Principal component

① Determine the population principal components for the covariance matrix

$$\Sigma = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}.$$

Also calculate the proportion of the total population variance explained by the I principal component.

Convert the covariance matrix to a correlation matrix ρ.

2) Data on X_1 = sales & X_2 = profits for 10 largest companies in the world give the sample mean & dispersions ~~vect~~ matrices as

$$\bar{X} = \begin{bmatrix} 155.60 \\ 14.70 \end{bmatrix}, \quad S = \begin{bmatrix} 7476.45 & 303.62 \\ 303.62 & 26.19 \end{bmatrix}$$

(i) Determine the sample principal components & their variances for these data.

(ii) Find the proportion of the total sample variance explained by the I PC.

(iii) Compute the correlation coeffs $r_{X_i Y_1}$, $i = 1, 2$. What interpretation, if any, can you give to the I PC.

① Consider the matrix of distances

$$\begin{array}{cc} & \begin{array}{cccc}1 & 2 & 3 & 4\end{array}\\ \begin{array}{c}1\\2\\3\\4\end{array} & \begin{bmatrix}0 & & & \\ 1 & 0 & & \\ 11 & 2 & 0 & \\ 5 & 3 & 4 & 0\end{bmatrix}\end{array}$$

1

Cluster the four items using each of the following procedures.

(a) Single linkage hierarchical procedure

(b) Complete " " "

Draw the dendrograms & compare the results in (a) & (b)

$$\begin{array}{cc} & \begin{array}{ccccc}1 & 2 & 3 & 4 & 5\end{array}\\ \begin{array}{c}1\\2\\3\\4\\5\end{array} & \begin{bmatrix}0 & & & & \\ 4 & 0 & & & \\ 6 & 9 & 0 & & \\ 1 & 7 & 10 & 0 & \\ 6 & 3 & 5 & 8 & 0\end{bmatrix}\end{array}$$

2

3

$$
\begin{array}{l}
\text{JP Morgan} \\
\text{Citibank} \\
\text{Wells Fargo} \\
\text{Royal Dutchshell} \\
\text{Exxon Mobil}
\end{array}
\begin{bmatrix}
1 & & & & \\
.63 & 1 & & & \\
.51 & .57 & 1 & & \\
.12 & .32 & .18 & 1 & \\
.16 & .21 & .15 & .68 & 1
\end{bmatrix}
$$

2.5.2023

Multivariate Analysis

Sample mean $\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$, $k = 1, 2, \dots, p$.

p is the no. of variables

n is the no. of observations on each of the p variables.

Sample variance $s_k^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2$

$s_{kk} = s_k^2$.

Sample covariance

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k),$$

$$i = 1, 2, \dots, p, \quad k = 1, 2, \dots, p$$

Sample correlation coeff (or Pearson's product-moment correlation coefficient)

is $$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

Sample mean array is $\bar{X}$, the sample variance - covariance array is S_n & the sample correlation array R

Problems

(1) Consider the seven pairs of measurements (x_1, x_2):

x_1	3	4	2	6	8	2	5
x_2	5	5.5	4	7	10	5	7.5

Calculate the sample means $\bar{x}_1$ & $\bar{x}_2$, the sample variances s_{11} & s_{22} & the sample covariance s_{12}.

(2) (ii) A morning newspaper lists the following used car prices for a foreign compact with age x_1 measured in yrs & selling price x_2 measured in thousands of dollars.

x_1	1	2	3	3	4	5	6	8	9	11
x_2	18.95	19.00	17.95	15.54	14	12.95	8.94	7.49	6	3.9

Interpret all the descriptive measures $\bar{x}_1, \bar{x}_2$, $s_{11}, s_{22}, s_{12}, r_{12}$.

(iii) The following are 5 measurements on the variables x_1, x_2 & x_3

x_1	9	2	6	5	8
x_2	12	8	6	4	10
x_3	3	4	0	2	1

Find the arrays $\bar{X}$, S_n & R.

__Note:__ $\cos(\theta_{ik}) = r_{ik}$

The unbiased sample Variance - covariance matrix of the pofm dispersion matrix Σ is

$$S = \frac{n}{n-1} S_n$$

$$= \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X})(X_j - \bar{X})'$$

i.e., $$S_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} (X_{ji} - \bar{X}_i)(X_{jk} - \bar{X}_k)$$

Generalised sample variance $= |S|$.

Generalised sample variance of the standardised variables is $|R|$

Note: trace of a matrix = sum of its eigen values.

determinant of a matrix is product of its eigen values

Total sample variance $= \sum_{i=1}^{p} s_{ii}$

$$\underset{p \times p}{D^{1/2}} = \cancel{\text{diag}} \begin{bmatrix} \sqrt{s_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{s_{22}} & 0 \cdots & 0 \\ 0 & 0 & \sqrt{s_{33}} & 0 \cdots 0 \\ \cdots & \cdots & \cdots & \sqrt{s_{pp}} \end{bmatrix}$$

$$(D^{1/2})^{-1} = D^{-1/2} = \begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{s_{22}}} & 0 \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix}$$

$$R = D^{-1/2} S D^{-1/2} \quad \&$$

$$S = D^{1/2} R D^{1/2}$$

The linear combinations

$b'X = b_1 X_1 + b_2 X_2 + \cdots + b_p X_p$ &

$c'X = c_1 X_1 + c_2 X_2 + \cdots + c_p X_p$ have

sample means, variances & covariances that are related to $\bar{x}$ & S by

Sample mean of $b'X = b'\bar{X}$

" " $c'X = c'\bar{X}$

" variance of $b'X = b'Sb$

" " $c'X = c'Sc$

Sample covariance of $b'X$ & $c'X = b'Sc$.

<u>Result</u> $A_{q \times p} X_{p \times 1}$

The q linear combinations AX have sample mean vector $A\bar{X}$

Factor Analysis

① **Attribute (correlation)**

$$
\begin{array}{lc}
 & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{array}{r} \text{Taste} \quad 1 \\ \text{Good buy for money} \quad 2 \\ \text{flavor} \quad 3 \\ \text{Suitable for snack} \quad 4 \\ \text{Provides lots of energy} \quad 5 \end{array} &
\begin{bmatrix} 1 & & & & \\ .02 & 1 & & & \\ .96 & .13 & 1 & & \\ .42 & .71 & .5 & 1 & \\ .01 & .85 & .11 & .79 & 1 \end{bmatrix}
\end{array}
$$

Hint: write $\rho = LL' + \Psi$.

② $$\rho = \begin{bmatrix} 1 & .63 & .45 \\ .63 & 1 & .35 \\ .45 & .35 & 1 \end{bmatrix}$$

Calculate the loading matrix L & matrix of specific variances Ψ using the principal component solution method.

What proportion of the total population variance is explained by the first common factor.

③ $$\Sigma = \begin{bmatrix} 1 & .4 & .9 \\ .4 & .1 & .7 \\ .9 & .7 & 1 \end{bmatrix}$$