Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 15, 2023, page 8-15

Covariance Matrix Method in Principal Component Analysis

Kathirvelu Amelia Sneha, Wan Heng Fong*
Department of Mathematics, Faculty of Science, UTM, Skudai, Johor Bahru, Malaysia
*Corresponding author: fwh@utm.my

Abstract
In recent years, increasing amounts of data are produced and collected. Principal Component Analysis (PCA) offers a valuable technique for reducing the dimensionality of such datasets. This research highlights the basic background needed to understand and implement the PCA technique. One of the methods that come from this interrelation is the Covariance Matrix Method. Furthermore, the crucial role of eigenvalues and eigenvectors in PCA is highlighted in this research. The objective of this research is to standardize the data to perform the covariance matrix method, calculate the Principal Component of the variables using the covariance matrix method, and create a scatter plot with standardized and transformed data for comparison. In this research, a numerical example with existing data from the World Happiness Record 2023 is illustrated to show how the PCA space is calculated. The results from this research can benefit students in relating their mathematical knowledge to the fundamental concepts in PCA. Moreover, by plotting the data using the first few Principal Components (PC), insights into the underlying structure or patterns present in the data can be gained.
Keywords: Principal Component Analysis (PCA), Covariance Matrix, Eigenvalues, Eigenvectors, Dimension Reduction.

1 Introduction

Linear algebra plays a fundamental role in Principal Component Analysis (PCA) [1]. Large datasets are increasingly common and are often difficult to interpret. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss [2]. The purpose of this project is to explore some basic linear algebra concepts used in building the foundation of PCA. PCA is a widely used statistical technique for dimensionality reduction and data exploration[3]. By studying the covariance matrix method, a deeper understanding of the fundamental principles of PCA can be gained. This includes understanding the relationship between the variables, identifying the principal components, and quantifying their contributions to the data variance. The basic mathematical operations in matrix algebra can be used in PCA since there is a relation between covariance matrices and PCA. Thus, this research will benefit students in understanding the fundamental knowledge of PCA.

This research aims to relate mathematical knowledge using technical applications in PCA, especially in using mathematical operations of linear algebra such as matrix operations and eigenvectors. The objectives of this research are to standardize the data to perform the covariance matrix method and calculate the Principal Component (PC) of the variables using the covariance matrix method. This research also aims to create a scatter plot with standardized data and transformed data for comparison.

2 Literature Review

PCA was invented in 1901 by Karl Pearson [5] as an analogy of the principal axis theorem in mechanics which was later independently developed and named by Harold Hotelling in the 1930s. Depending on the field of application, it is also named the discrete Kosambi-Karhunen–Loève transform (KLT) in signal processing, the Hotelling transform in multivariate quality control, Proper Orthogonal Decomposition (POD) in mechanical engineering, Singular Value Decomposition (SVD) of X, Eigenvalue

Decomposition (EVD) of X^TX in linear algebra, factor analysis, Eckart–Young theorem or Schmidt–Mirsky theorem in psychometrics, Empirical Orthogonal Functions (EOF) in meteorological science, empirical eigen function decomposition, empirical component analysis, quasiharmonic modes, spectral decomposition in noise and vibration, and empirical modal analysis in structural dynamics [5].

Numerous research studies have been conducted by various researchers in the mathematics field. PCA continues to be an active area of research, with new developments and applications being explored to address specific challenges and improve the analysis of high-dimensional data [4].

Principal Component Analysis (PCA) is a widely used technique in mathematics and various other fields for dimensionality reduction, data analysis, and visualization. In the mathematics field, PCA can be applied in several ways. For instance, PCA can be used to compress high-dimensional data into a lower-dimensional representation while retaining the most important information [5]. This is useful in fields such as signal processing, image compression, and data storage, where reducing the dimensionality of the data can save storage space and computational resources.

Other than that, PCA can be employed to visualize high-dimensional data in a lower-dimensional space. By projecting the data onto a few principal components, which capture the most significant variations in the data, one can create scatter plots or three-dimensional plots that provide insights into the data's structure and relationship.

Moreover, PCA has applications in image processing, such as face recognition and image denoising. In face recognition, PCA can be used to extract the most discriminative features from a set of face images, allowing for efficient classification or identification of individuals. PCA can also be utilized to remove noise from images by reconstructing the image using only the principal components that capture the essential information. These are a few examples of how PCA is utilized in the mathematics field. PCA's versatility and ability to extract essential information from high-dimensional data make it a valuable tool in various mathematical applications.

