**Department of Statistics,**

**Modern college of Arts, Science and Commerce, Pune-05**

**M.Sc.I (Statistics) Semester II**

**Date:**

**ST- 28 Submission date:**

**Practical No. 12**

**Title :PRINCIPAL COMPONENT**

Q.1Consider, ∑ = 5 2

2 2

1. Determine the population principal components y1& Y2. Also calculate the proportion of total population variance explained by the first principal component.
2. Convert matrix ∑ into a correlation matrix ρ and compute the proportion of total (standardized) population variance explained by first principle component.
3. Compare the components calculated in part (a) with those obtained in (b). Are they same? Should they be?
4. Compute the corr. ρ (Y1, Z1), ρ (Y2, Z1) and ρ (Y2, Z2) when Zi; i = 1,2 are the standardized variables of the original variables.

Q.2 ∑ = 2 0 0

0 4 0

0 0 4

Determine the principal components, Y1, Y2 and Y3. What can you say about

The eigenvectors and principal components associated with the eigenvalues

that are not distinct?

Q.3 Data on x1 : sales, x2 : profile for the ten largest U.S. industrial are as given below.

62.309 100.05 2.56

x = , s =

2.927 2.56 0.14

a. Determine the sample principal components and their var. for these data.

b. Find the proportion of total sample variance explained by 

c. Compute the correlation coefficient K = 1, 2. Comment on it.

Q.4 Convert the covariance matrix S given in Q.3 to a sample correlation matrix R :

a. Find the sample principal components  and their variances.

b. Compute the proportion of the total sample variance explained by.

c. Compute the correlation coefficient r are k = 1,2. Interpret.

d. Compare the components obtained in (a) with those obtained in Q.3 (a). Do you feel that it is better to determine Principal component from sample covariance matrix or sample correlation matrix?Explain.

Q5. Perform principal component analysis using the sample covariance matrix of the sweat data given below. Draw a scree plot and interpret it. Construct a Q-Q plot for each of the important principal components. Are there any suspect observation? Explain why?

|  |  |  |  |
| --- | --- | --- | --- |
| Individual | Sweat rate (X1) | Sodium (X2) | Potassium(X3) |
| 1 | 3.7 | 48.5 | 9.3 |
| 2 | 5.7 | 65.1 | 8 |
| 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 |
| 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 |

Process in minitab

1. Enter sigma matrix (Given) (Q1, Q2,Q3)

If sigma is unknown → Estimate from given sample.(Q5)

I = Identity matrix , E = matrix of ones

For Q. 4 use this

Convert the covariance matrix S into sample correlation matrix R:

Use Corrcov(S) in MATLAB only

Corrcov→ returns the correlation matrix R corresponding to the covariance matrix S.

1. Find out principal components Y1, Y2, … , Yp
2. Find eigen value and eigen vector of sigma matrix (variance-covariance matrix)

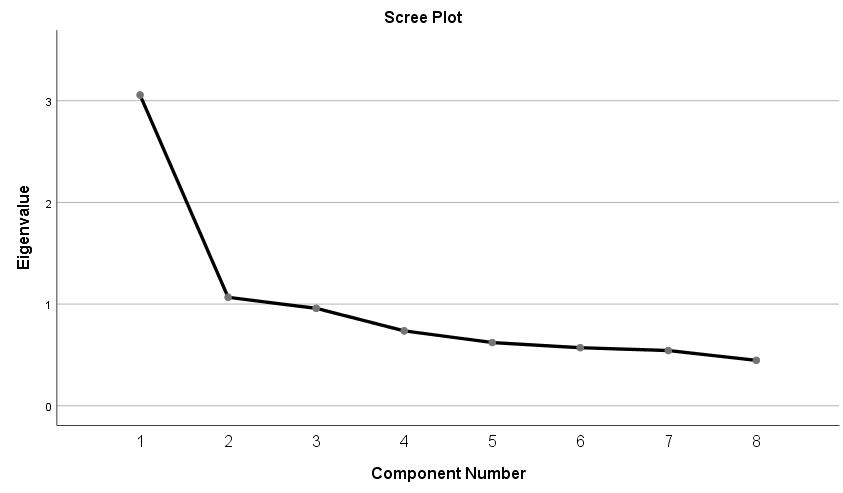
𝜆1≥𝜆2≥…≥𝜆p → eigen values

l1, l2, …, lp → eigen vectors.

