Data Science and Social Inquiry: HW2

R11323006 陳柏語 R11323015 張藝懷 B07303119 劉怡婷
B08303124 劉詠晴 B09303052 蔡尚恩

October 20, 2022

Question 1: PCA with non-diagonal covariance matrix

(a) (2 pts) What is the first principal component? Explain how you reach your answer carefully and write down its coefficients.

Sol.

$max_{b_1 \in \mathbb{R}^P}\ {b_1}' D b_1$ s.t.${b_1}' b_1 = 1$
$max_{b_1 \in \mathbb{R}^P}\ {b_1}' D b_1 = 0.855{b_{11}}^2 + 0.942{b_{12}}^2 + 0.738{b_{13}}^2 + 0.109{b_{14}}^2 + 2.024{b_{15}}^2$ s.t.${b_1}' b_1 = 1$

Since 2.024 is the biggest number in five coefficients, we can know $b_{15} = \pm 1, b_{11} = b_{12} = b_{13} = b14 = 0$

$$b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = P' a_i$$

$$a = Pb = \begin{bmatrix} -0.135 & -0.763 & 0.606 & -0.182 & 0.006 \\ -0.27 & -0.477 & -0.745 & -0.271 & 0.268 \\ -0.405 & -0.191 & -0.062 & 0.892 & -0.015 \\ -0.539 & 0.095 & -0.065 & -0.242 & -0.798 \\ -0.674 & 0.381 & 0.266 & -0.197 & 0.539 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.006 \\ 0.268 \\ -0.015 \\ -0.798 \\ 0.539 \end{bmatrix}$$

$$PC_1 = a'X = \begin{bmatrix} 0.006 & 0.268 & -0.015 & -0.798 & 0.539 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix}$$

$$= 0.006X_1 + 0.268X_2 - 0.015X_3 - 0.798X_4 + 0.539X_5$$

(*b*) (1 pt) Calculate the proportion of variance explained by each component and use them to make the scree plot.

Sol.

