25. 02. 2024

PCA

Multivariate

5

1. Dimension
2. Sorting & grouping
3. Dependence among the variables R / CL
4. Predictive — Time, Prob.
5. Hyp. Construction & Testing

PCM → Basic Sc.

EB → Language

SG → Social Sc.

Subjects.

7

G1 · 2 — P C M

G2 · 2 — E B

G3 · 2 — H G

Trans: (or) P, C

7 variables.

	P	C	M	R	E	S	G
1	✓	✓	✓	✓	✓	✓	✓
2	✓	✓	✓				
:	✓						
100	✓						

100 Rows   7 Cols.   100 × 7.

Step 1: Find mean & s.d. for each variable.

Step 2 (very Imp): Find the Covariance matrix.

Step 3: Find the Correlation matrix which is very simple to understand.

Covariance matrix:

$$\begin{array}{c} \\ v \end{array}\begin{pmatrix} x_1 & x_2 & x_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Data matrix.

$v(x_1)$, $v(x_2)$, $v(x_3)$

$C_{12} = C_{21}$

$C_{13} = C_{31}$

$C_{23} = C_{32}$

when there are 3 variables.

You want to compute

$({}^3C_1) = 3$ → variances

& To compute covariance $= ({}^3C_2) = 3$.

$$\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\begin{pmatrix} v(x_1) & C_{12} & C_{13} \\ C_{21} & v(x_2) & C_{23} \\ C_{31} & C_{32} & v(x_3) \end{pmatrix}_{3\times 3}$$

$$\left.\begin{aligned} v(x_1) &= 1 \\ v(x_2) &= 5 \\ v(x_3) &= 2 \end{aligned}\right\} \qquad \begin{aligned} C_{12} &= C_{21} = -2 \\ C_{13} &= C_{31} = 0 \\ C_{23} &= C_{32} = 0 \end{aligned}\Bigg\}$$

$$\begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}_{3\times 3}$$

$x_1 \quad x_2 \quad x_3 \quad x_3$

no of variances $= ({}^4C_1) = 4$.

no of covariances $= ({}^4C_2) = 6$

$$\begin{pmatrix} v(x_1) & C_{12} & C_{13} & C_{14} \\ C_{21} & v(x_2) & C_{23} & C_{24} \\ C_{31} & C_{32} & v(x_3) & C_{34} \\ C_{41} & C_{42} & C_{43} & v(x_4) \end{pmatrix}_{4\times 4}$$

$v(x_1) = 1 \quad v(x_2) = 5 \quad v(x_3) = 3$

$v(x_4) = 6$

$c_{12} = 2 \quad c_{13} = 0 \quad c_{14} = -2$

$c_{23} = 5 \quad c_{24} = -1 \quad c_{34} = 3$

$4 + 6 = 10$

$$\begin{pmatrix} 1 & 2 & 0 & -2 \\ 2 & 5 & 5 & -1 \\ 0 & 5 & 3 & 3 \\ -2 & -1 & 3 & 6 \end{pmatrix}$$

4×4

7×7 Correlation

$$\begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} & r_{15} & r_{16} & r_{17} \\ & 1 & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 1 & & \\ & & & & & 1 & \\ & & & & & & 1 \end{pmatrix}$$

7 variables

$C(7,2) = \frac{7 \times 6}{2} = 21$

P
C
M

G1. = no. of Correlation

$= \left({}^{3}C_{2}\right) = 3$

G2

B
Z

$= \left({}^{2}C_{2}\right) = 1$

G3

Sl
G

$= \left({}^{2}C_{2}\right) = 1$

$= 5$

7

G1 G2 G3

3 Dimension Reduction

Banking Cement Food

0 2

P
C
M

Q1. How many members in the group?

