Technical Note

Multivariate Statistical Data Analysis- Principal Component Analysis (PCA)

Sidharth Prasad Mishra*, Uttam Sarkar, Subhash Taraphder, Sanjay Datta, Devi Prasanna Swain[1], Reshma Saikhom, Sasmita Panda[2] and Menalsh Laishram[3]

Department of Animal Genetics and Breeding, Faculty of Veterinary and Animal Science, West Bengal University of Animal and Fishery Sciences, West Bengal, INDIA

[1]Department of Veterinary and Animal Husbandry Extension, Faculty of Veterinary and Animal Science, West Bengal University of Animal and Fishery Sciences, West Bengal, INDIA

[2, 3] Department of Livestock Production and Management, Faculty of Veterinary and Animal Science, West Bengal University of Animal and Fishery Sciences, West Bengal, INDIA

***Corresponding author:** sidpramishra44@gmail.com

Rec. Date:	Mar 17, 2017 11:40
Accept Date:	Apr 15, 2017 11:52
Published Online:	May 01, 2017
DOI	10.5455/ijlr.20170415115235

Abstract

Principal component analysis (PCA) is a multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent variables. Its goal is to extract the important information from the statistical data to represent it as a set of new orthogonal variables called principal components, and to display the pattern of similarity between the observations and of the variables as points in spot maps. Mathematically, PCA depends upon the eigen-decomposition of positive semi-definite matrices and upon the singular value decomposition (SVD) of rectangular matrices. It is determined by eigenvectors and eigenvalues. Eigenvectors and eigenvalues are numbers and vectors associated to square matrices. Together they provide the eigen-decomposition of a matrix, which analyzes the structure of this matrix such as correlation, covariance, or cross-product matrices. Performing PCA is quite simple in practice. Organize a data set as an m × n matrix, where m is the number of measurement types and n is the number of trials. Subtract of the mean for each measurement type or row x_i. Calculate the SVD or the eigenvectors of the co-variance. It was found that there were many interesting applications of PCA, out of which in day today life knowingly or unknowingly multivariate data analysis and image compression are being used alternatively.

Key words: Eigenvalue, Eigenvector, Linear Algebra, Matrix, Multivariate, PCA

How to cite: Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., & Saikhom, R. et al. (2017). Multivariate Statistical Data Analysis- Principal Component Analysis (PCA). International Journal of Livestock Research, 7(5), 60-78. http://dx.doi.org/10.5455/ijlr.20170415115235

DOI 10.5455/ijlr.20170415115235

Introduction

Principal component analysis is the oldest and best known technique of multivariate data analysis. It was first coined by Pearson (1901), and developed independently by Hotelling (1933). Like many other multivariate methods, it was not widely accepted nor used until the advent of electronic computers, but it is now well entrenched in virtually every statistical software packages. Principal Component Analysis (PCA) is the general name for a technique which uses sophisticated underlying mathematical principles to transforms a number of possibly correlated variables into a smaller number of variables called principal components. The origins of PCA lie in multivariate data analysis; however, it has a wide range of other applications. PCA has been called one of the most important results from applied linear algebra and perhaps its most common use is as the first step in trying to analyze large data sets. Some of the other common applications include; denoising signals, blind source separation and data compression. In general terms, PCA uses a vector space transform to reduce the dimensionality of large data sets. Using mathematical projection, the original data set, which may have involved many variables, can often be interpreted in just a few variables (i.e. the principal components). The central idea of principal component analysis is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables. Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem for a positive-semi-definite symmetric matrix. Thus, the definition and computation of principal components are straightforward but, as will be seen, this apparently simple technique has a wide variety of different applications, as well as a number of different derivations.

Definition

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.

"Or"

It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data.

