Data analysis

Chapter3 :
The Principal Components analysis (ACP)

Dr.Allaoui

06/10/2024

Table of contents

I The Principal Component Analysis (PCA)

In this course, we will explore Principal Component Analysis (PCA), a key technique for reducing the dimensionality of data while retaining as much information as possible. To effectively understand PCA, it's crucial to first grasp fundamental statistical concepts such as mean, variance, covariance, and correlation matrices. These concepts form the foundation for understanding how PCA works and how it simplifies complex data. A strong grasp of summary statistics will prepare you to engage with PCA and apply it to data analysis challenges.

1. Some prerequistes on a summary statistics.

Data representation

Data is collected in the form of a matrix, where rows represent observations and columns represent variables.

We define a ***rectangular data tables (matrix)***:

$$\begin{matrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{matrix}$$

1. This matrix represents n individuals and p variables, with each element X_{ij} being the observed value for individual i and variable j.
2. If the values of X_{ij} takes numbers, a ***quantitative variable is often expressed*** with a unit of measurement that serves as a reference.
3. Each random variable X_j, where $(X_{1,j}, \ldots, X_{n,j})$ has a mean $\bar{X}_j$ and a standard deviation σ_{X_j}.

🔔 *Reminder:Mean*

For a dataset X with n observations and p variables:

Mean of a variable X_j (for each $j = 1, 2, \ldots, p$) is given by:

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}.$$

where x_{ij} represents the i-th observation of the j-th variable

as example : x_{ij} is the value of an economic indicator i for the year j.

🔔 *Reminder:Variance and standard deviation*

- ***Variance*** of a variable X_j measures ***the spread of values around the mean*** and is calculated as:

$$S_{n,j}(X_j) = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2.$$

- $\sigma_j = \sqrt{\text{Var}(X_j)}$ is the standard deviation of X_j.
- The covariance between two variables X_{ij} and X_{ik} is calculated as follows:

$$\text{Cov}(X_{ij}, X_{ik}) = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_i)(X_{ik} - \bar{X}_k).$$

- The correlation between two variables X_i and X_j is calculated as follows:

$$\rho(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i} \cdot \sigma_{X_j}}$$

🔔 *Reminder:Covariance Matrix*

The covariance matrix is ***a square matrix*** that summarizes the covariance between pairs of variables in a dataset. Each element in the matrix represents the covariance between two variables.

For a dataset with n variables (features) and p observations, the covariance matrix Σ is defined as:

$$\Sigma = \begin{bmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Cov}(X_n, X_n) \end{bmatrix}.$$

Where:

- $\text{Cov}(X_i, X_j)$ is the covariance between variables X_i and X_j.
- The ***diagonal elements*** $\text{Cov}(X_i, X_i)$ represent ***the variance of each variable i*** $Var(X_i)$.

🔔 *Reminder:Correlation matrix*

The correlation matrix, denoted by $\mathbf{R}$, is ***a square matrix that summarizes the correlation between pairs of variables in a dataset.***

Each element in the matrix represents the ***correlation coefficient between two variables***, which standardizes the covariance on a scale of -1 to 1.

For a dataset with n variables (features) and p observations, the correlation matrix $\mathbf{R}$ is defined as:

$$\mathbf{R} = \begin{bmatrix} \rho(X_1, X_1) & \rho(X_1, X_2) & \cdots & \rho(X_1, X_n) \\ \rho(X_2, X_1) & \rho(X_2, X_2) & \cdots & \rho(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(X_n, X_1) & \rho(X_n, X_2) & \cdots & \rho(X_n, X_n) \end{bmatrix}$$

where:

- $\rho(X_i, X_j)$ is the correlation between variables X_i and X_j.
- The diagonal elements $\rho(X_i, X_i)$ represent the correlation of each variable with itself, which is always 1.

Reminder:Properties :

- ***Symmetry:*** The covariance (resp. correlation) are symmetric, meaning $\mathrm{Cov}(X_{ij}, X_{ik}) = \mathrm{Cov}(X_{ik}, X_{ij})$.
- ***Positive Semi-Definiteness***: The covariance (resp. correlation) matrix is always positive semi-definite, which means that ***all its eigenvalues are non-negative***.

