Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 12 NO: 2 PAGE: 299-306 YEAR: 2023
DOI: 10.17798/bitlisfen.1144360

Determining The Number of Principal Components with Schur's Theorem in Principal Component Analysis

Cihan KARAKUZULU[1], İbrahim Halil GÜMÜŞ [2*], Serkan GÜLDAL[3], Mustafa YAVAŞ[4]

[1]*Adıyaman University, Graduate Education Institute, Department of Mathematics, Adıyaman*
[2]*Adıyaman University, Faculty of Arts and Sciences, Department of Mathematics, Adıyaman*
[3]*Adıyaman University, Graduate Education Institute, Department of Physics, Adıyaman*
[4]*Adıyaman University, Vocational School of Technical Sciences, Department of Computer Technologies, Adıyaman*
(ORCID: 0000-0002-3027-6927) (ORCID: 0000-0002-3071-1159) (ORCID: 0000-0002-4247-0786) (ORCID: 0000-0002-9111-9095)

Keywords: Principal component analysis, Majorization theory, Schur's theorem, Positive semidefinite matrices, Eigenvalues.

Abstract
Principal Component Analysis is a method for reducing the dimensionality of datasets while also limiting information loss. It accomplishes this by producing uncorrelated variables that maximize variance one after the other. The accepted criterion for evaluating a Principal Component's (PC) performance is $\frac{\lambda_j}{tr(\boldsymbol{S})}$ where $tr(\boldsymbol{S})$ indicates the trace of the covariance matrix S. It is standard procedure to determine how many PCs should be maintained using a specified total variance. In this study, the diagonal elements of the covariance matrix are used instead of the eigenvalues to determine how many PCs need to be considered to obtain the defined threshold of the total variance. For this, an approach which uses one of the important theorems of majorization theory is proposed. Based on the tests, this approach lowers computational costs.

1. Introduction

In many disciplines, high-dimensional datasets are becoming more common. Although researchers intend to collect more detailed information with every added dimension, higher dimensional datasets have several drawbacks. They require more sophisticated methods to analyze, interpret, and visualize. Even processing is impractical or impossible in some cases. Additionally, storing the data and related costs, such as maintenance and security, are more expensive. However, these drawbacks are avoidable with no considerable information lost. One of the solutions is the reorganization of the dataset (a.k.a. dimension reduction). Thus, the dimensions can be described by a linear combination of newly defined dimensions since in high-dimensional data, dimensions are generally correlated and the data has a lower dimensional structure in essence.

Dimension reduction, taking the correlation of dimensions into account, is the process of obtaining a representation of the data that has lower dimensions. Dimensionality can be reduced by using the Principal Component Analysis (PCA) algorithm, which is suggested by [1] and [2]. Although more than 100 years have passed, it is still a widely used data reduction method. Its objective is to preserve as much variability as possible while lessening the dataset's dimensionality [3]. Namely, PCA extracts new variables that are linear functions of the variables in the original dataset for maximized variance. These new variables are called Principal Components (PCs). The spectral decomposition of the covariance matrix, which defines the PCs' variance by their eigenvalues and their directions by their eigenvectors, is the key to PCA. In other words, the process of obtaining PCs is mathematically an eigenvalue/eigenvector problem. Since the covariance matrix's eigenvectors and eigenvalues are used to define PCA, many matrix

*Corresponding author: igumus@adiyaman.edu.tr
Received: 17.07.2022, Accepted: 23.02.2023

analysis methods can be used to improve the quality of the newly defined dataset.

Despite being frequently used for unsupervised linear dimensionality reduction and visualization, PCA has also been employed to solve statistical problems like regression, clustering, and nonlinear dimensionality reduction [4-6]. Because the utility of PCA has been discovered in many different scientific fields, it is called by many different names today. In numerical analysis and matrix analysis, it is known as Singular Value Decomposition, Karhunen-Loẽve transforms in signal processing, and characteristic vector analysis in the physical sciences. Thurstone and other psychologists pioneered the development of Factor Analysis (FA) in the 1930s [7]. This is worth mentioning because FA and PCA are very related, and these two methods are sometimes confused. Incorrectly, these two names are used interchangeably.

