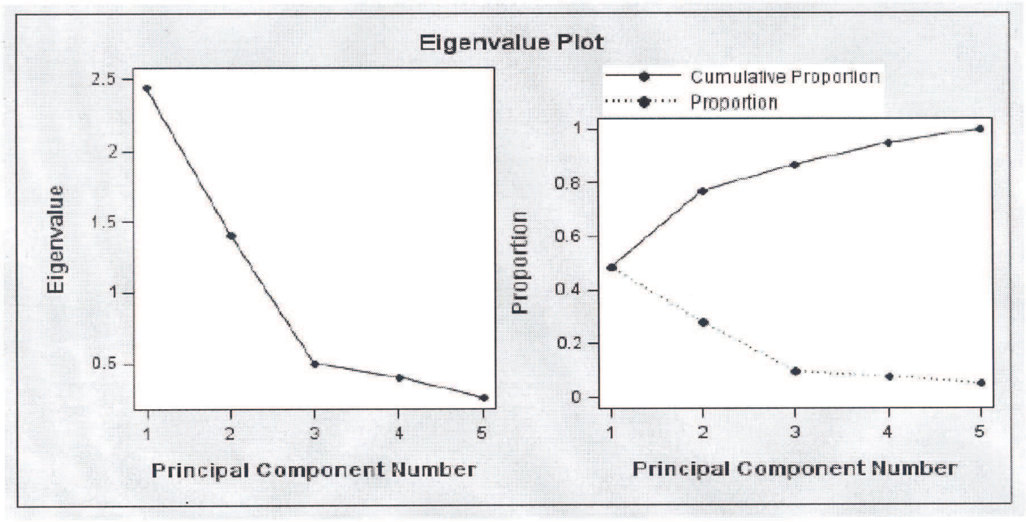


Figure 8.3.1: Scree plot: Visualization of the eigenvalues

8.3.2 Principal factor method

It is an iterative principal component method.

Johnson & Wichern recommend to use the principal component method & the maximum likelihood method!

8.3.3 Maximum likelihood method

Suppose that $\boldsymbol{F}_j$ & $\boldsymbol{\varepsilon}_j$ are jointly normal, then

$$\boldsymbol{X}_j - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F}_j + \boldsymbol{\varepsilon}_j \sim \text{independent } N_p(\mathbb{0}, \Sigma),\ j = 1, \cdots, n,$$

where $\Sigma = \boldsymbol{L}\boldsymbol{L}^\top + \Psi$.

That is,

$$\begin{aligned}&\boldsymbol{X}_j - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F}_j + \boldsymbol{\varepsilon}_j, \quad j = 1, \cdots, n\\ &\boldsymbol{F} \sim N(\mathbb{0}, \boldsymbol{I}),\ \boldsymbol{\varepsilon} \sim N(\mathbb{0}, \Psi),\ \Psi = \text{diag}(\psi_i)\ \&\\ &\boldsymbol{F} \amalg \boldsymbol{\varepsilon}.\end{aligned}$$

The likelihood of $\boldsymbol{\mu}$ & Σ is

$$L(\boldsymbol{\mu}, \Sigma) = \prod_{j=1}^{n} \left\{ \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{x}_j - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}) \right) \right\}$$

which depends on $\boldsymbol{L}$ & Ψ through $\Sigma = \boldsymbol{L}\boldsymbol{L}^\top + \Psi$.

We need a uniqueness condition that

$$\boldsymbol{L}^\top \Psi^{-1} \boldsymbol{L} = \Delta \text{ is a diagonal matrix.}$$

since $\boldsymbol{L}$ are determined only up to an orthogonal matrix $\boldsymbol{T}$.

Then use a numerical maximization of $L(\boldsymbol{\mu}, \Sigma)$.

<u>Remark</u>

- For a given data $\boldsymbol{x}_1 \cdots \boldsymbol{x}_n$, we may use $\boldsymbol{S}$ (or $\boldsymbol{R}$) instead of Σ (or $\boldsymbol{\rho}$)

- (ultra) Heywood case might occur, *i.e.*, Specific variance, $\widehat{\psi}_i = 0$ ($\widehat{\psi}_i < 0$). → See next example.

  However don′t worry about this. The software program obtains a feasible solution by slightly adjusting the loadings so that all specific variance estimates are nonnegative.

Example : Stock price data

- Factor solutions obtained by using maximum likelihood method and principal component analysis.

	Maximum likelihood			Principal components		
Variable	Factor 1	Factor 2	Specific variances	Factor 1	Factor 2	Specific variances
1. JP Morgan	.115	.755	.42	.732	-.437	.27
2. Citibank	.322	.788	.27	.831	-.280	.23
3. Wells Fargo	.182	.652	.54	.726	-.374	.33
4. Royal Dutch Shell	1.000	-.000	**.00**	.605	.694	.15
5. ExxonMobil	.683	-.032	.53	.563	.719	.17
Cumulative proportion of total sample variance explained	.323	.647		.487	.769	

$$\boldsymbol{R} - \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^\top - \widehat{\boldsymbol{\Psi}} = \begin{pmatrix} 0 & .001 & -.002 & -.000 & -.052 \\ .001 & 0 & .002 & .000 & -.033 \\ -.002 & .002 & 0 & .000 & .001 \\ .000 & .000 & .000 & 0 & .000 \\ .052 & -.033 & .001 & .000 & 0 \end{pmatrix}$$

"Better than the result by PC method"

Interpretation of stock price data.

- PC method
  Factor 1; general economic conditions "market factor"
  Factor 2; banking vs. oil "industry factor"
  With 2 factors, 77% of the variation explained.

- ML method
  Factor 1; "market factor"
  Factor 2; "banking factor" or "industry factor"
  Interpretation is not as clear as that by PC method
  With 2 factors, 65% of total sample variance explained.

8.4 Factor rotation

Recall that a rotation by an angle ϕ in the counterclockwise direction.

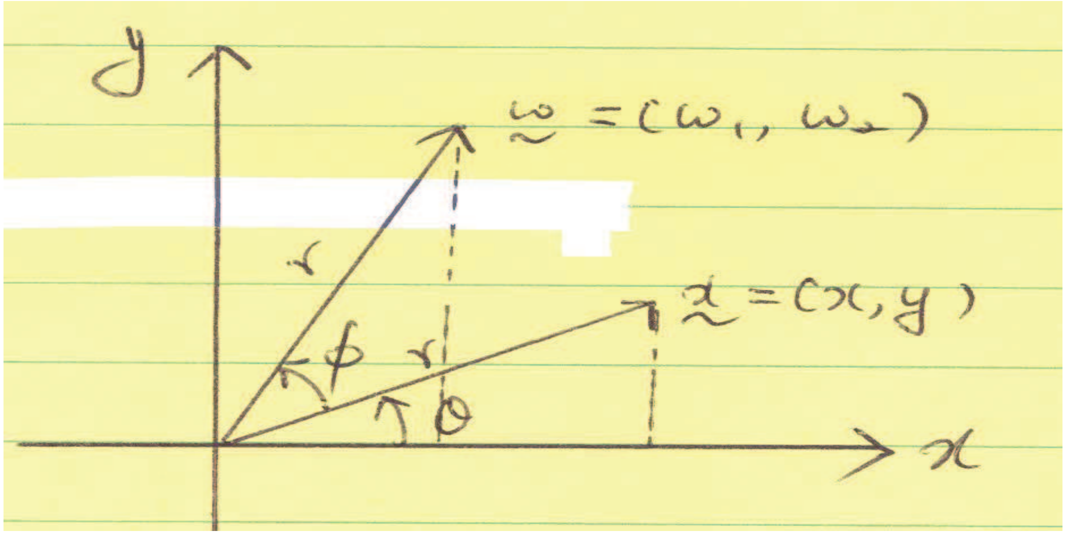

y
$\underset{\sim}{w} = (w_1, w_2)$
r
ϕ
r
θ
$\underset{\sim}{x} = (x, y)$
x

Figure 8.4.1: Rotation of a vector

$$\begin{aligned} x &= r\cos\theta, \quad y = r\sin\theta \\ w_1 &= r\cos(\theta+\phi) = r\cos\theta\cos\phi - r\sin\theta\sin\phi \\ &\quad = x\cos\phi - y\sin\phi \\ w_2 &= r\sin(\theta+\phi) = r\sin\theta\cos\phi + r\cos\theta\sin\phi \\ &\quad = y\cos\phi + x\sin\phi \end{aligned}$$

That is,

$$\begin{aligned} (w_1, w_2)^\top &= \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} (x \ y)^\top \\ \Leftrightarrow \quad \boldsymbol{w} = \boldsymbol{T}_1\boldsymbol{x}, &\quad \text{where } \boldsymbol{T}_1 = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}. \end{aligned}$$

Note that $\boldsymbol{T}_1$ is an orthonormal matrix.

Similarly a clockwise rotation is

$$\boldsymbol{w} = \boldsymbol{T}_2\boldsymbol{x}, \text{ where } \boldsymbol{T}_2 = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}.$$

That is, a rotation by an angle ϕ is the multiplication of a matrix.
Recall that the factor model is not unique, *i.e.*,

$$\begin{aligned} \boldsymbol{X} - \boldsymbol{\mu} &= \boldsymbol{LF} + \boldsymbol{\varepsilon} = \boldsymbol{LTT}^\top\boldsymbol{F} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{L}^*\boldsymbol{F}^* + \boldsymbol{\varepsilon} \quad \text{for any orthogonal } \boldsymbol{T}. \end{aligned}$$

The question is how to choose the optimal T.

Note that the row vectors of $\boldsymbol{LT}$ are obtained from $\boldsymbol{L}$ by rotating the row vectors of $\boldsymbol{L}$.
Also, the new factors $\boldsymbol{T}^\top\boldsymbol{F}$ are obtained by rotating the current factor $\boldsymbol{F}$.

In this sense, we call the problem of finding the optimal $\boldsymbol{T}$ as the factor rotation.

Note that rotation is not possible with 1 factor.

Motivational Example

- Consumer-preference Data: consumers' ratings for 5 attributes of a new product

- Variables: taste(X_1), good buy for money (X_2), flavor(X_3), suitable for snack(X_4), provides lots of energy(X_5)

- Two factor model

Variable	Estimated factor loadings Factor 1 (F_1)	 Factor 2 (F_2)
1. Taste	.56	.82
2. Good buy for money	.78	-.52
3. Flavor	.65	.75
4. Suitable for snack	.94	**-.10**
5. Provides lots of energy	.80	-.54
Cumulative proportion of total (standardized) sample variance explained	.571	.932

All the variables except variable 4 are significantly affected by the two factors.

If we rotate the factors so that some variables are affected only by one factor & the other variables are affected by the other factor, then we can interpret the factors more easily.

That is, for easier interpretation of "unobservable" factors

Determine the optimal factor rotation

- Object : Find the optimal factor rotation $\boldsymbol{T}$.

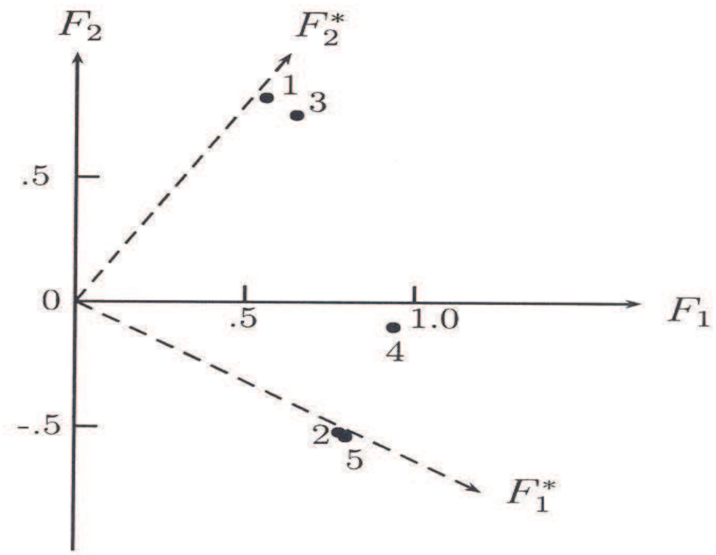


Figure 8.4.2: Graphical explanation of factor rotation

- Question : How to find it?

- Consider the m-factor model,

$$\boldsymbol{X}-\boldsymbol{\mu}=\boldsymbol{L}\boldsymbol{F}+\boldsymbol{\varepsilon}$$

Let $\boldsymbol{T}$ be an $m \times m$ orthogonal matrix, then

$$\begin{aligned}\boldsymbol{L} &= \begin{bmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & & \vdots \\ l_{p1} & \cdots & l_{pm} \end{bmatrix} \qquad \& \\ \boldsymbol{L}\boldsymbol{T} &= \begin{bmatrix} l_{11}^* & \cdots & l_{1m}^* \\ \vdots & & \vdots \\ l_{p1}^* & \cdots & l_{pm}^* \end{bmatrix}.\end{aligned}$$

Simple example)
Let $p=2$ & $m=2$. Consider the two factor model

$$\begin{aligned}\text{Model 1} \quad &: \quad \boldsymbol{X} - \boldsymbol{\mu} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \boldsymbol{F} + \boldsymbol{\varepsilon} \\ \text{Model 2} \quad &: \quad \boldsymbol{X} - \boldsymbol{\mu} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \boldsymbol{F} + \boldsymbol{\varepsilon}\end{aligned}$$

Note that model 1 has a simple structure than model 2 since X_1 is explained by F_1 & X_2 is explained by F_2.

If we compare the factor loadings of the two model we can see that

$$\text{variation}(l_{11}^2, l_{21}^2 | \text{Model 1}) \geq \text{variation}(l_{11}^2, l_{21}^2 | \text{Model 2})$$
$$\text{variation}(l_{12}^2, l_{22}^2 | \text{Model 1}) \geq \text{variation}(l_{12}^2, l_{22}^2 | \text{Model 2}).$$

Returning to the general m factor model, we can say that the optimal rotation $\boldsymbol{T}$ is the one which maximizes

$$\text{variation}\left(l_{11}^{*2}, \cdots, l_{p1}^{*2}\right), \cdots, \text{variation}\left(l_{1m}^{*2}, \cdots, l_{pm}^{*2}\right)$$

i) Varimax method
Find the orthogonal matrix $\boldsymbol{T}$ which maximizes

$$V = \sum_{j=1}^{m} \left[\sum_{i=1}^{p} l_{ij}^{*4} - \frac{(\sum_{i=1}^{p} l_{ij}^{*2})^2}{p} \right]$$

ii) Quartimax method
Find the orthogonal matrix $\boldsymbol{T}$ which maxmizes

$$\sum_{i=1}^{p} \left[\sum_{j=1}^{m} l_{ij}^{*4} - \frac{(\sum_{j=1}^{m} l_{ij}^{*2})^2}{p} \right].$$

iii) Equimax : Maximize the sum of varimax & Quartimax.

Remark

$$\begin{aligned} V &= \sum_{j=1}^{m}\left(\sum_{i=1}^{p}\left(l_{ij}^{*2} - (p\overline{l_j^{*2}}\;\right)^2\right) \\ &= \sum_{j=1}^{m}(p-1) \times (\text{sample variance of } l_{ij}^{*2}) \end{aligned}$$

$\because$

$$\begin{aligned} V &= \sum_{j=1}^{m}\left[\sum_{i=1}^{p}\left(l_{ij}^{*2}\right)^2 - \frac{\left(p\;\overline{l_j^{*2}}\right)^2}{p}\right] \\ &= \sum_{j=1}\left[\sum_{i=1} x_i^2 - p\;\bar{x}^2\right], \text{ where } x_i = l_{ij}^{*2} \\ &= \sum_{j=1}\sum_{i=1}(x_i - \bar{x})^2 \\ &= \sum_{j=1}\sum_{i=1}\left(l_{ij}^{*2} - \overline{l_j^{*2}}\right)^2 \\ &= \sum_{j=1}^{m}(p-1) \times (\text{sample variance of } l_{ij}^{*2}), \\ &\quad \text{where sample variance of } l_{ij}^{*2} = \frac{1}{p-1}\sum_{i=1}\left(l_{ij}^{*2} - \overline{l_j^{*2}}\right)^2 \end{aligned}$$

Return to the motivational example

- Factor loading after the varimax rotation

	Rotated estimated factor loading	
Variable	Factor 1 (F_1^*)	Factor 2 (F_2^*)
1. Taste	.02	.99
2. Good buy for money	.94	-.01
3. Flavor	.13	.98
4. Suitable for snack	.84	.43
5. Provides lots of energy	.97	-.02
Cumulative proportion of total (standardized) sample variance explained	.507	.932

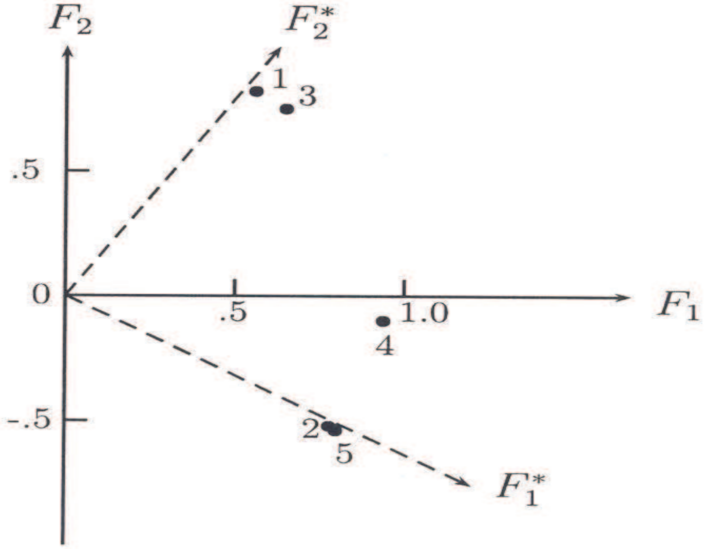


Figure 8.4.3: Factor rotation

Interpretation

- Factor 1 : a nutritional factor
- Factor 2 : a taste factor

Remark; Oblique rotations

These rotation are produced by multiplying $\boldsymbol{L}$ by a non-orthogonal matrix. That is, in two dimensions, we would rotate one axis through an angle θ_1 and the other through θ_2.

Such rotations are subject to much criticism and are not recommended.

8.5 Factor scores

We want to predict $\boldsymbol{F}$ using the data. The predicted values of $\boldsymbol{F}$, denoted by $\boldsymbol{f}$, are called "factor scores".

8.5.1 How to use

- Assess the validity of the assumption of

$$\boldsymbol{\varepsilon} \sim (\mathbb{0}, \Psi), \quad \text{where } \Psi = \text{diag}(\psi_i).$$

- Can be used as inputs of a subsequent analysis.
  For example, portfolio construction.

8.5.2 How to predict

Weighted Least Squares & Regression method

WLS method

$$\text{Model; } \boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon}, \quad \text{where } E\boldsymbol{\varepsilon} = \mathbb{0} \ \& \ V\boldsymbol{\varepsilon} = \Psi$$

Suppose $\boldsymbol{\mu}$, $\boldsymbol{L}$ and Ψ are known, we can estimate $\boldsymbol{F}$ by the weighted least squares method. That is,

$$\underset{\boldsymbol{f}}{min}(\boldsymbol{X} - \boldsymbol{\mu} - \boldsymbol{L}\boldsymbol{F})^\top \Psi^{-1} (\boldsymbol{X} - \boldsymbol{\mu} - \boldsymbol{L}\boldsymbol{F}).$$

The result is

$$\boldsymbol{f} = (\boldsymbol{L}^\top \Psi^{-1} \boldsymbol{L})^{-1} \boldsymbol{L}^\top \Psi^{-1} (\boldsymbol{X} - \boldsymbol{\mu}).$$

$(\because)$

$$\begin{aligned} & \Psi = \Psi^{\frac{1}{2}}\Psi^{\frac{1}{2}} \ \& \ \Psi^{-1} = \Psi^{-\frac{1}{2}}\Psi^{-\frac{1}{2}} \\ & \Psi^{-\frac{1}{2}}(\boldsymbol{X}-\boldsymbol{\mu}) = \Psi^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{F} + \Psi^{-\frac{1}{2}}\boldsymbol{\varepsilon} \\ \Leftrightarrow \quad & \boldsymbol{z} = \boldsymbol{W}\boldsymbol{F} + \boldsymbol{e}, \quad \boldsymbol{e} \ \sim \ (\mathbb{0}, \boldsymbol{I}) \end{aligned}$$

By the least squares method, we have

$$\begin{aligned} \boldsymbol{f} &= (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{z} \\ &= (\boldsymbol{L}^\top\Psi^{-\frac{1}{2}}\Psi^{-\frac{1}{2}}\boldsymbol{L})^{-1}\boldsymbol{L}^\top\Psi^{-\frac{1}{2}}\Psi^{-\frac{1}{2}}(\boldsymbol{X}-\boldsymbol{\mu}) \\ &= (\boldsymbol{L}^\top\Psi^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^\top\Psi^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \end{aligned}$$

□

Replacing $\boldsymbol{L}$ & Ψ with their estimates $\widehat{\boldsymbol{L}}$ & $\widehat{\Psi}$ and letting $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$, we have

$$\boldsymbol{f} = \left(\widehat{\boldsymbol{L}}^\top\widehat{\Psi}^{-1}\widehat{\boldsymbol{L}}\right)^{-1}\widehat{\boldsymbol{L}}^\top\widehat{\Psi}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}).$$

For each sample i, we have

$$\boldsymbol{f}_i = \left(\widehat{\boldsymbol{L}}^\top\widehat{\Psi}^{-1}\widehat{\boldsymbol{L}}\right)^{-1}\widehat{\boldsymbol{L}}^\top\widehat{\Psi}^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}), \quad i = 1, \cdots, n.$$

For the ML method with a constraint $\boldsymbol{L}^\top\Psi^{-1}\boldsymbol{L} = \Delta = \mathrm{d}iag(\Delta_i)$,

$$\boldsymbol{f}_i = \widehat{\Delta}^{-1}\widehat{\boldsymbol{L}}^\top\widehat{\Psi}^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}), \quad i = 1, \cdots, n.$$

Regression method

$$\boldsymbol{z} = \boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon}$$

Suppose that $\boldsymbol{\mu}, L$ and Ψ are known.
Assume that $\boldsymbol{F}$, $\boldsymbol{\varepsilon}$ are jointly normal, then

① $\boldsymbol{F} \sim N(\mathbb{0}, \boldsymbol{I})$

② $\boldsymbol{\varepsilon} \sim N(\mathbb{0}, \Psi)$

③ $\boldsymbol{F} \amalg \boldsymbol{\varepsilon}$.