Example

Consider a simple dataset with two variables (features) X and Y with observations:

1. Calculate the means $\bar{X}$ and $\bar{Y}$:

$$\bar{X} = \frac{2+4+6}{3} = 4$$

$$\bar{Y} = \frac{3+5+8}{3} \approx 5.33$$

2. Calculate the covariances and construct the covariance matrix:

- The variance of X is calculated using the formula:

$$\mathrm{Var}(X) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Calculating each term:

$$\text{For } X_1 = 2: \quad (2-4)^2 = (-2)^2 = 4$$

$$\text{For } X_2 = 4: \quad (4-4)^2 = (0)^2 = 0 \quad .$$

$$\text{For } X_3 = 6: \quad (6-4)^2 = (2)^2 = 4$$

Sum of squared differences:

$$4 + 0 + 4 = 8.$$

Thus, the variance is:

$$\mathrm{Var}(X) = \frac{8}{3-1} = \frac{8}{2} = 4$$

- For Variance of Y

Using the same formula:

$$\mathrm{Var}(Y) = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Calculating each term:

$$\text{For } Y_1 = 3: \quad (3-5.33)^2 \approx (-2.33)^2 \approx 5.4289$$

$$\text{For } Y_2 = 5: \quad (5-5.33)^2 \approx (-0.33)^2 \approx 0.1089$$

$$\text{For } Y_3 = 8: \quad (8-5.33)^2 \approx (2.67)^2 \approx 7.1289$$

Sum of squared differences:

$$5.4289 + 0.1089 + 7.1289 \approx 12.6667.$$

Thus, the variance is:

$$\text{Var}(Y) = \frac{12.6667}{2} \approx 6.3333.$$

3.The covariance is calculated as follows:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

Calculating each term:

$$\text{For Observation 1:} \quad (2-4)(3-5.33) \approx (-2)(-2.33) \approx 4.66$$

$$\text{For Observation 2:} \quad (4-4)(5-5.33) = (0)(-0.33) = 0$$

$$\text{For Observation 3:} \quad (6-4)(8-5.33) \approx (2)(2.67) \approx 5.34$$

Sum of products:

$$4.66 + 0 + 5.34 = 10.$$

Thus, the covariance is:

$$\text{Cov}(X, Y) = \frac{10}{2} = 5$$

4.Construct the Covariance Matrix

Now we can construct the covariance matrix :

$$\Sigma = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Var}(Y) \end{bmatrix} = \begin{bmatrix} 4 & 5 \\ 5 & 6.3333 \end{bmatrix}.$$

2. The Principal Component Analysis (PCA) method

Definition

Principal Component Analysis (PCA) is a statistical technique ***used to reduce the dimensionality of complex data***. It is applied to data tables where n ***rows represent individuals*** and p ***columns represent quantitative variables*** that are ***correlated***.

When analyzing a large number of variables, visual representation becomes challenging in low-dimensional spaces (such as a 2D plane). PCA addresses this issue by reducing the number of dimensions while preserving as much of the original data structure as possible.

PCA works by using the ***variance-covariance matrix (or correlation matrix),*** which captures the data's ***dispersion***. From this matrix, we ***transform correlated variables into a smaller set of uncorrelated variables called principal components.***

The i-th principal component C_{p_i} can be expressed as:

$$C_{p_i} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n, \quad i = 1, \ldots, n$$

with the coefficients satisfying the condition:

$$\sum_{i=1}^{n} a_{pi} \leq 1.$$

What is the importance of dimensionality reduction using PCA?

Dimensionality reduction involves transforming high-dimensional data into a lower-dimensional space while retaining all of its information. This allows us to understand the data more clearly. With PCA, interpreting the data becomes easier, and it provides better visual representations of the data.

2.1. Steps to calcul PCA

⚙ *Method:Step 1 : Explore the characteristics of your data (mean, variance, scales) before deciding which matrix to use for PCA.*

We know how to study each of these four ***variables individually***, by calculating summary statistics. Additionally, we can analyze the relationships between two variables (for example, Mathematics and French) by calculating the correlation coefficient.