This research involves the study of Matrix Algebra in Principal Component Analysis (PCA), eigenvectors, eigenvalues and important properties of matrices that are fundamental to PCA. The dimensions used for matrices in this research is 2 × 2. This research also includes the implementation of statistics which looks at distribution measurements using standard deviation, variance, and covariance. The aim of the statistical analysis of these data sets is to see if there is any relationship between the dimensions. The data for this study is taken from the World Happiness Report 2023. Mainly two factors that affect the happiness of eight countries namely Finland, the United States, Singapore, Malaysia, China, Cambodia, Lebanon and Afghanistan are selected as examples for this research namely Logged Gross Domestic Product (GDP) per Capita and Health Life Expectancy.

3 Principal Component Analysis

Matrix algebra is an essential mathematical tool used in Principal Component Analysis (PCA). PCA involves finding the eigenvectors and eigenvalues of a covariance matrix, which requires knowledge of linear algebra concepts such as matrix multiplication, eigen decomposition [6], and vector spaces. The eigenvectors and eigenvalues obtained from the covariance matrix are used to determine the principal components of the data, which are linear combinations of the original variables that capture the most significant variation in the data. Therefore, a solid understanding of linear algebra is crucial for understanding and implementing PCA.

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of a dataset while retaining as much of variation in the data as possible [6]. It does this by creating new variables, called Principal Components (PCs), that are linear combinations of the original variables. PCA is often used to identify patterns and relationships in data and to visualize high-dimensional datasets in a lower-dimensional space [6].

The goal of the PCA technique is to find a lower dimensional space or PCA space that is used to transform the data ($X = \{x_1, x_2, \ldots, x_N\}$) from a higher dimensional space to a lower dimensional space, where N represents the total number of samples or observations and x_i represents i^{th} sample, pattern, or observation [7]. The main aim of PCA is to find such principal components, which can describe the data points with a set of principal components.

The direction of the PCA space represents the direction of the maximum variance of the given data as shown in Figure 2.1 below.

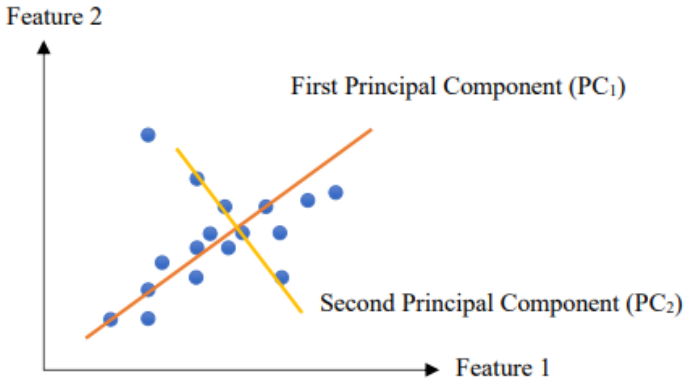

Figure 1 Example of the two-dimensional data ($x1$, $x2$) where two PCs are projected on the scatter

The figure shows an example of the two-dimensional data (x_1, x_2) where two PC are projected on the scatter plot. The PC are vectors. The first principal component is computed so that it explains the greatest amount of variance in the original features. The second component is orthogonal to the first, and it explains the greatest amount of variance left after the first principal component. In the small two-dimensional example above, we do not gain much by using PCA, since a feature vector of the form (Feature 1, Feature 2) will be very similar to a vector of the form (first principal component (PC_1), second principal component (PC_2)). But in very large datasets (where the number of dimensions can surpass 100 different variables), principal components remove noise by reducing a large number of features to just a couple of principal components. Principal components are orthogonal projections of data onto lower-dimensional space.

Dimensionality reduction is one of the pre-processing steps in many machine learning applications and it is used to transform the variables of data into a lower-dimension space [7]. The principal component analysis (PCA) technique is one of the most famous dimensionality reduction techniques. PCA technique has many goals including finding relationships between observations, extracting the most important information from the data, outlier detection and removal, and reducing the dimension of the data by keeping only the important information. All these goals are achieved by finding the PCA space, which represents the direction of the maximum variance of the given data [7]. This report shows how the covariance matrix is used to calculate the Principal Component (PC).