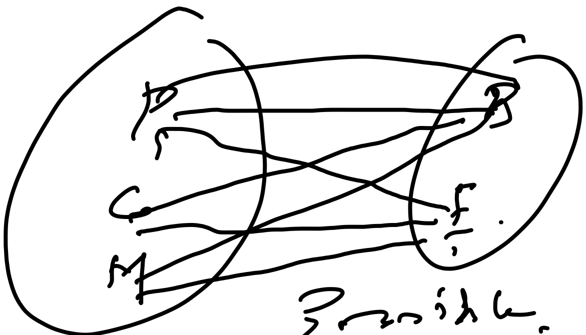
In G1. no of members = 3.

In G_2 no of members = 2

In G_3 " " " = 2.

Association may occur between

G_1 & G_2. G_1 & G_3 G_2 & G_3.



G_1 & G_3

6

no of a possible association = 6

G_2 & G_3

$[{}^7C_2]$

4

$6+6+4$
$= 16$

$$\boxed{21 = 5 + 16}$$

2

Total x y z.

Combination among x, y & z

3

xy. xz. yz.

$x_1 \quad x_2 \quad x_3$. $\qquad x_1 = DV$

x_2 & x_3 $= IV$.

$R_{1 \cdot 23}$. $\qquad$ C.

$r_{12} = 27$

Three types of model

1. Independent.
2. Hypothetical.
3. Saturated.

$x_1 \quad x_2 \quad x_3$.

They are completely ind.

$r_{12} = 0$. $\quad r_{13} = 0 \quad r_{23} = 0$

$$I = \begin{array}{c|ccc} & x_1 & x_2 & x_3 \\ \hline x_1 & 1 & 0 & 0 \\ x_2 & 0 & 1 & 0 \\ x_3 & 0 & 0 & 1 \end{array}$$

Identity or unit matrix.

$$\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1. \qquad |I| = 1.$$

$$
\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}
\begin{pmatrix} 1 & 0.99 & 0.98 \\ 0.99 & 1 & 0.90 \\ 0.97 & 0.90 & 1 \end{pmatrix}_{3\times 3}
$$

$r_{12} = 0.99$. $r_{13} = 0.97$ $r_{23} = .90$

$$\boxed{|A| \to 0.}$$

If p-value > 0.05

Then $\Rightarrow$ No linear association / correlation between two variables.

p-value ≤ 0.05

9 variables $=$ Iv.

1 Dv.

To get a model.

To improve R^2.

4 7

Step wise Regression

2 $R^2 = 0.743$

9 $R^2 = 0.791 \rightarrow$ Maximum

1 $R^2 = 0.49 \rightarrow$ Minimum

2 $R^2 = 0.633$

min 0.49 (1)	0.633 (2)	0.743 (2)	max 0.791 (9)

Statistically significant

PCA R^2

02.03.2024

Data matrix → 1 DV
>1 IV.

Regression

R^2, Adj R^2.

VIF — VIF for ind variables are very high among the ind.

Multi Collinearity

High Correlation variance.

Max variation occur in high VIF value

Regression failed.

Inputs / IV. → Apply PCA.

Factor Analysis.

Step. (1) For the theoretical concept we calculate covariance matrix.

(2) For applied part, we will correlation matrix

	x_1	x_2	x_3
1	0	0	0
2	0	0	0
.	0	0	0
100	0	0	0

100×3

3 × 10

mean of x_1

" " x_2

" " x_3

μ_1

μ_2

μ_3

$v(x_1)$ c_{12}

$v(x_2)$ c_{13}

$v(x_3)$ c_{23}

r_{12} r_{13} r_{23}.

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}_{3\times 3}$$

where $Row = Col \Rightarrow$ Diagonal.

$Row \neq Col \Rightarrow$ off diagonal

$$\rho = \begin{pmatrix} 1 & r_{12} = -0.893 & 0 \\ -0.893 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}_{3 \times 3}$$

$$r_{12} = \frac{C_{12}}{\sigma_1 \sigma_2} = \frac{-2}{\sqrt{1}\sqrt{5}}$$

$$= \frac{-2}{2.24}$$

$$= \underline{-0.893}$$

Step (3): we can apply PCA. [Principle Component Analysis] on Covariance matrix or we can apply on Correlation matrix]

For understanding PCA. you have to know polynomials & roots..