DOI 10.5455/ijlr.20170415115235

A Brief History on Principal Component Analysis

The origins of statistical techniques are often difficult to trace. Preisendorfer and Mobley (1988) noted that Beltrami (1873) and Jordan (1874) independently derived the singular value decomposition (SVD) in a form that underlies PCA. Fisher and Mackenzie (1923) used the SVD in the context of a two-way analysis of an agricultural trial. However, it is generally accepted that the earliest descriptions of the technique now known as PCA were given by Pearson (1901) and Hotelling (1933). Hotelling's paper is in two parts. The first, most important, part, together with Pearson's paper, is among the collection of papers edited by Bryant and Atchley (1975). The two papers adopted different approaches, with the standard algebraic derivation given above being close to that introduced by Hotelling (1933). Pearson (1901), on the other hand, was concerned with finding lines and planes that best fit a set of points in p-dimensional space, and the geometric optimization problems he considered also lead to PCs. Pearson's comments regarding computation, given over 50 years before the widespread availability of computers, are interesting. He states that his methods 'can be easily applied to numerical problems,' and although he says that the calculations become 'cumbersome' for four or more variables, he suggests that they are still quite feasible. In the 32 years between Pearson's and Hotelling's papers, very little relevant material seems to have been published, although Rao (1964) indicates that Frisch (1929) adopted a similar approach to that of Pearson. Also, a footnote in Hotelling (1933) suggests that Thurstone (1931) was working along similar lines to Hotelling, but the cited paper, which is also in Bryant and Atchley (1975), rather than PCA. Hotelling's approach towards PCA defined it as really rather different in character from factor analysis. Hotelling's motivation is that there may be a smaller 'fundamental set of independent variables which determine the values' of the original p variables. He notes that such variables have been called 'factors' in the psychological literature, but introduces the alternative term 'components' to avoid confusion with other uses of the word 'factor' in mathematics. Hotelling chooses his 'components' so as to maximize their successive contributions to the total of the variances of the original variables, and calls the components that are derived in this way the 'principal components.' The analysis that finds such components is then christened the 'method of principal components.' Hotelling's derivation of PCs is similar to that given above; using Lagrange multipliers and ending up with an eigenvalue/eigenvector problem, but it differs in three respects. A further paper by Hotelling (1936) gave an accelerated version of the power method for finding PCs; in the same year, Girshick (1936) provided some alternative derivations of PCs, and introduced the idea that sample PCs were maximum likelihood estimates of underlying population PCs. Girshick (1939) investigated the asymptotic sampling distributions of the coefficients and variances of PCs, but there appears to have been only a small amount of work on the development of different applications of PCA during the 25 years immediately following publication of Hotelling's paper. Since then, however, an explosion of new applications and further theoretical

DOI 10.5455/ijlr.20170415115235

developments has occurred. This expansion reflects the general growth of the statistical literature, but as PCA requires considerable computing power, the expansion of its use coincided with the widespread introduction of electronic computers. Despite Pearson's optimistic comments, it is not really feasible to do PCA by hand, unless p is about four or less. But it is precisely for larger values of p that PCA is most useful, so that the full potential of the technique could not be exploited until after the advent of computers. Four articles have explained about PCA precisely, the first of these, by Anderson (1963), is the most theoretical of the four. It discussed the asymptotic sampling distributions of the coefficients and variances of the sample PCs, building on the earlier work by Girshick (1939), and has been frequently cited in subsequent theoretical developments. Rao's (1964) paper is remarkable for the large number of new ideas concerning uses, interpretations and extensions of PCA that it introduced, and which will be cited at numerous points in the book. Gower (1966) discussed links between PCA and various other statistical techniques, and also provided a number of important geometric insights. Finally, Jeffers (1967) gave an impetus to the really practical side of the subject by discussing two case studies in which the uses of PCA go beyond that of a simple dimension reducing tool. To this list of important papers the book by Preisendorfer and Mobley (1988) should be added. Although it is relatively unknown outside the disciplines of meteorology and oceanography and is not an easy read, it rivals Rao (1964) in its range of novel ideas relating to PCA, some of which have yet to be fully explored.

Goals of PCA

The goals of PCA are to-

1. extract the most important information from the data table;
2. compress the size of the data set by keeping only this important information;
3. simplify the description of the data set; and
4. Analyze the structure of the observations and the variables.
5. Compress the data, by reducing the number of dimensions, without much loss of information.
6. This technique used in image compression

In order to analysis the data by Principal Component Analysis we have to be thorough in statistics and matrix algebra. So, we will discuss on Statistics which looks at distribution measurements, how the data is spread out and also on Matrix Algebra by calculating eigenvectors and eigenvalues which is the fundamental principle to determine PCA.

Statistics

Statistics has been defined differently by different authors from time to time. Some writers define it as 'statistical data', i.e. numerical statements of facts, while others define it as 'statistical methods', i.e. complete body of the principles and techniques used in collecting and analyzing the data. Webster

 DOI 10.5455/ijlr.20170415115235

defines statistics as "classified facts representing the conditions of the people in a state especially those facts which can be stated in numbers or in any other tabular or classified arrangement". This definition, since it confines statistics only to the data pertaining to state, is inadequate as the domain of statistics is much wider.

Standard Deviation

Standard Deviation of a set of observations of a series is the positive square root of the arithmetic mean of the squares of all the deviations from the arithmetic mean. Thus, in the calculation of standard deviation, first the arithmetic mean is calculated and the deviation of various items from the arithmetic mean is squared. The squared deviations are totaled and sum is divided by the number of items. The process of omitting the algebraic signs + and – of the deviations in the mean deviation is avoided in standard deviation. Hence standard deviation is a measure of dispersion of more mathematical significance. Standard deviation is generally denoted by σ. The square of standard deviation is known as variance. So variance is denoted by σ^2.

Formula for calculation of standard deviation for ungrouped data-

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (X_i - X`)(X_i - X`)}{(n-1)}}$$

Where, σ = Standard Deviation,
X = Arithmetic mean,
d = X – X = Deviation of individual observation from arithmetic mean,
n = Number of observations, and,
Σ d^2 = Summation of squares of deviations.