So

$$\begin{aligned} \begin{pmatrix} \boldsymbol{F} \\ \boldsymbol{\varepsilon} \end{pmatrix} &\sim N_{m+p}\left(\begin{pmatrix} \mathbb{0} \\ \mathbb{0} \end{pmatrix}, \begin{pmatrix} I & \mathbb{0} \\ \mathbb{0} & \Psi \end{pmatrix} \right) \ \& \\ \boldsymbol{z} = (\boldsymbol{L}\ \boldsymbol{I}) \begin{pmatrix} \boldsymbol{F} \\ \boldsymbol{\varepsilon} \end{pmatrix} &\sim N_p(\mathbb{0}, \Sigma), \quad \text{where } \Sigma = \boldsymbol{L}\boldsymbol{L}^\top + \Psi. \end{aligned}$$

Joint distribution of $\begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{F} \end{pmatrix}$ is a normal since

$$\begin{aligned} \boldsymbol{y} &= \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{F} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon} \\ \boldsymbol{F} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L} & \boldsymbol{I} \\ \boldsymbol{I} & \mathbb{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{F} \\ \boldsymbol{\varepsilon} \end{pmatrix} = \boldsymbol{A} \begin{pmatrix} \boldsymbol{F} \\ \boldsymbol{\varepsilon} \end{pmatrix} \\ &\sim N_{p+m}\left(\begin{pmatrix} \mathbb{0} \\ \mathbb{0} \end{pmatrix}, \begin{pmatrix} \Sigma & \boldsymbol{L} \\ \boldsymbol{L}^\top & \boldsymbol{I} \end{pmatrix} \right). \end{aligned}$$

Since the conditional expectation is optimal under the squared error loss (See the following Goal), we can predict $\boldsymbol{F}$ by $E(\boldsymbol{F}|\boldsymbol{z})$, which is

$$E(\boldsymbol{F}|\boldsymbol{z}) = \boldsymbol{L}^\top(\boldsymbol{L}\boldsymbol{L}^\top + \Psi)^{-1}\boldsymbol{z} = \boldsymbol{L}^\top(\boldsymbol{L}\boldsymbol{L}^\top + \Psi)^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$$

Replacing $\boldsymbol{L}, \Psi$, and $\boldsymbol{\mu}$ with their estimated ones, we have

$$\boldsymbol{f} = \widehat{\boldsymbol{L}}^\top(\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^\top + \widehat{\Psi})^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}).$$

For each sample i, it is obtained as

$$\boldsymbol{f}_i = \widehat{\boldsymbol{L}}^\top(\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^\top + \widehat{\Psi})^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}), \quad i = 1, \cdots, n.$$

We can replace $\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^\top + \widehat{\Psi}$ with the sample covariance of $\boldsymbol{x}$, *i.e.*

$$\boldsymbol{f}_i = \widehat{\boldsymbol{L}}^\top \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}), \quad i = 1, \cdots, n.$$

Remark

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}^\top\boldsymbol{F} + \boldsymbol{\varepsilon}$$

① If rotated loadings $\widehat{\boldsymbol{L}}^* = \widehat{\boldsymbol{L}}\boldsymbol{T}$ are used in place of the original loadings in both methods, the subsequent factor scores are related to $\boldsymbol{f}_i^*$ are related the $\boldsymbol{f}_j$ by

$$\boldsymbol{f}_i^* = \boldsymbol{T}^\top \boldsymbol{f}_i, \quad i = 1, \cdots, n.$$

② Which one (WLS or Regression methods) to be used?
None is recommended as uniformly superior.

<u>Goal:</u> Find $m(x)$ such that minimize $E_Y(Y - m(x))^2$, where $m(x)$ is any function of x. Note that $(Y - m(x))^2$ is the squared error loss and $E_Y(Y - m(x))^2$ is called as risk.

<u>Proof</u>

$$\begin{aligned}
& E_Y(Y - m(x))^2 = E_Y(Y - EY|x + EY|x - m(x))^2 \\
= \; & E_Y(Y - EY|x)^2 + E_Y(EY|x - m(x))^2 + 2E_Y(Y - EY|x)(EY|x - m(x)) \\
= \; & E_Y(Y - EY|x)^2 + E_Y(EY|x - m(x))^2 + 2(EY|x - m(x))E_Y(Y - EY|x) \\
& (\because \; EY|x - m(x) \text{ is independent of } Y) \\
= \; & E_Y(Y - EY|x)^2 + E_Y(EY|x - m(x))^2 \; (\because \; EEY|x = EY)
\end{aligned}$$

Hence

$$\begin{aligned}
E_Y(Y - m(x))^2 &= E_Y(Y - EY|x)^2 + (EY|x - m(x))^2 \\
& \quad (\because EY|x - m(x)) \text{ is independent of } Y. \\
&\geq E_Y(Y - EY|x)^2 (\because (EY|x - m(x))^2 \geq 0)
\end{aligned}$$

Hence $E_Y(Y - m(x))^2$ is minimized when $m(x) = EY|x$. □

<u>Remark</u>:

① $E_Y(Y - m(x))^2$ is also called as MSE.

② Conditional expectation is the best prediction, where 'best' means minimum mean squared error.

③ If X is random, $EY|X$ is a random function of X.

④ $E_Y Y = E_X E_Y Y|X$

<u>Example: Stock price data</u>

- Factor scores by the least squares and regression methods using stock price data
- The rotated loadings, specific variances and vector of standardized observations;

$$\hat{\boldsymbol{L}}^*_{\boldsymbol{z}} = \begin{pmatrix} .763 & .024 \\ .821 & .227 \\ .669 & .104 \\ .118 & .993 \\ .113 & .675 \end{pmatrix}, \hat{\Psi}_{\boldsymbol{z}} = \begin{pmatrix} .42 & 0 & 0 & 0 & 0 \\ 0 & .27 & 0 & 0 & 0 \\ 0 & 0 & .54 & 0 & 0 \\ 0 & 0 & 0 & .00 & 0 \\ 0 & 0 & 0 & 0 & .53 \end{pmatrix}, \boldsymbol{z} = \begin{pmatrix} .50 \\ -1.40 \\ -.20 \\ -.70 \\ 1.40 \end{pmatrix}$$

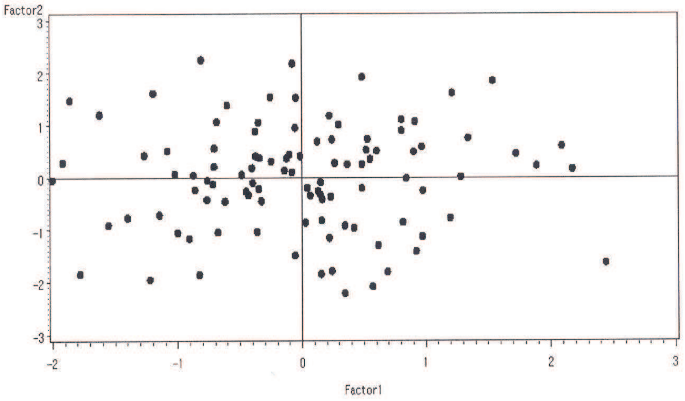
- Factor scores
  Weighted least squares;

$$\boldsymbol{f} = (\hat{\boldsymbol{L}}^*_{\boldsymbol{z}} \hat{\Psi}^{-1}_{\boldsymbol{z}} \hat{\boldsymbol{L}}_{\boldsymbol{z}})^{-1} \hat{\boldsymbol{L}}^*_{\boldsymbol{z}} \hat{\Psi}^{-1}_{\boldsymbol{z}} \boldsymbol{z} = \begin{pmatrix} -.61 \\ -.61 \end{pmatrix}$$

  Regression;

$$\begin{aligned} \boldsymbol{f} &= \hat{\boldsymbol{L}}^*_z \boldsymbol{R}^{-1} \boldsymbol{z} = \begin{pmatrix} .331 & .526 & .221 & -.137 & .011 \\ -.040 & -.063 & -.026 & 1.023 & -.001 \end{pmatrix} \begin{pmatrix} .50 \\ -1.40 \\ -.20 \\ -.70 \\ 1.40 \end{pmatrix} \\ &= \begin{pmatrix} -.50 \\ -.64 \end{pmatrix} \end{aligned}$$

- Factor scores using regression method for factor 1 and 2



Summary of factor analysis

$$\boldsymbol{X}_{p\times 1} - \boldsymbol{\mu}_{p\times 1} = \boldsymbol{L}_{p\times m}\boldsymbol{F}_{m\times 1} + \boldsymbol{\varepsilon}_{p\times 1}$$

$$\Leftrightarrow$$

$$\begin{pmatrix} X_1 - \mu_1 \\ \vdots \\ X_p - \mu_p \end{pmatrix} = \begin{pmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & & \vdots \\ l_{p1} & \cdots & l_{pm} \end{pmatrix} \begin{pmatrix} F_1 \\ \vdots \\ F_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

Assume that

1. $\boldsymbol{F} \sim (\mathbb{0}, \boldsymbol{I})$
2. $\boldsymbol{\varepsilon} \sim (\mathbb{0}, \Psi), \Psi = \text{diag}(\psi_i)$
3. $\boldsymbol{F} \amalg \boldsymbol{\varepsilon}$

Facts

1. $Cov(\boldsymbol{X}) = \Sigma = E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top = \ldots = \boldsymbol{L}\boldsymbol{L}^\top + \Psi$

So need to solve $\Sigma = \boldsymbol{L}\boldsymbol{L}^\top + \Psi$.

2. For any $m \times m$ orthogonal matrix $\boldsymbol{T}$ such that $\boldsymbol{T}^\top\boldsymbol{T} = \boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}$, we have $\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}^\top\boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{L}^*\boldsymbol{F}^* + \boldsymbol{\varepsilon}$.

   $\therefore$ Factor leading matrix $\boldsymbol{L}$ is determined only up to an orthogonal matrix $\boldsymbol{T}$

3. $$\Sigma = \boldsymbol{L}\boldsymbol{L}^\top + \Psi \Rightarrow VX_i = \sigma_{ii} = \underbrace{l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2}_{\text{communality}} + \underbrace{\psi_i}_{\text{specific variance}}$$

4. $cov(\boldsymbol{X}, \boldsymbol{F}) = \boldsymbol{L}$ i.e. factor loading matrix is covariances.

How to solve $\Sigma = \boldsymbol{L}\boldsymbol{L}^\top + \Psi$!

1. Principal component method
2. Principal factor method
3. Maximum likelihood method

- 2 & 3 are iterative methods

1. PC method

$$\begin{aligned} cov(\boldsymbol{X}) = \Sigma &= \lambda_1\boldsymbol{e}_1\boldsymbol{e}_1^\top + \cdots + \lambda_p\boldsymbol{e}_p\boldsymbol{e}_p^\top \\ \text{by spectral decomposition} & \\ &= [\sqrt{\lambda_1}\boldsymbol{e}_1 \ldots \sqrt{\lambda_p}\boldsymbol{e}_p] \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1^\top \\ \vdots \\ \sqrt{\lambda_p}\boldsymbol{e}_p^\top \end{bmatrix} \end{aligned}$$

If $m = p$, $\Sigma = \boldsymbol{L}\boldsymbol{L}^\top + \mathbb{0}$, where $\boldsymbol{L} = [\sqrt{\lambda_1}\boldsymbol{e}_1 \cdots \sqrt{\lambda_p}\boldsymbol{e}_p^]$.

When $\lambda_{m+1} \cdots \lambda_p$ are small, then neglect $\lambda_{m+1}\boldsymbol{e}_{m+1}\boldsymbol{e}_{m+1}^\top + \cdots + \lambda_p\boldsymbol{e}_p\boldsymbol{e}_p^\top$, then

$$\Sigma \cong [\sqrt{\lambda_1}\boldsymbol{e}_1 \ldots \sqrt{\lambda_m}\boldsymbol{e}_m] \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1^\top \\ \vdots \\ \sqrt{\lambda_m}\boldsymbol{e}_m^\top \end{bmatrix} = \boldsymbol{L}\boldsymbol{L}^\top.$$

Define $\psi_i = VX_i - \sum_{j=1}^m l_{ij}^2$, then $\Sigma \cong \boldsymbol{L}\boldsymbol{L}^\top + \Psi$, where $\Psi = \text{diag}(\psi_i)$.

<u>How to choose m !</u>

1. Use scree plot
   The number of factors is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size.

2. A large sample test also exists, but usually leading to a retention of more common factors.

<u>Factor rotation</u>

Goal: To find a simple structure to the factor leading matrix for interpretation.

Usually use the varimax rotation which is an orthogonal rotation.

<u>Factor scores</u>

We calculated principle component scores for further analyses, for example in a regression model.

Similarly, we obtain the factor scores using

1. weighted least squares method or

2. regression method

<u>Note</u>

For the standardized variables, $z_i = \dfrac{x_i - \mu_i}{\sqrt{\sigma_{ii}}}$, $cov(\boldsymbol{z}, \boldsymbol{F}) = \boldsymbol{L} = (l_{ij}) = corr(\boldsymbol{z}, \boldsymbol{F})$

pf) $cov(Z_i, F_j) = l_{ij}, VZ_i = 1, \ \& \ VF_j = 1$

$$\therefore \ corr(Z_i, F_j) = \frac{cov(Z_i, F_j)}{\sqrt{VZ_i}\sqrt{VF_j}} = cov(Z_i, F_j) = l_{ij} \qquad \square$$

<u>Remark</u>

- This $\{\boldsymbol{F} \sim (\mathbb{0}, \boldsymbol{I})\}$ is a crucial different point compared to confirmatory factor analysis (CFA).
- Factor analysis we learned so far is exploratory factor analysis.
- <u>Idea behind CFA</u>

ex) A correlation matrix $\boldsymbol{R}$ for 5 measurements of test scores,

$$\begin{cases} \text{paragraph comprehension } (X_1) \\ \text{sentence completion } (X_2) \\ \text{word meaning } (X_3) \\ \text{addition } (X_4) \\ \text{counting dots } (X_5). \end{cases}$$

OFM after standardizing X_1 through X_5, we have that

$$\begin{aligned} Z_1 &= l_{11}F_1 + l_{12}F_2 + \epsilon_1 \\ &\vdots \\ Z_5 &= l_{51}F_1 + l_{52}F_2 + \epsilon_5. \end{aligned}$$

$$\boldsymbol{F} \sim (\mathbb{0}, \boldsymbol{I}) \ \amalg \ \boldsymbol{\varepsilon} \sim (\mathbb{0}, \Psi), \Psi = \text{diag}(\psi_i)$$

However, we have a belief that two factors underlie test performance in students: verbal ability & quantitative ability.

These two factors are distinct but may be correlated.

X_1, X_2, X_3: measure verbal ability &
X_4, X_5: measure quantitative ability.

So a path diagram of two-factor model of psychological test performance is as follows:

$$\begin{array}{cccccc}
 & F_1 & & & F_2 & \\
\swarrow & \downarrow & \searrow & \swarrow & & \searrow \\
Z_1 & Z_2 & Z_3 & Z_4 & & Z_5 \\
\uparrow & \uparrow & \uparrow & \uparrow & & \uparrow \\
\varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 & & \varepsilon_5
\end{array}$$

$$\begin{array}{lll}
Z_1 = l_{11}F_1 & & +\,\varepsilon_1 \\
Z_2 = l_{21}F_1 & & +\,\varepsilon_2 \\
Z_3 = l_{31}F_1 & & +\,\varepsilon_3 \\
Z_4 = & l_{42}F_2 & +\,\varepsilon_4 \\
Z_5 = & l_{52}F_2 & +\,\varepsilon_5
\end{array}$$

$corr(F_1, F_2) = \rho_{12}$

Note that, in OFM, $F \sim (\mathbb{0}, \boldsymbol{I})$, i.e. factors are uncorrelated.

- Two important differences between the CFA & the (exploratory) factor analysis

  i) Imposed constraints on the factor loadings matrix associated with the confirmatory model.

  ii) The confirmatory model allows for correlation between the factors.

8.6 Factor analysis in R

First, read the data and then obtain correlation matrix of stock data. See Table 8.4 (Johnson and Wichern, 2007) of stock-price data (weekly rate of return). Before doing any multivariate methods, it will be better to check the following null hypothesis;

$$H_0 : Corr(\mathbf{X}_{n\times p}) = \mathbf{I}_p,$$

where $\mathbf{X}$ denote a data matrix.

```
> # Data reading
> #install.packages(psych)
> library(psych)
> data <- read.table("T8-4.dat")
> stock <- as.data.frame(data)
> names(stock) <- c("JPMorgan","Citibank","WellsFargo",
+  "RoyalDShell","ExxonMobil")
> len <- dim(stock)[1]
>
> # Correlation test
> cors <- cor(stock)
> round(cors,3)
            JPMorgan Citibank WellsFargo RoyalDShell ExxonMobil
JPMorgan       1.000    0.632      0.510       0.115      0.154
Citibank       0.632    1.000      0.574       0.322      0.213
WellsFargo     0.510    0.574      1.000       0.182      0.146
RoyalDShell    0.115    0.322      0.182       1.000      0.683
ExxonMobil     0.154    0.213      0.146       0.683      1.000
> cortest.bartlett(cors,n=len) #null hypo H_0 : cor(data) = I
$chisq
[1] 173.3102

$p.value
[1] 5.720169e-32

$df
[1] 10
```

```
> eigen(cors)$values
[1] 2.4372731 1.4070127 0.5005127 0.4000316 0.2551699
```

By Bartlett test, given data matrix is not an identity matrix with a p-value of $5.720169e - 32$. Hence further multivariate analyses are possible.

Second, a scree plot is plotted. Factor analysis is based on 2 factors since only two eigenvalues are greater than 1.

```
> # Scree plot
> pc <- princomp(stock,cor=TRUE)
> screeplot(pc,type="lines",main="Scree plot")
```

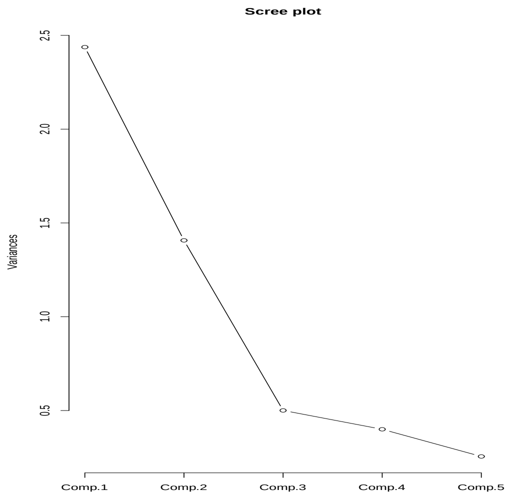


Figure 8.6.1: A scree plot for the stock data.

Third, do factor analysis without rotation with 2 factors. And then do factor rotation for easy interpretation of factors using `varimax`. At both analyses, factor scores are obtained using `regression` method.

```
> # Factor analysis
> fact1 <- factanal(stock,factors=2,rotation="none",
+ scores="regression")
> print(fact1$loadings, cutoff = 0) # print everything

Loadings:
            Factor1 Factor2
JPMorgan     0.121   0.754
Citibank     0.328   0.786
WellsFargo   0.188   0.650
RoyalDShell  0.997  -0.007
ExxonMobil   0.685   0.026

               Factor1 Factor2
SS loadings      1.622   1.610
Proportion Var   0.324   0.322
Cumulative Var   0.324   0.646
> fact2 <- factanal(stock,factors=2,rotation="varimax",
+ scores="regression")
```

Finally variables position and factor scores are plotted in Figure 8.6.2.

```
> # Positioning of variables and factor scores pattern
> par(mfrow=c(1,2))
>
> #postioning of variables
> plot(fact2$loadings[,1:2],type="n") #,xlim=c(-0.3,1.2))
> text(fact2$loadings[,1:2],names(stock))
> abline(v=0.45,h=0.5)
> grid()
> title("Positioning of Variables")
>
> #factor scores pattern
> #fact2$scores
> rownames(stock) <- c(seq(1:len))
> plot(fact2$scores[,1:2],type="n")
> text(fact2$scores[,1:2],rownames(stock))
> abline(v=0,h=0)
> grid()
> title("Factor Score")
```

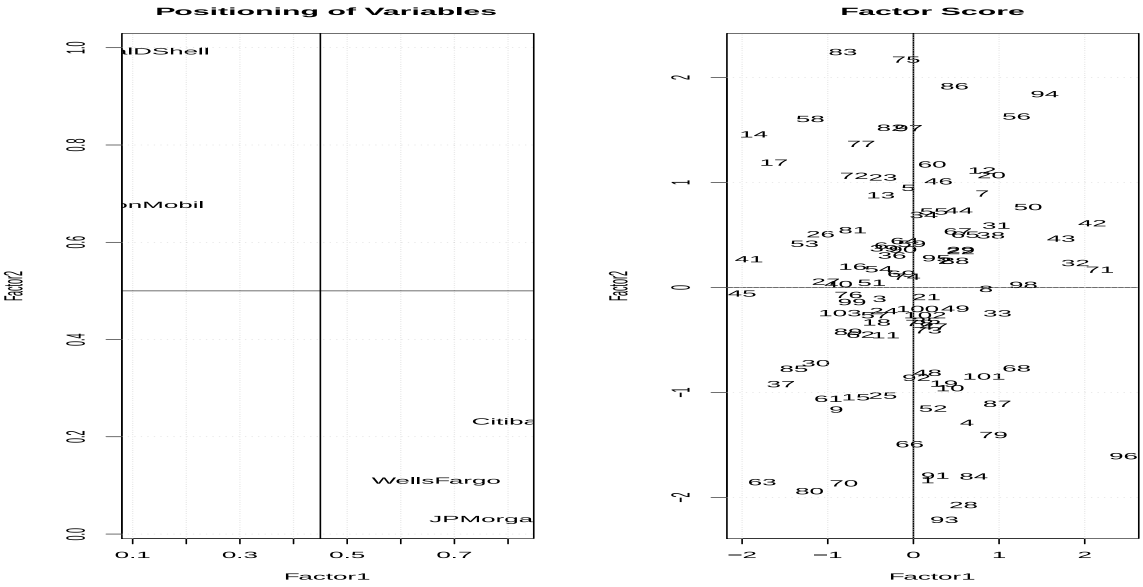


Figure 8.6.2: Positioning of variables and factor scores.