$\sum_{i=1} p\lambda_i = 0.855 + 0.942 + 0.738 + 0.109 + 2.024 = 4.668$

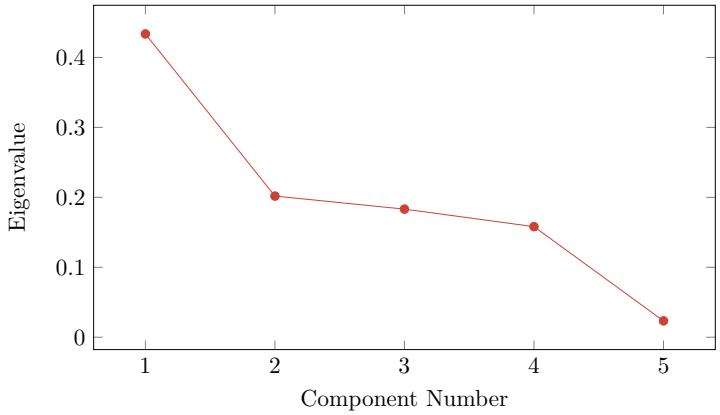
$\frac{2.024}{4.668} = 0.4335904$

$\frac{0.942}{4.668} = 0.20179949$

$\frac{0.855}{4.668} = 0.18316195$

$\frac{0.738}{4.668} = 0.15809769$

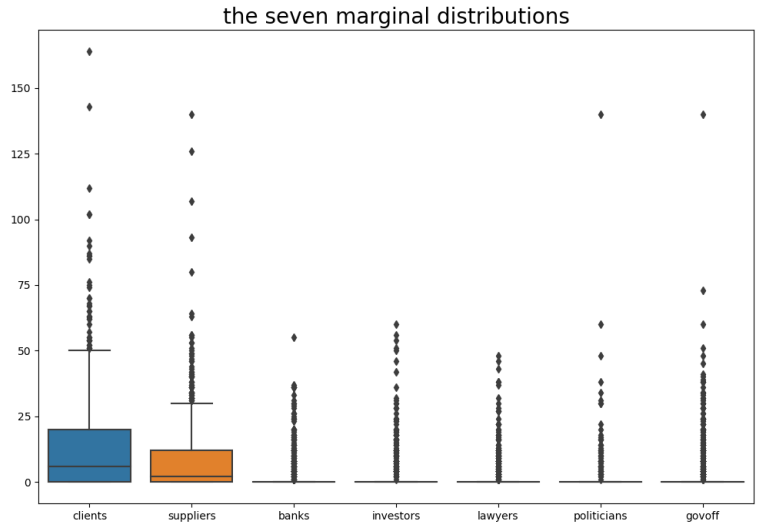
$\frac{0.109}{4.668} = 0.02335047$



Question 2: Prototyping CEO's behavior

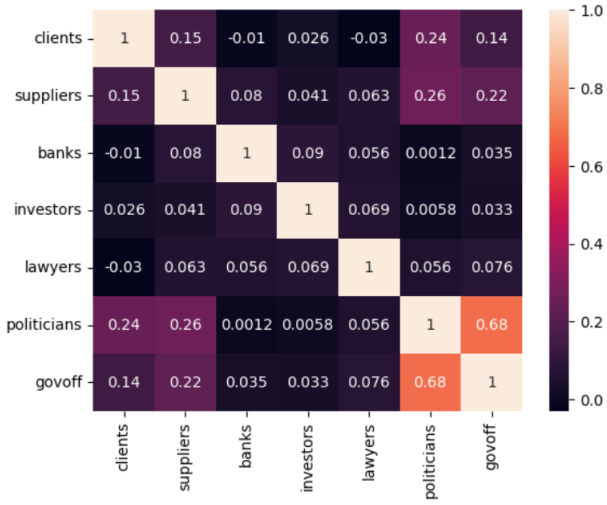
(*c*) (1 pt) Every data analysis should start with examining the raw data. Use a box plot to summarize the seven marginal distributions.

Sol.



(*d*) (1 pt) Use a heatmap to summarize the correlations between the number of activities. Which type correlates with type **politicians** most?

Sol.



We can observe that type govoff correlates with type politicians most.

(*e*) (1 pt) Run PCA. What is the first principal component?

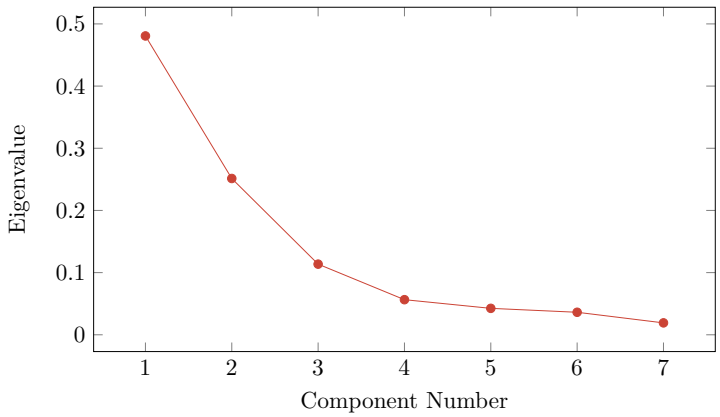
Sol.

$$PCA_1 = 0.954X_1 + 0.252X_2 + 0.002X_3 + 0.013X_4 - 0.003X_5 + 0.109X_6 + 0.115X_7$$

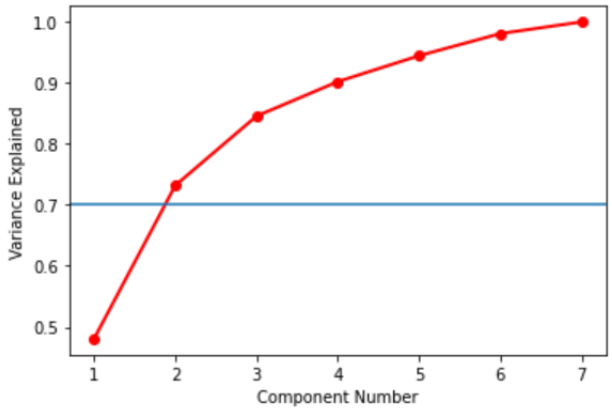
(*f*) (1 pt) Make the scree plot. How many principal components are needed to explain 70% of the variation?

Sol.

