Homework 3 - Applied Stochastic Processes

kipngeno koech - bkoech

October 17, 2024

Question 1: Random Vectors and Principal Component Analysis

Reading: Random vectors are fundamental constructs in probability and statistics, allowing researchers and practitioners to analyze relationships among multiple variables simultaneously. Each component of a random vector can represent a different feature or measurement, and the joint distribution encapsulates the uncertainty inherent in those variables.

For instance, consider a random vector X - $X_1, X_2, ..., X_n$ where each X_i is a random variable. The covariance matrix of X plays a crucial role in understanding the linear relationships among the components, guiding decisions in fields such as finance, machine learning, and signal processing. Sampling from random vectors introduces excitement in multivariate analyses, where one can explore properties like independence, marginal distributions, and conditional relationships. Moreover, techniques such as principal component analysis (PCA) leverage the variance structure of these vectors to reduce dimensionality while preserving essential information.

1. **5 points** let $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$, and $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$ are related by $Y - AX$ where

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

The joint PMF of X is given by:

$$P_X(X) = \begin{cases} (1-p)p^{x_3} & \text{if } X_1 < X_2 < X_3 \\ 0 & \text{otherwise} \end{cases}$$

where $x_1, x_2, x_3 \in \{0, 1, 2, \ldots\}$ and $0 < p < 1$.
Find the joint PMF $P_Y(y)$ of the transformed random vector Y.

$$Y = AX = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 - X_1 \\ X_3 - X_2 \end{pmatrix}$$

$$X_1 = Y_1, X_2 = Y_1 + Y_2, X_3 = Y_1 + Y_2 + Y_3$$

$$P_Y(y) = P_X(A^{-1}y) = P_X\left(\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}\right) = P_X\left(\begin{pmatrix} y_1 \\ y_1 + y_2 \\ y_1 + y_2 + y_3 \end{pmatrix}\right)$$

conditions for $X_1 < X_2 < X_3$ to hold:

$$y_1 < y_1 + y_2 < y_1 + y_2 + y_3$$

$$0 < y_2 < y_3$$

$$P_Y(y) = (1-p)p^{y_1+y_2+y_3} = (1-p)p^{y_1}p^{y_2}p^{y_3}$$

$$P_Y(y) = (1-p)p^{y_1}p^{y_2}p^{y_3}$$

2. You are working as a data analyst for a startup that collects various statistics from users' activities on its platform. The startup wants to reduce the dimensionality of its collected data without losing significant information. Your goal is to apply Principal Component Analysis (PCA) to the dataset to retain as much variance (information) as possible while reducing the dimensionality. This exercise will take you from the conceptual understanding of random vectors and covariance matrices to the practical application of PCA using Python.

Part 1: understanding the covariance Matrix of Random Vectors (12 points)
You are given a random vector $X = [X_1, X_2, X_3, X_4]^T$, representing four features of platforms users. The covariance matrix of this random vector is:

$$\Sigma_x = \begin{bmatrix} 5 & 1.2 & 0.8 & 0.6 \\ 1.2 & 4 & 0.5 & 0.3 \\ 0.8 & 0.5 & 3 & 0.2 \\ 0.6 & 0.3 & 0.2 & 2 \end{bmatrix}$$

Intepretation of the covariance matrix

(a) **(2 points)** What do the diagonal elements of the covariance matrix represent?

The diagonal elements of the covariance matrix represent the variance of the individual features.

(b) **(2 points)** What do the off-diagonal elements signify in terms of the relationship between the features?

The off-diagonal elements signify the covariance between the features.

Random Vector and Variance

(a) **(2 points)** Calculate the total variance of the random vector X.

$$\text{Total Variance} = \text{Trace}(\Sigma_x) = 5 + 4 + 3 + 2 = \mathbf{14}$$

(b) **(2 points)** How would you compute the variance captured by a single feature (e.g, the first feature X_1)?

$$\text{Variance of } X_1 = \Sigma_{11} = \mathbf{5}$$

Eigenvalues and Eigenvectors of the Covariance Matrix

(a) **(2 points)** Calculate the eigenvalues and eigenvectors of the covariance matrix Σ_x by hand
method used to calculate eigenvalues and eigenvectors is eigen decomposition:
The characteristic equation is given by:

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

$$\begin{vmatrix} 5-\lambda & 1.2 & 0.8 & 0.6 \\ 1.2 & 4-\lambda & 0.5 & 0.3 \\ 0.8 & 0.5 & 3-\lambda & 0.2 \\ 0.6 & 0.3 & 0.2 & 2-\lambda \end{vmatrix} = 0$$

$$\lambda^4 - 14\lambda^3 + 68.18\lambda^2 - 139.254\lambda + 101.356 = 0$$

$$\lambda_1 = \mathbf{6.20306}, \lambda_2 = \mathbf{3.20619}, \lambda_3 = \mathbf{2.71066}, \lambda_4 = \mathbf{1.88009}$$