In recent years, important research has been done using PCA. In [8], a method for investigating systematic co-variation of vowels has been presented ~~by~~ using PCA. In [9], an application of principal component analysis has been obtained to reduce the dimensionality of variables representing the speech signal. The obtained results have been used for the disturbed and fluent speech recognition processes. In [10], a new combination strategy based on PCA to increase the predictability of crude oil futures market returns has been proposed. In [11], the status of PCA in the area of ECG signal processing has been reviewed. The use of PCA for spectral data reduction and colorant estimation has been illustrated in [12]. For interested readers, there are many excellent works investigating the various facets of PCA [13-16]. In addition, there are works in the literature examining how many principal components should be in PCA. For example, see [17-20] and the references therein.

In this study, a method is proposed to help select the right number of dimensions in the newly defined dataset quickly and efficiently. In section 2, the method and its theoretical basis are given. In section 3, the proposed method is tested for various datasets and extreme cases. In the last section, concluding remarks and future works are presented.

2. Material and Method

As a data analysis tool, PCA involves a dataset with observations on p features for n samples. These data values define an $n \times p$ data matrix $\boldsymbol{X}$. The j^{th} column of $\boldsymbol{X}$ is the vector $\boldsymbol{x_j}$ of observations on the j^{th} feature. The purpose of PCA is to find a linear combination of the columns of the matrix $\boldsymbol{X}$ that has the optimal variance. These linear combinations are obtained by $\boldsymbol{Xa}$ such that $\boldsymbol{a} = \left[a_1, a_2, \dots, a_p\right]^T$ is a $p \times 1$ vector where T stands for transpose. Finding the linear combination which has the optimal variance is equivalent to the computation of a p-dimensional vector that maximizes the $\boldsymbol{a}^T\boldsymbol{Sa}$ where $\boldsymbol{S}$ is data covariance matrix, namely $Var(\boldsymbol{Xa}) = \boldsymbol{a}^T\boldsymbol{Sa}$. It is worth noting that increasing the magnitude of the vector arbitrarily increases variance. Therefore, $\|\boldsymbol{a}\| = 1$ is taken, resulting in a constrained optimization problem in which we look for the data in the most variable direction. This constrained optimization problem can be written in the following form

$$\begin{aligned} &\max_{\boldsymbol{a}} \boldsymbol{a}^T\boldsymbol{Sa} \\ &s.t. \|\boldsymbol{a}\|^2 = 1 \end{aligned} \qquad (1)$$

For solving this optimization problem, we write the Lagrangian

$$\boldsymbol{L} = \boldsymbol{a}^T\boldsymbol{Sa} + \lambda\left(1 - \boldsymbol{a}^T\boldsymbol{a}\right) \qquad (2)$$

By computing the partial derivative of $\boldsymbol{L}$ with respect to $\boldsymbol{a}$ and λ and equating these partial derivatives to $\boldsymbol{0}$, we get

$$\begin{aligned} &\boldsymbol{Sa} = \lambda\boldsymbol{a} \\ &\boldsymbol{a}^T\boldsymbol{a} = 1. \end{aligned} \qquad (3)$$

So, the Lagrange multiplier acts as the corresponding eigenvalue and must be an eigenvector of the data covariance matrix $\boldsymbol{S}$. By multiplying both sides of $\boldsymbol{Sa} = \lambda\boldsymbol{a}$ with $\boldsymbol{a}^T$ from the left side, the following equation is obtained.

$$Var(\boldsymbol{Xa}) = \boldsymbol{a}^T\boldsymbol{Sa} = \boldsymbol{a}^T\lambda\boldsymbol{a} = \lambda \qquad (4)$$

This means that the eigenvalue associated with the basis vector that spans this subspace is equal to the variance of the data projected onto a one-dimensional subspace. As a result, the selected basis was related to the greatest eigenvalue of the data covariance matrix to optimize the variance of the low-dimensional representation. Since $\boldsymbol{S}$ is a symmetric matrix, it has exactly p real eigenvalues. The eigenvectors corresponding to these eigenvalues can be constructed to create an orthonormal set of vectors. By adding restrictions of orthogonality of different coefficient vectors on the Lagrange multipliers method, we can obtain all eigenvectors of $\boldsymbol{S}$. These answers to the problem of producing up to p new linear combinations $\boldsymbol{Xa}_k = \sum_{j=1}^{p} a_{jk}\boldsymbol{x}_j$ and maximizing variance that is uncorrelated with earlier linear combinations [3].