Chapter 9

Canonical Correlation Analysis

Goal: To identify & quantify the associations between two sets of variables.

9.1 Canonical variates & canonical correlations

① Sample correlation coefficient between x & y

$$r_{xy} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}}$$

② Sample multiple correlation coefficient between $x_1, \cdots, x_p$ & $y_1 = y$.

$$R_{\boldsymbol{x}y} \; = \; \frac{\sqrt{S_{\boldsymbol{x}y}^{\top} S_{\boldsymbol{x}\boldsymbol{x}}^{-1} S_{\boldsymbol{x}y}}}{S_{yy}},$$

where $S_{yy} = \frac{1}{n-1}\sum(y_i - \bar{y})^2$,

$$\begin{aligned}
\boldsymbol{S} &= \begin{pmatrix} S_{\boldsymbol{x}\boldsymbol{x}} & S_{\boldsymbol{x}y} \\ S_{\boldsymbol{x}y}^{\top} & S_{yy} \end{pmatrix}; \text{ sample variance-covariance matrix,} \\
\boldsymbol{R} &= \begin{pmatrix} R_{\boldsymbol{x}\boldsymbol{x}} & r_{\boldsymbol{x}y} \\ r_{\boldsymbol{x}y}' & 1 \end{pmatrix}; \text{ sample correlation matrix,} \\
S_{\boldsymbol{x}y}^{\top} &= (S_{1y},\ S_{2y},\ \cdots,\ S_{py});\ \text{sample covariance vector between } x_i \text{ \& } y_1,\ i = 1, \cdots, p, \text{ \&} \\
r_{\boldsymbol{x}y}^{\top} &= (r_{1y},\ r_{2y},\ \cdots,\ r_{py});\ \text{sample correlation vector between } x_i \text{ \& } y_1,\ i = 1, \cdots, p.
\end{aligned}$$

③ What about

$$\begin{aligned}\boldsymbol{x}^{(1)} &= (X_1^{(1)}, X_2^{(1)}, \cdots, X_p^{(1)})^\top \quad \& \\ \boldsymbol{x}^{(2)} &= (X_1^{(2)}, X_2^{(2)}, \cdots, X_q^{(2)}) \quad (p \leqq q)?\end{aligned}$$

Two groups of variables $\boldsymbol{x}^{(1)}_{p\times 1}$ & $\boldsymbol{x}^{(2)}_{q\times 1}$ with

$$\begin{aligned}\underset{(p+q)\times 1}{\boldsymbol{x}} = \begin{pmatrix}\boldsymbol{x}^{(1)}\\ \cdots \\ \boldsymbol{x}^{(2)}\end{pmatrix}, \quad E\boldsymbol{x} &= \begin{pmatrix}\boldsymbol{\mu}^{(1)}\\ \cdots \\ \boldsymbol{\mu}^{(2)}\end{pmatrix} \quad \& \\ V\boldsymbol{x} &= \boldsymbol{\Sigma} = \begin{pmatrix}\underset{p\times p}{\boldsymbol{\Sigma_{11}}} & \underset{p\times q}{\boldsymbol{\Sigma_{12}}} \\ \underset{q\times p}{\boldsymbol{\Sigma_{21}}} & \underset{q\times q}{\boldsymbol{\Sigma_{22}}}\end{pmatrix}.\end{aligned}$$

Goal : To summarize the associations between $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ in terms of **a few carefully chosen covariances (or correlations) rather than the** pq **covariances in $\boldsymbol{\Sigma_{12}}$**

Idea : Linear combinations provide simple summary measures of a set of variables.

$$\begin{aligned}U &= \boldsymbol{a}^\top \boldsymbol{x}^{(1)}, \quad V = \boldsymbol{b}^\top \boldsymbol{x}^{(2)} \\ V(U) &= \boldsymbol{a}^\top Cov\left(\boldsymbol{x}^{(1)}\right)\boldsymbol{a} = \boldsymbol{a}^\top \boldsymbol{\Sigma}_{11}\boldsymbol{a} \quad V(V) = \boldsymbol{b}^\top \boldsymbol{\Sigma}_{22}\boldsymbol{b} \\ Cov(U,V) &= \boldsymbol{a}^\top Cov\left(\boldsymbol{x}^{(1)},\ \boldsymbol{x}^{(2)}\right)\boldsymbol{b} = \boldsymbol{a}^\top \boldsymbol{\Sigma}_{12}\boldsymbol{b}\end{aligned}$$

Find $\boldsymbol{a}$ & $\boldsymbol{b}$ such that

$$corr(U,V) = \frac{\boldsymbol{a}^\top \boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}^\top \boldsymbol{\Sigma}_{22}\boldsymbol{b}}} \tag{9.1.1}$$

is as large as possible.

Procedure:

① Find $U_1,\ V_1$ having unit variance which maximize (9.1.1).

② Find $U_2,\ V_2$ having unit variance which maximize (9.1.1) among all choices that are uncorrelated with $U_1,\ V_1$.

③ At the k^{th} step, find $U_k,\ V_k$ having unit variance which maximize (9.1.1) among all choices uncorrelated with the previous $k-1$ canonical variable pairs $(U_i, V_i), i = 1, \ldots, k-1$.

$Corr(U_k, V_k);\ k^{th}$ canonical correlation.

Result

$$p \le q, \qquad Cov\left(\underset{p\times 1}{\boldsymbol{x}}^{(1)}\right) = \boldsymbol{\Sigma}_{11}, Cov\left(\underset{q\times 1}{\boldsymbol{x}^{(2)}}\right) = \boldsymbol{\Sigma}_{22}$$
$$Cov\left(\boldsymbol{x}^{(1)},\ \boldsymbol{x}^{(2)}\right) = \boldsymbol{\Sigma}_{12},\ \boldsymbol{\Sigma};\ \text{full rank.}$$

Let $U = \boldsymbol{a}^\top \boldsymbol{x}^{(1)}$ and $V = \boldsymbol{b}^\top \boldsymbol{x}^{(2)}$, then

① $\max\limits_{\boldsymbol{a},\boldsymbol{b}} Corr(U, V) = \rho_1^*$ is attained by

$$U_1 = \underbrace{\boldsymbol{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}}_{\boldsymbol{a}_1^\top}\boldsymbol{x}^{(1)} \quad \& \quad V_1 = \underbrace{\boldsymbol{f}_1^\top \boldsymbol{\Sigma}_{22}^{-1/2}}_{\boldsymbol{b}_1^\top}\boldsymbol{x}^{(2)}$$

②

$$U_k = \underbrace{\boldsymbol{e}_k^\top \boldsymbol{\Sigma}_{11}^{-1/2}}_{\boldsymbol{a}_k^\top}\boldsymbol{x}^{(1)} \quad \& \quad V_k = \underbrace{\boldsymbol{f}_k^\top \boldsymbol{\Sigma}_{22}^{-1/2}}_{\boldsymbol{b}_k^\top}\boldsymbol{x}^{(2)}$$

maximizes $Corr(U_k, V_k) = \rho_k^*$ among those linear combinations uncorrelated with the preceding $1, 2, \cdots, k-1$ canonical variables. Here $\rho_1^{*2} \ge \rho_2^{*2} \ge \cdots \ge \rho_p^{*2}$ are the eigenvalues of

$$\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}},$$

and $\boldsymbol{e}_1,\ \boldsymbol{e}_2, \ldots,\ \boldsymbol{e}_p$ are the associated eigenvectors &

$$\boldsymbol{f}_i \propto \boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{e}_i.$$

③

$$\begin{aligned} VU_k &= VV_k = 1 \\ Cov(U_k, U_l) &= Corr(U_k, U_l) = 0,\ k \ne l \\ Cov(V_k, V_l) &= Corr(V_k, V_l) = 0,\ k \ne l \\ Cov(U_k, V_l) &= Corr(U_k, V_l) = 0,\ k \ne l \end{aligned}$$

for $k, l = 1, 2, \cdots, p$.

pf) See the class homepage for the proof. □

Remark

i) $\rho_1^{*2}, \cdots, \rho_p^{*2}$ are also the p largest eigenvalues of $\mathbf{\Sigma}_{22}^{-\frac{1}{2}}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-\frac{1}{2}}$ with corresponding eigenvectors $\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_p$.

ii) Standardized variables $\boldsymbol{z}_{p\times 1}^{(1)}$ & $\boldsymbol{z}_{q\times 1}^{(2)}$

$$\begin{aligned} U_k &= \boldsymbol{a}_k^\top \boldsymbol{z}^{(1)} = \boldsymbol{e}_k^\top \boldsymbol{\rho}_{11}^{-\frac{1}{2}} \boldsymbol{z}^{(1)} \\ V_k &= \boldsymbol{b}_k^\top \boldsymbol{z}^{(2)} = \boldsymbol{f}_k^\top \boldsymbol{\rho}_{22}^{-\frac{1}{2}} \boldsymbol{z}^{(2)} \end{aligned}$$

Here,

$$\begin{aligned} Cov\boldsymbol{z}^{(1)} &= \boldsymbol{\rho}_{11},\ Cov\boldsymbol{z}^{(2)} = \boldsymbol{\rho}_{22},\ Cov(\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}) = \boldsymbol{\rho}_{12} = \boldsymbol{\rho}_{21}^\top \\ \boldsymbol{e}_k \ \&\ \boldsymbol{f}_k &\ ;\quad \text{eigenvectors of } \boldsymbol{\rho}_{11}^{-\frac{1}{2}}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-\frac{1}{2}} \\ &\quad\ \ \&\ \boldsymbol{\rho}_{22}^{-\frac{1}{2}}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-\frac{1}{2}},\ \text{respectively} \\ Corr(U_k, V_k) &= \rho_k^*,\ k = 1, 2, \cdots, p, \end{aligned}$$

where $\rho_1^{*2} \ge \rho_2^{*2} \ge \cdots \ge \rho_p^{*2}$ are the nonzero eigenvalues of $\boldsymbol{\rho}_{11}^{-\frac{1}{2}}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-\frac{1}{2}}$. Note that $Corr(U_k, V_k) = \rho_k^*,\ k = 1, 2, \cdots, p$ are not changed by standardization. See comment at pg. 542.

ex) Example 10.1 for calculating canonical variates & canonical correlations for standardized variables.

Hard to calculate, so we'll depend on computer.

9.2 Interpreting the population canonical variables

9.2.1 Identifying

$$\begin{aligned} U_1 &= \boldsymbol{a}_1^\top \boldsymbol{x}^{(1)} = a_{11}x_1^{(1)} + a_{12}x_2^{(1)} + \cdots + a_{1p}x_p^{(1)} \\ U_2 &= \boldsymbol{a}_2^\top \boldsymbol{x}^{(1)} = a_{21}x_1^{(1)} + a_{22}x_2^{(1)} + \cdots + a_{2p}x_p^{(1)} \\ &\ \ \vdots \\ U_p &= \boldsymbol{a}_p^\top \boldsymbol{x}^{(1)} = a_{p1}x_1^{(1)} + a_{p2}x_2^{(1)} + \cdots + a_{pp}x_p^{(1)} \end{aligned}$$

$$\begin{aligned}
V_1 &= \boldsymbol{b}_1^\top \boldsymbol{x}^{(2)} = b_{11}x_1^{(2)} + b_{12}x_2^{(2)} + \cdots + b_{1q}x_q^{(2)} \\
V_2 &= \boldsymbol{b}_2^\top \boldsymbol{x}^{(2)} = b_{21}x_1^{(2)} + b_{22}x_2^{(2)} + \cdots + b_{2q}x_q^{(2)} \\
&\vdots \\
V_q &= \boldsymbol{b}_q^\top \boldsymbol{x}^{(2)} = b_{q1}x_1^{(2)} + b_{q2}x_2^{(2)} + \cdots + b_{qq}x_q^{(2)}
\end{aligned}$$

$\Leftrightarrow$

Let $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1^\top \\ \boldsymbol{a}_2^\top \\ \vdots \\ \boldsymbol{a}_p^\top \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_1^\top \\ \boldsymbol{b}_2^\top \\ \vdots \\ \boldsymbol{b}_q^\top \end{bmatrix}$, then

$$\begin{aligned}
\underset{p\times 1}{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{x}^{(1)} \quad \& \quad \underset{q\times 1}{\boldsymbol{v}} &= \boldsymbol{B}\boldsymbol{x}^{(2)} \\
Cov(\boldsymbol{u},\ \boldsymbol{x}^{(1)}) &= Cov(\boldsymbol{A}\boldsymbol{x}^{(1)},\ \boldsymbol{x}^{(1)}) = \boldsymbol{A}\Sigma_{11}
\end{aligned}$$

<u>Note</u>

$$\begin{aligned}
VU_i = 1, \quad VX_k^{(1)} &= \sigma_{kk} \ \left(kk^{th} \text{ element of } \boldsymbol{\Sigma}_{11}\right) \\
\therefore\ Corr\left(U_i,\ X_k^{(1)}\right) &= Cov\left(U_i, X_k^{(1)}\right) / \left\{\sqrt{VU_i}\sqrt{VX_k^{(1)}}\right\} \\
&= Cov\left(U_i, X_k^{(1)}\right) / \sigma_{kk}^{\frac{1}{2}} \\
&= Cov\left(U_i, \sigma_{kk}^{-\frac{1}{2}} X_k^{(1)}\right)
\end{aligned}$$

Let $\boldsymbol{V}_{11}^{-\frac{1}{2}}$ be $diag\left\{\sigma_{11}^{-\frac{1}{2}}, \sigma_{22}^{-\frac{1}{2}}, \cdots, \sigma_{pp}^{-\frac{1}{2}}\right\}$, then

$$\begin{aligned}
\boldsymbol{\rho}_{\boldsymbol{u},\boldsymbol{x}^{(1)}} &= Corr(\boldsymbol{u}, \boldsymbol{x}^{(1)}) = Corr(\boldsymbol{u}, \boldsymbol{V}_{11}^{-\frac{1}{2}}\boldsymbol{x}^{(1)}) \\
&= Cov\left(\boldsymbol{A}\boldsymbol{x}^{(1)}, \boldsymbol{V}_{11}^{-\frac{1}{2}}\boldsymbol{x}^{(1)}\right) = \boldsymbol{A}\boldsymbol{\Sigma}_{11}\boldsymbol{V}_{11}^{-\frac{1}{2}}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\underset{p\times p}{\boldsymbol{\rho}_{\boldsymbol{u},\boldsymbol{x}^{(1)}}} &= \boldsymbol{A}\boldsymbol{\Sigma}_{11}\boldsymbol{V}_{11}^{-\frac{1}{2}}, \qquad \underset{q\times q}{\boldsymbol{\rho}_{\boldsymbol{v},\boldsymbol{x}^{(2)}}} = \boldsymbol{B}\boldsymbol{\Sigma}_{22}\boldsymbol{V}_{22}^{-\frac{1}{2}} \\
\underset{p\times q}{\boldsymbol{\rho}_{\boldsymbol{u},\boldsymbol{x}^{(2)}}} &= \boldsymbol{A}\boldsymbol{\Sigma}_{12}\boldsymbol{V}_{22}^{-\frac{1}{2}}, \qquad \underset{q\times p}{\boldsymbol{\rho}_{\boldsymbol{v},\boldsymbol{x}^{(1)}}} = \boldsymbol{B}\boldsymbol{\Sigma}_{21}\boldsymbol{V}_{11}^{-\frac{1}{2}},
\end{aligned}$$

where $\boldsymbol{V}_{22}^{-\frac{1}{2}} = diag\left\{\left[V\left(X_1^{(2)}\right)\right]^{-1/2}, \left[V\left(X_2^{(2)}\right)\right]^{-1/2}, \cdots, \left[V\left(X_q^{(2)}\right)\right]^{-1/2}\right\}$.

Note that $\boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{x}^{(1)}} = \boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{z}^{(1)}}$ and so forth since

$$\boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{x}^{(1)}} = \boldsymbol{A}\boldsymbol{\Sigma}_{11}\boldsymbol{V}_{11}^{-\frac{1}{2}} = \boldsymbol{A}\boldsymbol{V}_{11}^{\frac{1}{2}}\boldsymbol{V}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{11}\boldsymbol{V}_{11}^{-\frac{1}{2}} = \boldsymbol{A}_{\boldsymbol{z}}\boldsymbol{\rho}_{11} = \rho_{\boldsymbol{U},\boldsymbol{z}^{(1)}},$$

where $\boldsymbol{z}^{(1)}$is the standardized variable and see comment at p.542 for $\boldsymbol{A}\boldsymbol{V}_{11}^{\frac{1}{2}} = \boldsymbol{A}_{\boldsymbol{z}}$. **The correlations are unaffected by the standardization**.

Remark

- The correlations $\boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{x}^{(1)}}$ and $\boldsymbol{\rho}_{\boldsymbol{V},\boldsymbol{x}^{(2)}}$ can help supply meanings for the canonical variables.
- The spirit is the same as in principal component analysis when the correlations between the principal components and their associated variables may provide subject-matter interpretations for the components.

9.2.2 Canonical correlations as generalizations of other correlation coefficients

$\boldsymbol{x}^{(1)};\ p \times 1\ ,\ \boldsymbol{x}^{(2)};\ q \times 1$

① $p = q = 1$

$$\left|Corr\left(X_1^{(1)}, X_1^{(2)}\right)\right| = \left|Corr\left(aX_1^{(1)}, bX_1^{(2)}\right)\right|^{\forall} a, b \neq 0$$

$\therefore U_1 = X_1^{(1)}$ & $V_1 = X_1^{(2)}$ have correlation

$$\rho_1^* = \left|Corr\left(X_1^{(1)}, X_1^{(2)}\right)\right|.$$

② $\boldsymbol{a}^\top = (0\ \cdots\ 0\ \underset{i^{th}}{1}\ 0\ \cdots\ 0),\ \boldsymbol{b}^\top = (0\ \cdots\ 0\ \underset{k^{th}}{1}\ 0\ \cdots\ 0)$

$$\begin{aligned}\left|Corr\left(X_i^{(1)}, X_k^{(2)}\right)\right| &= \left|Corr\left(\boldsymbol{a}^\top\boldsymbol{x}^{(1)}, \boldsymbol{b}^\top\boldsymbol{x}^{(2)}\right)\right| \\ &\leq \max_{\boldsymbol{a},\boldsymbol{b}} Corr\left(\boldsymbol{a}^\top\boldsymbol{x}^{(1)}, \boldsymbol{b}^\top\boldsymbol{x}^{(2)}\right) = \rho_1^*\end{aligned}$$

That is $\rho_1^* \geq |\rho_{ij}|$, where $\boldsymbol{\rho} = \{\rho_{ij}\} = \boldsymbol{V}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{V}_{22}^{-\frac{1}{2}}$.

③ Multiple correlation coefficient is a special case of a canonical correlation when $\boldsymbol{x}^{(1)} = X_1^{(1)}(p = 1)$. That is,

$$\rho_{1\boldsymbol{x}^{(2)}} = \max_{\boldsymbol{b}} Corr\left(X_1^{(1)}, \boldsymbol{b}^\top \boldsymbol{x}^{(2)}\right) = \rho_1^*$$

for $p = 1$ from the Result 7.12 (pg. 402 of textbook).

Note Correlation is location invariant, but not scale invariant. That is,

$$Corr(aX + b, cY + d) \neq Corr(X, Y).$$

Since

$$Corr(aX + b, cY + d) = \frac{acCov(X, Y)}{|a||c|\sqrt{VX}\sqrt{VY}} = \frac{ac}{|ac|} Corr(X, Y).$$

As long as $ac > 0$, correlation is location-scale invariant.

