BADJI MOKHTAR UNIVERSITY
FACULTY OF ECONOMICS AND MANAGEMENT SCIENCES
Department: Financial Sciences

Module: Data Analysis
Level: Bachlor 3rd Level
Specialization: Marketing and E-Marketing
Instructor: Dr. Soumaya Allaoui

Exam: 5th Semester
Date: January 20th, 2025
Duration: 1 hour 30 minutes

Exercise1:

1. What are the objectives of factor analysis methods?
2. Explain the differences between the methods:
    a. Principal Component Analysis (PCA)
    b. Correspondence Analysis (CA)
    c. Multiple Correspondence Analysis (MCA).
3. When should we apply the operations of centering and redaction to the data in PCA?
4. What is the indicator of association between quantitative variables?
5. Explain the use of the χ^2-square test for Correspondence Analysis (CA).

Exercise2:

During a survey conducted on a sample of size 60, the following contingency table was obtained:

	J_1	J_2	Total
I_1	20	20	40
I_2	75	15	90
I_3	45	35	80
Total	140	70	210

1. Calculate the frequency table associated with the given contingency table. (Tip: Use fractions instead of decimal numbers!)
2. Calculate the χ^2-Square statistic. For $\alpha = 5\%$, what can you conclude about the relationship between the parameters? Deduce the inertia.
3. Calculate the row-profile matrix and deduce the mean row profile.
4. Calculate the diagonal matrix.
5. Calculate the χ^2-distance between the row profiles $d(L_1, L_3)$ and interpret the results

Exercise3:

As a marketing manager, you aim to understand how the climatic characteristics of cities influence consumer behavior and local needs. The following data represents precipitation (P, in cm), maximum temperatures (t_{max}), and minimum temperatures (t_{min}), in (°C) recorded in various cities in 2024:

	p	tmax	tmin
Ajaccio	12.04	23.7	5.9
Brest	17.18	15.5	−1.8
Dunkerque	11.83	13.1	2.8
Nancy	6.23	13.5	−2.4
Nice	16.99	21.1	7.2
Toulouse	3.87	20.3	−0.9

Question

1. Calculate the means and standard deviations of P, t_{max}, and t_{min}.
2. Calculate the dispersion matrix (covariance or correlation) corresponding to this problem statement, and then comment on it.
3. Calculate the variances (eigenvalues) of the principal components for this problem statement.
4. What is the percentage of variance explained by each principal component? And deduce which dimensions to retain.
5. Find the normalized eigenvector corresponding to the first principal component, and calculate the coordinates of the individuals in the space spanned by the eigenvectors.
6. Find the principal component matrix.

Exam solution

Exercise 1: (5pts)

1- **What are the objectives of factor analysis methods?**

To reduce the dimensionality of data by identifying underlying factors or components that explain the observed variability in the data.

2- Explain the differences between the methods: Principal Component Analysis (PCA), Correspondence Analysis (CA), and Multiple Correspondence Analysis (MCA).

PCA (Principal Component Analysis): Primarily used for quantitative (continuous) variables to reduce dimensionality and identify the most important components that explain the variance in the data.

3- When should we apply the operations of centering and scaling to the data in PCA?

centering is always applied, and scaling is applied when variables are on different scales or have different variances

4- Explain the use of the Chi-square test for Correspondence Analysis (CA).

The Chi-square test is used to assess the independence of rows and columns in a contingency table, helping to interpret the associations between categorical variables in CA.

5- What is the indicator of association between quantitative variables?

Covariance shows the relationship between variables, while correlation standardizes it, making it easier to interpret (ranges from -1 to 1).

Exercise 2:(8pts)

1. Frequency Table: *The total sample size* $n = 210$. ***(1pt)***

	J_1	J_2	Total
I_1	$\frac{20}{210}$	$\frac{20}{210}$	$\frac{40}{210}$
I_2	$\frac{75}{210}$	$\frac{15}{210}$	$\frac{90}{210}$
I_3	$\frac{45}{210}$	$\frac{35}{210}$	$\frac{80}{210}$
Total	$\frac{140}{210}$	$\frac{70}{210}$	1

2. χ***-Square Statistic (3pts).***

The formula for the Chi-Square statistic is:. $\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

where O_{ij} are the observed frequencies, and E_{ij} are the expected frequencies.

The expected frequency for each cell is calculated as:

$$1 E_{ij} = \frac{(\text{row total}_i \times \text{column total}_j)}{\text{grand total}}$$

.