The linear combinations $\boldsymbol{X}\boldsymbol{a}_k$ are the Principal Components (PCs) of the dataset. Sometimes, many researchers also use the term PCs when mentioning to the eigenvectors $\boldsymbol{a}_k$. The variance associated with the set of retained PCs can be used to assess the quality of any q-dimensional subspace. The trace of the covariance matrix $\boldsymbol{S}$ is the sum of variances of the p original variables. It is simple to prove that this value is exactly the sum of the variances of all PCs. As a result, the accepted gauge of a PC's quality is $\frac{\lambda_j}{tr(\boldsymbol{S})}$ where $tr(\boldsymbol{S})$ denotes the trace of matrix $\boldsymbol{S}$. Determination of how many PCs should be preserved is usual practice to utilize a predetermined percentage of the total variance. This predetermined percentage is commonly 70% [21]. As noticed, in order to obtain the predetermined percentage of the total variance, we need to find all the eigenvalues of the matrix. Then, it is necessary to identify the $\frac{\lambda_j}{tr(\boldsymbol{S})}$ values whose sum exceeds this predetermined percentage of the total variance.

It is known that the eigenvalues of a matrix are obtained by finding the roots of the characteristic polynomial of that matrix. It is not possible to obtain these roots analytically for matrices larger than 4×4. Based on the Abel-Ruffini theorem, for polynomials of degree 5 or more, there is no algebraic solution. As a result, eigenvalues are obtained by using numerical methods. However, this means extra time for a data analyst who decides how many PCs to take based on finding the $\frac{\lambda_j}{tr(\boldsymbol{S})}$.

Now let's introduce the concept of majorization, which allows us to compare two vectors and observe which has "less spread out" components. Comparison of two vectors frequently leads to inequalities that can be expressed as majorization relations. Let $z = (z_1, z_2, \dots, z_n) \in \mathbb{R}^n$ and $z^{\downarrow}$ be the vector obtained by repositioning the coordinates of z in decreasing order. Thus if $z^{\downarrow} = \left(z_1^{\downarrow}, z_2^{\downarrow}, \dots, z_n^{\downarrow}\right) \in \mathbb{R}^n$, then $z_n^{\downarrow} \leq \cdots \leq z_2^{\downarrow} \leq z_1^{\downarrow}$.

For $x, y \in \mathbb{R}^n$, y majorizes x (or x is majorized by y), written as $x \prec y$, if

$$\sum_{j=1}^{k} x_j^{\downarrow} \leq \sum_{j=1}^{k} y_j^{\downarrow} \qquad (5)$$

for $1 \leq k < n$ and

$$\sum_{j=1}^{n} x_j^{\downarrow} = \sum_{j=1}^{n} y_j^{\downarrow}. \qquad (6)$$

If inequality is put in place of the equality

$$\sum_{j=1}^{n} x_j^{\downarrow} \leq \sum_{j=1}^{n} y_j^{\downarrow}, \qquad (7)$$

we state that y weakly majorizes x, and is indicated by $x \prec_w y$.

Majorization theory is a crucial tool that enables us to solve problems in various disciplines. One of these disciplines is matrix theory. Let's write the diagonal elements and eigenvalues of a $n \times n$ symmetric matrix $\boldsymbol{X}$ as vectors, respectively, by

$$d(\boldsymbol{X}) = \left(d_1(\boldsymbol{X}), d_2(\boldsymbol{X}), \dots, d_n(\boldsymbol{X})\right) \qquad (8)$$

and

$$\lambda(\boldsymbol{X}) = \left(\lambda_1(\boldsymbol{X}), \lambda_2(\boldsymbol{X}), \dots, \lambda_n(\boldsymbol{X})\right) \qquad (9)$$

Note that the components of these vectors are always arranged in decreasing order throughout the paper. The following theorem which is known as Schur's Theorem has a key role in our study, which can be found in [22].