Variance

Variance is another measure of the spread of data in a data set. It is defined as the mean of the deviation of each term from its arithmetic mean of whole data. In fact it is almost identical to the standard deviation. The formula is this-

$$Var(X) = \frac{\sum_{i=1}^{n} (X_I - X`)(X_i - X`)}{(n-1)}$$

σ^2 = Variance,
X` = Arithmetic mean
X – X` = Deviation of individual observation from arithmetic mean,
n = Number of observations.

σ^2 and the formula (there is no square root in the formula for variance). σ^2 is the usual symbol for variance of a sample. Both these measurements are measures of the spread of the data. Standard deviation

DOI 10.5455/ijlr.20170415115235

is the most common measure, but variance is also used. The reason why I have introduced variance in addition to standard deviation is to provide a solid platform from which covariance, can be defined.

Covariance

The last two measures we have looked at are purely 1-dimensional. Data sets like this could be heights of all the people in the room, marks for the class statistics exam etc. However many data sets have more than one dimension, and the aim of the statistical analysis of these data sets is usually to see if there is any relationship between the dimensions. For example, we might have as our data set both the height of all the students in a class and the mark they received on genetics paper. We could then perform statistical analysis to see if the height of a student has any effect on their mark. Standard deviation and variance only operate on 1 dimension, so that we could only calculate the standard deviation for each dimension of the data set independently of the other dimensions. However, it is useful to have a similar measure to find out how much the dimensions vary from the mean with respect to each other. Covariance is such a measure. Covariance is always measured between 2 dimensions. If we calculate the covariance between one dimension and itself, we will get the variance. So, if we had a 3-dimensional data set (A, B, C), then we could measure the covariance between the A and B dimensions, the B and C dimensions, and the A and C dimensions. Measuring the covariance between A and A, or B and B, or C and C would give us the variance of the A, B and C dimensions respectively. The formula for covariance is very similar to the formula for variance. The formula for covariance with respect to variance could also be written like this-

$$\text{Var}(X) = \frac{\sum_{i=1}^{n} (X_i - X`)\,(X_i - X`)}{(n-1)}$$

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

Where,

X`= Arithmetic mean of data X
Y`= Arithmetic mean of data Y
n = Number of observation

The Covariance Matrix

Covariance is always measured between 2 dimensions. If we have a data set with more than 2 dimensions, there is more than one covariance measurement that can be calculated. For example, from a 3 dimensional data set (dimensions x, y, z) we could calculate the cov(x,y), cov(y,z) and cov(x,z). In fact, for an n-dimensional data set, we can calculate $\frac{n!}{(n-2)!*2}$ different covariance values.

Hosted@www.ijlr.org DOI 10.5455/ijlr.20170415115235

A useful way to get all the possible covariance values between all the different dimensions is to calculate them all and put them in a matrix. So, by definition the covariance matrix for a set of data with n-dimensions is-

$$C^{M*N} = (c_{i,j}, c_{i.\,j} = \text{cov}(\text{Dim}_i, \text{Dim}_j)$$

Where,
C^{M*N} is a matrix with n rows and n columns, and
Dim_x is the xth dimension.

This typical formula says that if you have an n-dimensional data set, then the matrix has n rows and columns (so is square) and each entry in the matrix is the result of calculating the covariance between two separate dimensions. E.g. The entry on row 2, column 3, is the covariance value calculated between the 2nd dimension and the 3rd dimension.

An Example

We'll make up the covariance matrix for an imaginary 3 dimensional data set, using the usual dimensions x, y and z. Then, the covariance matrix has 3 rows and 3 columns, and the values are this-

$$\begin{bmatrix} \text{cov}(x,y) & \text{cov}(y,z) & \text{cov}(x,z) \\ \text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\ \text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z) \end{bmatrix}$$

Some points that taken into consideration is that the diagonal elements are the covariance value between one of the dimensions with itself. These are the variances for that particular dimension. The other point is that the $\text{cov}(x,y) = \text{cov}(y,x)$ as the matrix is symmetrical about the main diagonal.

Orthogonal or Orthonormal Basis of a Vector

An orthogonal basis of an vector space V with an inner product, is a set of basis vectors whose elements are mutually orthogonal and of magnitude 1 (i.e. unit vector). Elements in an orthogonal basis do not have to be unit vectors, but must be mutually perpendicular. It is easy to change the vectors in an orthogonal basis by scalar multiples to get an orthogonal basis, and indeed this is a typical way that an orthogonal basis is constructed. Two vectors are orthogonal if they are perpendicular, i.e., they form a right angle between each other, i.e. if their inner product is zero. In mathematical notation,

$$a^T \cdot b = \sum_{i=1}^{n} a_i b_i = 0 \Rightarrow a \text{ perpendicular to } b$$

The standard basis of the n-dimensional euclidean space R^n is an example of orthogonal (and ordered) basis.