9.3 The sample Canonical variables and correlations

A random sample of size n from $\underset{p\times 1}{\boldsymbol{x}^{(1)}}$, $\underset{q\times 1}{\boldsymbol{x}^{(2)}}$

$$\begin{aligned}
\boldsymbol{X} &= \left[\boldsymbol{X}^{(1)} \ \vdots \ \boldsymbol{X}^{(2)}\right] \\
&= \begin{bmatrix} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1p}^{(1)} & \vdots & x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1q}^{(2)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2p}^{(1)} & \vdots & x_{21}^{(2)} & x_{22}^{(2)} & \cdots & x_{2q}^{(2)} \\ \vdots & & & & \vdots & & & & \vdots \\ x_{n1}^{(1)} & x_{n2}^{(1)} & \cdots & x_{np}^{(1)} & \vdots & x_{n1}^{(2)} & x_{n2}^{(2)} & \cdots & x_{nq}^{(2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^{(1)'} & \vdots & \boldsymbol{x}_1^{(2)'} \\ \boldsymbol{x}_2^{(1)'} & \vdots & \boldsymbol{x}_2^{(2)'} \\ \vdots & \vdots & \vdots \\ \boldsymbol{x}_n^{(1)'} & \vdots & \boldsymbol{x}_n^{(2)'} \end{bmatrix}
\end{aligned}$$

$$\bar{\boldsymbol{x}} = \begin{bmatrix} \bar{\boldsymbol{x}}^{(1)} \\ \cdots \\ \bar{\boldsymbol{x}}^{(2)} \end{bmatrix}, \text{ where } \begin{array}{l} \bar{\boldsymbol{x}}^{(1)} = \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_j^{(1)} \\ \bar{\boldsymbol{x}}^{(2)} = \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_j^{(2)} \end{array}$$

$$\underset{(p+q)\times(p+q)}{\boldsymbol{S}} = \begin{bmatrix} \underset{p\times p}{\boldsymbol{S}_{11}} & \vdots & \underset{p\times q}{\boldsymbol{S}_{12}} \\ \cdots & \cdots & \cdots \\ \underset{q\times p}{\boldsymbol{S}_{21}} & \vdots & \underset{q\times q}{\boldsymbol{S}_{22}} \end{bmatrix}$$

$$\boldsymbol{S}_{kl} = \frac{1}{n-1}\sum_{j=1}^n \left(\boldsymbol{x}_j^{(k)} - \bar{\boldsymbol{x}}^{(k)}\right)\left(\boldsymbol{x}_j^{(l)} - \bar{\boldsymbol{x}}^{(l)}\right)^\top, \quad k = 1, 2.$$

The linear combinations

$$\widehat{U} = \widehat{\boldsymbol{a}}^\top \boldsymbol{x}^{(1)}; \quad \widehat{V} = \widehat{\boldsymbol{b}}^\top \boldsymbol{x}^{(2)}$$

have sample correlation

$$r_{\widehat{U},\widehat{V}} = \frac{\widehat{\boldsymbol{a}}^\top \boldsymbol{S}_{12}\widehat{\boldsymbol{b}}}{\sqrt{\widehat{\boldsymbol{a}}^\top \boldsymbol{S}_{11}\widehat{\boldsymbol{a}}}\sqrt{\widehat{\boldsymbol{b}}^\top \boldsymbol{S}_{22}\boldsymbol{b}}}.$$

Idea and procedure are the same as those of population case.

<u>Result</u> $\widehat{\rho}_1^{*2} \geq \widehat{\rho}_2^{*2} \geq \cdots \geq \widehat{\rho}_p^{*2}$; ordered eigenvalues of

$$\boldsymbol{S}_{11}^{-\frac{1}{2}}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-\frac{1}{2}}$$

with corresponding eigenvectors

$$\widehat{\boldsymbol{e}}_1, \ \widehat{\boldsymbol{e}}_2, \ \cdots, \ \widehat{\boldsymbol{e}}_p \ \& \ p \leq q$$

$\widehat{\boldsymbol{f}}_1, \ \widehat{\boldsymbol{f}}_2, \ \cdots, \ \widehat{\boldsymbol{f}}_p$; eigenvectors of $\boldsymbol{S}_{22}^{-\frac{1}{2}}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-\frac{1}{2}}$ where the first p $\boldsymbol{f}'s$ may be obtained from

$$\widehat{\boldsymbol{f}}_k = (1/\widehat{\rho}_k^*)\boldsymbol{S}_{22}^{-\frac{1}{2}}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-\frac{1}{2}}\widehat{\boldsymbol{e}}_k, \ k = 1, \cdots, p.$$

$\Rightarrow$

$\widehat{U}_k = \boldsymbol{e}_k^\top \boldsymbol{S}_{11}^{-\frac{1}{2}}\boldsymbol{x}^{(1)}, \ \widehat{V}_k = \boldsymbol{f}_k^\top \boldsymbol{S}_{22}^{-\frac{1}{2}}\boldsymbol{x}^{(2)}$; canonical variables &

$r_{\widehat{U}_k,\widehat{V}_k} = \widehat{\rho}^*_k$; the largest possible correlation among linear combinations uncorrelated with the preceding $k-1$ sample canonical variates.

$\widehat{\rho}^*_1,\ \widehat{\rho}^*_2, \cdots,\ \widehat{\rho}^*_p$; sample canonical correlations.

pf) The proof follows the pf of Result 10.1, with $\boldsymbol{S}_{kl}$ substituted for $\boldsymbol{\Sigma}_{kl},\ k,l = 1,2$. □

Remark

① The sample canonical variates have unit sample variances.

$$\boldsymbol{S}_{\widehat{U}_k,\widehat{U}_k} = \boldsymbol{S}_{\widehat{V}_k,\widehat{V}_k} = 1$$

② Sample correlations are

$$r_{\widehat{U}_k,\widehat{U}_l} = r_{\widehat{V}_k,\widehat{V}_l} = r_{\widehat{U}_k,\widehat{V}_l} = 0,\ k \neq l$$

③

$$\begin{aligned}
&\underset{p\times p}{\widehat{\boldsymbol{A}}} = (\widehat{\boldsymbol{a}}_1\ \widehat{\boldsymbol{a}}_2 \cdots \widehat{\boldsymbol{a}}_p)^\top,\ \underset{q\times q}{\widehat{B}} = (\widehat{\boldsymbol{b}}_1\ \widehat{\boldsymbol{b}}_2 \cdots \widehat{\boldsymbol{b}}_q)^\top \\
\Rightarrow\ & \\
&\underset{p\times 1}{\widehat{\boldsymbol{u}}} = \widehat{\boldsymbol{A}}\boldsymbol{x}^{(1)},\quad \underset{q\times 1}{\widehat{\boldsymbol{v}}} = \widehat{\boldsymbol{B}}\boldsymbol{x}^{(2)} \\
\Rightarrow\ & \\
&\boldsymbol{R}_{\widehat{\mathbf{u}},\boldsymbol{x}^{(1)}} = \widehat{\boldsymbol{A}}\boldsymbol{S}_{11}\boldsymbol{D}_{11}^{-\frac{1}{2}};\ \text{matrix of sample correlation of }\ \widehat{\boldsymbol{u}}\text{ with }\boldsymbol{x}^{(1)} \\
&\boldsymbol{R}_{\widehat{\mathbf{v}},\boldsymbol{x}^{(2)}} = \widehat{\boldsymbol{B}}\boldsymbol{S}_{22}\boldsymbol{D}_{22}^{-\frac{1}{2}};\ \text{matrix of sample correlation of }\ \widehat{\boldsymbol{v}}\text{ with }\boldsymbol{x}^{(2)} \\
&\boldsymbol{R}_{\widehat{\mathbf{u}},\boldsymbol{x}^{(2)}} = \widehat{\boldsymbol{A}}\boldsymbol{S}_{12}\boldsymbol{D}_{22}^{-\frac{1}{2}};\ \text{matrix of sample correlation of }\ \widehat{\boldsymbol{u}}\text{ with }\boldsymbol{x}^{(2)} \\
&\boldsymbol{R}_{\widehat{\mathbf{v}},\boldsymbol{x}^{(1)}} = \widehat{\boldsymbol{B}}\boldsymbol{S}_{21}\boldsymbol{D}_{11}^{-\frac{1}{2}};\ \text{matrix of sample correlation of }\ \widehat{\boldsymbol{v}}\text{ with }\boldsymbol{x}^{(1)},
\end{aligned}$$

where

$$\boldsymbol{D}_{11}^{-\frac{1}{2}} = diag\left\{\left[\text{sample variance of } x_i^{(1)}\right]^{-\frac{1}{2}}\right\}_{p\times p}$$

and

$$\boldsymbol{D}_{22}^{-\frac{1}{2}} = diag\left\{\left[\text{sample variance of } x_i^{(2)}\right]^{-\frac{1}{2}}\right\}_{q\times q}.$$

④ If the observations are standardized,

$$\boldsymbol{Z} \;=\; [\boldsymbol{Z}^{(1)} \vdots \boldsymbol{Z}^{(2)}] = \begin{bmatrix} \boldsymbol{z}_1^{(1)'} & \vdots & \boldsymbol{z}_1^{(2)'} \\ \vdots & \vdots & \vdots \\ \boldsymbol{z}_n^{(1)'} & \vdots & \boldsymbol{z}_n^{(2)'} \end{bmatrix}$$

⇒
Sample canonical variates;

$$\underset{p\times 1}{\widehat{\boldsymbol{u}}} = \widehat{\boldsymbol{A}}_z \boldsymbol{z}^{(1)}, \; \underset{q\times 1}{\widehat{\boldsymbol{v}}} = \widehat{\boldsymbol{B}}_z \boldsymbol{z}^{(2)}, \text{ where}$$
$$\widehat{\boldsymbol{A}}_z = \widehat{\boldsymbol{A}}\boldsymbol{D}_{11}^{\frac{1}{2}} \;\&\; \widehat{\boldsymbol{B}}_z = \widehat{\boldsymbol{B}}\boldsymbol{D}_{22}^{\frac{1}{2}}.$$

Note that **the correlations at ③ remain unchanged & the sample canonical correlations are unaffected by standardization**. For example,

$$\begin{aligned} \boldsymbol{R}_{\widehat{\boldsymbol{u}},\boldsymbol{x}^{(1)}} &= \widehat{\boldsymbol{A}}\boldsymbol{S}_{11}\boldsymbol{D}_{11}^{-\frac{1}{2}} = \widehat{\boldsymbol{A}}\boldsymbol{D}_{11}^{\frac{1}{2}}\boldsymbol{D}_{11}^{-\frac{1}{2}}\boldsymbol{S}_{11}\boldsymbol{D}_{11}^{-\frac{1}{2}} \\ &= \widehat{\boldsymbol{A}}_z\boldsymbol{R}_{11} = \boldsymbol{R}_{\widehat{\boldsymbol{u}},\boldsymbol{z}^{(1)}}. \end{aligned}$$

ex) Example 10.4 (Canonical correlation analysis of the chicken-bone data)

$$\begin{array}{rclll} \widehat{\rho}_1^* & = & 0.631 & \widehat{U}_1 = 0.781 z_1^{(1)} + 0.345 z_2^{(1)} \\ & & & \widehat{V}_1 = 0.060 z_1^{(2)} + 0.944 z_2^{(2)} \\ \widehat{\rho}_2^* & = & 0.057 & \widehat{U}_2 = -0.856 z_1^{(1)} + 1.106 z_2^{(1)} \\ & & & \widehat{V}_2 = -2.648 z_1^{(2)} + 2.475 z_2^{(2)} \end{array}$$

For example,

$$\boldsymbol{R}_{\widehat{\boldsymbol{u}},\boldsymbol{x}^{(1)}} \;=\; \begin{array}{cc} \begin{bmatrix} 0.9548 & -0.2974 \\ 0.7388 & 0.6739 \end{bmatrix} & \begin{array}{c} z_1^{(1)} \\ z_2^{(1)} \end{array} \\ \begin{array}{cc} U_1 & \qquad U_2 \end{array} & \end{array}$$

See also Panel 10.1.

9.4 Additional Sample Descriptive Measures

9.4.1 Matrices of Errors of Approximations

Given the matrices $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{B}}$ before, we know that $\hat{\boldsymbol{U}} = \hat{\boldsymbol{A}}\boldsymbol{x}^{(1)}$ and $\hat{\boldsymbol{V}} = \hat{\boldsymbol{B}}\boldsymbol{x}^{(2)}$. Hence, it is the same as

$$\underset{(p\times 1)}{\boldsymbol{x}^{(1)}} = \underset{(p\times p)}{\hat{\boldsymbol{A}}^{-1}}\underset{(p\times 1)}{\hat{\boldsymbol{U}}} \qquad \underset{(q\times 1)}{\boldsymbol{x}^{(2)}} = \underset{(q\times q)}{\hat{\boldsymbol{B}}^{-1}}\underset{(q\times 1)}{\hat{\boldsymbol{V}}}.$$

Since sample $Cov(\hat{\boldsymbol{U}}, \hat{\boldsymbol{V}}) = \hat{\boldsymbol{A}}\boldsymbol{S_{12}}\hat{\boldsymbol{B}}^T$, sample $Cov(\hat{\boldsymbol{U}}) = \hat{\boldsymbol{A}}\boldsymbol{S_{11}}\hat{\boldsymbol{A}}^T = \boldsymbol{I}_p$ and sample $Cov(\hat{\boldsymbol{V}}) = \hat{\boldsymbol{B}}\boldsymbol{S_{22}}\hat{\boldsymbol{B}}^T = \boldsymbol{I}_q$,

$$\begin{aligned}
\boldsymbol{S}_{12} &= \hat{\boldsymbol{A}}^{-1} diag(\hat{\boldsymbol{\rho}}^*)(\hat{\boldsymbol{B}}^{-1})^T \\
&= \hat{\rho}_1^*\hat{\boldsymbol{a}}^{(1)}\hat{\boldsymbol{b}}^{(1)'} + \hat{\rho}_2^*\hat{\boldsymbol{a}}^{(2)}\hat{\boldsymbol{b}}^{(2)'} + \ \ldots\ + \hat{\rho}_p^*\hat{\boldsymbol{a}}^{(p)}\hat{\boldsymbol{b}}^{(p)'} \\
\boldsymbol{S}_{11} &= \hat{\boldsymbol{A}}^{-1}(\hat{\boldsymbol{A}}^{-1})^T \\
&= \hat{\boldsymbol{a}}^{(1)}\hat{\boldsymbol{a}}^{(1)'} + \hat{\boldsymbol{a}}^{(2)}\hat{\boldsymbol{a}}^{(2)'} + \ \ldots\ + \hat{\boldsymbol{a}}^{(p)}\hat{\boldsymbol{a}}^{(p)'} \\
\boldsymbol{S}_{22} &= \hat{\boldsymbol{B}}^{-1}(\hat{\boldsymbol{B}}^{-1})^T \\
&= \hat{\boldsymbol{a}}^{(1)}\hat{\boldsymbol{a}}^{(1)'} + \hat{\boldsymbol{b}}^{(2)}\hat{\boldsymbol{b}}^{(2)'} + \ \ldots\ + \hat{\boldsymbol{b}}^{(p)}\hat{\boldsymbol{b}}^{(p)'},
\end{aligned}$$

where $\hat{\boldsymbol{a}}^{(i)}$ and $\hat{\boldsymbol{b}}^{(i)}$ denote the ith column of $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{B}}$ respectively. If we use only first r canonical pairs,

$$\tilde{\boldsymbol{x}}^{(1)} = \begin{bmatrix} \hat{\boldsymbol{a}}^{(1)} & \hat{\boldsymbol{a}}^{(2)} & \ldots & \hat{\boldsymbol{a}}^{(r)} \end{bmatrix} \begin{bmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \vdots \\ \hat{U}_r \end{bmatrix}$$

and

$$\tilde{\boldsymbol{x}}^{(2)} = \begin{bmatrix} \hat{\boldsymbol{b}}^{(1)} & \hat{\boldsymbol{b}}^{(2)} & \ldots & \hat{\boldsymbol{b}}^{(r)} \end{bmatrix} \begin{bmatrix} \hat{V}_1 \\ \hat{V}_2 \\ \vdots \\ \hat{V}_r \end{bmatrix}.$$

Then, we can approximate $\boldsymbol{S}_{12}$ with using $Cov(\tilde{\boldsymbol{x}}^{(1)}, \tilde{\boldsymbol{x}}^{(2)})$. Hence the *matrics of errors of approximation* becomes

$$\begin{aligned}
\boldsymbol{S}_{12} - (\hat{\rho}_1^*\hat{\boldsymbol{a}}^{(1)}\hat{\boldsymbol{b}}^{(1)'} + \hat{\rho}_2^*\hat{\boldsymbol{a}}^{(2)}\hat{\boldsymbol{b}}^{(2)'} + \ \ldots\ + \hat{\rho}_r^*\hat{\boldsymbol{a}}^{(r)}\hat{\boldsymbol{b}}^{(r)'}) = \hat{\rho}_{r+1}^*\hat{\boldsymbol{a}}^{(r+1)}\hat{\boldsymbol{b}}^{(r+1)'} + \\
\hat{\rho}_{r+2}^*\hat{\boldsymbol{a}}^{(r+2)}\hat{\boldsymbol{b}}^{(r+2)'} + \ \ldots\ + \hat{\rho}_p^*\hat{\boldsymbol{a}}^{(p)}\hat{\boldsymbol{b}}^{(p)'}
\end{aligned}$$

Same for $\boldsymbol{S}_{11}$ and $\boldsymbol{S}_{22}$.

9.4.2 Proportions of Explained Sample Variance

Recall that

$$\underset{p\times 1}{\widehat{\boldsymbol{u}}} = \underset{p\times p}{\widehat{\boldsymbol{A}}_z}\underset{p\times 1}{\boldsymbol{z}}^{(1)},\ \underset{q\times 1}{\widehat{\boldsymbol{v}}} = \underset{q\times q}{\widehat{\boldsymbol{B}}_z}\underset{q\times 1}{\boldsymbol{z}}^{(2)}$$

$$Cov(\boldsymbol{z}^{(1)}, \widehat{\boldsymbol{u}}) \;=\; Cov(\widehat{\boldsymbol{A}}_z^{-1}\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{u}}) = \widehat{\boldsymbol{A}}_z^{-1}$$

since $Cov(\widehat{\boldsymbol{u}}) = \boldsymbol{I}$. Furthermore, $VZ_i^{(1)} = 1$ by standardization. Hence

$$Cov(\boldsymbol{z}^{(1)}, \widehat{\boldsymbol{u}}) = \widehat{\boldsymbol{A}}_z^{-1} = Corr(\boldsymbol{z}^{(1)}, \widehat{\boldsymbol{u}})$$

and

$$\underset{p\times p}{\widehat{\boldsymbol{A}}_z^{-1}} = (\widehat{\boldsymbol{a}}_z^{(1)}\ \widehat{\boldsymbol{a}}_z^{(2)}\ \cdots\ \widehat{\boldsymbol{a}}_z^{(p)}) = \{r_{z_i^{(1)}, \widehat{u}_j}\}.$$

Similarly,

$$\begin{aligned} Cov(\boldsymbol{z}^{(2)}, \widehat{\boldsymbol{v}}) &= Cov(\widehat{\boldsymbol{B}}_z^{-1}\widehat{\boldsymbol{v}}, \widehat{\boldsymbol{v}}) = \widehat{\boldsymbol{B}}_z^{-1} \\ \widehat{\boldsymbol{B}}_z^{-1} &= (\widehat{\boldsymbol{b}}_z^{(1)}\ \widehat{\boldsymbol{b}}_z^{(2)}\ \cdots\ \widehat{\boldsymbol{b}}_z^{(q)}) = \{r_{z_i^{(2)}, \widehat{v}_j}\}_{q\times q}. \end{aligned}$$

Note that

$$Cov(\widehat{\boldsymbol{u}}) = Cov(\widehat{\boldsymbol{A}}_z\boldsymbol{z}^{(1)}) = \widehat{\boldsymbol{A}}_z\boldsymbol{R}_{11}\widehat{\boldsymbol{A}}_z^{\top} = \boldsymbol{I}$$

by the definition of $\widehat{\boldsymbol{u}}$. Hence

$$\begin{aligned} \boldsymbol{R}_{11} &= \widehat{\boldsymbol{A}}_z^{-1}\left(\widehat{\boldsymbol{A}}_z^{-1}\right)^{\top} \\ &= \left(\widehat{\boldsymbol{a}}_z^{(1)}\ \widehat{\boldsymbol{a}}_z^{(2)}\ \cdots\ \widehat{\boldsymbol{a}}_z^{(p)}\right)\left(\widehat{\boldsymbol{a}}_z^{(1)}\ \widehat{\boldsymbol{a}}_z^{(2)}\ \cdots\ \widehat{\boldsymbol{a}}_z^{(p)}\right)^{\top} \\ &= \sum_{i=1}^{p}\widehat{\boldsymbol{a}}_z^{(i)}\widehat{\boldsymbol{a}}_z^{(i)\top} \end{aligned}$$

Total sample variance in the first set $= tr(\underset{p\times p}{\boldsymbol{R}_{11}}) = p$.

Similarly, total sample variance in the second set $= tr(\underset{p\times p}{\boldsymbol{R}_{22}}) = q$.