However, how can we analyze all four ***variables simultaneously***, even to create a single visual representation? The challenge lies in the fact that the individuals (students) are no longer represented in a two-dimensional space, but in a higher-dimensional space (in this case, four dimensions).

⚙ *Method:Calcul of Standard Deviation Vector :*

- Let σ be a vector defined as:

$$\sigma = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_p \end{bmatrix}$$

where σ_j is the standard deviation of the X_j variable (or column j) of the data matrix X.

- The vector $\frac{1}{\sigma}$ is constructed by taking the reciprocal of each element in the standard deviation vector σ:

$$\frac{1}{\sigma} = \begin{bmatrix} \frac{1}{\sigma_1} \\ \frac{1}{\sigma_2} \\ \vdots \\ \frac{1}{\sigma_p} \end{bmatrix}.$$

✎ *Note*

While dividing by is standard for obtaining unbiased estimates of variance in descriptive statistics, using n in PCA is common practice, ***as the focus is on maximizing the variance*** explained by the principal components rather than estimating population parameters.

So in what follows, ***the variance-covariance matrix*** is typically calculated by dividing by n, which simplifies matrix calculations and centers the data around its mean.

⚙ *Method:Step2: Standardization (Z-score Normalization)*

In PCA, these values are often ***standardized***, this is particularly recommended when the variables are measured in different units, such as kilograms, kilometers, centimeters, etc.; otherwise, ***the results of the analysis will be***

significantly affected. The ***objective is to make the variables comparable***. Generally, the variables are standardized in such a way that they have:

- When ***variables*** are on the ***same scales***, we ***center the data*** (subtracting the mean)
  We obtain ***a centered version of the matrix*** $\mathbf{X}$ as:
  $\mathbf{Z} = \mathbf{X} - \mathbf{1}_n \mathbf{g}^T$,
  where $\mathbf{1}_n$ is ***a column vector of ones of size*** n, and $\mathbf{g}^T$ is the transpose of the vector $\mathbf{g}$.
  This means that for each variable, we ***standardize*** it to obtain Z_j:
  $Z_{ij} = X_{ij} - \bar{X}_j$.
  If we calculate the mean of each variable from the centered matrix, we will find it to be 0, and thus the center of the new point cloud will be: $\mathrm{G}_{xc}(0, 0, 0)$.
- ***When your variables are on different scales or units*** (e.g., one variable in dollars and another in kilograms), a ***centered and redused version*** of the matrix $\mathbf{X}$ is used and given by :
  $\mathbf{Z}^* = \mathbf{X} - \mathbf{1}_n \mathbf{g}^T \frac{1}{\sigma} I_p = \mathbf{Z} \frac{1}{\sigma} I_p$,
  We can write $\frac{1}{\sigma} I_p = D\left(\frac{1}{\sigma}\right)$ where $D\left(\frac{1}{\sigma}\right)$ denotes a ***weigted matrix*** (a diagonal matrix formed from the vector of σ).
  This means that for each variable X_j, standardize it to obtain Z_j^*:
  $Z_{ij}^* = \frac{X_{ij} - \bar{X}_j}{\sigma_j}$.

✎ *Note*

After performing these steps for all variables, the ***standardized data*** matrix can be represented as:

$$\mathbf{Z} = \begin{pmatrix} Z_{11} & Z_{12} & \cdots & Z_{1p} \\ Z_{21} & Z_{22} & \cdots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{np} \end{pmatrix}$$

- Each variable will have: A mean of 0, i.e., $\bar{Z}_j = 0$. A standard deviation of 1, i.e., $\mathrm{Var}(Z_j) = 1$.
- This ***standardized matrices*** is then used for PCA ***to ensure that the analysis is not biased by difference values or the scales of the original variables***.