⁞

1. Now how to determine an approximate number of principal components.
2. By scree plot
3. Percentage variation explained
4. Scree plot

Plot of 𝜆i vs i



Consider all principal components before elbow i.e. consider 1st  PCA

1. Percentage variation explained.

Percentage of variation explained by 1st P.C. =

Percentage of variation explained by 2nd P.C. =

Percentage of variation explained by 1st two P.C. =

1. Var(Yi) = 𝜆i

Variance of ith P.C. = 𝜆i

1. Correlation between Yi and Xk

Pik = Measurable importance of the kth variable to ith P.C.

1. We ca standardize the variables X1, X2, … , Xp i.e., Z1, Z2, … , Zp

Where,

We can find out P.C. for this standardized variable. Where cov(Z) = 𝜌

i.e. correlation matrix

Var(Zi) = 1 and

Variation explained by ith P.C. of Z is,

Q1)

a) Determine the population principal components y1& Y2. Also calculate the proportion of total population variance explained by the first principal component.

>> sigma = [5 2;2 2]

sigma =

5 2

2 2

>> [v d] = eig(sigma)

v =

0.4472 -0.8944

-0.8944 -0.4472

d =

Diagonal Matrix

1 0

0 6

>> # Y1= -0.8944(X1)-0.4472(X2) .......Population Principal Component

>> # Y2= 0.4472(X1)-0.8944(X2)

>> # % variation explains by first principal component

>> # and is given by (L1/L) \*100% .... L1 is 1st eigen value and L is sum of all eigen values

>> # and is given by PC1=(L1/L) \*100% .... L1 is 1st eigen value and L is sum of all eigen values

>> PC1= (6/7) \*100

**PC1 = 85.714 %**

The first principal component explains **85.714 %** of the total variation

b) Convert matrix ∑ into a correlation matrix ρ and compute the proportion of total (standardized) population variance explained by first principle component.

x=matrix (c (5,2,2,2), nrow = 2, ncol = 2)

{> x

[,1] [,2]

[1,] 5 2

[2,] 2 2

> cov2cor(x)

[,1] [,2]

[1,] 1.0000000 0.6324555

[2,] 0.6324555 1.0000000

}

>> R=[1 0.6324555;0.6324555 1]

R =

1.0000 0.6325

0.6325 1.0000

>> [v d] = eig(R)

v =

-0.7071 0.7071

0.7071 0.7071

d =

Diagonal Matrix

0.3675 0

0 1.6325

>> L11 = d(2,2) #............. L11 is 1st eigen value

L11 = 1.6325

>> L22 = d(1,1) #..............L22 is 2st eigen value

L22 = 0.3675

>> P11\_ = v (: ,2) #............... orthonormal eigen vector of corresponding eigen value

P11\_ =

0.7071

0.7071

>> P22\_ = v (: ,1) #............... orthonormal eigen vector of corresponding eigen value

P22\_ =

-0.7071

0.7071

>> # first principle component of correlation matrix is given by PC1\_R=(L11/L)\*100% .... L1 is 1st eigen value and L is sum of all eigen values

>> PC1\_R = (1.6325/ (1.6325+0.3675)) \*100

**PC1\_R = 81.625**

The first principal component explains **81.625 %** of the total variation

**c)** Compare the components calculated in part (a) with those obtained in (b). Are they same? Should they be?

% Variation explains by first principal component by covariance matrix is **PC1 = 85.714 %**

is greater than by correlation matrix is **PC1\_R = 81.625 %**

And it is not compulsory both have too same

d)Compute the corr. ρ (Y1, Z1), ρ (Y2, Z1) and ρ (Y2, Z2) when Zi; i = 1,2 are the standardized variables of the original variables.