Theorem: Let $\boldsymbol{X}$ be a $n \times n$ symmetric matrix. Then

$$d(\boldsymbol{X}) \prec \lambda(\boldsymbol{X}). \qquad (10)$$

This theorem says that

$$\begin{aligned} d_1(\boldsymbol{X}) &\leq \lambda_1(\boldsymbol{X}) \\ d_1(\boldsymbol{X}) + d_2(\boldsymbol{X}) &\leq \lambda_1(\boldsymbol{X}) + \lambda_2(\boldsymbol{X}) \\ d_1(\boldsymbol{X}) + d_2(\boldsymbol{X}) + \cdots + d_n(\boldsymbol{X}) &= \lambda_1(\boldsymbol{X}) + \lambda_2(\boldsymbol{X}) \\ &\quad + \cdots + \lambda_n(\boldsymbol{X}) \\ &= tr(\boldsymbol{X}) \end{aligned} \qquad (11)$$

for symmetric matrix $\boldsymbol{X}$. Namely, the total of the largest k eigenvalues of a symmetric matrix is bounded below by the total of the largest k diagonal elements of that matrix. Considering the fact that the covariance matrix is a symmetric matrix, if we want to use a predetermined percentage of the total variance, instead of calculating the total of the k largest eigenvalues of the covariance matrix that exceed this value, it will be sufficient to calculate the total of the k largest diagonal elements of the covariance matrix that exceeds this value. Using diagonal elements instead of calculating eigenvalues will provide us processing speed and convenience.

In this study, the diagonal elements of the covariance matrix are used instead of the eigenvalues. For this, we will make use of the majorization theory, which is commonly used to obtain inequalities. Thus, before starting the PCA process, the predetermined

percentage of the total variance will be obtained more quickly with the help of the diagonal elements. The approach is validated by numerical simulations (See **Hata! Başvuru kaynağı bulunamadı.**). Let's illustrate this fact with examples in the following section.

