*Information Technology and Management Science*
ISSN 2255-9094 (online)
2020, vol. 23, pp. 41–44
https://doi.org/10.7250/itms-2020-0006
*https://itms-journals.rtu.lv*

# Initial Dataset Dimension Reduction Using Principal Component Analysis

Oļegs Užga-Rebrovs[1], Gaļina Kulešova[2]
[1] *Rezekne Academy of Technologies, Rezekne, Latvia*
[2] *Riga Technical University, Riga, Latvia*

***Abstract* – Any data in an implicit form contain information of interest to the researcher. The purpose of data analysis is to extract this information. The original data may contain redundant elements and noise, distorting these data to one degree or another. Therefore, it seems necessary to subject the data to preliminary processing. Reducing the dimension of the initial data makes it possible to remove interfering factors and present the data in a form suitable for further analysis. The paper considers an approach to reducing the dimensionality of the original data based on principal component analysis.**

***Keywords* – Data labels in the space of principal components, data recovery in a space of lower dimension, data transformation into a space of principal components, eigenvectors and eigenvalues of variance/covariance matrix, variance/covariance matrix of data.**

## I. Introduction

In its most general form, *data* are defined as a set of entities related to a given domain and characterised by a set of attribute values [1]. In this definition, the concept of entities is used as a generalizing concept at a high level of abstraction. When it comes to specific data, the concept of entities is replaced by certain relevant content. The typical presentation of data is in the form of a matrix (table) whose rows represent objects, and the columns represent sets of attribute values.

However, there are many other types of data that cannot be presented in tabular form. As such data, one can mention time sequences, for example, data on changes in securities prices or data on changes in populations of biological species. Another type is data streams, for example, recording sensor readings or recordings of negotiations between aircraft pilots and the dispatch service, audio and video data. A specific data type is represented by network data, for example, data in social and information networks, as well as data from a global Web network.

In statistics, powerful methods have long been developed to test the truth or falsity of hypotheses and theories put forward. As noted in [1], in the theory of experimental developments, hypothesis testing and model building are some of the greatest contributions of statistics. Such a data analysis is called confirmatory data analysis (CD). At its core, such an analysis is a deductive reasoning.

On the other hand, in the second half of the 20th century, in connection with the rapid development of information technologies, huge volumes of various kinds of data began to accumulate. The data themselves are a statement of the current state of affairs in a certain area. Users are highly interested in the knowledge hidden in those data. It follows that such an analysis of the data is necessary that makes it possible to identify the relevant features of their structure, patterns and/or dependencies, which are presented in data in an implicit form. Such an analysis is called *exploratory* data analysis (EDA). More detailed information on various aspects of data analysis can be found in [2]–[9].

The huge volume and complex structure of initial data require special methods for their analysis. Some of the widespread methods of this kind are principal component analysis (PCA), singular value decomposition (SVD), correspondence analysis, (CA), factor analysis (FA) and linear discriminant analysis (LDA). Each of these methods uses one or another approach to reducing the dimension of the original data.

Principal component analysis is the most common data reduction method. The objectives of this paper are as follows: to present the theoretical foundations and formal procedures of the PCA, to demonstrate PCA procedures with an illustrative example and to analyse the potential benefits of PCA in the context of data analysis.

## II. Theoretical Foundations of Principal Component Analysis

Principal component analysis (PCA) can be defined as a multidimensional approach that analyses a data table in which observations are described through various interdependent variables [10]. Its purpose is to extract important information from statistical data in order to present them as a multitude of new orthogonal uncorrelated variables called principal components and to display patterns of similarity between observations and variables as points in a new space.

The general idea of the PCA is to transform a set of initial data into a new space in which the directions of the coordinate axes correspond to the directions of the greatest variability of the initial data.

Let a matrix (table) $\mathbf{A}$ of initial data of size $m \times n$ be given. Each row of this matrix represents one object and a vector of attribute values that characterise this object. Each column of the matrix corresponds to the set of values of this attribute for all objects. Let us sequentially consider PCA procedures.