# We know that

# Rik=Pik \* sqrt (Li) ……Pikisist unit of kth row of orthonormal vector of corresponding eigen value and Li is ith eigen value

**#1) ρ (Y1, Z1)**

>> R\_Y1Z1 = P11\_ (1,1) \* sqrt(L11)

R\_Y1Z1 = 0.9035

**#2)ρ (Y2, Z1)**

>> R\_Y2Z1 = P22\_ (2,1) \*sqrt(L22)

R\_Y2Z1 = 0.4287

**#3)ρ (Y2, Z2)**

>> R\_Y2Z1 = P22\_ (1,1) \*sqrt(L22)

R\_Y2Z1 = -0.4287

Q2) Determine the principal components, Y1, Y2 and Y3. What can you say about

The eigenvectors and principal components associated with the eigenvalues

that are not distinct?

sigma =

2 0 0

0 4 0

0 0 4

>> [v d] = eig(sigma)

v =

1 0 0

0 1 0

0 0 1

d =

Diagonal Matrix

2 0 0

0 4 0

0 0 4

>>#Principal Components are

>> #Y1 = 1(X3)

>> #Y2 = 1(X2)

>> #Y3 = 1(X1)

**Q.3** Data on x1 : sales, x2 : profile for the ten largest U.S. industrial are as given below.

62.309 100.05 2.56

x = , s =

2.927 2.56 0.14

a. Determine the sample principal components and their var. for these data.

b. Find the proportion of total sample variance explained by 

c. Compute the correlation coefficient K = 1, 2. Comment on it.

**Ans.**

>> s=[100.05 2.56;2.56 0.14]

s =

100.0500 2.5600

2.5600 0.1400

>> [v d]=eig(s)

v =

0.025598 -0.999672

-0.999672 -0.025598

d =

Diagonal Matrix

7.4448e-02 0

0 1.0012e+02

**a)**

>> #sample principal components are

>> #Y1 = -0.9996(X1)-0.0255(X2)

>> #Y2= 0.0256(X1)-0.9997(X2)

>> # Variance of sample principal components

>> #var (Yi)=Li ......Li is ith eigen value

>> var\_Y1=d (2,2)

var\_Y1 = 100.12

>> var\_Y2=d(1,1)

var\_Y2 = 0.074448

**b)**

>> #proportion of total sample variance explained by Y1

>> L=d (1,1) +d (2,2) …...sum of eigen values

L = 100.19

>> L1=d (2,2) ……first eigen value

L1 = 100.12

>> PC=(L1/L) \*100

Ans = 99.926

The first principal component explains **99.926 %** of the total variation

**C)** Compute the correlation coefficient ,k= 1, 2. Comment on it.

ryk, y1 = {Pik \* sqrt (Li)} / (sqrt (σkk))

…………Pik iskth variable to ith PC

Li is ith eigen value

σ kk iss (k, k)

1)

r\_y1x1=P1\_ (1,1) \*sqrt(L1)/sqrt (s (1,1))

**r\_y1x1 = -1.0000**

2)

r\_y1x2=P1\_ (2,1) \*sqrt(L1)/sqrt(s(2,2))

r\_y1x2 = -0.6845

**Q4)** Convert the covariance matrix S given in Q.3 to a sample correlation matrix R :

a. Find the sample principal components  and their variances.

b. Compute the proportion of the total sample variance explained by.

c. Compute the correlation coefficient r are k = 1,2. Interpret.

d. Compare the components obtained in (a) with those obtained in Q.3 (a). Do you feel that it is better to determine Principal component from sample covariance matrix or sample correlation matrix? Explain.