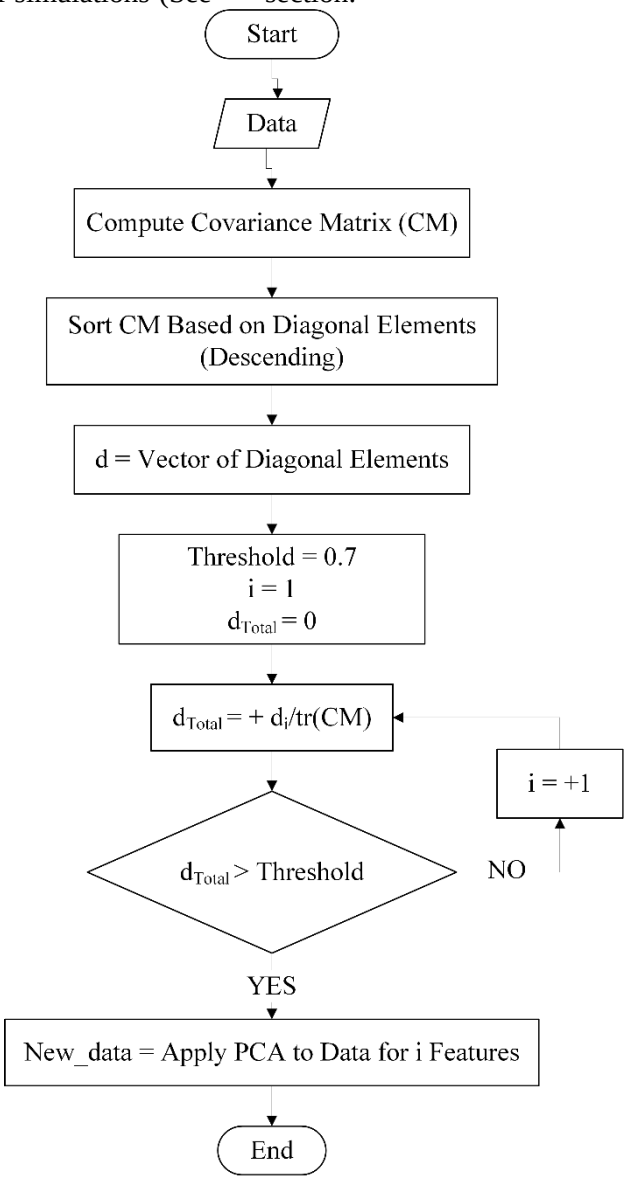

Figure 1. Flowchart of the proposed approach

3. Results and Discussion

The proposed approach is tested for 5 different datasets. Also, the method is tested for symmetric matrices up to 100×100. Yeast dataset is selected to exemplify the approach [23, 24]. Since this dataset has 8 features, the covariance matrix is an 8×8 which is obtained as

$$\boldsymbol{Y} = \begin{pmatrix} 355.953 & 82.8840 & -21.243 & 21.7097 & 0.44443 & -0.8670 & 4.14396 & -17.0248 \\ 82.8840 & 325.978 & -14.622 & 22.9344 & 0.33581 & -4.3622 & 7.31506 & -7.0348 \\ -21.243 & -14.622 & 258.660 & -0.8992 & -0.1108 & -2.1924 & -2.49194 & 3.07863 \\ 21.7097 & 22.9344 & -0.8992 & 206.712 & -0.0527 & -0.8870 & -10.7392 & -2.3823 \\ 0.44443 & 0.33581 & -0.1108 & -0.0527 & 0.23378 & -0.0296 & 0.34835 & -0.2966 \\ -0.8670 & -4.3622 & -2.1924 & -0.8870 & -0.0296 & 50.7703 & -1.0152 & -1.6376 \\ 4.14396 & 7.31506 & -2.4919 & -10.739 & 0.34835 & -1.0152 & 227.085 & 8.58366 \\ -17.024 & -7.0348 & 3.07863 & -2.3823 & -0.2966 & -1.6376 & 8.58366 & 117.485 \end{pmatrix}$$

When we computed the components of the diagonal elements and eigenvalues of this covariance matrix in descending order, we get the following two vectors

$$d(\boldsymbol{Y}) = (355.953, 325.978, 258.660, 227.085, 206.712, 117.485, 50.7703, 0.23378) \quad (12)$$

and

$$\lambda(\boldsymbol{Y}) = (434.575, 259.07, 253.696, 231.774, 197.33, 115.579, 50.6234, 0.231774) \quad (13)$$

As noticed,

$$355.953 \leq 434.575$$

$$355.953 + 325.978 \leq 434.575 + 259.07$$

$$355.953 + 325.978 + 258.660 \leq 434.575 + 259.07 + 253.696 \quad (14)$$

$$\vdots$$

$$355.953 + 325.978 + 258.660 + 227.085 + 206.712 + 117.485 + 50.7703 + 0.23378 = 434.575 + 259.07 + 253.696 + 231.774 + 197.33 + 115.579 + 50.6234 + 0.231774$$

Both sides of the last equality give us the trace of the covariance matrix, $tr(\boldsymbol{Y}) = 1542{,}879$. Cumulative percentage values of sums of $\frac{\lambda_j(\mathbf{Y})}{\mathrm{tr}(\mathbf{Y})}$ and $\frac{d_j(\mathbf{Y})}{\mathrm{tr}(\mathbf{Y})}$ for $j = 1,2,\ldots,8$ are depicted in **Hata! Başvuru kaynağı bulunamadı.**.

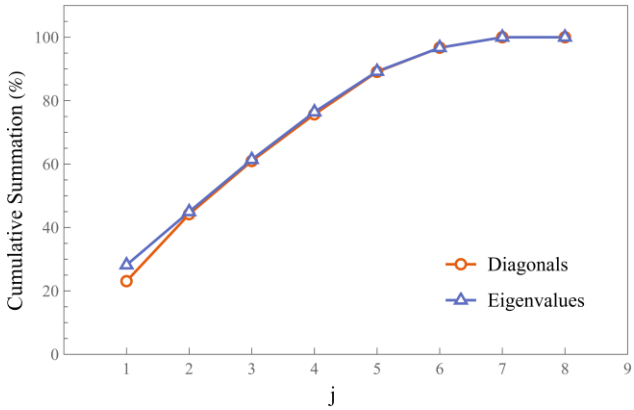

Figure 2. Cumulative summation of eigenvalues and diagonals are shown in percentage for Yeast dataset.