Scree Plot



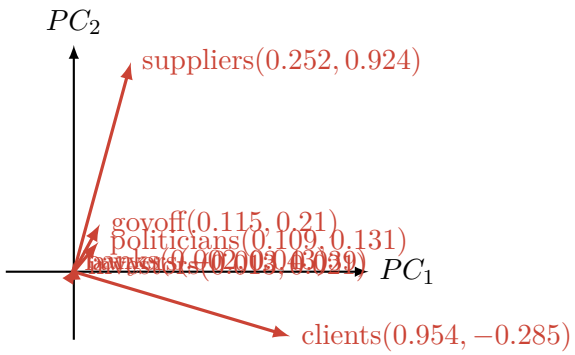
Cummulative Explained Variance



Two principal components are needed to explain 70% of the variation.

(*g*) (1 pt) Put the first component on the x-axis and the second component on the y-axis. Plot the coefficients of each variable. How would you interpret the first two components?

Sol.



PC_1 implies the strong relationship with the clients variable, since the coefficient is the highest.As for PC_2, it is the suppliers variable.

Question 3: Practicing the hierarchical clustering algorithm

(*h*) (1 pt) Perform the hierarchical clustering with the average linkage. Clearly indicate which observations are pooled in each step.

Sol.

STEP1

$$distance_1 = \begin{bmatrix} 0 & 3\sqrt{2} & \sqrt{10} & \sqrt{10} & \sqrt{10} \\ 3\sqrt{2} & 0 & 2\sqrt{10} & 2\sqrt{10} & 2 \\ \sqrt{10} & 2\sqrt{10} & 0 & 2 & 6 \\ \sqrt{10} & 2\sqrt{10} & 2 & 0 & 2\sqrt{10} \\ \sqrt{10} & 2 & 6 & 2\sqrt{10} & 0 \end{bmatrix}$$

$C_1 : (X_1, Y_1)$
$C_2 : (X_2, Y_2), (X_5, Y_5)$
$C_3 : (X_3, Y_3), (X_4, Y_4)$

STEP2

$$distance_2 = \begin{bmatrix} 0 & \frac{3\sqrt{2}+10}{2} & \sqrt{10} \\ \frac{3\sqrt{2}+10}{2} & 0 & \frac{3\sqrt{10}+3}{2} \\ \sqrt{10} & \frac{3\sqrt{10}+3}{2} & 0 \end{bmatrix}$$

$C_1 : (X_1, Y_1), (X_3, Y_3), (X_4, Y_4)$
$C_2 : (X_2, Y_2), (X_5, Y_5)$

STEP3

$C_1 : (X_1, Y_1), (X_3, Y_3), (X_4, Y_4), (X_2, Y_2), (X_5, Y_5)$

(i) (1 pt) Perform the hierarchical clustering with the complete linkage.

Sol.

STEP1

$$distance_1 \begin{bmatrix} 0 & 3\sqrt{2} & \sqrt{10} & \sqrt{10} & \sqrt{10} \\ 3\sqrt{2} & 0 & 2\sqrt{10} & 2\sqrt{10} & 2 \\ \sqrt{10} & 2\sqrt{10} & 0 & 2 & 6 \\ \sqrt{10} & 2\sqrt{10} & 2 & 0 & 2\sqrt{10} \\ \sqrt{10} & 2 & 6 & 2\sqrt{10} & 0 \end{bmatrix}$$

$C_1 : (X_1, Y_1)$
$C_2 : (X_2, Y_2), (X_5, Y_5)$
$C_3 : (X_3, Y_3), (X_4, Y_4)$

STEP2

$$distance_2 \begin{bmatrix} 0 & 3\sqrt{2} & \sqrt{10} \\ 3\sqrt{2} & 0 & 2\sqrt{10} \\ \sqrt{10} & 2\sqrt{10} & 0 \end{bmatrix}$$

$C_1 : (X_1, Y_1), (X_3, Y_3), (X_4, Y_4)$
$C_2 : (X_2, Y_2), (X_5, Y_5)$

STEP3

$C_1 : (X_1, Y_1), (X_3, Y_3), (X_4, Y_4), (X_2, Y_2), (X_5, Y_5)$