Contributions of the first r canonical $\left(\widehat{U_1}\cdots\widehat{U_r}\right)$ variates is

$$tr\left(\sum_{i=1}^{r}\widehat{\boldsymbol{a}}_z^{(i)}\widehat{\boldsymbol{a}}_z^{(i)'}\right)=tr\left(\widehat{\boldsymbol{a}}_z^{(1)}\widehat{\boldsymbol{a}}_z^{(1)'}+\cdots+\widehat{\boldsymbol{a}}_z^{(r)}\widehat{\boldsymbol{a}}_z^{(r)'}\right)=\sum_{i=1}^{r}\sum_{k=1}^{p}r^2_{z_k^{(1)},\widehat{U}_i},$$

where

$$\begin{aligned}\widehat{\boldsymbol{A}}_z^{-1} &= (\widehat{\boldsymbol{a}}_z^{(1)}\ \widehat{\boldsymbol{a}}_z^{(2)}\ \cdots\ \widehat{\boldsymbol{a}}_z^{(p)})=\{r_{z_i^{(1)},\widehat{U}_j}\}\\ &= \begin{bmatrix} r_{z_1^{(1)},\widehat{U}_1} & r_{z_1^{(1)},\widehat{U}_2} & \cdots & r_{z_1^{(1)},\widehat{U}_p}\\ r_{z_2^{(1)},\widehat{U}_1} & r_{z_2^{(1)},\widehat{U}_2} & \cdots & r_{z_2^{(1)},\widehat{U}_p}\\ \vdots & \vdots & \ddots & \vdots\\ r_{z_p^{(1)},\widehat{U}_1} & r_{z_p^{(1)},\widehat{U}_2} & \cdots & r_{z_p^{(1)},\widehat{U}_p}\end{bmatrix}.\end{aligned}$$

Similarly, contributions of the first r canonical $\left(\widehat{V}_1\cdots\widehat{V}_r\right)$ variates is

$$tr\left(\sum_{i=1}^{r}\widehat{\boldsymbol{b}}_z^{(i)}\widehat{\boldsymbol{b}}_z^{(i)'}\right)=tr\left(\widehat{\boldsymbol{b}}_z^{(1)}\widehat{\boldsymbol{b}}_z^{(1)'}+\cdots+\widehat{\boldsymbol{b}}_z^{(r)}\widehat{\boldsymbol{b}}_z^{(r)'}\right)=\sum_{i=1}^{r}\sum_{k=1}^{p}r^2_{z_k^{(2)},\widehat{V}_i}.$$

Proportion of total standardized sample variance in first set explained by $\widehat{U}_1,\widehat{U}_2,\cdots,\widehat{U}_r$ is

$$R^2_{\boldsymbol{z}^{(1)}|\widehat{U}_1,\cdots,\widehat{U}_r}=\frac{\sum_{i=1}^{r}\sum_{k=1}^{p}r^2_{z_k^{(1)},\widehat{U}_i}}{p}$$

Proportion of total standardized sample variance in second set explained by $\widehat{V}_1,\widehat{V}_2,\cdots,\widehat{V}_r$ is

$$R^2_{\boldsymbol{z}^{(2)}|\widehat{V}_1,\cdots,\widehat{V}_r}=\frac{\sum_{i=1}^{r}\sum_{k=1}^{p}r^2_{z_k^{(2)},\widehat{V}_i}}{q}.$$

ex) See example 10.7 (Calculating proportions of sample variance explained by canonical variates)

9.5 Large Sample inference

$$\boldsymbol{\Sigma}_{12} = \mathbf{0} \ \Rightarrow \ Cov\left(\boldsymbol{a}'\boldsymbol{x}^{(1)}, \boldsymbol{b}'\boldsymbol{x}^{(2)}\right) = \boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b} = 0 \ \ \forall \boldsymbol{a}, \boldsymbol{b}$$

No need to pursue a canonical analysis.

Result

$$\boldsymbol{x}_j = \begin{bmatrix} \boldsymbol{x}_j^{(1)} \\ \boldsymbol{x}_j^{(2)} \end{bmatrix}, j = 1, \cdots, n \text{ a random sample from } \ N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

LRT of $H_0 : \boldsymbol{\Sigma}_{12} = \mathbf{0}$ $v.s.$ $H_1 : \boldsymbol{\Sigma}_{12} \neq \mathbf{0}$ rejects H_0 for large values of

$$-2\ln\Lambda = n\ln\left(\frac{|\boldsymbol{S}_{11}||\boldsymbol{S}_{22}|}{|\boldsymbol{S}|}\right) = -n\ln\prod_{i=1}^{p}\left(1 - \widehat{\rho}_i^{*2}\right),$$

where

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{bmatrix}$$

is an unbiased estimator of $\boldsymbol{\Sigma}$.

For large n, $-2\ln\Lambda \ \dot{\sim}\ \chi^2(pq)$.

Bartlett's modification;

$$-\left(n - 1 - \frac{1}{2}(p+q+1)\right)\sum_{i=1}^{p}\ln\left(1 - \widehat{\rho}_i^{*2}\right) \ \dot{\sim}\ \chi^2_{pq}$$

Remark

① $H_0 : \boldsymbol{\Sigma}_{12} = \mathbf{0}$ $(\rho_1^* = \rho_2^* = \cdots = \rho_p^* = 0)$

② After $H_0 : \boldsymbol{\Sigma}_{12} = \mathbf{0}$ is rejected, we want to test $H_0^{(k)} : \rho_1^* \neq 0, \cdots, \rho_k^* \neq 0, \rho_{k+1}^* = \cdots = \rho_p^* = 0$.

Bartlett; Reject $H_0^{(k)}$ if

$$-\left(n - 1 - \frac{1}{2}(p+q+1)\right)\sum_{i=k+1}^{p}\ln\left(1 - \widehat{\rho}_i^{*2}\right) > \chi^2_{(p-k)(q-k)}(\alpha)$$

9.6 Canonical Correlation in R

There are several packages we can use in R. For example,

(1) Package 'CCA' : https://cran.r-project.org/web/packages/CCA/CCA.pdf

(2) Package 'candisc' : ftp://cran.r-project.org/pub/R/web/packages/candisc/candisc.pdf

But the main defect of these package is that we have to use only data matrix for input arguments.

Some book like Johnson and Wichern (2007) only gives the correlation matrix. Hence here all procedures are done without using those packages.

Example 9.6.1 *(Calculating canonical variates and canonical correlation for standardized variables)*

Suppose $\boldsymbol{Z}^{(1)} = [Z_1^{(1)}, Z_2^{(1)}]^T$ and $\boldsymbol{Z}^{(2)} = [Z_1^{(2)}, Z_2^{(2)}]^T$ are standardized variables of $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$. Let $\boldsymbol{Z} = [\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}]^T$ and

$$Cov(\boldsymbol{Z}) = Corr(\boldsymbol{x}) = \left[\begin{array}{c|c} \boldsymbol{\rho}_{11} & \boldsymbol{\rho}_{12} \\ \hline \boldsymbol{\rho}_{21} & \boldsymbol{\rho}_{22} \end{array}\right] = \left[\begin{array}{cc|cc} 1.0 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1.0 & 0.3 & 0.4 \\ \hline 0.5 & 0.3 & 1.0 & 0.2 \\ 0.6 & 0.4 & 0.2 & 1.0 \end{array}\right].$$

```
# example 10.1
rho11 <- matrix(c(1,0.4,0.4,1),nrow=2)
rho12 <- matrix(c(0.5,0.3,0.6,0.4),nrow=2)
rho21 <- t(rho12)
rho22 <- matrix(c(1,0.2,0.2,1),nrow=2)

rho <- rbind(cbind(rho11,rho12),cbind(rho21,rho22))

> rho
[,1] [,2] [,3] [,4]
[1,]  1.0  0.4  0.5  0.6
[2,]  0.4  1.0  0.3  0.4
[3,]  0.5  0.3  1.0  0.2
```

```
[4,]  0.6  0.4  0.2  1.0
```

Then

$$\boldsymbol{\rho}_{11}^{-1/2} = \begin{bmatrix} 1.0681 & -0.2229 \\ -0.2229 & 1.0681 \end{bmatrix}$$

$$\boldsymbol{\rho}_{22}^{-1} = \begin{bmatrix} 1.0417 & -0.2083 \\ -0.2083 & 1.0417 \end{bmatrix}$$

In R,

```
# inverse and sqrt inverse matrix of rho11
inv.rho11 <- solve(rho11) # inverse matrix of rho_11

eigen(inv.rho11)
V11 <- eigen(inv.rho11)$vectors
S11 <- sqrt(diag(eigen(inv.rho11)$values))

sqrt.inv.rho11 <- V11%*%S11%*%solve(V11)
# sqrt inverse matrix of rho_11

# inverse and sqrt inverse matrix of rho22
inv.rho22 <- solve(rho22) # inverse matrix of rho_22

eigen(inv.rho22)
V22 <- eigen(inv.rho22)$vectors
S22 <- sqrt(diag(eigen(inv.rho22)$values))

sqrt.inv.rho22 <- V22%*%S22%*%solve(V22)
# sqrt inverse matrix of rho_22

> sqrt.inv.rho11
[,1]       [,2]
[1,]  1.0680744 -0.2229201
[2,] -0.2229201  1.0680744

> inv.rho22
[,1]       [,2]
```

```
[1,]  1.0416667 -0.2083333
[2,] -0.2083333  1.0416667
```

- We have the another way to get square root matrix using the package ‘expm’.

```
> sqrtm(inv.rho11)
           [,1]       [,2]
[1,]  1.0680744 -0.2229201
[2,] -0.2229201  1.0680744
```

- The Spectral decomposition is used for getting the square root inverse matrix but this package uses Schur decomposition. By the way the results are the same.

And

$$\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2} = \begin{bmatrix} 0.4371 & 0.2178 \\ 0.2178 & 0.1096 \end{bmatrix}.$$

The eigenvalues, ρ_1^{*2} and ρ_2^{*2}, are obtained from solving

$$\begin{aligned} 0 = \begin{vmatrix} 0.4371-\lambda & 0.2178 \\ 0.2178 & 0.1096-\lambda \end{vmatrix} &= (0.4371-\lambda)(0.1096-\lambda)-(0.2178)^2 \\ &= \lambda^2 - 0.5467\lambda + 0.0005. \end{aligned}$$

They are $\rho_1^{*2} = 0.5458$ and $\rho_2^{*2} = 0.0009$. The eigenvector $\boldsymbol{e}_1$ follows from the vector equation

$$\begin{bmatrix} 0.4371 & 0.2178 \\ 0.2178 & 0.1096 \end{bmatrix} \boldsymbol{e}_1 = (0.5458)\boldsymbol{e}_1.$$

Thus, $\boldsymbol{e}_1 = [0.8947,\ 0.4466]^T$ and in R,

```
## calculation of rho star vector ##
rho.star <- sqrt.inv.rho11%*%rho12%*%inv.rho22%*%rho21%*%sqrt.inv.rho11
rho.star1 <- t(sqrt.inv.rho22%*%rho21%*%sqrt.inv.rho11)%*%
               (sqrt.inv.rho22%*%rho21%*%sqrt.inv.rho11)

eigen(rho.star)$values # rho star squared (eigenvalues)

eigen(rho.star)$vectors # eigenvectors
```

```
> eigen(rho.star)$values
[1] 0.5457180317 0.0009089525

> eigen(rho.star)$vectors
           [,1]       [,2]
[1,] -0.8946536  0.4467605
[2,] -0.4467605 -0.8946536
```

- As you can see before, values can be different because eigenvectors are not unique. Even though its length is 1.

And

$$\boldsymbol{a}_1 = \boldsymbol{\rho}_{12}^{-1/2}\boldsymbol{e}_1 = \begin{bmatrix} 0.8561 \\ 0.2776 \end{bmatrix}.$$

Note that $\boldsymbol{f}_1 \propto \boldsymbol{\rho}_{22}^{-1/2}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{e}_1$ and $\boldsymbol{b}_1 = \boldsymbol{\rho}_{21}^{-1/2}\boldsymbol{f}_1$. Consequently,

$$\boldsymbol{b}_1 \propto \boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{a}_1 = \begin{bmatrix} 0.3959 & 0.2292 \\ 0.5209 & 0.3542 \end{bmatrix} \begin{bmatrix} 0.8561 \\ 0.2776 \end{bmatrix} = \begin{bmatrix} 0.4026 \\ 0.5443 \end{bmatrix}.$$

We have to scale $\boldsymbol{b}_1$ to 1 ($\boldsymbol{b}_1^T\boldsymbol{b}_1 = 1$). So

$$\boldsymbol{b}_1 = \frac{1}{0.7389}\begin{bmatrix} 0.4026 \\ 0.5443 \end{bmatrix} = \begin{bmatrix} 0.5448 \\ 0.7366 \end{bmatrix}.$$

```
## calculation of canonical variates a1 ##

eigen(rho.star)$vectors

a1 <- sqrt.inv.rho11%*%eigen(rho.star)$vectors[,1]

## calculation of canonical variates b1 ##

b1 <- inv.rho22%*%rho21%*%a1

scale <- 1/sqrt(t(b1)%*%rho22%*%b1)

b1 <- b1%*%scale

> a1
```

```
          [,1]
[1,] -0.8559647
[2,] -0.2777371

> b1
          [,1]
[1,] -0.5448119
[2,] -0.7366455
```

- As the eigenvector is not unique, the absolute value is the same, but the sign is different.

Hence, the first pair of canonical variables is

$$\begin{aligned} U_1 &= \boldsymbol{a}_1^T \boldsymbol{Z}^{(1)} = 0.86 Z_1^{(1)} + 0.28 Z_2^{(1)} \\ V_1 &= \boldsymbol{b}_1^T \boldsymbol{Z}^{(2)} = 0.54 Z_1^{(2)} + 0.74 Z_2^{(2)} \end{aligned}$$

from Johnson and Wichern (2007) and

$$\begin{aligned} U_1 &= \boldsymbol{a}_1^T \boldsymbol{Z}^{(1)} = -0.86 Z_1^{(1)} - 0.28 Z_2^{(1)} \\ V_1 &= \boldsymbol{b}_1^T \boldsymbol{Z}^{(2)} = -0.54 Z_1^{(2)} - 0.74 Z_2^{(2)} \end{aligned}$$

from the R function. Their canonical correlation is

$$\rho_1^* = \sqrt{\rho_1^{*2}} = \sqrt{0.5458} = 0.74.$$

This is the largest correlation possible between linear combinations of variables from the $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ sets.

The second canonical correlation, $\rho_2^* = \sqrt{0.0009} = 0.03$ is very small, and it means that the second pair of canonical variables, although uncorrelated with members of the first pair, conveys very little information about the association between sets.

Example 9.6.2 *(Canonical correlation analysis of the chicken-bone data)*

- As the R process is same as before, the function called 'can.cor.func' is developed.

- The function returns out $\boldsymbol{\rho}^*$, $\boldsymbol{U}$ and $\boldsymbol{V}$.

- The usage and the function code is included at the end of this section.

So, for given sample correlation $\boldsymbol{R}$,

$$\boldsymbol{R} = \left[\begin{array}{c|c} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \hline \boldsymbol{R}_{21} & \boldsymbol{R}_{22} \end{array}\right] = \left[\begin{array}{cc|cc} 1.0 & 0.505 & 0.569 & 0.602 \\ 0.505 & 1.0 & 0.422 & 0.467 \\ \hline 0.569 & 0.422 & 1.0 & 0.926 \\ 0.602 & 0.467 & 0.926 & 1.0 \end{array}\right].$$

In R,

```
# example 10.4
rho11 <- matrix(c(1,0.505,0.505,1),nrow=2)
rho12 <- matrix(c(0.569,0.422,0.602,0.467),nrow=2)
rho21 <- t(rho12)
rho22 <- matrix(c(1,0.926,0.926,1),nrow=2)

rho <- rbind(cbind(rho11,rho12),cbind(rho21,rho22))

> rho
      [,1]  [,2]  [,3]  [,4]
[1,] 1.000 0.505 0.569 0.602
[2,] 0.505 1.000 0.422 0.467
[3,] 0.569 0.422 1.000 0.926
[4,] 0.602 0.467 0.926 1.000
```

So, using the function ‘ca.cor.func’, we get

$$\begin{array}{c} \hat{\rho}_1^* = 0.631 \\ \hat{\rho}_2^* = 0.057 \end{array}.$$

Corresponding canonical variates are

$$\begin{array}{c} \hat{U}_1 = \boldsymbol{a}_1^T \boldsymbol{Z}^{(1)} = -0.781 Z_1^{(1)} - 0.345 Z_2^{(1)} \\ \hat{V}_1 = \boldsymbol{b}_1^T \boldsymbol{Z}^{(2)} = -0.060 Z_1^{(2)} - 0.944 Z_2^{(2)} \end{array}$$

and

$$\begin{array}{c} \hat{U}_2 = \boldsymbol{a}_2^T \boldsymbol{Z}^{(1)} = 0.856 Z_1^{(1)} - 1.106 Z_2^{(1)} \\ \hat{V}_2 = \boldsymbol{b}_2^T \boldsymbol{Z}^{(2)} = 2.648 Z_1^{(2)} - 2.475 Z_2^{(2)} \end{array}.$$

And result from Johnson and Wichern (2007) is

$$\begin{aligned} \hat{\rho}_1^* &= 0.631 \\ \hat{\rho}_2^* &= 0.057 \end{aligned}$$

and corresponding canonical variates are

$$\begin{aligned} \hat{U}_1 &= \boldsymbol{a}_1^T \boldsymbol{Z}^{(1)} = 0.781 Z_1^{(1)} + 0.345 Z_2^{(1)} \\ \hat{V}_1 &= \boldsymbol{b}_1^T \boldsymbol{Z}^{(2)} = 0.060 Z_1^{(2)} + 0.944 Z_2^{(2)} \end{aligned}$$

and

$$\begin{aligned} \hat{U}_2 &= \boldsymbol{a}_2^T \boldsymbol{Z}^{(1)} = -0.856 Z_1^{(1)} + 1.106 Z_2^{(1)} \\ \hat{V}_2 &= \boldsymbol{b}_2^T \boldsymbol{Z}^{(2)} = -2.648 Z_1^{(2)} + 2.475 Z_2^{(2)} \end{aligned}$$

- We can see the sign of coefficients are different. But it is because eigenvector is not unique.

- So we can say that result is the same.

In R,

```
can.cor.func(R=rho,p=2)

> can.cor.func(R=rho,p=2)
$rho.star
[1] 0.63108502 0.05679406

$U
         [,1]       [,2]
U1 -0.7807924 -0.3445068
U2  0.8559732 -1.1061835

$V
          [,1]      [,2]
V1 -0.06025088 -0.943949
V2  2.64815634 -2.474939
```

Example 9.6.3 *(Calculating matrices of errors of approximation)*

For given sample correlation $\boldsymbol{R}$,

$$\boldsymbol{R} = \left[\begin{array}{c|c} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \hline \boldsymbol{R}_{21} & \boldsymbol{R}_{22} \end{array}\right] = \left[\begin{array}{cc|cc} 1.0 & 0.505 & 0.569 & 0.602 \\ 0.505 & 1.0 & 0.422 & 0.467 \\ \hline 0.569 & 0.422 & 1.0 & 0.926 \\ 0.602 & 0.467 & 0.926 & 1.0 \end{array}\right]$$

In R,

```
# example 10.6
rho11 <- matrix(c(1,0.505,0.505,1),nrow=2)
rho12 <- matrix(c(0.569,0.422,0.602,0.467),nrow=2)
rho21 <- t(rho12)
rho22 <- matrix(c(1,0.926,0.926,1),nrow=2)

rho <- rbind(cbind(rho11,rho12),cbind(rho21,rho22))

> rho
      [,1]  [,2]  [,3]  [,4]
[1,] 1.000 0.505 0.569 0.602
[2,] 0.505 1.000 0.422 0.467
[3,] 0.569 0.422 1.000 0.926
[4,] 0.602 0.467 0.926 1.000
```

By using the function 'can.cor.func', we get

$$\begin{aligned} \hat{\rho}_1^* &= 0.631 \\ \hat{\rho}_2^* &= 0.057 \end{aligned}$$

and corresponding canonical variates are

$$\begin{aligned} \hat{U}_1 &= \boldsymbol{a}_1^T \boldsymbol{Z}^{(1)} = -0.781 Z_1^{(1)} - 0.345 Z_2^{(1)} \\ \hat{V}_1 &= \boldsymbol{b}_1^T \boldsymbol{Z}^{(2)} = -0.060 Z_1^{(2)} - 0.944 Z_2^{(2)} \end{aligned}$$

and

$$\begin{aligned} \hat{U}_2 &= \boldsymbol{a}_2^T \boldsymbol{Z}^{(1)} = 0.856 Z_1^{(1)} - 1.106 Z_2^{(1)} \\ \hat{V}_2 &= \boldsymbol{b}_2^T \boldsymbol{Z}^{(2)} = 2.648 Z_1^{(2)} - 2.475 Z_2^{(2)} \end{aligned} .$$

So $\hat{\boldsymbol{A}}^{-1}$ and $\hat{\boldsymbol{B}}^{-1}$ are

$$\hat{\boldsymbol{A}}^{-1} = \begin{bmatrix} -0.781 & -0.345 \\ 0.856 & -1.106 \end{bmatrix}^{-1} = \begin{bmatrix} -0.955 & 0.297 \\ -0.739 & -0.674 \end{bmatrix}$$

$$\hat{\boldsymbol{B}}^{-1} = \begin{bmatrix} -0.060 & -0.944 \\ 2.648 & -2.475 \end{bmatrix}^{-1} = \begin{bmatrix} -0.934 & 0.356 \\ -1.000 & -0.023 \end{bmatrix}.$$

So the matrices of errors of approximation created by the 1st canonical pair are

$$\begin{aligned} \boldsymbol{R}_{12} - Corr(\tilde{\boldsymbol{x}}^{(1)}, \tilde{\boldsymbol{x}}^{(2)}) &= (0.057) \begin{bmatrix} 0.297 \\ -0.674 \end{bmatrix} \begin{bmatrix} 0.356 & -0.023 \end{bmatrix} \\ &= \begin{bmatrix} 0.006 & -0.000 \\ -0.014 & 0.001 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \boldsymbol{R}_{11} - Corr(\tilde{\boldsymbol{x}}^{(1)}) &= \begin{bmatrix} 0.297 \\ -0.674 \end{bmatrix} \begin{bmatrix} 0.297 & -0.674 \end{bmatrix} \\ &= \begin{bmatrix} 0.088 & -0.200 \\ -0.200 & 0.454 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \boldsymbol{R}_{22} - Corr(\tilde{\boldsymbol{x}}^{(2)}) &= \begin{bmatrix} 0.356 \\ -0.023 \end{bmatrix} \begin{bmatrix} 0.356 & -0.023 \end{bmatrix} \\ &= \begin{bmatrix} 0.127 & -0.008 \\ -0.008 & 0.001 \end{bmatrix}. \end{aligned}$$

And the proportions of sample variance explained by 1st canonical variates are

$$\frac{1}{2} tr[\boldsymbol{R}_{11} - \hat{\boldsymbol{a}}^{(1)} \hat{\boldsymbol{a}}^{(1)\prime}] = 0.27$$

$$\frac{1}{2} tr[\boldsymbol{R}_{22} - \hat{\boldsymbol{b}}^{(1)} \hat{\boldsymbol{b}}^{(1)\prime}] = 0.06$$