 DOI 10.5455/ijlr.20170415115235

Eigenvectors and Eigenvalues

Eigenvectors and eigenvalues are numbers and vectors associated to square matrices. Together they provide the eigen-decomposition of a matrix, which analyzes the structure of a matrix. Even though the eigen-decomposition does not exist for all square matrices, it has a particularly simple expression for matrices such as correlation, covariance, or cross-product matrices. The eigen-decomposition of this type of matrices is important because it is used to find the maximum (or minimum) of functions involving these matrices. Specifically PCA is obtained from the eigen-decomposition of a covariance or a correlation matrix.

Matrix Algebra

Matrix algebra serves to provide a background that is required in PCA especially when eigenvectors and eigenvalues of a given matrix are taken into consideration. Here shown an assumption on basic knowledge of matrices.

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 11 \\ 5 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

Fig 1: Example of one non-eigenvector and one eigenvector

$$2 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 24 \\ 16 \end{bmatrix} = 4 \times \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

Fig 2: Example of how a scaled eigenvector is still and eigenvector

Eigenvectors

Two matrices would be multiplied together if both are of compatible size in terms of rows and columns. Eigenvectors are a special case of this concept. Consider the two multiplications between a matrix and a vector as mention in above two figures (Fig.1 and Fig. 2). In the first example, the resulting vector is not an integer multiple of the original vector, whereas in the second example, the example is exactly 4 times the vector we began with. Why is this? Well, the vector is a vector in 2 dimensional space. The vector $\begin{bmatrix} 3 \\ 2 \end{bmatrix}$ (from the second example multiplication) represents an arrow pointing from the origin (0, 0), to the point (3, 2). The other matrix, the square one, can be thought of as a transformation matrix. If you multiply this matrix on the left of a vector, the answer is another vector that is transformed from its original position. It is the nature of the transformation that the eigenvectors arise from. Imagine a transformation matrix that,

DOI 10.5455/ijlr.20170415115235

when multiplied on the left, reflected vectors in the line y = x. Then you can see that if there were a vector that lay on the line y = x, it's reflection it itself. This vector (and all multiples of it, because it wouldn't matter how long the vector was), would be an eigenvector of that transformation matrix. What properties do these eigenvectors have? We should first know that eigenvectors can only be found for square matrices. And, not every square matrix has eigenvectors. And, given an n×n matrix that does have eigenvectors, there are n of them. Given 3×3 square matrix will have 3 eigenvectors. Another property of eigenvectors is that even if we scale the vector by some amount before multiplying it, it will still get the same multiple of it as a result, as shown in Fig. 2. This is because if you scale a vector by some amount, all you are doing is making it longer, not changing its direction. Lastly, all the eigenvectors of a matrix are perpendicular, i.e. at right angles to each other, no matter how many dimensions you have. By the way, another word for perpendicular, in statistical language, is orthogonal. This is important because it means that you can express the data in terms of these perpendicular eigenvectors, instead of expressing them in terms of the x and y axes. Another important thing to know is that when mathematicians find eigenvectors, they like to find the eigenvectors whose length is exactly one. This is because, as we know that, the length of a vector doesn't affect whether it's an eigenvector or not, whereas the direction does. So, in order to keep eigenvectors standard, whenever we find an eigenvector we usually scale it to make it have a length of 1, so that all eigenvectors have the same length. Here's a demonstration from our example above-

$$\begin{bmatrix}3\\2\end{bmatrix}$$

is an eigenvector, and the length of that vector is

$$\sqrt{(3^2+2^2)} = \sqrt{13}$$

so we divide the original vector by this much to make it have a length of 1.

$$\begin{bmatrix}3\\2\end{bmatrix} \div \sqrt{13} = \begin{bmatrix}\frac{3}{\sqrt{13}}\\ \frac{2}{\sqrt{13}}\end{bmatrix}$$

Eigenvalues

Eigenvalues are closely related to eigenvectors, in fact, we saw an eigenvalue in Fig.1. Notice how, in both those examples, the amount by which the original vector was scaled after multiplication by the square matrix was the same? In that example, the value was 4. 4, is the eigenvalue associated with that eigenvector. No matter what multiple of the eigenvector we took before we multiplied it by the square matrix, we would always get 4 times the scaled vector as our result (as in Table 1). So we can see that eigenvectors and eigenvalues always come in pairs.

 DOI 10.5455/ijlr.20170415115235

Notations and Definition

There are several ways to define eigenvectors and eigenvalues, the most common approach defines an eigenvector of the matrix A as a vector u that satisfies the following equation-

$$Au = \lambda u$$

When rewritten, the equation becomes-

$$(A - \lambda I)\, u = 0,$$

Where, λ is a scalar called the eigenvalue associated to the eigenvector.

In a similar manner, we can also say that a vector u is an eigenvector of a matrix A if the length of the vector (but not its direction) is changed when it is multiplied by A.