**Ans.**

Convert the covariance matrix S given in Q.3 to a sample correlation matrix R :

> s=matrix (c (100.05,2.56,2.56,0.14), nrow = 2, ncol = 2)

> s

[,1] [,2]

[1,] 100.05 2.56

[2,] 2.56 0.14

> R=cov2cor(s1) # ……. By using R software

[,1] [,2]

[1,] 1.0000000 0.6840178

[2,] 0.6840178 1.0000000

**a)** sample principal components  and their variances.

>> R= [1.0000000 0.6840178;0.6840178 1.0000000]

R =

1.0000 0.6840

0.6840 1.0000

>> [v d] =eig(R)

v =

-0.7071 0.7071

0.7071 0.7071

d =

Diagonal Matrix

0.3160 0

0 1.6840

>> L1= d (2,2)

L1 = 1.6840

>> L2=d (1,1)

L2 = 0.3160

P1\_ =

0.7071

0.7071

P2\_ =

-0.7071

0.7071

>> #a) sample principal components Y1 & Y2 and their variances

>> #Y1 = 0.7071(X1) + 0.7071(X2)

>> #Y2 = -0.7071(X1) + 0.7071(X2)

>> #Var (Yi) = Li …...Li is ith eigen value

>> Var\_Y1=L1

Var\_Y1 = 1.6840

>> Var\_Y2=L2

Var\_Y2 = 0.3160

**b)** Compute the proportion of the total sample variance explained by.

#L1=first eigen value & L =sum of eigen values

PC=L1/L\*100 #.........% of Variation explain by First Principal Component

PC = 84.201

**The first principal component explains 84.201 % of the total variation**

**c)** Compute the correlation coefficient r are k = 1,2. Interpret.

ry1, zk = Pik \* sqrt (Li) #……...Pik measures the importance of the kth variable to ith PC

1) )

>> r\_y1z1=P1\_ (1,1) \*sqrt(L1) #...... P1\_ = eigen vector corresponding to eigen value

r\_y1z1 = 0.9176

2)

>> r\_y1z2=P1\_ (2,1) \*sqrt(L1) #...... P1\_ = eigen vector corresponding to eigen value

r\_y1z2 = 0.9176

**d)**Compare the components obtained in (a) with those obtained in Q.3 (a). Do you feel that it is better to determine Principal component from sample covariance matrix or sample correlation matrix? Explain.

**Q3.a)**

>> var\_Y1=d (2,2)

var\_Y1 = 100.12

>> var\_Y2=d(1,1)

var\_Y2 = 0.074448

**Q4.a)**

>> Var\_Y1=L1

Var\_Y1 = 1.6840

>> Var\_Y2=L2

Var\_Y2 = 0.3160

Variance of first principal component in Q.3 is **100.12** which is **greater** than variance of first principal component in Q.4 which is **1.684**

But variance of second principal component in Q3.(**0.07444**) is **less** than Q4.(**0.3160**)

Q5)

>> X=[X1 X2 X3]

X =

3.7000 48.5000 9.3000

5.7000 65.1000 8.0000

3.8000 47.2000 10.9000

3.2000 53.2000 12.0000

3.1000 55.5000 9.7000

4.6000 36.1000 7.9000

2.4000 24.8000 14.0000

7.2000 33.1000 7.6000

6.7000 47.4000 8.5000

5.4000 54.1000 11.3000

3.9000 36.9000 12.7000

4.5000 58.8000 12.3000

3.5000 27.8000 9.8000

4.5000 40.2000 8.4000

1.5000 13.5000 10.1000

8.5000 56.4000 7.1000

4.5000 71.6000 8.2000

6.5000 52.8000 10.9000

4.1000 44.1000 11.2000

5.5000 40.9000 9.4000

>> s=1/19\*X'\*(eye (20)-1/20\*ones (20)) \*X

s =

2.8794 10.0100 -1.8091

10.0100 199.7884 -5.6400

-1.8091 -5.6400 3.6277

>> [v d] =eig(s)