1. For each column of the matrix $\mathbf{A}$ , the average value of the attribute $\overline{a}_j$ , $j=1,...,n$ is calculated.

2. The calculated average values $\overline{a}_j$ are subtracted from the actual attribute values $a_{ij}$ .

$$d_{ij}=a_{ij}-\overline{a}_j\,,\; i=1,\ldots,m\,,\; j=1,\ldots,n\,. \qquad (1)$$

As a result, we have a data matrix $\mathbf{D}$ with centred attribute values.

3. For each $j$ -th column of the matrix $\mathbf{D}$ , the variance value $s_j^2$ is calculated. For each pair of attribute values, the covariance value $s_{jl}$ , $j,l=1,\ldots,n$ , $j\neq l$ is calculated. The results are aggregated in a variance/covariance matrix

$$\mathbf{S}=\begin{pmatrix} s_1^2\ldots & s_{1j}\ldots & s_{1n}\\ s_{j1}\ldots & s_j^2\ldots & s_{jn}\\ s_{n1}\ldots & s_{nj} & s_n^2 \end{pmatrix}.$$

4. For the matrix $\mathbf{S}$ , its eigenvectors and eigenvalues are calculated. As a result, we have two matrices:

$$\mathbf{V}=\begin{pmatrix} v_{11}\ldots v_{21}\ldots\ldots v_{n1}\\ v_{12}\ldots v_{22}\ldots\ldots v_{n2}\\ \ldots\ldots\ldots\ldots\ldots\ldots\\ v_{1n}\ldots v_{2n}\ldots\ldots v_{nn} \end{pmatrix},\; \mathbf{\Lambda}=\begin{pmatrix} \lambda_1\ldots\ldots\ldots\ldots\ldots\\ \ldots\ldots\lambda_j\ldots\ldots\\ \ldots\ldots\ldots\ldots\ldots\lambda_n \end{pmatrix}.$$

Each eigenvector of the matrix $\mathbf{S}$ is represented by a column vector of the matrix $\mathbf{V}$ . Each eigenvalue of the matrix $\mathbf{S}$ is represented by a diagonal element of the matrix $\mathbf{\Lambda}$ .

5. The transformation of the initial data into the space of principal components is carried out. In the general case, the matrix of labels of the initial data in the space of principal coordinates can be calculated by the expression

$$\mathbf{PC_n}=\mathbf{V_n^T}\mathbf{D^T}, \qquad (2)$$

where $\mathbf{V_n^T}$ – the transposed matrix of eigenvectors of the variance/covariance matrix $\mathbf{S}$ of size $n\times n$ ; $\mathbf{d^T}$ – a transposed matrix of centred attribute values.

In (2), the labels of the initial data are their coordinates in the space of principal components. Let us consider a simple practical example.

**Example 1**. The matrix $\mathbf{A}$ represents the original data set.

$$\mathbf{A}=\begin{pmatrix} 7....3...1\\ 3....5...1.5\\ 10..1...2\\ 5...3...1\\ 4...4...2\\ 9...2...1.5\\ 6...2...1\\ 8...1...2 \end{pmatrix}.$$

It is necessary to transform these initial data into the space of principal components and reduce the dimension of the initial data.

We centre the initial attribute values from the matrix $\mathbf{A}$ . The results are presented in the matrix $\mathbf{D}$ .

$$\mathbf{D}=\begin{pmatrix} 0.5000.......0.3750....-0.5000\\ -3.5000.....2.3750.......0\\ 3.5000....-1.6250.......0.5000\\ -1.5000.....0.3750....-0.5000\\ -2.5000.....1.3750.......0.5000\\ 2.5000....-0.6250.......0\\ -0.5000..-0.6250....-0.5000\\ 1.5000.....-1.6250.......0.5000 \end{pmatrix}$$