To calculate the eigen vectors, we substitute the eigen values into the equation:
The eigen vectors are:
for eigen value $\lambda_1 = 6.20306$

$$\begin{bmatrix} -1.20306 & 1.2 & 0.8 & 0.6 \\ 1.2 & -2.20306 & 0.5 & 0.3 \\ 0.8 & 0.5 & -3.20306 & 0.2 \\ 0.6 & 0.3 & 0.2 & -4.20306 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The augmented matrix is:

$$\left[\begin{array}{cccc|c} -1.20306 & 1.2 & 0.8 & 0.6 & 0 \\ 1.2 & -2.20306 & 0.5 & 0.3 & 0 \\ 0.8 & 0.5 & -3.20306 & 0.2 & 0 \\ 0.6 & 0.3 & 0.2 & -4.20306 & 0 \end{array}\right]$$

$$R_1 = \frac{1}{-1.20306} R_1 = \left[\begin{array}{cccc|c} 1 & -0.997 & -0.66497 & -0.4987 & 0 \\ 1.2 & -2.20306 & 0.5 & 0.3 & 0 \\ 0.8 & 0.5 & -3.20306 & 0.2 & 0 \\ 0.6 & 0.3 & 0.2 & -4.20306 & 0 \end{array}\right]$$

$$R_2 = R_2 - 1.2R_1 = \left[\begin{array}{cccc|c} 1 & -0.997 & -0.66497 & -0.4987 & 0 \\ 0 & -1.00611 & 1.297965 & 0.89847 & 0 \\ 0.8 & 0.5 & -3.20306 & 0.2 & 0 \\ 0.6 & 0.3 & 0.2 & -4.20306 & 0 \end{array}\right]$$

(b) **(2 points)** List the eigenvalues in descending order and explain what they represent in terms of variance

$$\lambda_1 = \mathbf{6.20306}, \lambda_2 = \mathbf{3.20619}, \lambda_3 = \mathbf{2.71066}, \lambda_4 = \mathbf{1.88009}$$

The eigen values represent the variance of the data along the principal components. The first eigen value $\lambda_1 = 6.20306$ represents the variance of the data along the first principal component, the second eigen value $\lambda_2 = 3.20619$ represents the variance of the data along the second principal component, the third eigen value $\lambda_3 = 2.71066$ represents the variance of the data along the third principal component, and the fourth eigen value $\lambda_4 = 1.88009$ represents the variance of the data along the fourth principal component.

Part 2: Principal Component Analysis (PCA) (8 points)

Now that you have a grasp of the covariance matrix and its eigenvalues, you will apply PCA to a random vector

Principal Component Directions

(a) **(2 points)** Using the eigenvectors, describe the principal component directions. What do these directions represent in terms of variance in the data?

(b) **(2 points)** Explain the concept of orthogonality in PCA and why is it important?

Orthogonality in PCA means that the principal components are perpendicular to each other. This is important because it ensures that the principal components are independent of each other. This means that the variance of the data is maximized along the principal components.

Transformation of Random Vector

Let the eigenvector matrix be P and defined the transformed random vector Y by $Y = P^T X$

(a) **(2 points)** What is the covariance matrix of the vector Y?

$$\Sigma_Y = P^T \Sigma_x P$$

(b) **(2 points)** How does this Transformation affect the correlation between the transformed features?
The transformation affects the correlation between the transformed features by making them uncorrelated. The covariance matrix of the transformed random vector Y is a diagonal matrix, which means that the transformed features are uncorrelated.c

Part 3: Performing PCA by Hand on a Simple Dataset (8 points) consider a simple dataset represented by the following 2-dimensional random vector $Y = [Y_1, Y_2]^T$:

$$Y = \begin{bmatrix} 1.2 & 2.8 \\ 0.8 & 2.4 \\ 1.6 & 3.2 \\ 1.4 & 2.9 \end{bmatrix}$$

Mean Centering

(a) **(2 points)** Calculate the mean of the dataset for each feature Y_1 and Y_2

$$\text{Mean of } Y_1 = \frac{1.2 + 0.8 + 1.6 + 1.4}{4} = \mathbf{1.25}$$

$$\text{Mean of } Y_2 = \frac{2.8 + 2.4 + 3.2 + 2.9}{4} = \mathbf{2.825}$$

(b) **(2 points)** Subtract the mean from each feature to center the data

$$\text{Centered Data} = \begin{bmatrix} 1.2 - 1.25 & 2.8 - 2.825 \\ 0.8 - 1.25 & 2.4 - 2.825 \\ 1.6 - 1.25 & 3.2 - 2.825 \\ 1.4 - 1.25 & 2.9 - 2.825 \end{bmatrix} = \begin{bmatrix} -0.05 & -0.025 \\ -0.45 & -0.425 \\ 0.35 & 0.375 \\ 0.15 & 0.075 \end{bmatrix}$$

Covariance Matrix (2 points)

3. Calculate the covariance matrix of the centered data

$$\text{Covariance Matrix} = \frac{1}{n-1} \text{Centered Data}^T \text{Centered Data}$$

$$\text{Covariance Matrix} = \frac{1}{4-1} \begin{bmatrix} -0.05 & -0.025 \\ -0.45 & -0.425 \\ 0.35 & 0.375 \\ 0.15 & 0.075 \end{bmatrix}^T \begin{bmatrix} -0.05 & -0.025 \\ -0.45 & -0.425 \\ 0.35 & 0.375 \\ 0.15 & 0.075 \end{bmatrix}$$

$$\text{Covariance Matrix} = \frac{1}{3} \begin{bmatrix} -0.05 & -0.45 & 0.35 & 0.15 \\ -0.025 & -0.425 & 0.375 & 0.075 \end{bmatrix} \begin{bmatrix} -0.05 & -0.025 \\ -0.45 & -0.425 \\ 0.35 & 0.375 \\ 0.15 & 0.075 \end{bmatrix}$$

$$\text{Covariance Matrix} = \frac{1}{3} \begin{bmatrix} 0.35 & 0.335 \\ 0.335 & 0.328 \end{bmatrix}$$

$$\text{Covariance Matrix} = \begin{bmatrix} 0.1167 & 0.1117 \\ 0.1117 & 0.1093 \end{bmatrix}$$