It is observed that the cumulative sums of eigenvalues and diagonals are almost the same. In the specified example, the ratio of the sum of the four largest eigenvalues to the trace of the covariance matrix is approximately 76% which is higher than 70%. This result shows that the first 4 largest eigenvalues (i.e. 4 PCs) should be taken for no considerable information lost. Additionally, corresponding diagonal values are approximately 76%. Therefore, instead of computing how many eigenvalues provide the predetermined percentage of the total variance (70%), as claimed in the previous sections, we can utilize the information from the diagonal elements of the covariance matrix.

To highlight the value of the proposed approach, computational times of eigenvalue and diagonal calculations are compared for 5 datasets. The list of datasets is given in Table 1. Here, Yeast is a medical dataset and consists of a protein-protein network [23, 24]. Accelerometer is used to estimate the engine failure time. Data obtained from the vibrations of the cooling fan with weights on its blades. It can be used for classification and other purposes for situations requiring vibration analysis [25]. Breast Cancer Coimbra is a medical dataset, anthropometric data and parameters that can be collected in routine blood analysis. These data indicate the presence and absence of cancer and are all quantitative data [26]. Cardiotocography is also a medical dataset, consisting of measurements of fetal heart rate and uterine contraction properties in cardiotocograms. This dataset is classified and labeled by expert obstetricians [27]. Combined Cycle Power Plant dataset is the data collected from a power plant operating at full load for six years. It is aimed to estimate the hourly net electrical energy output of the facility from the hourly average Temperature, Ambient Pressure, Relative Humidity and Exhaust

Vacuum characteristics [28, 29]. All features of datasets consist of numeric values. They are frequently used in machine learning classification and regression studies. The datasets are collected from different fields. Variation in the number of features requires different computation times. The computation efforts are shown in Figure 1. The results show the computation with eigenvalues requires more time than diagonals in every case. This difference is as great as 30 times for Cardiotocography datasets at maximum because of the higher number of features.

Table 1. 5 different datasets are selected from various subjects

Datasets	Number of features	Number of samples
Yeast **[23, 24]**	8	1484
Accelerometer **[25]**	5	153000
Breast Cancer Coimbra **[26]**	10	116
Cardiotocography **[27]**	23	2126
Combined Cycle Power Plant **[28, 29]**	4	9568