It means that the first sample covariates $\hat{U}_1$ accounts for 27% of the set's total sample variance and the first sample covariates $\hat{V}_1$ accounts for 6% of the set's total sample variance. In R,

```
# rho.star
rho.star.vec <- can.cor.func(R=rho,p=2)$rho.star

> rho.star.vec
[1] 0.63108502 0.05679406

# getting hat{a}, hat{b}
a.mat <- can.cor.func(R=rho,p=2)$U
b.mat <- can.cor.func(R=rho,p=2)$V

> a.mat
         [,1]       [,2]
U1 -0.7807924 -0.3445068
U2  0.8559732 -1.1061835

> b.mat
          [,1]      [,2]
V1 -0.06025088 -0.943949
V2  2.64815634 -2.474939

# getting inverse matrix
inv.a.mat <- solve(a.mat)
inv.b.mat <- solve(b.mat)

> inv.a.mat
             U1         U2
[1,] -0.9547684  0.2973505
[2,] -0.7388070 -0.6739171

> inv.b.mat
             V1          V2
[1,] -0.9343476  0.35636292
[2,] -0.9997413 -0.02274612

# matrix error calculation
R12error <- rho.star.vec[2]*inv.a.mat[,2]%*%t(inv.b.mat[,2])
R11error <- inv.a.mat[,2]%*%t(inv.a.mat[,2])
```

```
R22error <- inv.b.mat[,2]%*%t(inv.b.mat[,2])

> R12error
             [,1]          [,2]
[1,]  0.006018166 -0.0003841307
[2,] -0.013639608  0.0008705962

> R11error
            [,1]       [,2]
[1,]  0.08841733 -0.2003896
[2,] -0.20038958  0.4541642

> R22error
             [,1]          [,2]
[1,]  0.126994534 -0.0081058748
[2,] -0.008105875  0.0005173861

## total sample proportion of variance
proportion.u1 <- sum(diag(rho11-inv.a.mat[,1]%*%t(inv.a.mat[,1])))/2
proportion.v1 <- sum(diag(rho22-inv.b.mat[,1]%*%t(inv.b.mat[,1])))/2

> proportion.u1
[1] 0.2712908

> proportion.v1
[1] 0.06375596
```

9.6.1 R function ; can.cor.func

```
#####################################################################
#           Canonical correlation analysis function                 #
#####################################################################
#                     Input Arguments                               #
#####################################################################
# R : Full correlation matrix                                       #
# p : column length of rho11 matrix                                 #
#####################################################################
```

```
can.cor.func <- function(R,p){
as.matrix(R)
total <- dim(R)[1]
# matrix setting (partition)
rho11 <- R[1:p,1:p]
rho12 <- R[1:p,(p+1):total]
rho21 <- t(rho12)
rho22 <- R[(p+1):total,(p+1):total]

# calculation
## inverse and sqrt inverse matrix of rho11
inv.rho11 <- solve(rho11) # inverse matrix of rho_11

eigen(inv.rho11)
V11 <- eigen(inv.rho11)$vectors
S11 <- sqrt(diag(eigen(inv.rho11)$values))

sqrt.inv.rho11 <- V11%*%S11%*%solve(V11)
# sqrt inverse matrix of rho_11

## inverse and sqrt inverse matrix of rho22
inv.rho22 <- solve(rho22) # inverse matrix of rho_22

eigen(inv.rho22)
V22 <- eigen(inv.rho22)$vectors
S22 <- sqrt(diag(eigen(inv.rho22)$values))

sqrt.inv.rho22 <- V22%*%S22%*%solve(V22)
# sqrt inverse matrix of rho_22

## calculation of rho star vector ##
rho.star.mat <- sqrt.inv.rho11
                %*%rho12%*%inv.rho22%*%rho21%*%sqrt.inv.rho11
rho.star.vector <- sqrt(eigen(rho.star.mat)$values) # rho star vectors

## calculation of canonical variates ##
a.mat <- matrix(0,nrow=p,ncol=p)
b.mat <- matrix(0,nrow=p,ncol=(total-p))

for (i in 1:p){
a.mat[i,] <- sqrt.inv.rho11%*%eigen(rho.star.mat)$vectors[,i]
b.mat[i,] <- inv.rho22%*%rho21%*%a.mat[i,]
scale <- 1/sqrt(t(b.mat[i,])%*%rho22%*%b.mat[i,])
b.mat[i,] <- b.mat[i,]%*%scale
}
# index
u.idx <- paste0("U",c(1:p))
v.idx <- paste0("V",c(1:p))
# naming
rownames(a.mat) <- c(u.idx)
rownames(b.mat) <- c(v.idx)
###########################################
return(list(rho.star=rho.star.vector,U=a.mat,V=b.mat))
}
```

Chapter 10

Discrimination and classification

10.1 Introduction

Discrimination & classification are multivariate techniques concerned with separating distinct sets of objects (or observations) & allocating new objects to previously defined groups.

For example,
Group: good credit, bad credit
Input variables: age, education level, income, $\cdots$

The objectives are

- to make a rule, a function of input variables, which discriminate the two groups based on the given data,
- and, classify a new customer, who has only input variables, based on the rules constructed.

Examples:

	Populations (Groups)	Measurements on the individual
1.	Graduate school admission categories e.g. admit, deny	Scores on exams, GPA.
2.	Safe or unsafe recipients of a drug	Age, blood pressure, etc.
3.	Species of a flower	Measurements on a new specimen
4.	Hamilton or Madison as authors of the Federalist Papers	Frequency of words, length of sentences
5.	Loan applications, good or bad risk	Income, credit history

10.2 Classification with two populations

Suppose we make an observation, $\boldsymbol{x}$ on an individual & it is known that the individual comes from one of two p-variate populations.

We denote these populations by π_1 & π_2.

We want to develop a rule to help us to conclude that the individual comes form either of these populations.

That is, we want to specify a region, say R_1 such that if $\boldsymbol{x}$ is in R_1, we conclude that it came from π_1 &
if not, $\boldsymbol{x}$ is in the complement, $R_2 = R_1^c$, we conclude that it came from π_2.

For example, suppose $p = 1$ and $\mu_1 < \mu_2$, where μ_i is the expected value of i^{th} population, $i = 1, 2$.

The region R_1 would be defined by $x \leq c$ & R_2 defined by $x > c$.

The problem is the choice of c.

10.2.1 Fisher's method

Two populations with common variance, i.e.

$$\pi_1 : (\mu_1, \Sigma), \qquad \pi_2 : (\mu_2, \Sigma)$$

Goal: Observe $\boldsymbol{x}$ either from π_1 or π_2, and specify a classification rule (R_1, R_2), where

$\boldsymbol{x}$ is from π_1 if $\boldsymbol{x} \in R_1$,

$\boldsymbol{x}$ is from π_2 if $\boldsymbol{x} \in R_2$

with $R_1 \cup R_2 = \Re^p$ and $R_1 \cap R_2 = \phi$.

Idea: Transform the multivariate observation $\boldsymbol{x}$ to univariate observations y such that y's derived from population π_1 and π_2, were separated as much as possible. That is,

$$y_{1i} = \boldsymbol{a}^\top \boldsymbol{x}_{1i}, \ i = 1, \cdots, n_1$$
$$y_{2i} = \boldsymbol{a}^\top \boldsymbol{x}_{2i}, \ i = 1, \cdots, n_2$$

and maximize

$$\frac{|\bar{y}_1 - \bar{y}_2|}{S_y},$$

where

$$\begin{aligned} S_y^2 &= \frac{\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \text{ and} \\ S_j^2 &= \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_{ji} - \bar{y}_j)^2. \end{aligned}$$

How to find $\boldsymbol{a}$!

Note: No distributional assumption, but equal variances are assumed.

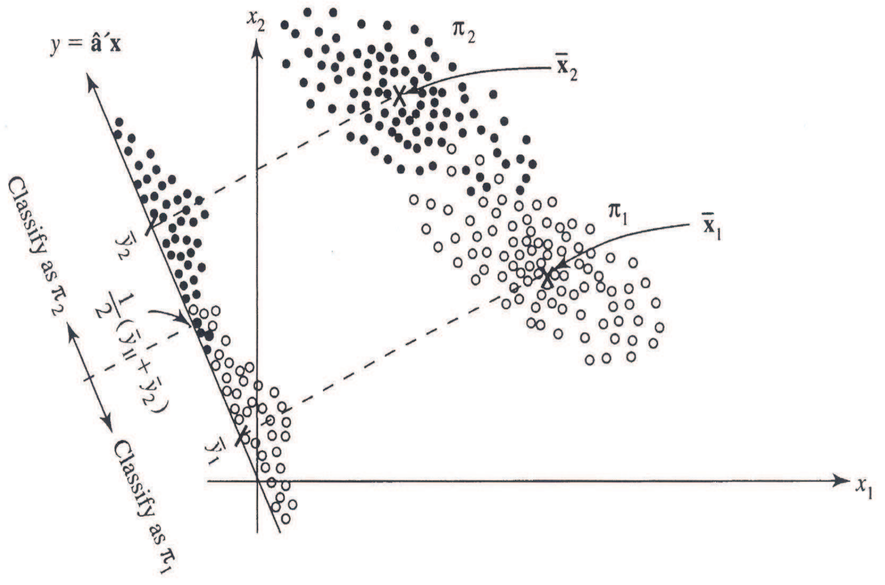


Figure 10.2.1: A pictorial representation of Fisher's procedure for two populations with $p = 2$.

Result (classification with linear discriminant function)

$$\hat{y} = \hat{\boldsymbol{a}}^\top \boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\, \boldsymbol{x} \text{ maximizes}$$

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{S_y^2} = \frac{(\hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_1 - \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_2)^2}{\hat{\boldsymbol{a}}^\top \boldsymbol{S}_{pooled}\, \hat{\boldsymbol{a}}} = \frac{(\hat{\boldsymbol{a}}^\top \boldsymbol{d})^2}{\hat{\boldsymbol{a}}^\top \boldsymbol{S}_{pooled}\, \hat{\boldsymbol{a}}}$$

over all possible $\hat{\boldsymbol{a}}$, where $\boldsymbol{d} = \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2$.

The maximum of the above ratio is

$$D^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\, (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2).$$

Remark

$$\boldsymbol{S}_{pooled} = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{(n_1 - 1) + (n_2 - 1)},$$

where

$$\boldsymbol{S}_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1)(\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1)^\top$$

and

$$\boldsymbol{S}_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(\boldsymbol{x}_{2j} - \bar{\boldsymbol{x}}_2)(\boldsymbol{x}_{2j} - \bar{\boldsymbol{x}}_2)^\top.$$

pf) Use the maximization lemma, that is,

Maximization lemma
$\boldsymbol{B}_{p\times p}$: *p.d.* and $d_{p\times 1}$: a given vector
$max_{\boldsymbol{x}\neq \mathbb{0}} \dfrac{(\boldsymbol{x}^\top \boldsymbol{d})^2}{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}} = \boldsymbol{d}^\top \boldsymbol{B}^{-1}\boldsymbol{d}$ with the maximum attained when $\boldsymbol{x} = c\boldsymbol{B}^{-1}\boldsymbol{d}$ for any $c \neq 0$.

Take $\boldsymbol{x} = \boldsymbol{a}$ and $\boldsymbol{B} = \boldsymbol{S}_{pooled}$, then obvious. □

An allocation rule:

$\boldsymbol{x}_0 \in \pi_1$ if $\hat{y}_0 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\, \boldsymbol{x}_0 \geq \hat{m}$ and

$\boldsymbol{x}_0 \in \pi_2$ if $\hat{y}_0 < \hat{m}$, where $\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)$

Remark:

$$\begin{aligned}
\hat{m} &= \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}\left\{\hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_1 + \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_2\right\} \\
&= \frac{1}{2}\left\{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\, \bar{\boldsymbol{x}}_1 + (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\, \bar{\boldsymbol{x}}_2\right\} \\
&= \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\, (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)
\end{aligned}$$

ex) To construct a procedure for detecting hemophilia A carriers, blood samples were assayed for two groups of women and measurements on the two variables,

$$\begin{aligned} X_1 &= \log_{10}(\text{AHF activity}) \\ X_2 &= \log_{10}(\text{AHF-like antigen}) \end{aligned}$$

recorded ("AHF" denotes antihemophilic factor).

For normal group, $n_1 = 30$, $\bar{\boldsymbol{x}}_1 = \begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix}$.

For hemophilia A carriers group, $n_2 = 22$, $\bar{\boldsymbol{x}}_2 = \begin{pmatrix} -0.2483 \\ 0.0262 \end{pmatrix}$ and $\boldsymbol{S}^{-1}_{pooled} = \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix}$.

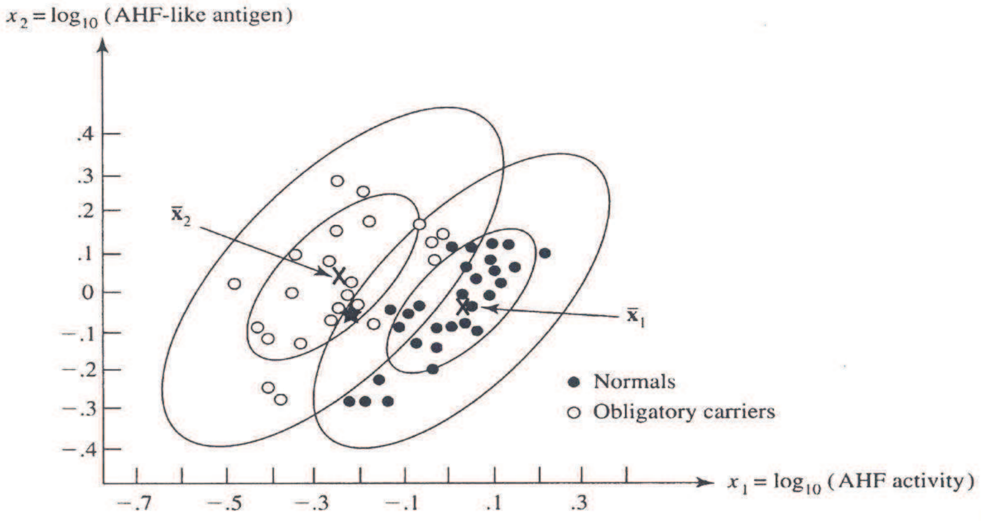


Figure 10.2.2: Scatter plots of [$\log_{10}$ (AHF activity), $\log_{10}$ (AHF-like antigen)] for the normal group and obligatory hemophilia A carriers.

$$
\begin{aligned}
\hat{y} = \hat{\boldsymbol{a}}^\top \boldsymbol{x} &= (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\, \boldsymbol{x} \\
&= \begin{pmatrix} 0.2418 & -0.0652 \end{pmatrix} \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.149 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= 37.61x_1 - 28.92x_2 \\
\bar{y}_1 = \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_1 &= \begin{pmatrix} 37.61 & -28.92 \end{pmatrix} \begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix} = 0.88 \\
\bar{y}_2 = \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_2 &= \begin{pmatrix} 37.61 & -28.92 \end{pmatrix} \begin{pmatrix} -0.2483 \\ -0.0262 \end{pmatrix} = -10.10 \\
\hat{m} &= \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = -4.61 \\
D^2 &= (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = 10.98
\end{aligned}
$$

When $\boldsymbol{x}_0 = \begin{pmatrix} -0.210 \\ -0.044 \end{pmatrix}$,

$$
\hat{y_0} = \hat{\boldsymbol{a}}^\top \boldsymbol{x}_0 = \begin{pmatrix} 37.61 & -28.92 \end{pmatrix} \begin{pmatrix} -0.21 \\ -0.044 \end{pmatrix} = -6.62 < -4.61 = \hat{m}
$$

So $\boldsymbol{x}_0 \in \pi_2$ (hemophilia A carriers group).

10.2.2 Bayes classification

Two populations with known pdf's

$$
\begin{aligned}
\pi_1 &: \text{ pdf } f_1(\boldsymbol{x}) \\
\pi_2 &: \text{ pdf } f_2(\boldsymbol{x}).
\end{aligned}
$$

Assume $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ are of continuous type.

Define

$$P(2|1) = P(\boldsymbol{x} \in R_2|\pi_1) = \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x}$$

= probability of classifying an object as π_2 when, in fact, it is from π_1 and

$$P(1|2) = P(\boldsymbol{x} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}$$

= probability of mis-classifying objects from π_2.

Let p_1 = prior of π_1 and p_2 = prior of π_2 with $p_1 + p_2 = 1$.

Overall probilities are:

1. P(correct classification as π_1) $= p_1 \cdot P(\boldsymbol{x} \in R_1|\pi_1) = p_1 \cdot P(1|1)$,
2. P(misclassification as π_1) $= p_2 \cdot P(\boldsymbol{x} \in R_1|\pi_2) = p_2 \cdot P(1|2)$,
3. P(correct classification as π_2) $= p_2 \cdot P(\boldsymbol{x} \in R_2|\pi_2) = p_2 \cdot P(2|2)$, and
4. P(misclassification as π_2) $= p_1 \cdot P(\boldsymbol{x} \in R_2|\pi_1) = p_1 \cdot P(2|1)$.

Costs of misclassification are:

		Classify as	
		π_1	π_2
True Pop.	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

One possible approach is as follows:

Define ECM (Expected Cost of Misclassification) as

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2, \text{ then}$$

R_1 and R_2 minimize the ECM based on the following rule, where

$$\begin{array}{cccc} R_1: & \dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} & \geq & \left(\dfrac{c(1|2)}{c(2|1)}\right) & \left(\dfrac{p_2}{p_1}\right) \\ & \text{density ratio} & & \text{cost ratio} & \text{prior ratio} \\ R_2: & \dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} & < & \left(\dfrac{c(1|2)}{c(2|1)}\right) & \left(\dfrac{p_2}{p_1}\right) \\ & \text{density ratio} & & \text{cost ratio} & \text{prior ratio} \end{array}.$$

pf)

$$\begin{aligned} ECM &= c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \\ &= c(2|1)p_1 \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + c(1|2)p_2 \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \end{aligned}$$

Noting that $\Omega = R_1 \cup R_2$ and $R_2 = R_1^c$,

$$1 = \int_{\Omega} f_1(\boldsymbol{x})d\boldsymbol{x} = \int_{R_1} f_1(\boldsymbol{x})d\boldsymbol{x} + \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x}.$$

$$\begin{aligned} \therefore\ ECM &= c(2|1)p_1 \left[1 - \int_{R_1} f_1(\boldsymbol{x})d\boldsymbol{x}\right] + c(1|2)p_2 \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_{R_1} \{c(1|2)p_2 f_2(\boldsymbol{x}) - c(2|1)p_1 f_1(\boldsymbol{x})\}\, d\boldsymbol{x} + c(2|1)p_1 \end{aligned}$$

Note that $p_1, p_2, c(1|2), c(2|1), f_1(\boldsymbol{x})$, and $f_2(\boldsymbol{x})$ are all nonnegative.

ECM is minimized if R_1 consists of $\boldsymbol{x}$ such that $c(1|2)p_2 f_2(\boldsymbol{x}) - c(2|1)p_1 f_1(\boldsymbol{x}) \leq 0$ and excludes $\boldsymbol{x}$ for which this quantity is positive.
Hence the result follows. □

10.3 Classification with two multi. normal populations

Bayes classification with $N_p(\boldsymbol{\mu}_i, \Sigma_i),\ i = 1, 2$.

10.3.1 Equal variance

That is, Case I: $\Sigma_1 = \Sigma_2 = \Sigma$

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\}, i = 1, 2,$$

where $\boldsymbol{x} = (x_1, \cdots, x_p)^\top$.

Result (minimize the ECM)

Allocate $\boldsymbol{x}_0$ to π_1 if

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (10.3.1)$$

Allocate $\boldsymbol{x}_0$ to π_2 otherwise.

Note that $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and Σ are known.

pf) Allocate $\boldsymbol{x}_0$ to π_1 if $\dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \left(\dfrac{c(1|2)}{c(2|1)}\right)\left(\dfrac{p_2}{p_1}\right)$

$\Leftrightarrow \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\} \geq \left[\left(\dfrac{c(1|2)}{c(2|1)}\right)\left(\dfrac{p_2}{p_1}\right)\right]$

Note that

$$\begin{aligned} &-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \\ &= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \end{aligned}$$

Therefore
$R_1 : (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \geq \ln\left[\left(\dfrac{c(1|2)}{c(2|1)}\right)\left(\dfrac{p_2}{p_1}\right)\right]$ □

In general, we don't know $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and Σ. So plug in $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2$, and $\boldsymbol{S}_{pooled}$ based on independent random samples. That is,

$\boldsymbol{x}_{1i}$: r.s. from $N_p(\boldsymbol{\mu}_1, \Sigma), i = 1, \cdots, n_1$

$\boldsymbol{x}_{2i}$: r.s. from $N_p(\boldsymbol{\mu}_2, \Sigma), i = 1, \cdots, n_2$

$$\begin{aligned} \bar{\boldsymbol{x}}_1 &= \frac{1}{n_1}\sum_{j=1}^{n_1} \boldsymbol{x}_{1j}, \boldsymbol{S}_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\boldsymbol{x}_{1j}-\bar{\boldsymbol{x}}_1)(\boldsymbol{x}_{1j}-\bar{\boldsymbol{x}}_1)^\top \\ \bar{\boldsymbol{x}}_2 &= \frac{1}{n_2}\sum_{j=1}^{n_2} \boldsymbol{x}_{2j}, \boldsymbol{S}_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\boldsymbol{x}_{2j}-\bar{\boldsymbol{x}}_2)(\boldsymbol{x}_{2j}-\bar{\boldsymbol{x}}_2)^\top \end{aligned}$$

and $\boldsymbol{S}_{pooled} = \dfrac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2}{(n_1-1)+(n_2-1)}$.