4 Calculating PCA

This section serves to provide a background for the mathematical concepts used in PCA. Specifically, this section looks at matrix algebra, and eigenvectors and eigenvalues of a given matrix. Eigenvectors can only be found for square matrices [8]. Not every square matrix has eigenvectors. Given an n × n matrix that does have eigenvectors, there are n of them. Given a 3 × 3 matrix, there are 3 eigenvectors. Another property of the eigenvector is that even if the vector is scaled by some amount before being multiplied, the same multiple of it is obtained as a result. This is because if the vector is scaled by some amount, only the length is increased, but the direction remains unchanged. Lastly, all the eigenvectors of a matrix are perpendicular regardless of the number of dimensions of the matrix. Another word for perpendicular is orthogonal [8].

4.1 Mathematical Background

Statistics is closely related to Principal Component Analysis (PCA) as PCA is a statistical technique used for data analysis. PCA is used to identify patterns in high-dimensional data by reducing the dimensionality of the data while retaining most of its variability.

The Standard Deviation (SD) of a data set is a measure of how spread out the data is. The way to calculate it is to compute the squares of the distance from each data point to the mean of the set, add them all up, divide by $n - 1$, and take the positive square root. It is written as,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n\ -\ 1)}}$$

where s is the usual symbol for standard deviation of a sample.

Variance is another measure of the spread of data in a data set. It is almost identical to the standard deviation with formula as below,

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

where σ^2 is the symbol for variable of a sample.

Next, standardization is presented. Standardization requires the knowledge of mean, standard deviation, and variance. The formula for standardization is,

$$Z\ = \frac{X-\bar{X}}{\sigma}.$$

Covariance is always measured between 2 dimensions. When the covariance between one dimension and itself is calculated, the variance is obtained. For example, if you had a two-dimensional data set (x, y) then the covariance between the x and y dimensions can be measured. The formula for covariance is very similar to the formula for variance. The formula for covariance could also be written as,

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

Following this two main steps needed to perform a Principal Components Analysis on a set of data is discussed. One of which is covariance matrix and the other which is calculating eigenvalues and eigenvectors. Covariance matrix is a symmetric matrix and is always a positive semi-definite matrix [9]. The diagonal values of the covariance matrix represent the variance of the variable, x_i , i = 1, …, M, while the off-diagonal entries represent the covariance between two different variables as shown in equation. A positive value in covariance matrix means a positive correlation between the two variables, while the negative value indicates a negative correlation and a zero value indicate that the two variables are uncorrelated or statistically independent [9].

The eigenvalues are scalar values, while the eigenvectors are non-zero vectors, which represent the PCs, and each eigenvector represents one PC. The eigenvectors represent the directions of the PCA space, and the corresponding eigenvalues represent the scaling factor, length, magnitude, or the robustness of the eigenvectors [9]. The eigenvector with the highest eigenvalue represents the first PC and it has the maximum variance.

4.2 Calculating Feature Vector

Now that the eigenvectors and eigenvalues are calculated, dimensionality reduction can be done.

Final data = Row Feature Vector × Row Data Adjust.

Row Feature Vector is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvector at the top, and Row Data Adjust is the mean-adjusted data transposed, the data items are in each column, with each row holding a separate dimension. Final Data is the final data set, with data items in columns, and dimensions along rows. It will give the original data solely in terms of the vectors chosen [10]. The original data set has two axes, x and y, so the data is in terms of them. It is possible to express data in terms of any two axes. If these axes are perpendicular, then the expression is the most efficient [10]. This was why it is important that eigenvectors are always perpendicular to each other. The data is being changed in terms

of the axes *x* and *y*, and now they are in terms of two eigenvectors. In this case the new data set has reduced dimensionality.

4.3 Application of the Principal Component Analysis

The discussion focuses on the result analysis from existing data when Principal Component Analysis (PCA) is applied. The sample data used for this research is from the World Happiness Record 2023 [11]. For simplicity of the research, eight countries are selected namely Finland, United States, Singapore, Malaysia, China, Cambodia, Lebanon, and Afghanistan. These countries are listed descending according to their ladder score with the highest score 7.804 for Finland and 1.859 for Afghanistan. Two factors affecting the happiness of the countries namely Gross Domestic Product (GDP) Per Capita (*X*) and Healthy Life Expectancy (*Y*) are chosen [11]. Table 1 shows two factors affecting the happiness of eight countries from the year 2023. The first factor is GDP per capita is represented by variable *X* and the second factor which is Health life Expectancy is represented by factor *Y.*

Table 1: Factors affecting happiness of the countries

Factor / Country	Gross Domestic Product (GDP) Per Capita (*X*)	Healthy Life Expectancy (%) (*Y*)
Finland	10.792	71.150
United States	11.048	65.850
Singapore	11.571	73.800
Malaysia	10.169	65.662
China	9.738	68.689
Cambodia	8.385	61.900
Lebanon	9.478	66.149
Afghanistan	7.324	54.712

Standardizing the data is an important step before performing PCA [10]. The reason why it is important to perform standardization before PCA is that the latter is extremely sensitive regarding the variances of the variables. In order to find the standardized data, firstly the mean, sum, standard deviation, and variance are calculated. Once the values of mean, sum, standard deviation, and variance of X, and Y are obtained, the standardized data, Z can be calculated. Table 2 shows the new data obtained after standardization where the new value for *X* is written as Z_X and *Y* as Z_Y.