For example, the matrix

$$A = \begin{bmatrix} 0 & 1 \\ -2 & 3 \end{bmatrix}$$

Then,

$$|A \quad \lambda . I| = \left| \begin{bmatrix} 0 & 1 \\ -2 & 3 \end{bmatrix} - \quad \lambda . \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = \left| \begin{bmatrix} 0 & 1 \\ -2 & 3 \end{bmatrix} - \quad \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right|$$

$$= \begin{vmatrix} -\lambda & 1 \\ -2 & -3-\lambda \end{vmatrix} = (-\lambda \times (-3-\lambda)) - (-2 \times 1) = \lambda^2 + 3\lambda + 2$$

And setting the determinant to 0, we obtain 2 eigenvalues-

$$\lambda_1 = -1 \quad \text{and} \quad \lambda_2 = -2$$

Linear Algebra

The inverse of an orthogonal matrix will gives the transpose of the matrix. The importance of the concept is that if A is an orthogonal matrix, then

$A^{-1} = A^T$

Let A be an m × n matrix

$A = [a_1\ a_2\ \ldots\ a_n]$

Where, a_i is the i^{th} column vector.

We now show that,

$A^T A = I$

Where, I is the identity matrix

 DOI 10.5455/ijlr.20170415115235

Let us examine the ijth element of the matrix $A^T A$. The ijth element of $A^T A$ is

$(A^T A)_{ij} = a_i^T a_j$.

So, it should be memorized that the columns of an orthonormal matrix are orthonormal to each other. In other words, the dot product of any two column matrix will be zero. The only exception is a dot product of one particular column with itself, which will be equals to one. $A^T A$ is the exact description of the identity matrix. The definition of A^{-1} is $A^{-1}A = I$. Therefore,
because $A^T A = I$, it follows that $A^{-1} = A^T$

If A is any matrix, the matrices $A^T A$ and AA^T are both symmetric. Let's us examine the transpose of each.

$$(AA^T)^T = A^{TT} A^T = AA^T$$
$$(A^T A)^T = A^T A^{TT} = A^T A$$

The equality of the quantity ion its transpose satisfies the concept.

A matrix will be symmetrical if it is orthogonally diagonalizable. As the statement is bi-directional, it requires a two-way derivation. One needs to prove both the forward and the backward case.
Let us examine the forward case. If A is orthogonally diagonalizable, then A is a symmetric matrix. By hypothesis, orthogonally diagonalizable means that there exists some E such that-

$A = EDE^T$

Where, D is a diagonal matrix and
E is some special matrix which diagonalize A.
Let us compute for A^T.
$A^T = (EDE^T)\, T = E^{TT} D^T E^T = EDE^T = A$

Evidently, if A is orthogonally diagonalizable, it must also be symmetric.
A symmetric matrix is diagonalized by a matrix of its orthonormal eigenvectors. Let A be a square n × n symmetric matrix with associated eigenvectors $\{e_1, e_2 \ldots e_n\}$. Let $E = [e_1\ e_2 \ldots e_n]$ where the ith column of E is the eigenvector e_i. This theorem asserts that there exists a diagonal matrix D where

$A = EDE^T$

This theorem is an extension of the previous theorem. It provides a description to determine the matrix E, the "diagonalizer" for a symmetric matrix. It establishes that the special diagonalizer is in fact a matrix of the original matrix's eigenvectors. This proof is in two parts. The first part, describes that any matrix can

DOI 10.5455/ijlr.20170415115235

be orthogonally diagonalized if and only if it that matrix eigenvectors are all linearly independent. While second part proves that a symmetric matrix has the special property that all its eigenvectors are not just linearly independent but also orthogonally.

In the first part of the proof, let A be just some matrix, not necessarily symmetric, and let it have independent eigenvectors (i.e. no degeneracy). Furthermore, let $E = [e_1\ e_2\ \ldots\ e_n]$ be the matrix of eigenvectors placed in the columns. Let D be a diagonal matrix where the i^{th} eigenvalue is placed in the ii^{th} position. We will now show that AE = ED. We can examine the columns of the right-hand and left-hand sides of the equation.

$$\begin{aligned} &\text{Left hand side :} \quad AE = [Ae_1\ Ae_2\ \ldots\ Ae_n] \\ &\text{Right hand side :} \quad ED = [\lambda_1 e_1\ \lambda_2 e_2\ \ldots\ \lambda_n e_n] \end{aligned}$$

Eventually, if AE = ED, then $Ae_i = \lambda_i e_i$ for all i^{th} elements. This equation is the definition of the eigenvalue equation. Therefore, it must be that AE = ED. A little rearrangement provides $A = EDE^{-1}$, thus, proves the first derivation.