Let us define the variance/covariance matrix for the data in the matrix $\mathbf{A}$ . Since all procedures require calculations of large volume, they must be performed using appropriate software tools. In this paper, all relevant calculations are performed in the Matlab software environment.

$$\mathbf{S}=\begin{pmatrix} 6.0000......-3.0714......0.2857\\ -3.0714.......1.9821...-0.1429\\ 0.2857......-0.1429......0.2147 \end{pmatrix}$$

Let us define the eigenvalues and the matrix of eigenvectors of the matrix $\mathbf{S}$ .

$$\lambda_1=7.6748\;\lambda_2=0.3219\;\lambda_3=0.1998$$

$$\mathbf{V}=\begin{pmatrix} 0.8788......-0.4708......-0.0776\\ -0.4752....-0.8782......-0.0539\\ 0.0428......-0.0842.......-0.9955 \end{pmatrix}$$

To perform PCA, the eigenvalues of the matrix $\mathbf{S}$ , which are the diagonal elements of the matrix $\mathbf{\Lambda}$ , must be ordered in decreasing order of their absolute values. The eigenvectors of the matrix $\mathbf{S}$ , which are the column vectors of the matrix $\mathbf{V}$ , must be ordered in the same order. In our example, the eigenvalues, $\lambda_j$ , $j=1,2,3$ , and column vectors of the matrix $\mathbf{V}$ satisfy this requirement.

Now we have all the necessary data to transform the initial data in Example 1 into the space of principal components. If we perform this procedure, we get the initial data labels in the space of three principal axes.

We will not perform this procedure in order not to increase the volume of the article. Instead, we immediately turn to procedures for reducing the dimension of the set of initial data. The main idea of such a reduction is to remove a certain number of eigenvalues from the matrix $\mathbf{\Lambda}$ and the corresponding eigenvectors from the matrix $\mathbf{V}$ . Therefore, a criterion is needed to determine the number of discarded eigenvalues and eigenvectors. A large number of criteria of this kind have been developed [13]. In this paper, we use a simple and widely used criterion. Its essence is as follows. The sum of variations in the values of individual attributes is calculated using (3)

$$Tr\left(\mathbf{S}\right) = s_1^2 + s_2^2 + \ldots + s_n^2 . \qquad (3)$$

Assessment $Tr\left(\mathbf{S}\right)$ is called the trace $\mathbf{S}$ . Having available estimates of variations $s_j^2$ , $j = 1,\ldots,n$ and a general estimate $Tr\left(\mathbf{S}\right)$ , the proportion of variability of attribute $a_j$ values in overall variability can be determined from (4)

$$\frac{s_j^2}{Tr\left(\mathbf{S}\right)},\ j = 1,\ldots,n. \qquad (4)$$

Then, the boundary value of the total variability of the attributes left for further analysis is specified. Typically, this limit value is 0.9 or 90 %.

Let us perform the necessary calculations. According to (3),

$$Tr\left(\mathbf{S}\right) = 6.0000 + 1.9821 + 0.2147 = 8.1968 .$$

By (4)

$$a_1 : \frac{6.0000}{8.1968} \approx 73\% ;\ a_2 : \frac{1.9821}{8.1968} \approx 24\% ;\ a_3 : \frac{0.2143}{8.1968} \approx 3\% .$$

Since the total variability of the values of attributes $a_1$ and $a_2$ is 97 %, the transformation of the initial attribute values into the space of principal components can be successfully performed based on the first two eigenvectors of the matrix $\mathbf{S}$ .

Imagine a reduced version of the matrix $\mathbf{V}$ :

$$\mathbf{V_k} = \begin{pmatrix} 0.8788......-0.4708 \\ -0.4752....-0.8782 \\ 0.0428......-0.0842 \end{pmatrix} .$$

The transformation of the initial data into the space of principal components based on a reduced version of the matrix $\mathbf{V}^*$ is performed by (5)

$$\mathbf{PC_k} = \mathbf{V_k^T D^T} , \qquad (5)$$

where $k$ – the number of the remained eigenvectors in the matrix $\mathbf{V}_k^T$ .