	J_1	J_2	Total
I_1	$\frac{80}{3}$	$\frac{40}{3}$	40
I_2	60	30	90
I_3	$\frac{160}{3}$	$\frac{80}{3}$	80
Total	140	70	210

i	j	O_{ij}	E_{ij}	$O_{ij} - E_{ij}$	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
1	1	20	$\frac{80}{3} \approx 26.67$	−6.67	44.49	1.67
1	2	20	$\frac{40}{3} \approx 13.33$	6.67	44.49	3.34
2	1	75	60	15	225	3.75
2	2	15	30	−15	225	7.50
3	1	45	$\frac{160}{3} \approx 53.33$	−8.33	69.39	1.30
3	2	35	$\frac{80}{3} \approx 26.67$	8.33	69.39	2.60

Now, compute the Chi-Square statistic:

$$\chi^2 = 1.67 + 3.34 + 3.75 + 7.50 + 1.30 + 2.60 = 20.16$$

3. Conclusion for Chi-Square (1pt)

For $\alpha = 5\%$, if the Chi-Square statistic exceeds the critical value for 2 degree of freedom (since we have a 3x2 table), we reject the null hypothesis. The critical value is $\chi^2_{0.05,2} = 5.99.$

$\chi^2 > 5.99$, we reject the null hypothesis that the two variables I and J are ***independent.***

4. Row-Profile Matrix and Mean Row Profile:(1pts)

The ***row-profile matrix*** is created by normalizing the rows (i.e., dividing each cell by the row total). For the row-profile matrix, each row corresponds to a probability distribution across the columns.

Thus, the row-profile matrix is:

	J_1	J_2
I_1	0.5	0.5
I_2	0.8333	0.1667
I_3	0.5625	0.4375

The ***mean row profile*** is calculated by taking the average of each column across the rows

Thus, the mean row profile is:

$$\text{Mean Row Profile} = (0.6253, 0.3681).$$

5. Diagonal Matrix : (0.5pts)

4. χ^2-Distance between Row Profiles $d(L_1, L_3)$ (1pts)

$$d(L_1, L_3) = \frac{1}{2} \sum_j \frac{(L_{1j} - L_{3j})^2}{L_{1j} + L_{3j}}$$

For J_1: $\frac{(0.5 - 0.5625)^2}{0.5 + 0.5625} = \frac{(-0.0625)^2}{1.0625} = \frac{0.00390625}{1.0625} \approx 0.003673$

For J_2:

$$\frac{(0.5 - 0.4375)^2}{0.5 + 0.4375} = \frac{(0.0625)^2}{0.9375} = \frac{0.00390625}{0.9375} \approx 0.004167.$$

Now sum these components and multiply by :

$$d(L_1, L_3) = \frac{1}{2}(0.003673 + 0.004167) = \frac{1}{2}(0.007840) \approx 0.00392$$

.

Interpretation of Results : (0.5pts)

Chi-Square distance between the row profiles of I_1 and I_3 is very small (approximately 0.00392), suggesting that the profiles are ***quite similar***. This indicates that the distribution of observations in the first and third rows is relatively close in terms of their relative proportions across J_1 and \(J_2 .

Exercise 3 (7pts)

Means of the Variables (0.75pts)

The mean of precipitation P is:

$$\bar{P} = \frac{12.04 + 17.18 + 11.83 + 6.23 + 16.99 + 3.87}{6} = \frac{68.14}{6} = 11.3567$$

The mean of maximum temperature $t_{\max}$ is:

$$\bar{t}_{\max} = \frac{23.7 + 15.5 + 13.1 + 13.5 + 21.1 + 20.3}{6} = \frac{107.2}{6} = 17.8667$$

The mean of minimum temperature $t_{\min}$ is:

$$\bar{t}_{\min} = \frac{5.9 + (-1.8) + 2.8 + (-2.4) + 7.2 + (-0.9)}{6} = \frac{10.8}{6} = 1.8.$$

Variance and standard deviation of the Variables (1.5pts)

$$\text{Var}(P) = \frac{1}{6}\sum_{i=1}^{6}(P_i - \bar{P})^2 \text{Var}(P) = \frac{1}{6}\left((12.04 - 11.69)^2 + (17.18 - 11.69)^2 + (11.83 - 11.69)^2 + (6.23 - 11.69)^2 + (16.99 - 11.69)^2 + (3.87 - 11.69)^2\right)$$

$$\text{Var}(P) = \frac{1}{6} \cdot 149.27 = 24.88\,\text{cm}^2 \sigma_P = \sqrt{\text{Var}(P)} = \sqrt{24.88} = 4.99$$

$$\text{Var}(t_{\max}) = \frac{1}{6}\left((23.7 - 17.02)^2 + (15.5 - 17.02)^2 + (13.1 - 17.02)^2 + (13.5 - 17.02)^2 + (21.1 - 17.02)^2 + (20.3 - 17.02)^2\right)$$

$$\text{Var}(t_{\max}) = \frac{1}{6}(44.91 + 2.34 + 15.33 + 12.45 + 16.72 + 10.88)$$