For the second part of the proof, we show that a symmetric matrix always has orthogonal eigenvectors. For some symmetric matrix, let λ_1 and λ_2 be distinct eigenvalues for eigenvectors e_1 and e_2.

$$\begin{aligned} & \lambda_1 e_1 \cdot e_2 = (\lambda_1 e_1)^T e_2 \\ \Rightarrow\ & \lambda_1 e_1 \cdot e_2 = (Ae_1)^T e_2 \\ \Rightarrow\ & \lambda_1 e_1 \cdot e_2 = e_1^T A^T e_2 \\ \Rightarrow\ & \lambda_1 e_1 \cdot e_2 = e_1^T Ae_2 \\ \Rightarrow\ & \lambda_1 e_1 \cdot e_2 = e_1^T (\lambda_2 e_2) \\ & \lambda_1 e_1 \cdot e_2 = \lambda_2 e_1 \cdot e_2 \end{aligned}$$

Thus by the last equation we can equate that,

$(\lambda_1 - \lambda_2)\, e_1 \cdot e_2 = 0$

Since it was conjectured that the eigenvalues are in unique, it must be the case that $e_1 \cdot e_2 = 0$. Therefore, the eigenvectors of a symmetric matrix are orthogonal. Thus, it proves that the eigenvectors of matrix A are all orthonormal provided that eigenvectors should be normalized. This means that E is an orthogonal matrix by theorem 1, $E^T = E^{-1}$ thereby, we can rewrite the final result as-

$A = EDE^T$

Thus, a symmetric matrix is diagonalized by a matrix of its eigenvectors.

For any arbitrary m × n matrix X, the symmetric matrix $X^T X$ has a set of orthonormal eigenvectors of $\{v_1, v_2. \ldots v_n\}$ and a set of associated eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. The set of vectors $\{Xv_1, Xv_2. \ldots Xv_n\}$ then form an orthogonal basis, where each vector Xv is square root times the length of λ_i. All of these properties arise from the dot product of any two vectors as given below-

DOI 10.5455/ijlr.20170415115235

$$(Xv_i) \cdot (Xv_j) = (Xv_i)^T (Xv_j)$$

$= v_i^T X^T X v_j = v_i^T (\lambda_j v_j)$

$= \lambda_j v_i \cdot v_j (Xv_i) \cdot (Xv_j)$

$= \lambda_j \delta_{ij}$

The last relation arises because the eigenvectors of X is orthogonally resulting in the Kronecker delta. This equation states that any two vectors in the set are orthogonal. The second property arises from the above equation by realizing that the length squared of each vector is defined as-

$$kX\hat{}v_i k^2 = (X\hat{}v_i) \cdot (X\hat{}v_i) = \lambda_i$$

Methodology

We are going to analysis a own designed data set by applying PCA. The steps are as follows-

Step 1: Get Some Data

In this simple example, we are going to use our own made-up data set. It's only got 2 dimensions, and the reason why we have chosen this is so that we can provide plots of the data to show what the PCA analysis is doing at each step. The data used is given in Table 1, along with a plot of that data.

Table 1: Original data on the left, data with the subtracted mean taken on the right

Variable X	Variable Y	Deviation from Mean for X	Deviation from Mean for Y
2.5	2.4	0.69	0.49
0.5	0.7	-1.31	-1.21
2.2	2.9	0.39	0.99
1.9	2.2	0.09	0.29
3.1	3.0	1.29	1.09
2.3	2.7	0.49	0.79
2	1.6	0.19	-0.31
1	1.1	-0.81	-0.81
1.5	1.6	-0.31	-0.31
1.1	0.9	-.71	-1.01

Step 2: Subtract the Mean

For PCA to work properly, mean of the entire data was subtracted from each of the data dimensions. The mean subtracted is the average across each dimension. So, all the x values have x' (the mean of the x values of all the data points) subtracted, and all the y values have y' subtracted from them. This produces a data set whose mean is zero.

DOI 10.5455/ijlr.20170415115235

Step 3: Calculate the Covariance Matrix

This is done in exactly the same way as we have discussed previously. Since the data is 2 dimensional, the covariance matrix will be 2×2. There is no complicacy here, so the result will be-

$$\text{cov} = \begin{bmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555558 \end{bmatrix}$$

As the non-diagonal elements in this covariance matrix are positive, it would be expected that both the x and y variable increase together.

Step 4: Calculate the Eigenvectors and Eigenvalues of the Covariance Matrix

Since the covariance matrix is square, the eigenvectors and eigenvalues for the matrix can be calculated. This step is most important, as it helps to get useful information about the data. By this we can determine the eigenvectors and eigenvalues are given below-

$$\text{Eigenvalue} = \begin{bmatrix} 0.0490833989 \\ 1.28402771 \end{bmatrix}$$

$$\text{Eigenvector} = \begin{bmatrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{bmatrix}$$

It is important to notice that these eigenvectors are both unit eigenvectors as it is very important for PCA. Also, in most of the statistical software packages develop a unit eigenvectors.