v =

-0.050841 -0.817484 -0.573704

-0.998284 0.024877 0.053020

0.029072 -0.575415 0.817345

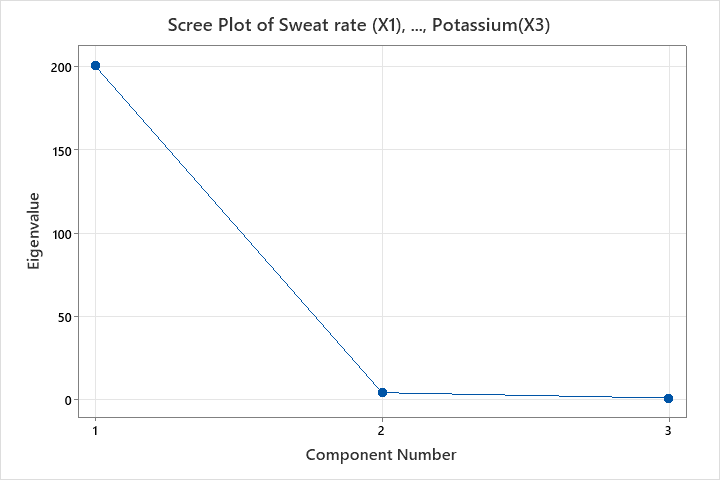
d =

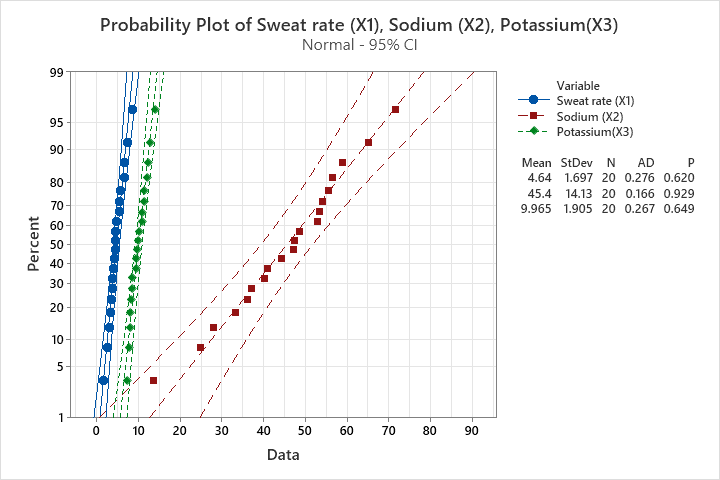
Diagonal Matrix

200.4625 0 0

0 1.3014 0

0 0 4.5316





Conclusion: All the three variables follows normality

**#Li = eigen values, i=1,2,3**

>> L1=d(1,1)

L1 = 200.46

>> L2=d(3,3)

L2 = 4.5316

>> L3=d(2,2)

L3 = 1.3014

**#Pi\_ = orthonormal eigen vector corresponding to eigen value (Li) respectively, i=1,2,3**

>> P1\_=v(:,1)

P1\_ =

-0.050841

-0.998284

0.029072

>> P2\_=v(:,3)

P2\_ =

-0.573704

0.053020

0.817345

>> P3\_=v(:,2)

P3\_ =

-0.817484

0.024877

-0.575415

# Sample Principal Component Y1, Y2 and Y3

>> #Y1= -0.050841(X1) - 0.998284(X2) + 0.029072(X3)

>> #Y2= -0.573704(X1) + 0.053020(X2) +0.817345(X3)

>> #Y3= -0.817484(X1) + 0.024877(X2) - 0.575415(X3)

#Percentage of variation explain by ith PC

# And is given by (Li/L) \*100%........Li ith eigen value and L is sum of eigen value

PC\_1=