By expression (5), we transform the initial centred data from the matrix $\mathbf{D}$ into the space of principal components using the appropriate Matlab commands. The transformation results are presented in the following matrix

$$\mathbf{PC} = \begin{pmatrix} 0.2398...-4.2046....3.8695.-1.5178..-2.8291....2.4941.-0.1638...2.1119 \\ -0.5226.-0.4379.-0.2628...0.4190..-0.0726.-0.6281....0.8264...0.6788 \end{pmatrix}$$

The results presented in the matrix $\mathbf{PS}_k$ can be used to perform the necessary data analysis. In this paper, the goal is to reduce the dimensionality of the initial data based on PCA. To do this task, you need to restore the original data using row vectors from the matrix $\mathbf{PS}$ . The calculated expression for our task is

$$\mathbf{A}^* = \mathbf{PC} * \mathbf{V_k^T} + \overline{\mathbf{a}}_j , \qquad (6)$$

where $\mathbf{PC}_k$ – the matrix of principal components (matrix $\mathbf{PS}_2$ in our example); $\mathbf{V}_k^T$ – the transposed reduced matrix of eigenvectors of the matrix $\mathbf{S}$ ; $\overline{a}_j$ , $j = 1,\ldots,k$ – average attribute values in the initial data table.

We will restore the original data in our example based on the principal components $PC_1$ , and $PC_2$ . Recovery results are presented in the form of the matrix $\mathbf{A}^*$ .

$$\mathbf{A}^* = \begin{pmatrix} 6.7108...2.8050...9.9006...5.1661...4.0137...8.6918...6.3561...8.3559 \\ 2.5110...4.6231...0.7861...3.3463...3.9696...1.4397...2.7028...1.6214 \end{pmatrix}$$

Regarding the results obtained, two remarks should be made. Firstly, the matrix $\mathbf{A}^*$ is presented in the form in which it is displayed at the output of the corresponding Matlab procedure. If necessary, for example, for visual comparison with the initial data matrix $\mathbf{A}$ , this matrix can be transposed. Secondly, any label of the initial data in the space of principal components is a linear combination of all the components of the vector of attribute values for a given object. Therefore, when restoring the original data, the third row corresponding to the values of attribute $a_3$ is presented in the matrix $\mathbf{A}^*$ . Since our goal is to reduce the dimension of the original data set, we simply ignore this third row. As a result, we have the matrix $\mathbf{A}^*$ in the form presented above.

How can the degree of proximity of the restored and original data be estimated? The following estimate directly measures the degree of distortion of the restored attribute values regarding their original values.

$$\delta\left(a_j, a_j^*\right) = \sum_{i=1}^{m} \left(a_{ij} - a_{ij}^*\right)^2 , \qquad (7)$$

where $a_{ij}$ – the actual value of attribute $a_j$ for the object $o_i$ ; $a_j^*$ – the restored value of this attribute.

$$\delta\left(a_1, a_1^*\right) = \sum \left(a_{i1} - a_{i1}^*\right)^2 = 0.5078.$$

Such a small error value indicates a high-quality reduction of the PCA-based initial data in the present example.

## III. Conclusion

1. The goal of PCA is to transform the initial data into the space of uncorrelated principal components. With the correct choice of the number of eigenvectors of the variation/covariance matrix, the inverse transformation of data from the space of principal components can significantly reduce the dimension of the initial data.

2. Data transformation is based on estimates of the variability of these data, which are displayed in the form of a matrix of variation/covariance of attribute values.

3. Transformation procedures use the matrix of ordered eigenvectors $\mathbf{V}$ of the variance/covariance matrix $\mathbf{S}$ . The ordering of eigenvectors is based on an ordered set of eigenvalues of the matrix $\mathbf{S}$ in the matrix $\mathbf{\Lambda}$ .