Hence, allocate $\boldsymbol{x}_0$ to π_1 if

$$(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\ \boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\ (\bar{\boldsymbol{x}}_1+\bar{\boldsymbol{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right].$$

Remark

- Fisher's classification rule is equivalent to the minimum ECM rule with equal priors and equal costs of misclassification.
- LHS of (10.3.1) is called as sample linear discrimination (classification) function.

10.3.2 Unequal variance

That is, Case 2: $\Sigma_1 \neq \Sigma_2$

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\}, i = 1, 2,$$

where $\boldsymbol{x} = (x_1, \cdots, x_p)^\top$.

No simplification in $\dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})}$ is available.

<u>**Result (minimize the ECM)**</u>

Allocate $\boldsymbol{x}_0$ to π_1 if

$$-\frac{1}{2}\boldsymbol{x}_0^\top\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)\boldsymbol{x}_0 + \left(\boldsymbol{\mu}_1^\top\Sigma_1^{-1} - \boldsymbol{\mu}_2^\top\Sigma_2^{-1}\right)\boldsymbol{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (10.3.2)$$

Allocate $\boldsymbol{x}_0$ to π_2 otherwise.

Remark

1. $k = \frac{1}{2} \ln \left(\frac{|\Sigma_1|}{|\Sigma_2|} \right) + \frac{1}{2} \left(\boldsymbol{\mu}_1^\top \Sigma_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \Sigma_2^{-1} \boldsymbol{\mu}_2 \right)$

2. LHS of (10.3.2) is called as quadratic classification (discrimination) function.

3. The classification regions are defined by quadratic functions of $\boldsymbol{x}$. When $\Sigma_1 = \Sigma_2$, the quadratic term, $-\frac{1}{2}\boldsymbol{x}_0^\top \left(\Sigma_1^{-1} - \Sigma_2^{-1} \right) \boldsymbol{x}_0$, disappears, and the regions reduce to linear functions of $\boldsymbol{x}$. See (10.3.1).

4. In practice, we use the sample quantities $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \boldsymbol{S}_1$ and $\boldsymbol{S}_2$ for $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1$ and Σ_2, respectively.

10.4 Evaluating classification functions

TPM (Total Probability of Misclassification)

$$\begin{aligned} &= p_1 P(2|1) + p_2 P(1|2) \\ &= p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} \end{aligned}$$

OER (Optimum Error Rate) $= \min_{R_1, R_2} TPM$

OER can be calculated for some special situations but, in practice, $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ are unknown. For example, certain population parameters appearing in allocation rules must be estimated from the sample.

Need to evaluate sample classification functions,

AER (Actual Error Rate) $= p_1 \int_{\hat{R}_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{\hat{R}_1} f_2(\boldsymbol{x}) d\boldsymbol{x}$.

Similar to OER, it can't, in general, be calculated. Because it depends on the unknown density functions $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$.

There is a measure of performance that **does not depend on the form of the parent populations** and that can be calculated for any classification procedure.

- APER (APparent Error Rate)

- Jackknifing or cross-validation

10.4.1 APER

		Predicted	membership	
		π_1	π_2	
Actual	π_1	n_{1c}	$n_{1M} = n_1 - n_{1c}$	n_1
membership	π_2	$n_{2M} = n_2 - n_{2c}$	n_{2c}	n_2

The table is called as confusion matrix.

APER $= \dfrac{n_{1M} + n_{2M}}{n_1 + n_2}$; proportion of items in the data set that are misclassified.

ex) Discriminating owners from nonowners of riding mower

Two groups in a city:

π_1: riding-mower owners
π_2: those without riding mowers

A riding-mower manufacturer is interested in classifying families as prospective owners or nonowners on the basis of x_1 = income and x_2 = lot size.

Random samples of $n_1 = 12$ current owners and $n_2 = 12$ current nonowners.

Using Fisher's linear classification, we have the following plot.

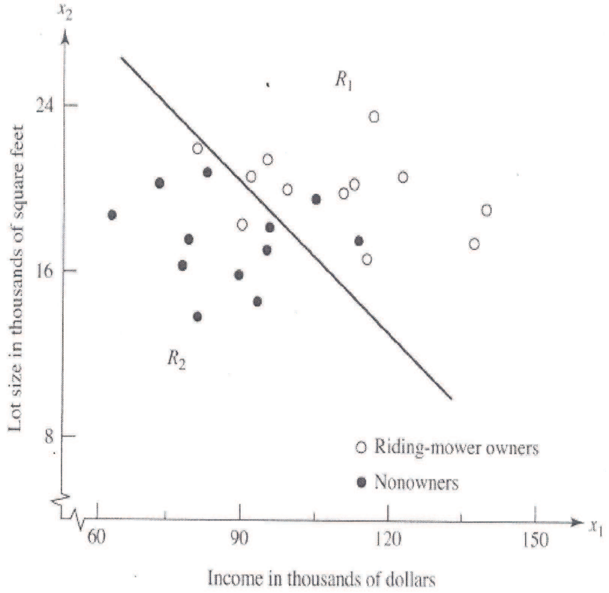


Figure 10.4.1: Income and lot size for riding-mower owners and nonowners

			Predicted	membership	
			π_1	π_2	
Actual	π_1	riding-mower owners	$n_{1c} = 10$	$n_{1M} = 2$	$n_1 = 12$
membership	π_2	nonowners	$n_{2M} = 2$	$n_{2c} = 10$	$n_2 = 12$

$$\therefore \text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2} = \frac{2+2}{12+12} = \frac{4}{24} = 16.7\%$$

Remark

1. The APER is intuitively appealing and easy to calculate

2. Unfortunately, it tends to <u>underestimate</u> the AER. $(\because)$ Data used to build the classification function are also used to evaluate it.

10.4.2 Cross-validation (Lachenbruch's holdout procedure)

1. Hold 1 observation from π_1 and develop a classification function based $n_1 - 1$, n_2 observations.

2. Classify the holdout observation, using the function constructed in Step 1.

3. Repeat Step 1 and 2 until all of the π_1 observations are classified. Let $n_{1M}^{(H)}$ be the number of holdout observations misclassified in this group.

4. Repeat Step 1 to 3 for π_2 observations.

$$\hat{P}(2|1) = \frac{n_{1M}^{(H)}}{n_1}, \hat{P}(1|2) = \frac{n_{2M}^{(H)}}{n_2}$$
$$\therefore\ \hat{E}(AER) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2}$$

ex) Evaluating Fisher's LD

$$\pi_1 : (\boldsymbol{\mu}_1, \Sigma),\ \pi_2 : (\boldsymbol{\mu}_2, \Sigma)$$

Data: $\boldsymbol{X}_1^\top = \begin{pmatrix} 2 & 4 & 3 \\ 12 & 10 & 8 \end{pmatrix}$, $\boldsymbol{X}_2^\top = \begin{pmatrix} 5 & 3 & 4 \\ 7 & 9 & 5 \end{pmatrix}$

Summary statistics:

$\bar{\boldsymbol{x}}_1 = \begin{pmatrix} 3 \\ 10 \end{pmatrix}$, $\bar{\boldsymbol{x}}_2 = \begin{pmatrix} 4 \\ 7 \end{pmatrix}$,

$\boldsymbol{S}_1 = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}$, $\boldsymbol{S}_2 = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}$, and $\boldsymbol{S}_{pooled} = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix} = \frac{1}{4}(2\boldsymbol{S}_1 + 2\boldsymbol{S}_2)$

Fisher's LD classification:

$\hat{R}_2 : (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1)^\top \boldsymbol{S}^{-1}\boldsymbol{x} > \frac{1}{2}(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1)^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_2 + \bar{\boldsymbol{x}}_1)$

$\Leftrightarrow (\boldsymbol{x} - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_2) < (\boldsymbol{x} - \bar{\boldsymbol{x}}_1)^\top \boldsymbol{S}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_1)$

Leave-one-out estimation:

leave-out	$(\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_H^{-1}(\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_2)$	$\left(\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1^{(-j)}\right)^\top \boldsymbol{S}_H^{-1} \left(\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1^{(-j)}\right)$	
$\boldsymbol{x}_{11} = (2, 12)^\top$	10.3	4.5	(O)
$\boldsymbol{x}_{12} = (4, 10)^\top$	2.8	4.5	(X)
$\boldsymbol{x}_{13} = (3, 8)^\top$	0.8	4.5	(X)

$\therefore n_{1M}^{(H)} = 2,\ n_1 = 3;\ \widehat{P(2|1)}^{H} = 2/3$

Similarly, $n_{2M}^{(H)} = 1, n_2 = 3; \widehat{P(1|2)}^{H} = 1/3$

$\therefore \hat{E}(AER) = \dfrac{2+1}{3+3} = \frac{1}{2}$

Remark

- Mahalanobis distance $= (\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i),\ i = 1, 2.$ We choose π_1 if $MD_1 < MD_2$.
- In practice, we replace $\boldsymbol{\mu}_i, \Sigma$ by $\bar{\boldsymbol{x}}_i, \boldsymbol{S}$.

10.5 Final comments

1) Classification with more than two populations

- Assign the new observation to π_1 if
$$\frac{f_1(\boldsymbol{x})}{f_j(\boldsymbol{x})} > k_{1j} = \frac{c(1|j)p_j}{c(j|1)p_1} \text{ for all } j \neq 1.$$

  In general, assign to π_i if
$$\frac{f_i(\boldsymbol{x})}{f_j(\boldsymbol{x})} > k_{ij} = \frac{c(i|j)p_j}{c(j|i)p_i} \text{ for all } j \neq i.$$

- Assign the new observation to π_1 if $MD_1 < MD_j$ for all $j \neq 1$.

  In general, assign to π_i if $MD_i < MD_j$ for all $j \neq i$

2) So far, all classification functions are based on quantitative variables. We can apply logistic regression to classification where some or all of the variables are qualitative.

3) Variable selection

The methods for doing variable selection in discriminant analysis are similar to those used in regression.

The SAS program STEPDISC will do this.

4) Canonical discrimination functions
As an option to eliminating variables, we might consider using linear combinations of variables to reduce the number of discriminators.

The Fisher linear discriminant functions produce results that are identical to using all canonical discriminant functions.

The SAS program CANDISC uses a method due to R.A. Fisher.

5) Categorical variables

Introduce indicator variables. For example, we have three categories, 'high school diploma', 'undergraduate degree' or 'masters degree'.

In this case, introduce two indicator variables, say
$Z_1 = 1$ if high school
$Z_1 = 0$ otherwise and
$Z_2 = 1$ if undergraduate and
$Z_2 = 0$ otherwise.

Using these variables in the analysis is acceptable, but there are situations where it is questionable.

Care must be taken when using the STEPDISC procedure. Note that if Z_2 is eliminated, then the grouping is high school or not. **The safe procedure is to either keep all of the indicators or none.**

10.6 Discrimination and Classification in R

The data were recorded for two species of irises. The objective is to develop a rule for classifying a new flower based on the four variables sl (Sepal Length), sw (Sepal Width), pl (Petal Length), pw (Petal Width).

Fisher(1936) Iris Data

Obs	SL	SW	PL	PW	Species
1	65	28	46	15	Versicolor
2	62	22	45	15	Versicolor
3	59	32	48	18	Versicolor
4	61	30	46	14	Versicolor
5	60	27	51	16	Versicolor
⋮					
49	68	28	48	14	Versicolor
50	67	30	50	17	Versicolor
51	64	28	56	22	Virginica
52	67	31	56	24	Virginica
⋮					
98	57	25	50	20	Virginica

10.6.1 Mahalanobis Distance

The allocation rule, consisting in assigning a unit to the group whose average is closest to, will assign unit $\mathbf{x_0}$ to group 1 if $(\mathbf{x_0} - \bar{\mathbf{x}}_\mathbf{1})^T\mathbf{S}^{-1}(\mathbf{x_0} - \bar{\mathbf{x}}_\mathbf{1}) < (\mathbf{x_0} - \bar{\mathbf{x}}_\mathbf{2})^T\mathbf{S}^{-1}(\mathbf{x_0} - \bar{\mathbf{x}}_\mathbf{2})$ and to group 2 if the opposite inequality holds.

```
> library(WMDB)
> # data read
> irises <- read.table("T11-5.dat")
> irises <- data.frame(irises)
> names(irises) <- c("SL","SW","PL","PW","cla")
>
> irises$cla <- as.numeric(irises$cla)
> G <- as.factor(irises[,5])
> clasts <- wmd(irises[,-5],G,TstX=irises[,-5],var.equal=F)
      1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
blong 1 1 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
      29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53
blong  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  2  2
      54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78
blong  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  3  2  2  2  3  2  2  2  2  3
      79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102
blong  2  2  2  2  2  3  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2   2   3   3
      103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121
blong   3   3   3   3   2   3   3   3   3   3   3   3   3   3   3   3   3   3   3
      122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
blong   3   3   3   3   3   3   3   3   2   3   3   3   2   2   3   3   3   3   3
      141 142 143 144 145 146 147 148 149 150
blong   3   3   3   3   3   3   3   3   3   3
[1] "num of wrong judgement"
[1]  69  73  78  84 107 130 134 135
[1] "samples divided to"
[1] 3 3 3 3 2 2 2 2
[1] "samples actually belongs to"
[1] 2 2 2 2 3 3 3 3
Levels: 1 2 3
[1] "percent of right judgement"
[1] 0.9466667
```

10.6.2 Linear Discriminant Analysis

```
> # Linear Discriminant Analysis
> library(MASS)
> index <- sample(2,size = nrow(irises),replace = TRUE, prob = c(0.7,0.3))
> train <- irises[index == 1,]
> test <- irises[index == 2,]
> flda <- lda(cla˜.,data=train)
> fts1 <- predict(flda, newdata=test)
> tab1 <- table(test$cla, fts1$class)
> mclda <- 1-sum(diag(tab1))/sum(tab1)
> flda
Call:
lda(cla ˜ ., data = train)

Prior probabilities of groups:
        1         2         3
```

```
0.3362832 0.3274336 0.3362832

Group means:
        SL       SW       PL        PW
1 4.986842 3.410526 1.473684 0.2315789
2 5.981081 2.767568 4.291892 1.3243243
3 6.628947 2.976316 5.571053 2.0473684

Coefficients of linear discriminants:
          LD1        LD2
SL  0.8869015  0.1601621
SW  1.3416127 -2.2062498
PL -2.1452432  0.9988982
PW -3.0757217 -3.0905493

Proportion of trace:
   LD1    LD2
0.9903 0.0097
> tab1

     1  2  3
  1 12  0  0
  2  0 12  1
  3  0  0 12
> mclda
[1] 0.02702703
```

10.6.3 Quadratic Discriminant Analysis

```
> # Quadratic Discriminant Analysis
> fqda <- qda(cla˜.,data=train)
> fts2 <- predict(fqda,newdata=test)
> tab2 <- table(test$cla,fts2$class)
> mcqda <- 1-sum(diag(tab2))/sum(tab2)
> fqda
Call:
qda(cla ˜ ., data = train)
```

```

Prior probabilities of groups:
        1         2         3 
0.3362832 0.3274336 0.3362832 

Group means:
        SL       SW       PL        PW
1 4.986842 3.410526 1.473684 0.2315789
2 5.981081 2.767568 4.291892 1.3243243
3 6.628947 2.976316 5.571053 2.0473684
> tab2
   
     1  2  3
  1 12  0  0
  2  0 11  2
  3  0  1 11
> mcqda
[1] 0.08108108
```

10.6.4 Regularized Discriminant Analysis

```
rda
```

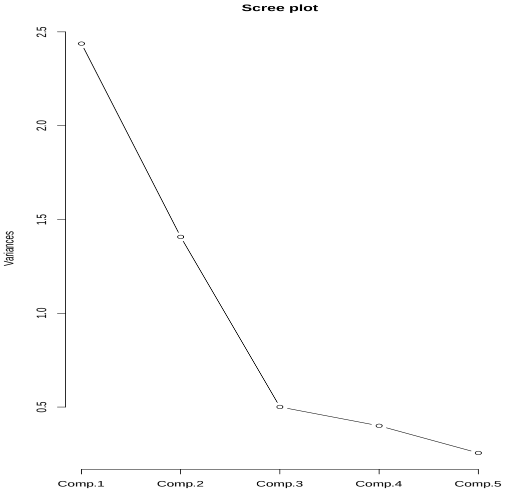


Figure 10.6.1: A scree plot for the stock data.

Chapter 11

Clustering

11.1 Introduction

Goal: Find natural groupings of the items (or variables) using some measures of association.

Note: Different from classification because number of groups is unknown.

11.2 Similarity measures

11.2.1 distances and similarity coefficients for pairs of item

Given two $p-$dimensional vectors $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^\top$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_p)^\top$,

- Euclidean distance;

$$\begin{aligned} d(\boldsymbol{x}, \boldsymbol{y}) &= \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y})} \\ &= \left[(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2\right]^{1/2} \end{aligned}$$

- Statistical (Mahalanobis) distance; For some symmetric matrix $\boldsymbol{A}$, $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\top \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{y})}$.

Usually $\boldsymbol{A} = \boldsymbol{S}^{-1}$, where $\boldsymbol{S}$ is sample covariance matrix.

Remark Some distance measures

- Minkowski matric; $d(\boldsymbol{x}, \boldsymbol{y}) = [\sum_{i=1}^{p} |x_i - y_i|^m]^{1/m}$
- Canberra matric; $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} \dfrac{|x_i - y_i|}{x_i + y_i}$
- Czekanowski coefficient; $d(\boldsymbol{x}, \boldsymbol{y}) = 1 - \dfrac{2\sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p}(x_i + y_i)}$

Note

$$\boldsymbol{x}_1 = \begin{pmatrix} 1 \\ 96 \end{pmatrix}, \boldsymbol{x}_2 = \begin{pmatrix} 2 \\ 108 \end{pmatrix} \Rightarrow d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{1^2 + 12^2} = \sqrt{145}.$$

If the second variable was measured in inches and we change to feet, then $\boldsymbol{x}_1 = \begin{pmatrix} 1 \\ 8 \end{pmatrix}$, $\boldsymbol{x}_2 = \begin{pmatrix} 2 \\ 9 \end{pmatrix} \Rightarrow d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{1^2 + 1^2} = \sqrt{2}$.

That is, there is a problem introduced by scales that are not comparable.

$\therefore$ The "standardized" variables are commonly used.

Remark

- Whenever possible, it is advisable to use "true" distance.
- When meaningful $p-$dimensional measurements are not available, use the binary variables.
- In some applications, the variable that we observe is simply the answer to a question, say yes or no.

This is usually coded as a 'one' or 'zero', so that we can apply our standard results for clustering.

Let

$$x_{ij} = \begin{cases} 1 & \text{if the } j^{th} \text{ binary characteristic is present in item } i, \\ 0 & \text{otherwise} \end{cases}$$

and

$$x_{kj} = \begin{cases} 1 & \text{if the } j^{th} \text{ binary characteristic is present in item } k, \\ 0 & \text{otherwise,} \end{cases}$$

then

$$(x_{ij} - x_{kj})^2 = \begin{cases} 0 & \text{if } x_{ij} = x_{kj} = 1 \text{ or } 0 \\ 1 & \text{if } x_{ij} \neq x_{kj} \end{cases}$$

and $\sum_{j=1}^{p}(x_{ij} - x_{kj})^2 =$ the number of mismatches.

ex)

	Variables				
	1	2	3	4	5
item i	1	0	0	1	1
item k	1	1	0	1	0

$\sum_{j=1}^{5}(x_{ij} - x_{kj})^2 = (1-1)^2 + (0-1)^2 + \cdots + (1-0)^2 = 2 =$ number of mismatches.

Note that this measure treats $1-1$ and $0-0$ matches equally.

As a measure of similarity, this distance may be misleading. That is, a $1-1$ match may imply a stronger indication of similarity than a $0-0$ match.

For example, suppose the question is whether the individual can speak a particular foreign language, say Greek. If both respondents answer yes, that is, a $1-1$ match, it gives a stronger indication of similarity than if they both answer no.

To allow for differential treatment of $1-1$ and $0-0$, let us make a contingency table.

		item k		
		1	0	Totals
item i	1	a	b	a+b
	0	c	d	c+d
Totals		a+c	b+d	p=a+b+c+d

ex) Continued

		item k		
		1	0	Totals
item i	1	2	1	3
	0	1	1	2
Totals		3	2	5

Several measures of similarity to be used are suggested.

For example,

1. $\dfrac{a+d}{p}$; equal weights for $1-1$ and $0-0$

2. $\dfrac{a+d}{a+d+2(b+c)}$; double weights for unmatched pairs

3. $\dfrac{a}{p}$; ignore $0-0$ matches in the numerator

4. $\dfrac{a}{b+c}$; ratio of matches to mismatches with $0-0$ matches excluded.

$$\vdots$$

For our example, the similarity measure for the two items using the first idea is $\dfrac{3}{5} = \dfrac{a+d}{p}$.

Example (Calculating the values of a similarity coefficient)

Suppose five individuals possess the following characteristics:

			Eye	Hair		
	Height	Weight	color	color	Handedness	Gender
Individual 1	68 in	140 lb	green	blond	right	female
Individual 2	73 in	185 lb	brown	brown	right	male
Individual 3	67 in	165 lb	blue	blond	right	male
Individual 4	64 in	120 lb	brown	brown	right	female
Individual 5	76 in	210 lb	brown	brown	left	male

Define six binary variables $X_1, X_2, X_3, X_4, X_5, X_6$ as

$$X_1 = \begin{cases} 1 & \text{height} \geq 72 \text{ inch} \\ 0 & \text{height} < 72 \text{ inch} \end{cases} X_4 = \begin{cases} 1 & \text{blond hair} \\ 0 & \text{not blond hair} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{weight} \geq 150 \text{ lb} \\ 0 & \text{weight} < 150 \text{ lb} \end{cases} X_5 = \begin{cases} 1 & \text{right handed} \\ 0 & \text{left handed} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{brown eyes} \\ 0 & \text{otherwise} \end{cases} X_6 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

The scores for individuals 1 and 2 on the $p = 6$ binary variable variables are

		X_1	X_2	X_3	X_4	X_5	X_6
Individual	1	0	0	0	1	1	1
	2	1	1	1	0	1	0

and the number of matches and mismatches are indicated in the two-way array

		Individual 2		
		1	0	Total
Individual 1	1	1	2	3
	0	3	0	3
	Totals	4	2	6

Employing similarity coefficient 1, which gives equal weight to matches, we compute

$$\frac{a+d}{p} = \frac{1+0}{6} = \frac{1}{6}$$

Continuing with similarity coefficient 1, we calculate the remaining similarity numbers for pairs of individuals. These are displayed in the 5×5 symmetric matrix

		Individual				
		1	2	3	4	5
	1	1				
Individual	2	1/6	1			
	3	4/6	3/6	1		
	4	4/6	3/6	2/6	1	
	5	**0**	**5/6**	2/6	2/6	1

Based on the magnitudes of the similarity coefficient, we should conclude that individuals 2 and 5 are most similar and individuals 1 and 5 are least similar.