⚙ *Method:Calcul covariance matrix or correlation matrix for PCA*

In the context of Principal Component Analysis (PCA),

- When the variables in your ***dataset are measured on the same scale***, we use the ***covariance matrix*** of the obtained standardized matrix $\mathbf{Z}$ (a centered version of the matrix $\mathbf{X}$ defined before),
- When the variables in your ***dataset are measured on the different scales***, we use the ***correlation matrix*** of the obtained standardized matrix $\mathbf{Z}^*$ (a centered and redused version of the matrix $\mathbf{X}$ defined above).
- We define the ***variance-covariance matrix*** associated with the vector $\left(Z_1, \ldots, Z_p\right)$ as follows:

$$\mathbf{\Sigma} = \frac{1}{n} \mathbf{Z}^T \mathbf{Z}.$$

- We define the ***correlation matrix*** associated with the vector $\left(Z_1^*, \ldots, Z_p^*\right)$ as follows:

$$\mathbf{R} = \frac{1}{n}\mathbf{Z}^{*T}\mathbf{Z}^{*}.$$

We note that both matrices $\mathbf{\Sigma}$ and $\mathbf{R}$ are symmetric:

$$\mathbf{\Sigma}^T = \frac{1}{n}\left(\mathbf{X}^T\mathbf{X}\right)^T = \frac{1}{n}\mathbf{X}^T\mathbf{X} = \mathbf{\Sigma}.$$

and $\mathbf{R}^T = \frac{1}{n}\left(\mathbf{Z}^T\mathbf{Z}\right)^T = \frac{1}{n}\mathbf{Z}^T\mathbf{Z} = \mathbf{R}.$

⚙ *Method:Compute the eigenvalues and eigenvectors of the covariance or correlation matrix :*

After calculating the covariance or correlation matrix in Principal Component Analysis (PCA), you should ***compute the eigenvalues and eigenvectors*** of the ***covariance or correlation*** matrix:

The eigenvectors represent ***the directions of maximum variance*** in the data (principal components), and ***the eigenvalues*** indicate the ***magnitude of variance along each of those directions***.

Note that eigenvectors corresponding to the largest eigenvalues are chosen to capture the most variance. The number of eigenvectors you select will determine the dimensionality of the transformed feature space.

⚙ *Method:Decide how many principal components*

Sort the eigenvalues in descending order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

Decide how many principal components to retain based on the sorted eigenvalues using one of these methods:

1. ***Variance Explained Criterion : (most recommended approach in theory)***
   - Calculate Total Variance : The total variance in the dataset is given by: $\text{Total Variance} = \sum_{i=1}^{p} \lambda_i$.
   - Calculate Proportion of Variance Explained: The proportion of variance explained by each principal component i is:

     $$\text{Proportion of Variance for PC}_i = \frac{\lambda_i}{\text{Total Variance}} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}.$$
   - Calculate Cumulative Explained Variance : The cumulative explained variance for the first \(k\) principal components is:

     $$\text{Cumulative Explained Variance}_k = \sum_{i=1}^{k} \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}.$$

     This can also be expressed as:

     $$\text{Cumulative Explained Variance}_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j}$$
   - Choose the number of components where the cumulative explained variance reaches a satisfactory threshold (commonly 80% to 90%).
2. ***Kaiser Criterion :*** Retain components with eigenvalues greater than 1. This rule is based on the idea that each component should explain at least as much variance as a single original variable. Components with eigenvalues less than 1 contribute less information than one of the original variables and are often discarded.

Method:Project the Original Data onto the New Feature Space

1. Create a matrix $\mathbf{V}$ of the selected eigenvectors (principal components). This matrix will be used to transform the original data.
2. ***Project the original data onto the new feature space*** by multiplying the centered data matrix by the matrix of selected eigenvectors. This gives you the principal component scores.

   $\mathbf{C} = \mathbf{ZV}.$

   where:

   - $\mathbf{C}$ is the matrix of principal component scores.
   - $\mathbf{Z}$ is the centered (or centred and reduced) data matrix.
   - $\mathbf{V}$ is the matrix of selected eigenvectors.
3. Evaluate Component Contributions: Look at the loadings of each variable on the principal components to understand which variables contribute most to each component.
4. Perform Further Analysis: Depending on your goals, you can use the principal component scores for various analyses, such as clustering, regression, or visualization, while potentially reducing the dimensionality of the dataset.