4. Due to the large volume of necessary calculations, PCA procedures require the use of specialized software tools.

5. Reducing the initial data set allows getting rid of non-informative attributes and noise in the data. This is clearly demonstrated in the example discussed in the present paper.

## References

[1] J. De Mast and B. P. H. Kemper, "Principles of Exploratory Data Analysis in Problem Solving: What Can We Learn from a Well-Known Case?" *Quality Engineering*, vol. 21, no. 4, pp. 366–375, 2009. https://doi.org/10.1080/08982110903188276

[2] J. W. Tukay, "Analysing data: Sanctification or defective work?" *American Psychologist*, vol. 24, no. 2, pp. 83–91, 1969. https://doi.org/10.1037/h0027108

[3] J. W. Tukay, *Exploratory data analysis*. MA: Addison-Wesley, 1977.

[4] J. W. Tukay, "We Need both Exploratory and Confirmatory", *The American Statistician*, vol. 34, no. 1, pp. 23–25, 1980. https://doi.org/10.1080/00031305.1980.10482706

[5] J. W. Tukay, "The future of data analysis", *The Annals of Mathematical Statistics*, vol. 3, no. 1, pp. 1–67, 1962. https://doi.org/10.1214/aoms/1177704711

[6] H. Y. Cheng, "Exploratory data analysis in the context of data mining and resampling", *Int. Journal of Psychological Research*, vol. 3, no. 1, pp. 9–22, 2010. https://doi.org/10.21500/20112084.819

[7] J. T. Behrens, "Principles and procedures of exploratory data analysis", *Psychological Methods*, vol. 2, no. 2, pp. 131–160, 1997. https://doi.org/10.1037/1082-989X.2.2.131

[8] H. Y. Cheng "Abduction? Deduction? Induction? Is There Logic of Exploratory Data Analysis?" *Annual Meeting of the American Educational Research Association*, New Orleans, LA, April 4–8, 28 p., 1991.

[9] B. D. Haig, "An abductive theory of scientific method", *Psychological Methods*, vol. 10, no. 4, pp. 371–388, 2005. https://doi.org/10.1037/1082-989X.10.4.371

[10] I. T. Jolliffe, *Principal Component Analysis* (Second Edition). Springer-Verlag, New York, Berlin Heidelberg, 2002.

[11] J. D. Jackson, *A User's Guide to Principal Component Analysis*. John Willey & Sons, Inc., 1991.

[12] L. Ch. Paul, A. Suman and N. Sultan, "Methodological Analysis of Principal Component Analysis (PCA) Method", *Int. Journal of Computational Engineering & Management*, vol. 16, issue 2, pp. 32–38, 2013.

[13] P. R. Peres-Neto, D. A. Jackson and K. M. Somers, "How many principal components? stopping rules for determining the numbers of non-trivial axes revisited", *Computational Statistics & Data Analysis*, vol. 49, no. 4, pp. 974–997, 2005. https://doi.org/10.1016/j.csda.2004.06.015

**Oļegs Užga-Rebrovs** is a Leading Researcher of the Information and Communication Technologies Research Centre at Rezekne Academy of Technologies (Latvia). He received his Doctoral degree in Information Systems from Riga Technical University in 1994. His research interests include different approaches to processing incomplete, uncertain and fuzzy information, in particular, fuzzy sets theory, rough set theory as well as fuzzy classification and fuzzy clustering techniques and their application in bioinformatics. Currently he focuses on the problems of data analysis.
E-mail: rebrovs@tvnet.lv

**Gaļina Kulešova** is a Researcher of the Faculty of Computer Science and Information Technology at Riga Technical University (Latvia). She received her M. Sc. degree in Decision Support Systems from Riga Technical University. Current research interests include artificial neural networks, data mining, ontology engineering, classification methods, data analysis and bioinformatics.
E-mail: galina.kulesova@rtu.lv