$2x + 3 = 0$ => 1st degree polynomial

$x = -3/2$ => $(-3/2)$ root

RHS = 0.

$LHS = 2x+3$

$= 2\left(\frac{-3}{2}\right) + 3 = 0 = RHS$

Zero

$x^2 - 5x + 6 = 0$

$(x-2)(x-3) = 0$

$x = 2$ or 3. Roots

Degree of Polynomial = 2/3/4

no of roots = 2/3/4

P C M B E H G (7)

P C M

B E

H G

No of variables = 7

we can construct a seven degree polynomial

⇓

we get seven roots

⇓

Eigen values.

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}_{3\times 3}$$

PCA → Diagonal

How much variance can be explained?

$$|\Sigma - \lambda I| = 0 \Rightarrow \text{Polynomial}$$

$$\left| \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right| = 0$$

$$\begin{array}{ccc} a_{11} & a_{12} & a_{13} \\ \left|\begin{matrix} 1-\lambda \\ -2 \\ 0 \end{matrix}\right. & \begin{matrix} -2 \\ 5-\lambda \\ 0 \end{matrix} & \left.\begin{matrix} 0 \\ 0 \\ 2-\lambda \end{matrix}\right| \end{array} = 0$$

$$(1-\lambda)\left[(5-\lambda)(2-\lambda)\right] - (-2)\left[-2(2-\lambda)\right] = 0$$

$$(1-\lambda)(5-\lambda)(2-\lambda) - 4(2-\lambda) = 0$$

Polynomial

3rd degree

$$\left.\begin{aligned} \lambda_1 &= 5.83 \\ \lambda_2 &= 2.00 \\ \lambda_3 &= 0.17 \end{aligned}\right\} \quad \lambda_1 + \lambda_2 + \lambda_3 = 8 = v(\lambda_1) + v(\lambda_2) + v(\lambda_3)$$

$$= \begin{array}{ccc} + & - & + \\ a_{11} & a_{12} & a_{13} \\ \left|\begin{matrix} a_{21} \\ a_{31} \end{matrix}\right. & \begin{matrix} a_{22} \\ a_{32} \end{matrix} & \left.\begin{matrix} a_{23} \\ a_{33} \end{matrix}\right| \end{array} = 8$$

$\Delta = a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{33}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$

In cases of correlation matrix, of [illegible] variables.

$$\begin{pmatrix} 1 & -0.89 & 0 \\ -0.89 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}_{3\times 3}$$

$\lambda_1 + \lambda_2 + \lambda_3 = 3$

Objective of PCA.

Dimension Reduction.

Why? No. of Eigen values > 1 [illegible]

= no. of [illegible]

Sir

$\lambda_1 > 1$

$\lambda_2 > 1$

Two Eigen values > 1.

λ_3

λ_4

λ_5 < 1.

λ_6

Four Eigen values < 1.

$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6 = 6$

Contribution of 1st Eigen values.
" " 2nd " "

Contribution of other four eigen values (Balancing figure).

$\lambda_1 = 3.253$. $\quad 3.253/6 = 54\%$

$\lambda_2 = 1.654$. $\quad 1.654/6 = 27\%$

Total contribution by two eigen values $= 81\%$ explained by the two eigen values.

Remaining. $(100 - 81)\% = 19\%$ also explained by the four eigen values < 1.

x_1 x_2 x_3

x_4 x_5 x_6 ⇒ 6 dimension

Two or three

Correlation matrix → 6x6.

PCA. He helps ?

determined

x_1 x_2 . . x_6

x_1 > x_2 > ?

x_n <> -

How much explained by 1ST Eigenvalue

good.

> 80%.

20%.