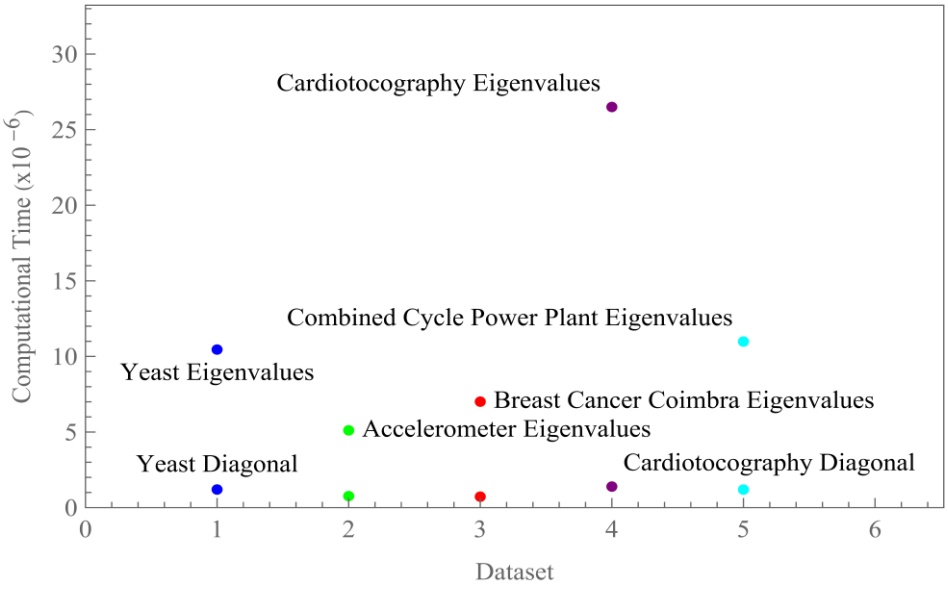

Figure 1. Computation time varieties for datasets, but the computation of eigenvalues requires more time than the computation of diagonal elements.

The covariance matrix of the listed datasets changes from 4×4 to 25×25. Computation of eigenvalues is relatively efficient. However, it is a known fact that increasing the size of the matrices will complicate the calculation of their eigenvalues. To illustrate this situation, let's take a randomly chosen positive semi-definite matrix (for an $n \times n$ matrix M, $M^T M$ is always positive semidefinite, and elements in [1,5]). Increment in the size of this matrix regularly up to 100×100 shows that computation with diagonals has linear complexity and eigenvalues have exponential complexity.

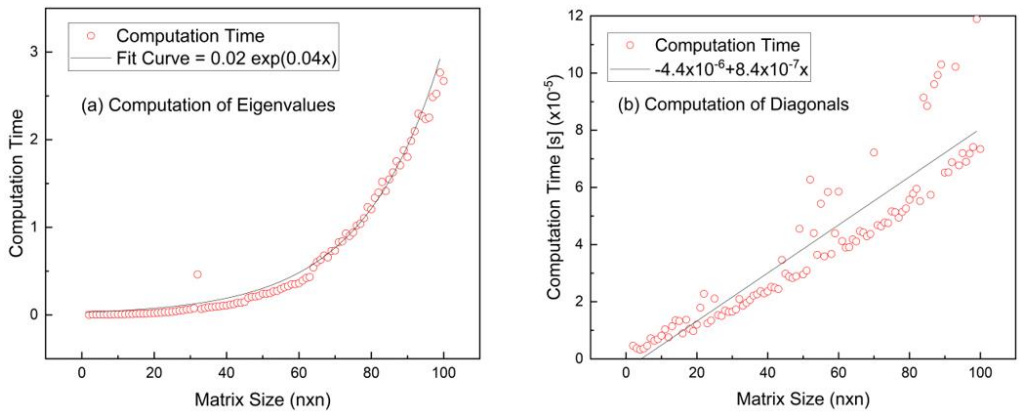

Figure 2. Required computational time is shown for eigenvalues (a), and diagonals (b).

4. Conclusion

In Principal Component Analysis, it is a standard procedure to determine how many Principal Components should be retained using a predetermined percentage of the total variance. For this, it is necessary to calculate all the eigenvalues of the covariance matrix. Then the necessary step is to identify how many of the largest eigenvalues we need so that the cumulative sum of the $\frac{\lambda_j}{tr(\boldsymbol{S})}$ exceeds the specified threshold. However, calculating the eigenvalues of the covariance matrix brings a computational cost. In this study, this process was done by using the diagonal elements of the covariance matrix instead of the eigenvalues of the covariance matrix. For this, Schur's theorem, which is well known in majorization theory, was used. The time savings of using diagonal elements was demonstrated using five different datasets. In addition, the increase in time required for computations as a result of increasing the matrix size is illustrated using randomly taken positive semi-definite matrices. As a result, it is seen that it is advantageous to use diagonal elements instead of calculating eigenvalues.

Contributions of the Authors

The authors confirm that the contribution is equally for this paper.

Conflict of Interest Statement

There is no conflict of interest between the authors of the article.

Statement of Research and Publication Ethics

The study complies with research and publication ethics.

References

[1] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559-572, 1901.

[2] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[3] I. T. Jolliffe, "Graphical representation of data using principal components," *Principal component analysis*, pp. 78-110, 2002.

[4] T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised learning," in *The elements of statistical learning*: Springer, pp. 485-585, 2009.