$$\text{Var}(t_{\max}) = \frac{1}{6} \cdot 102.63 = 17.10\,^\circ\text{C}^2$$

$$\sigma_{t_{\max}} = \sqrt{\text{Var}(t_{\max})} = \sqrt{17.10} = 4.14$$

Variance of $t_{\min}$ (Minimum Temperature)

$$\text{Var}(t_{\min}) = \frac{1}{6}\left((5.9 - 1.80)^2 + (-1.8 - 1.80)^2 + (2.8 - 1.80)^2 + (-2.4 - 1.80)^2 + (7.2 - 1.80)^2 + (-0.9 - 1.80)^2\right).$$

$$\text{Var}(t_{\min}) = \frac{1}{6}(16.81 + 13.69 + 1.00 + 17.64 + 29.16 + 7.29)$$

$$\text{Var}(t_{\min}) = \frac{1}{6} \cdot 85.59 = 14.27\,^\circ\text{C}^2$$

$$\sigma_{t_{\min}} = \sqrt{\text{Var}(t_{\min})} = \sqrt{14.27} = 3.78$$

Dispersion Matrix (1pts)

From our previous calculations:

Exercises solution

- $\sigma_P = 4.99, \sigma_{t_{\max}} = 4.14, \sigma_{t_{\min}} = 3.78.$

- The variances and covariances are:

$$\text{Var}(P) = 24.88, \quad \text{Var}(t_{\max}) = 17.10, \quad \text{Var}(t_{\min}) = 14.27$$

$$\text{Cov}(P, t_{\max}) = 10.57, \quad \text{Cov}(P, t_{\min}) = 8.43, \quad \text{Cov}(t_{\max}, t_{\min}) = 12.04$$

$$\Sigma = \begin{bmatrix} 24.88 & 10.57 & 8.43 \\ 10.57 & 17.10 & 12.04 \\ 8.43 & 12.04 & 14.27 \end{bmatrix}$$

The correlation matrix is then written as:

$$R = \begin{bmatrix} 1 & 0.51 & 0.45 \\ 0.51 & 1 & 0.77 \\ 0.45 & 0.77 & 1 \end{bmatrix}$$

Correlation Coefficients

$$\text{Corr}(P, t_{\max}) = \frac{\text{Cov}(P, t_{\max})}{\sigma_P \cdot \sigma_{t_{\max}}} = \frac{10.57}{4.99 \cdot 4.14} = 0.51$$

$$\text{Corr}(P, t_{\min}) = \frac{\text{Cov}(P, t_{\min})}{\sigma_P \cdot \sigma_{t_{\min}}} = \frac{8.43}{4.99 \cdot 3.78} = 0.45.$$

$$\text{Corr}(t_{\max}, t_{\min}) = \frac{\text{Cov}(t_{\max}, t_{\min})}{\sigma_{t_{\max}} \cdot \sigma_{t_{\min}}} = \frac{12.04}{4.14 \cdot 3.78} = \frac{12.04}{15.65} = 0.77.$$

Comments : (0.75pts)

$\text{Corr}(P, t_{max}) = 0.51$: There is a moderate positive correlation between precipitation (P) and maximum temperature (t_{max}), implying that an increase in precipitation is moderately associated with an increase in maximum temperature.

$Corr(P, t_{min}) = 0.45$: There is also a moderate positive correlation between precipitation (P) and minimum temperature (t_{min}), suggesting that higher precipitation is moderately linked to higher minimum temperatures.

$Corr(t_max, t_{min}) = 0.77$: There is a strong positive correlation between maximum and minimum temperatures, meaning that as the maximum temperature increases, the minimum temperature tends to increase as well.

Eigenvalues and Principal Components Calculation*: *(1.25pts)

To find the eigenvalues, we solve the characteristic equation:

$$\det(R - \lambda I) = 0$$

$$\lambda_1 \approx 2.09, \quad \lambda_2 \approx 0.77, \quad \lambda_3 \approx 0.14.$$

The total sum of the eigenvalues is:

$$\sum \lambda_i = 2.09 + 0.77 + 0.14 = 2.99$$

Now, the percentage of variance explained by each component is calculated as: *(0.75Pts)*

$$\text{Percentage of Variance for } \lambda_1 = \frac{2.09}{2.99} \times 100 \approx 69.9\%$$

$$\text{Percentage of Variance for } \lambda_2 = \frac{0.77}{2.99} \times 100 \approx 25.8\%$$

$$\text{Percentage of Variance for } \lambda_3 = \frac{0.14}{2.99} \times 100 \approx 4.7\%$$

We retain the first principal component as it explains 69.9% of the variance.

Normalized Eigenvector for the First Principal Component (1pts)

The normalized eigenvector corresponding to λ_1 is:

$$v_1 = \begin{bmatrix} 0.57 \\ 0.56 \\ 0.61 \end{bmatrix}.$$

10