G_1 $\quad$ G_2

λ_1 $\quad$ λ_2

$\lambda_1 > 1$

$\lambda_2 > 1$

$\lambda_3 \quad \lambda_4 \quad \lambda_5 \quad \lambda_6$

Explained $\quad$ Explained

$$\frac{\lambda_1}{\lambda_1 + \cdots + \lambda_6} + \frac{\lambda_2}{\lambda_1 + \lambda_2 \cdots + \lambda_6} + \frac{\lambda_3 + \lambda_4 + \lambda_5 + \lambda_6}{\lambda_1 + \cdots + \lambda_6} = 1$$

54% $\quad$ 27% $\quad$ 19%

81% + 19% = 100%

Explained $\quad$ Unexplained

Dimension is reduced

$6 \longrightarrow 2$

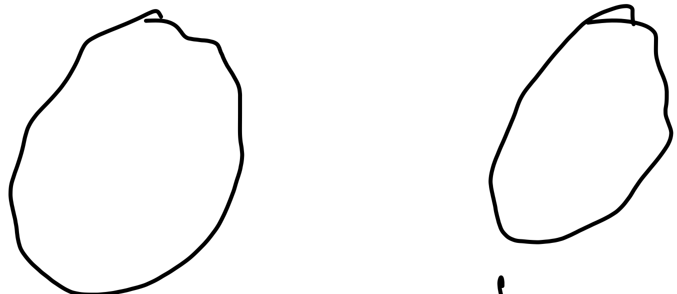
Explained $\geq$ 70%

ML Explained is 90%

PCA - helped us to determine the number of possible homogeneous groups with the help of eigen value ; G_1 G_2.

BIG7

members

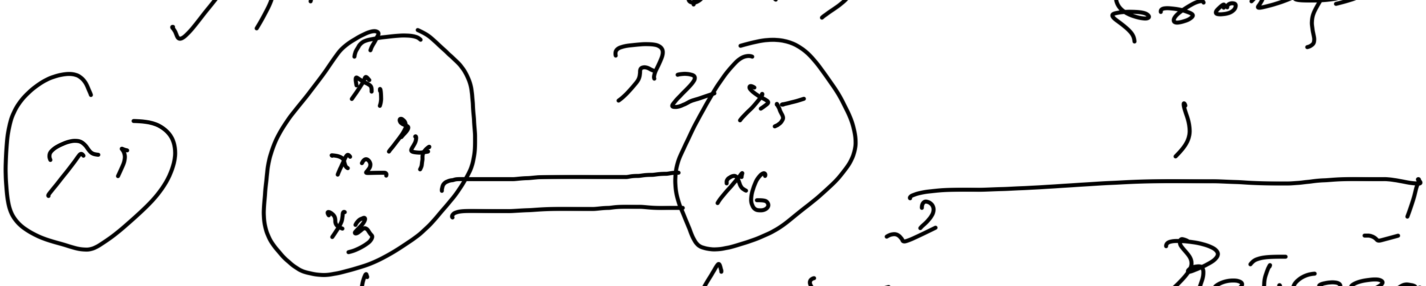


Since Eigen value based on Correlation matrix & correlates pairwise so any group occurs contains minimum two variables

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6$

$\lambda_1 > 1, \quad \lambda_2 > 1$

✓G_1 ✓G_2 Groups



Correlation of within groups = increase.

Correlation of Between group

G1 = internal

	x_1	x_2	x_3	x_4
x_1	1	r_{12}	r_{13}	r_{14}
x_2	r_{21}	1	r_{23}	r_{24}
x_3	r_{31}	r_{32}	1	r_{34}
x_4	r_{41}	r_{42}	r_{43}	1

x_1, x_2, x_3, x_4

4 variables

4×4

r_{12} High

r_{14} "

r_{23} "

r_{34} "

r_{13} "

r_{24}

1 2 3 4

12 13 14 23 24 34

$4C_2 = 6$ pairs

1 2 3 4 5

12 13 14 15 23 24 25 34 35 45

$(5C_2)$

$= 10$

"Minimize Risk.

Total Risk = Systematic + unsystematic.

Diversification.

Investors

Loss.

s_1

s_2

.

s_5

x_1

x_2

.

x_5

Sp 5

$x_1 \ldots x_5$

x_1

x_2

x_3

Factor Analysis.