Table 2: Data obtained after standardization

Standardized data / Country	Gross Domestic Product (GDP) Per Capita (Z_X)	Healthy Life Expectancy (Z_Y)
Finland	0.691	0.883
United States	0.872	-0.0238
Singapore	1.241	1.337
Malaysia	0.251	-0.056
China	-0.053	0.462
Cambodia	-1.008	-0.700
Lebanon	-0.237	0.027
Afghanistan	-1.757	-1.930

The covariance matrix defines both the spread (variance) and the orientation (covariance) of the data. The covariance matrix C is already obtained as,

$$C = \begin{pmatrix} 1.7566401 & 6.518466 \\ 6.518466 & 29.872518 \end{pmatrix}$$

The calculation of the eigenvector begins with the formula below,

$$Cv = nv .$$

The eigenvectors, *v* are those vectors that travel in the same direction when multiplied by matrix *C*, and eigenvalues *λ* are the scalar of the respective eigenvectors. Hence, after calculation, eigenvalues obtained are, $\lambda_1 = 31.310257$ and $\lambda_2 = 0.318901$.

Now that the eigenvalues are calculated, the next step is to calculate two-dimensional vectors v_1 and v_2 with the knowledge of the eigenvalues. The eigenvector, v_1 is calculated with the value of λ_1 whereas v_2 is calculated using the value of λ_2. Thus, each eigenvector has a correspondent eigenvalue. Upon calculation, $v_1 = \begin{bmatrix} 0.215387 \\ 0.976529 \end{bmatrix}$ for $\lambda_1 = 31.310257$, and $v_2 = \begin{bmatrix} -0.215387 \\ 0.976529 \end{bmatrix}$ for $\lambda_2 = 0.318901$ is obtained.

4.4 Calculating the Final Data

The original data can be represented as feature vectors. PCA allows the representation of the data as linear combinations of Principal Components (PC).

Feature Vector, *V* is obtained,

$$V = \begin{bmatrix} 0.976529 & -0.215387 \\ 0.215387 & 0.976529 \end{bmatrix}.$$

Now, the transformed data, *D* is calculated.

$$D = \begin{bmatrix} 0.69085971 & 0.88328727 \\ 0.8715366 & -0.0237894 \\ 1.24065383 & 1.33682559 \\ 0.25116557 & -0.0559649 \\ -0.0530209 & 0.46209565 \\ -1.0079265 & -0.6998182 \\ -0.2365209 & 0.02738345 \\ -1.7567475 & -1.9300195 \end{bmatrix}.$$

Once the transformed data is calculated, the final data, DV can be calculated.

Hence, final data calculated as shown below,

$$DV = \begin{bmatrix} 0.864893 & 0.713753 \\ 0.845957 & -0.210949 \\ 1.49947 & 1.03823 \\ 0.233216 & -0.108749 \\ 0.0477529 & 0.46267 \\ -1.135 & -0.466299 \\ -0.225071 & 0.0776843 \\ -2.13122 & -1.50634 \end{bmatrix}.$$

To visualize the outcome of standardized data, a scatter plot between *X* and *Y* is shown in Figure 2.

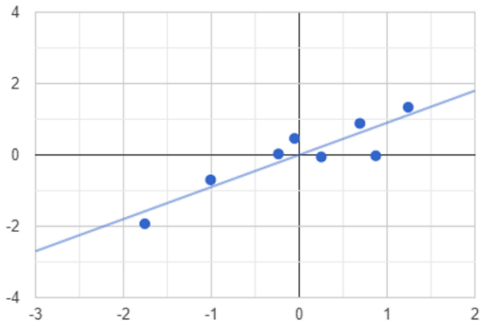

Figure 2 Scatter plot of Standardized Data, Z

Each point in the scatter plot in Figure 2 represents the data after being standardized which means it is centered at zero. This is required for PCA. This means a lot of information is removed by removing the first PC. The variance for Z_X is 1.757 and 229.873 for Z_Y whereas the mean for both is zero and the standard deviation is one.