Step 5: Choosing Components and Forming a Feature Vector

Here is the notion where data compression and reduced dimensionality comes into play. If we look at the eigenvectors and eigenvalues from the previous section, we will notice that the eigenvalues are quite different values. In fact, it turns out that the eigenvector with the highest eigenvalue is the principle component of the data set. The eigenvector with the largest eigenvalue was the one that pointed down the middle of the data. It is the most significant relationship between the data dimensions. In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives you the components in order of significance. After that components with lesser significance were ignored. If the eigenvalues are small, the information will be much more accurate. If some the components were leaves out, the final data set will have fewer dimensions than the original. To be precise, if we originally have dimensions in our data, we would have calculated eigenvectors and eigenvalues, and then we choose only the first eigenvectors, then the final data set has only dimensions. What needs to be done now is we need to form a feature vector, which is just a fancy name for a matrix of

DOI 10.5455/ijlr.20170415115235

vectors. This is constructed by taking the eigenvectors that we want to keep from the list of eigenvectors, and forming a matrix with these eigenvectors in the columns.

$$\text{Feature Vector} = (\text{eig}_1\text{eig}_2\text{eig}_3.....\text{eig}_n)$$

Given with the example set of data, and the fact that here will be 2 eigenvectors, that means there will be two choices. A feature vector can be form with either of the eigenvectors-

$$\begin{bmatrix} 0.677873399 & -0.735178656 \\ -0.735178656 & -0.677873399 \end{bmatrix}$$

or, the smaller one can be chosen to leave out, less significant component and only have a single column vector-

$$\begin{bmatrix} 0.677873399 \\ -0.735178656 \end{bmatrix}$$

The result of each of this step is used to develop a new data set in the next step.

Step 5: Deriving the New Data Set

Once the components (eigenvectors) chosen to keep in the data and a feature vector have been formed, then the transpose of the vector was calculated and multiply it on the left over of the original data set, transposed.

Final Data = Row Feature Vector × Row Data Adjust

Where, 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, i.e. the data items are in each column, with each row holding a separate dimension. The transpose of the feature vector and the data were calculated first. Then the final data set, with data items in columns, and dimensions along rows were determined. It will give the original data solely in terms of the vectors that was taken into consideration. The original data set had two axes, x and y, so the new data set was in terms of them. It is possible to express data in terms of any two axes that depends on the statistician. If these axes are perpendicular, then the expression is the most efficient. That's why eigenvectors are always perpendicular to each other. Here the original data have been changed in terms of the axes x and y to two newly developed eigenvectors. So, the new data set has reduced dimensionality, in terms of the vectors that satisfy the original data the most and leaving the unimportant eigenvectors. The new eigenvectors of the taken example is given below-

Table 2: The table of data by applying the PCA analysis using both eigenvectors

DOI 10.5455/ijlr.20170415115235

Variable X	Variable Y
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287

To show this on the data, the final transformation with each of the possible feature vectors has to be done. The transpose of each of the result was carried out, to bring the data back to the nice table-like format. By this it is understandable that no information had been lost in this decomposition and there is a strong correlation newly developed data with the original one. The other transformation can be made by taking only the eigenvector with the largest eigenvalue. The table of data resulting from that is given below-

Table 3: The data after transforming using only the most significant eigenvector

Transformed Data (Single eigenvector) for Variable X
-0.8279 70186
1.77758033
-0.992197494
-0.274210416
-1.67580142
-0.912949103
0.0991094375
1.14457216
0.438046137
1.22382056

As expected, it only has a single dimension. If we compare this data set with the one resulting from using both eigenvectors, we will notice that this data set is exactly the first column of the other. So, if we were to plot this data, it would be 1-dimensional, and would be points on a line in exactly the x positions of the points in the plot in Table 2. We have effectively thrown away the whole other axis, which is the other eigenvector. Basically we have transformed our data so that is expressed in terms of the patterns between them, where the patterns are the lines that most closely describe the relationships between the data. This is helpful because we have now classified our data point as a combination of the contributions from each of those lines. Initially we had the simple x and y axes. This is fine, but the x and y values of each data point

DOI 10.5455/ijlr.20170415115235

don't really tell us exactly how that point relates to the rest of the data. Now, the values of the data points tell us exactly where (i.e. above/below) the trend lines the data point sits. In the case of the transformation using both eigenvectors, we have simply altered the data so that it is in terms of those eigenvectors instead of the usual axes (Table 3). But the single-eigenvector decomposition has removed the contribution due to the smaller eigenvector and left us with data that is only in terms of the other.

Geometrical Interpretation

- PCA projects the data along the directions where the data varies the most.
- These directions are determined by the eigenvectors of the covariance matrix corresponding to the largest eigenvalues.
- The magnitude of the eigenvalues corresponds to the variance of the data along the eigenvector directions.

Conclusion

One benefit of PCA is that we can examine the variances associated with the principle components. Often one finds that large variances associated with the first $k < m$ principal components, and then a precipitous drop up. One can conclude that most interesting dynamics occur only in the first k dimensions. Both the strength and weakness of PCA is that it is a non-parametric analysis. There are no parameters to tweak and no coefficients to adjust based on user experience the answer are unique and independent of the user. This same strength can also be viewed as a weakness. If one knows a priori some features of the dynamics of a system, then it makes sense to incorporate these assumptions into a parametric algorithm or an algorithm with selected parameters. Thus, the appropriate parametric algorithm is to first convert the data to the appropriately centered polar coordinates and then compute PCA. Performing PCA is quite simple in practice. Organize a data set as an m×n matrix, where m is the number of measurement types and n is the number of trials. Subtract of the mean for each measurement type or row x_i. Calculate the SVD or the eigenvectors of the co-variance. It was found that there were many interesting applications of PCA, out of which in day today life knowingly or unknowingly multivariate data analysis and image compression are being used alternatively.