Other pairs fall between these extremes.

If we were to divide the individuals into two relatively homogeneous subgroups on the basis of the similarity numbers, we might form the subgroups (1 3 4) and (2 5).

Note Relationship between similarity matrix and distance matrix

i) Any distance measure $d(P,Q)$ between two points P and Q is valid if
① $d(P,Q) = d(Q,P)$
② $d(P,Q) > 0$, if $P \neq Q$
③ $d(P,Q) = 0$, if $P = Q$
④ $d(P,Q) \leq d(P,R) + d(R,Q)$ (triangle inequality), where R is any other intermediate point.

ii) Distance that must satisfy i) can't always be constructed from similarity matrix.

iii) It can be shown that
① similarity matrix, $\{\tilde{S}_{ij}\}$, is n.n.d.
② $\max_{i,j}\tilde{S}_{ij} = 1$
$\Rightarrow d_{ij} = \sqrt{2(1-\tilde{S}_{ij})}$ has the properties of a distance.

iv) Under Euclidean distance, if d_{ij} is a distance, then $d^{'}_{ij} = \max_{i,j}\{d_{ij}\} - d_{ij}$ is a similarity measure.

11.2.2 Similarities and association measures for pairs of variables

Usually sample correlation coefficient.

For binary variables, there are n items categorized in the table for each pair of variable, i.e.,

		Variable k		
		1	0	Totals
Variable i	1	a	b	a+b
	0	c	d	c+d
	Totals	a+c	b+d	n=a+b+c+d

Hence, it can be shown that usual sample correlation coefficient to the binary variables is

$$r = \frac{ad - bc}{[(a+b)(c+d)(a+c)(b+d)]^{1/2}}.$$

11.2.3 Concluding Comments on Similarity

Most practitioners use distances to cluster items and correlations to cluster variables.

However, at times, inputs to clustering algorithm may be simple frequencies.

ex) Numerals in 11 languages

Table 11.2.1: Numerals in 11 languages

Engl-	Norwe-	Dani-	Dutch	German	French	Span-	Italian	Polish	Hung-	Finnish
ish(E)	gian(N)	sh(Da)	(Du)	(G)	(Fr)	ish(Sp)	(I)	(P)	arian(H)	(Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neljä
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

The words for 1 in French, Spanish, and Italian all begin with u. For illustrative purposes, we might compare languages by looking at the *first letter* of the numbers.

We call the words for the same number in two different languages *concordant* if they have the same first letter and *discordant* if they do not.

From Table 11.2.1, the tables of concordances (frequencies of matching first initials) for the numbers $1 - 10$ is given in Table 11.2.2.

We see that English and Norwegian have the same first letter for 8 of the 10 word pairs. The remaining frequencies were calculated in the same manner.

The result in Table 11.2.2 confirm our initial visual impression of Table 11.2.1.

That is, English, Norwegian, Danish, Dutch, and German seem to form a group. French, Spanish, Italian, and Polish might be grouped together, whereas Hungarian and Finnish appear to stand alone.

So far, we have used our visual impression of similarity or distance measures to form groups.

We now discuss less subjective schemes for creating clusters.

Table 11.2.2: Concordant First Letters for Numbers in 11 Languages

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	10										
N	8	10									
Da	8	9	10								
Du	3	5	4	10							
G	4	6	5	5	10						
Fr	4	4	4	1	3	10					
Sp	4	4	5	1	3	8	10				
I	4	4	5	1	3	9	9	10			
P	3	3	4	0	2	5	7	6	10		
H	1	2	2	2	1	0	0	0	0	10	
Fi	1	1	1	1	1	1	1	1	1	2	10

11.3 Hierarchical clustering methods

A series of successive merges or a series of successive divisions.

- Agglomerative hierarchical methods;
  individual objects (# of clusters = # of objects)
  ⇒ grouping and eventually 1 cluster

- Divisive hierarchical methods:
  single group (# of clusters = 1)
  ⇒ subgrouping and eventually each object form a group.

We'll concentrate on agglomerative hierarchical procedures and, in particular, linkage methods.

1. Single linkage (minimum distance or nearest neighbor)

2. Complete linkage (maximum distance or farthest neighbor)

3. Average linkage (average distance)

4. Centroid linkage (distance of each centers given by averages of each clusters)

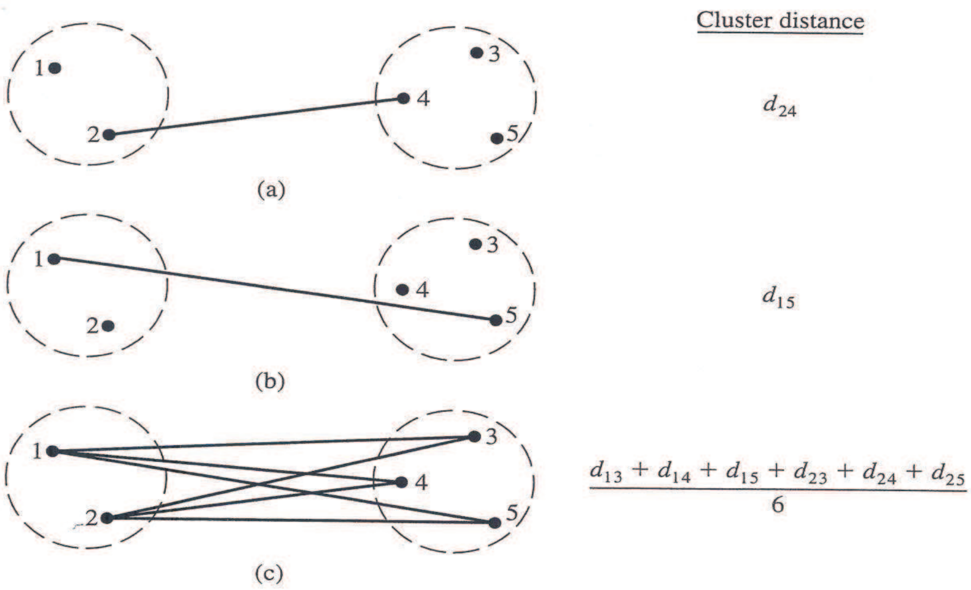

Cluster distance
1
2
3
4
5
d_{24}
(a)
d_{15}
(b)
$\frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$
(c)

Figure 11.3.1: Intercluster distance (dissimilarity) for (a) single linkage, (b) complete linkage, and (c) average linkage.

Algorithm for grouping N objects (items or variables)

1. Start with N clusters, each containing a single entity and an $N \times N$ symmetric matrix of distances (or similarities), $\boldsymbol{D} = \{d_{ik}\}$.

2. Search the distance matrix for the nearest (most similar) pairs of clusters. Let the distance between most similar clusters u and v be d_{uv}

3. Merge clusters u and v and label it (uv). Update the distance matrix.

4. Repeat step 2 and 3 a total of $N - 1$ times (All objects will be in a single cluster after the algorithm terminates.)

ex) Single linkage (minimum distance)

$$\boldsymbol{D} = \{d_{ik}\} = \begin{array}{cc} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{pmatrix} \end{array} \quad ;$$

The hypothetical distances between pairs of five objects.

$\min_{i,k}(d_{ik}) = d_{53} = 2 \Rightarrow (3,5)$

$d_{(3,5)1} = \min(d_{31}, d_{51}) = \min(3, 11) = 3$
$d_{(3,5)2} = \min(d_{32}, d_{52}) = \min(7, 10) = 7$
$d_{(3,5)4} = \min(d_{34}, d_{54}) = \min(9, 8) = 8$

$$\boldsymbol{D} = \begin{array}{cc} & \begin{array}{cccc} (3,5) & 1 & 2 & 4 \end{array} \\ \begin{array}{c} (3,5) \\ 1 \\ 2 \\ 4 \end{array} & \begin{pmatrix} 0 & & & \\ \textcircled{3} & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{pmatrix} \end{array}$$

minimum distance= $d_{(3,5)1} = 3 \Rightarrow (3,5,1)$

$d_{(1,3,5)2} = \min(d_{(3,5)2}, d_{12}) = \min(7, 9) = 7$
$d_{(1,3,5)4} = \min(d_{(3,5)4}, d_{14}) = \min(8, 6) = 6$

$$\boldsymbol{D} = \begin{array}{cc} & \begin{array}{ccc} (1,3,5) & 2 & 4 \end{array} \\ \begin{array}{c} (1,3,5) \\ 2 \\ 4 \end{array} & \begin{pmatrix} 0 & & \\ 7 & 0 & \\ 6 & \textcircled{5} & 0 \end{pmatrix} \end{array}$$

minimum distance=$d_{42} = 5 \Rightarrow (2,4)$

$d_{(1,3,5)(2,4)} = \min(d_{(1,3,5)2}, d_{(1,3,5)4}) = \min(7, 6) = 6$

$$D = \begin{array}{cc} & \begin{array}{cc} (1,3,5) & (2,4) \end{array} \\ \begin{array}{c} (1,3,5) \\ (2,4) \end{array} & \begin{pmatrix} 0 & \\ \textcircled{6} & 0 \end{pmatrix} \end{array} \quad \Rightarrow (1,2,3,4,5)$$

Dendrogram

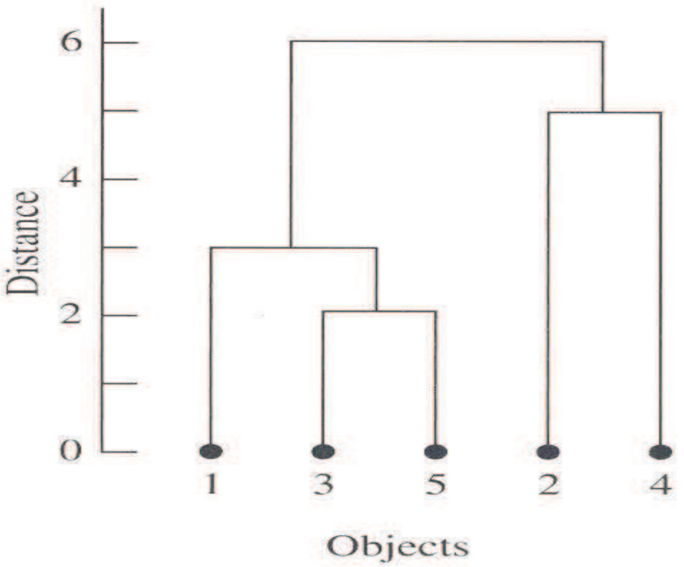


Figure 11.3.2: Single linkage dendrogram for distances between five objects.

It seems clear that, in this example, two clusters stand out, but that is not always the case.

ex) (Single linkage clustering of 11 languages)

To develop a matrix of distances, we subtract the concordances from the perfect agreement figure of 10 that each language has with itself.

That is, $d_{ij} = \max_{i,j}\{\tilde{S}_{ij}\} - \tilde{S}_{ij}$.

$$\begin{array}{cccccccccccc}
 & E & N & Da & Du & G & Fr & Sp & I & P & H & Fi \\
E & 0 \\
N & 2 & 0 \\
Da & 2 & 1 & 0 \\
Du & 7 & 5 & 6 & 0 \\
G & 6 & 4 & 5 & 5 & 0 \\
Fr & 6 & 6 & 6 & 9 & 7 & 0 \\
Sp & 6 & 6 & 5 & 9 & 7 & 2 & 0 \\
I & 6 & 6 & 5 & 9 & 7 & 1 & 1 & 0 \\
P & 7 & 7 & 6 & 10 & 8 & 5 & 3 & 4 & 0 \\
H & 9 & 8 & 8 & 8 & 9 & 10 & 10 & 10 & 10 & 0 \\
Fi & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 8 & 0
\end{array}$$

Dendrogram

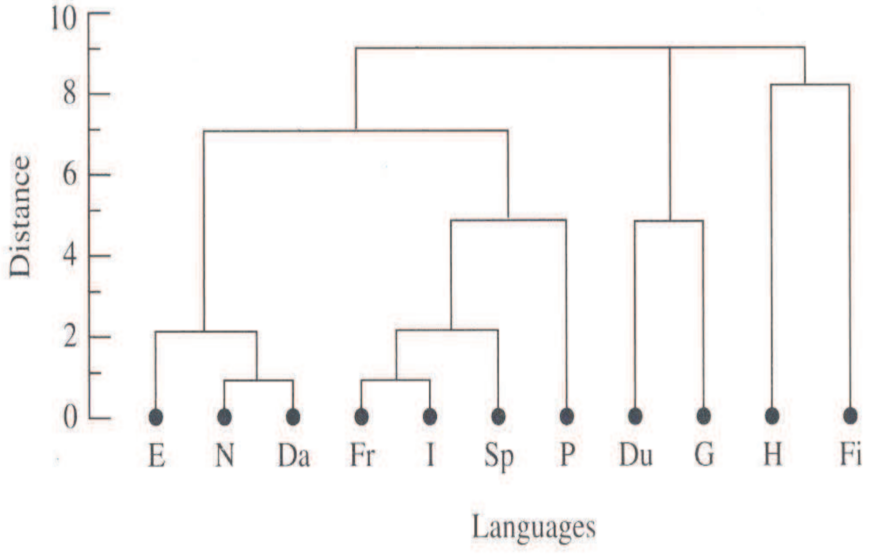


Figure 11.3.3: Single linkage dendrogram for distances between numbers in 11 languages.

Remark

1. Complete linkage use maximum distance instead of minimum distance

of single linkage.

2. There is Ward's Hierarchical clustering method.

11.4 Non-Hierarchical methods

K-Means method

Step 1: Partition the data into K initial clusters. This may be done at random.

Step 2: Determine the centroid (that is the mean) for each cluster.

For each observation, reassign it to the cluster, that is closest. That is, compute the distance to each centroid and assign it to the one that is smallest.

Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.

Remark: It is best to use the standardized data and normally uses the Euclidean distance.

Step 3: Repeat Step 2 until no more reassignments have been made.

Remark: We could begin the procedure by specifying K initial centroids and proceed as in step 2.

ex)

	observations	
Item	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Goal: Divide the four items into $K = 2$ clusters

Step 1: Random partition, such as (AB) and (CD)

Step 2:

	centroid	
Cluster	$\bar{x}_1$	$\bar{x}_2$
(AB)	$\frac{5-1}{2} = 2$	2
(CD)	-1	$\frac{-2-2}{2} = -2$

$d^2(A, (AB)) = (5-2)^2 + (3-2)^2 = 10$
$d^2(A, (CD) = (5+1)^2 + (3+2)^2 = 61$ if A is not moved
$d^2(A, (B)) = (5+1)^2 + (3-1)^2 = 40$
$d^2(A, (ACD) = (5-1)^2 + (3+.33)^2 = 27.09$ if A is moved to the (CD) group

Since A is closer to (AB) than (ACD), it is not reassigned.

$d^2(B, (AB)) = (-1-2)^2 + (1-2)^2 = 10$
$d^2(B, (CD)) = (-1+1)^2 + (1+2)^2 = 9$ if B is not moved
$d^2(B, (A)) = (-1-5)^2 + (1-3)^2 = 40$
$d^2(B, (BCD)) = (-1+1)^2 + (1+1)^2 = 4$ if B is moved to the (CD) group

Since B is closer to (BCD) than (AB), it is reassigned and we update the centroid as follows.

	centroid	
Cluster	$\bar{x}_1$	$\bar{x}_2$
(A)	5	3
(BCD)	-1	-1

We check C for reassignment.
$d^2(C, (A)) = (1-5)^2 + (-2-3)^2 = 41$
$d^2(C, (BCD)) = (1+1)^2 + (-2+1)^2 = 5$ if C is not moved
$d^2(C, (AC)) = (1-3)^2 + (-2-.5)^2 = 10.25$
$d^2(C, (BD)) = (1+2)^2 + (-2+.5)^2 = 11.25$ if C is moved to the (A) group

Since C is closer to (BCD) than (AC), it is not reassigned.

Continuing, we find that no more assignment and the final $K = 2$ clusters are (A) and (BCD).

For this final clusters, we have $d^2(A, A) = 0, d^2(A, BCD) = (5 + 1)^2 + (3 + 1)^2 = 52$, $d^2(B, A) = (-1 - 5)^2 + (1 - 3)^2 = 40$. The remaining (squared) distances are as follows:

	Item			
Cluster	A	B	C	D
(A)	0	40	41	89
(BCD)	52	4	5	5

Equivalently, we can determine the $K = 2$ clusters by using the criterion

$$min(E = \sum d^2_{i,\ c(i)})$$

where $d^2_{i,\ c(i)}$ is the squared distance of case i from the centroid of the assigned cluster. In this example, we have 7 possibilities for $K = 2$ clusters:

A, (BCD)
B, (ACD)
C, (ABD)
D, (ABC)
(AB), (CD)
(AC), (BD)
(AD), (BC)

For the A, (BCD) pair:

A $d^2_{A,\ c(A)} = 0$

(BCD) $d^2_{B,\ c(B)} + d^2_{C,\ c(C)} + d^2_{D,\ c(D)} = 4 + 5 + 5 = 14$

Consequently, $\sum d^2_{i,\ c(i)} = 0 + 14 = 14$. For the remaining pairs, you may verity that

B, (ACD) $\sum d^2_{i,\ c(i)} = 48.7$
C, (ABD) $\sum d^2_{i,\ c(i)} = 27.7$
D, (ABC) $\sum d^2_{i,\ c(i)} = 31.3$

$$\begin{gathered}(\mathrm{AB}),\ (\mathrm{CD})\ \sum d^2_{i,\ c(i)} = 28 \\ (\mathrm{AC}),\ (\mathrm{BD})\ \sum d^2_{i,\ c(i)} = 27 \\ (\mathrm{AD}),\ (\mathrm{BC})\ \sum d^2_{i,\ c(i)} = 51.3\end{gathered}$$

Since the smallest $\sum d^2_{i,\ c(i)}$ occurs for (A) and (BCD), this is the final partition.

Remark:

- Depend on initial choice of clusters. Hence the procedure should be repeated for a different choice of initials. In particular, the specification of K could lead to unusual clustering. Outlying observations can produce unusual clusters.
- Useful when the data set is large.
- The SAS procedure FASCLUS uses this method.

11.5 Clustering based on statistical models

Previous methods are intuitive but no statistical model to explain how the observations were produced.

If there are K clusters, then the observation vector for a single object is modeled as arising from the mixing distribution

$$f_{Mix}(\mathbf{x}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}),$$

where each $p_k \geq 0$ and $\sum_{i=1}^{K} p_k = 1$. This distribution $f_{Mix}(\mathbf{x})$ is called a mixture of the K distributions $f_k(\mathbf{x}), k = 1, \cdots, K$ because the observation is generated from the component distribution $f_k(\mathbf{x})$ with probability p_k.

The normal mixture model for one observation is

$$f_{Mix}(\mathbf{x}|\mu_1, \Sigma_1, \cdots, \mu_K, \Sigma_K) = \\ \sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_k)^\top \Sigma_k^{-1}(\mathbf{x}-\mu_k)\right).$$

Inferences are based on the likelihood, which for N objects and a fixed number of clusters K, is

$$L(p_1, \cdots, p_K, \mu_1, \Sigma_1, \cdots, \mu_K, \Sigma_K) = \\ \prod_{j=1}^{N} \left(\sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x}_j - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}_j - \mu_k) \right) \right).$$

<u>Remark</u>:

- The most common mixture model is a mixture of multivariate normal distributions.
- For a given K, the number of parameters for a mixture of multivariate normal distributions $(N_p(\mu_k, \Sigma_k))$ is $K(p+1)(p+2)/2 - 1$ since $K - 1$ for probabilities (sum to 1), $K \times p$ for means, and $K \times p(p+1)/2$ for covariances.
- Because of too many parameters, assumed forms for Σ_k are $\eta \mathbf{I}, \eta_k \mathbf{I}$, and $\eta_k Diag(\lambda_1, \cdots, \lambda_p)$. Additional structures are also possible.
- MCLUST is available in the R software library for an analysis.

11.6 Number of clusters; not well defined

- Dendrogram
- CCC (Cubic Clustering Criteria) > 2; good clusters
  CCC (Cubic Clustering Criteria) < 2; suggests potential clusters
- Hotelling's T^2 statistic

11.7 Extension & Ideas!

A procedure called ACECLUS forms canonical variables based on the original data using the computations similar to CANDISC (Canonical Discrimination).

The clustering is then done on these variables and often produces better results.

Appendix A

References

Azzalini, A. and Capitanio, A (2014). *The Skew-Normal and Related Families*, New York: Cambridge University Press.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, 2nd edition, Pacific Grove: Duxbury Advance Series.

Hogg, R. V., McKean, J. W. and Craig, A. T. (2013). *Introduction to Mathematical Statistics*, 7th edition, New Jersey: Prentice Hall.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6th edition, New Jersey: Prentice Hall.

Schott, J. R. (2005). *Matrix Analysis for Statistics*, 2nd edition, New Jersey: John Wiley & Sons, Inc.