Figure 3 shows the scatter plot of transformed data, D. With the transformed data, D, the plot is shown as below.

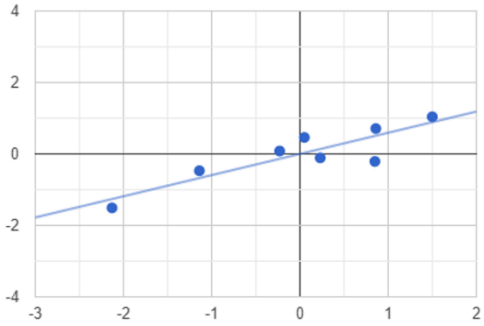

Figure 3 Scatter plot of transformed data, D

Each point in the scatter plot in Figure 3 represents the transformed data.

The variance for PC_1 is 31.310 whereas, 0.319 for PC_2. Both scatter plots reveal clusters and patterns within the data. In the standardized data scatter plot, clusters are formed based on the relationships between the X and Y variables. In the transformed data scatter plot, clusters arise based on the relationships between the principal components. It is important to note that the transformed data scatter plot shows different patterns or groupings compared to the standardized data scatter plot, as PCA reorients the data along the axes of maximum variance.

5 Conclusion

The covariance matrix approach is essential to PCA because it makes it possible to calculate the main components, which are responsible for the majority of the data's variability. It is essential to the PCA algorithm because it makes dimensionality reduction, variance analysis, and reconstruction easier. The relationships between various characteristics or variables in the dataset are quantified by the covariance matrix. It evaluates the relationship between changes in one variable and those in another.

PCA finds the direction and size of the largest variance in the dataset by computing the covariance matrix. This knowledge is crucial for identifying the eigenvectors of the covariance matrix known as the principal components, or PCs. PCA is the best orthogonal transformation for a set of vectors since it produces the fewest possible non-correlated PCs with the highest concentration of energy from the initial set. Each PC has an eigenvalue of one, a standard deviation of 1, and a mean of zero. Finding the factors from the original data that are most closely connected with each component,

such as PCs that are big in magnitude and furthest from zero in both directions, is necessary for interpreting the PCs. The calculation of PCA and eigenvectors depends on a very large, in-size, covariance matrix. As a result, the computational technique becomes more advanced, especially for huge datasets.

The covariance matrix provides the necessary information to reconstruct the data by multiplying the selected eigenvectors by the corresponding eigenvalues. With the aid of this reconstruction process, it is possible to assess the effects of dimensionality reduction on the data and determine how much information is lost in exchange for dimensionality reduction.

Acknowledgment

I wish to express my sincere gratitude to all who have contributed throughout the course of this work.

References

[1] Tharwat, A. Principal component analysis – a tutorial: Int. J. Applied Pattern Recognition, 2016. 3(3): 197-240.

[2] Lay, D.C. Linear Algebra and Its Applications. 4th ed. Boston, MA: Pearson Education, Inc. 2012.

[3] Hasan, B. and Abdulazeez, A. A Review of Principal Component Analysis Algorithm for Dimensionality Reduction: Journal of Soft Computing and Data Mining (JSCDM), 2021. 2(1): 20-30.

[4] Salem, N. and Hussein, S. Data dimensional reduction and principal components analysis: 16th International Learning & Technology Conference 2019, 2019. 163: 292-299.

[5] Gewers, F. Principal Component Analysis: A Natural Approach to Data Exploration: ACM Computing Surveys, 2021. 54(4): 70:1-34.

[6] Ikeuchi, K. Computer Vision: A Reference Guide. New York, USA: Springer. 2014.

[7] Kherif, F. and Latypova, A. Machine Learning: Methods and Applications to Brain Disorders. Switzerland: Elsevier Inc. 2020.

[8] Axler, S. Linear Algebra Done Right. 2nd ed. San Francisco, USA: Springer, 1997.

[9] Zagidullina, A. High-Dimensional Covariance Matrix Estimation: An Introduction to Random Matrix Theory. Cham, Switzerland: Springer. 2019.

[10] Bjorklund, M. Be careful with your principal components: *Evolution*, 2019. 73(10): 2151-2158.

[11] Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., and Wang, S. World Happiness Report 2023. New York, USA: Sustainable Development Solutions Network. 2023.

[12] Kherif, F. and Latypova, A. Machine Learning: Methods and Applications to Brain Disorders. Switzerland: Elsevier Inc. 2020.