[5] C. Hafemeister and R. Satija, "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression," *Genome biology*, vol. 20, no. 1, pp. 1-15, 2019.

[6] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[7] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020.

[8] J. Wilson Black, J. Brand, J. Hay, and L. Clark, "Using principal component analysis to explore co-variation of vowels," *Language and Linguistics Compass*, vol. 17, no. 1, p. e12479, 2023.

[9] I. Świetlicka, W. Kuniszyk-Jóźkowiak, and M. Świetlicki, "Artificial Neural Networks Combined with the Principal Component Analysis for Non-Fluent Speech Recognition," *Sensors*, vol. 22, no. 1, p. 321, 2022.

[10] Y. Zhang and Y. Wang, "Forecasting crude oil futures market returns: A principal component analysis combination approach," *International Journal of Forecasting*, vol. 39, no. 2, pp. 659-673, 2023.

[11] F. Castells, P. Laguna, L. Sörnmo, A. Bollmann, and J. M. Roig, "Principal Component Analysis in ECG Signal Processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 074580, 2007.

[12] D.-Y. Tzeng and R. S. Berns, "A review of principal component analysis and its applications to color technology," *Color Research & Application*, vol. 30, no. 2, pp. 84-98, 2005.

[13] O. H. J. Christie, "Introduction to multivariate methodology, an alternative way?," *Chemometrics and Intelligent Laboratory Systems*, vol. 29, no. 2, pp. 177-188, 1995.

[14] M. Ghil *et al.*, "Advanced Spectral Methods for Clımatic Time Series," *Reviews of Geophysics*, vol. 40, no. 1, pp. 3-1-3-41, 2002.

[15] J. Hwang *et al.*, "Fast and sensitive recognition of various explosive compounds using Raman spectroscopy and principal component analysis," *Journal of Molecular Structure*, vol. 1039, pp. 130-136, 2013.

[16] P. Federolf, R. Reid, M. Gilgien, P. Haugen, and G. Smith, "The application of principal component analysis to quantify technique in sports," *Scandinavian Journal of Medicine & Science in Sports*, vol. 24, no. 3, pp. 491-499, 2014.

[17] L. Ferré, "Selection of components in principal component analysis: A comparison of methods," *Computational Statistics & Data Analysis*, vol. 19, no. 6, pp. 669-682, 1995.

[18] E. Saccenti and J. Camacho, "Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 99-116, 2015.

[19] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers, "How many principal components? stopping rules for determining the number of non-trivial axes revisited," *Computational Statistics & Data Analysis*, vol. 49, no. 4, pp. 974-997, 2005.

[20] D. A. Jackson, "Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches," *Ecology*, vol. 74, no. 8, pp. 2204-2214, 1993.

[21] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[22] F. Zhang, *Matrix theory: basic results and techniques*. Springer, 2011.

[23] K. Nakai and M. Kanehisa, "Expert system for predicting protein localization sites in gram-negative bacteria," (in eng), *Proteins*, vol. 11, no. 2, pp. 95-110, 1991.

[24] K. Nakai and M. Kanehisa, "A knowledge base for predicting protein localization sites in eukaryotic cells," (in eng), *Genomics*, vol. 14, no. 4, pp. 897-911, Dec 1992.

[25] G. Scalabrini Sampaio, A. R. d. A. Vallim Filho, L. Santos da Silva, and L. Augusto da Silva, "Prediction of Motor Failure Time Using An Artificial Neural Network," *Sensors*, vol. 19, no. 19, p. 4342, 2019.

[26] M. Patrício *et al.*, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, vol. 18, no. 1, p. 29, 2018.

[27] D. Ayres-de Campos, J. Bernardes, A. Garrido, J. Marques-de-Sá, and L. Pereira-Leite, "SisPorto 2.0: a program for automated analysis of cardiotocograms," (in eng), *J Matern Fetal Med*, vol. 9, no. 5, pp. 311-8, Sep-Oct 2000.

[28] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *International Journal of Electrical Power & Energy Systems*, vol. 60, pp. 126-140, 2014.

[29] H. Kaya and P. Tufekci, *Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine*. 2012.