References

1. Abdi H, Valentin D and Edelman B. 1999. Neural Networks. Thousand Oaks, CA: Sage.
2. Abdi H and Valentin D. 2007. Multiple factor analysis (mfa). In: Salkind NJ, ed. Encyclopedia of Measurement and Statistics. Thousand Oaks, CA: Sage Publications. 657-663.
3. Abdi H and Williams LJ. 2010. Barycentric discriminant analyis (BADIA). In: Salkind NJ, ed. Encyclopedia of Research Design. Thousand Oaks, CA: Sage.
4. Abdi H and Williams LJ. 2010. Correspondence analysis. In: Salkind NJ, ed. Encyclopedia of Research Design. Thousand Oaks: Sage Publications.

DOI 10.5455/ijlr.20170415115235

5. Abdi H. 2009. Centroid. *Wiley Interdisciplinary Reviews: Computational Statistics*. 1: 259-260.
6. Abdi H. 2003. Factor Rotations. In Lewis-Beck M, Bryman A., Futimg T., eds. Encyclopedia for Research Methods for the Social Sciences. Thousand Oaks, CA: Sage Publications. 978-982.
7. Abdi H. 2010. Partial least square regression, Projection on latent structures Regression, PLS-Regression. *Wiley Interdisciplinary Reviews: Computational Statistics*. 97-106.
8. Abdi H. 2007. Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition(GSVD). In: Salkind NJ, ed. Encyclopedia of Measurement and Statistics. Thousand Oaks: Sage Publications. 907-912.
9. Bell A and Sejnowski T. 1997. "The Independent Components of Natural Scenes are Edge Filters". *Vision Research*. 37(23): 3327-3338.
10. Benzecri´ JP. 1973. L'analyse des donnees´, Vols. 1 and 2. Paris: Dunod.
11. Bishop C. 1996. Neural Networks for Pattern Recognition. Clarendon, Oxford, UK.
12. Boyer C and Merzbach U. 1989. A History of Mathematics. 2nd ed. New York: John Wiley and Sons.
13. Diamantaras KI and Kung SY. 1996. Principal Component Neutral Networks: Theory and Applications. New York: John Wiley and Sons.
14. Dray S. 2008. On the number of principal components: a test of dimensionality based on measurements of similarity between matrices. *Comput Stat Data Anal*. 52: 2228-2237.
15. Eastment HT and Krzanowski WJ. 1982. Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*. 24: 73-77.
16. Eckart C and Young G. 1936. The approximation of a matrix by another of a lower rank. *Psychometrika*. 1: 211-218.
17. Escofier B and Pages` J. 1994. Multiple factor analysis. *Comput Stat Data Anal*. 18: 121-140.
18. Good I. 1969. Some applications of the singular value decomposition of a matrix. *Technometrics*. 11: 823-831.
19. Gower J. 1971. Statistical methods of comparing different multivariate analyses of the same data. In: Hodtson F, Kendall D, Tautu P, eds. Mathememtics in the Archæological and Historical Sciences. Edinburgh: Edingburh University Press. 138-149.
20. Grattan-Guinness I. The Rainbow of Mathematics. New York: Norton. 1997.
21. Greenacre MJ. 2007. Correspondence Analysis in Practice. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.
22. Harris RJ. 2001. A Primer of Multivariate Statistics. Mahwah, NJ: Lawrence Erlbaum Associates.
23. Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*. 25: 417-441.
24. Hwang H, Tomiuk MA and Takane Y. 2009. Correspondence analysis, multiple correspondence analysis and recent developments. In: Millsap R, Maydeu-Olivares A, eds. Handbook of Quantitative Methods in Psychology. London: Sage Publications. 243-263.
25. Jackson JE. 1991. A User's Guide to Principal Components. New York: John Wiley & Sons.
26. Jolliffe IT. 2002. Principal Component Analysis. New York: Springer.
27. Jordan C. 1874. Memoire´ sur les formes bilineaires´. *J Math Pure Appl*. 19: 35-54.
28. Lay D. 2000. Linear Algebra and It's Applications. Addison-Wesley, New York.
29. Mitra P and Pesaran B. 1999. "Analysis of Dynamic Brain Imaging Data". *Biophysical Journal*. 76: 691-708.
30. Pearson K. 1901. On lines and planes of closest fit to systems of points in space. *Philos Mag A*. 6: 559-572.
31. Peres-Neto PR, Jackson DA and Somers KM. 2005. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput Stat Data Anal*. 49: 974-997.
32. Quenouille M. 1956. Notes on bias and estimation. *Biometrika*. 43: 353-360.
33. Saporta G and Niang N. 2006. Correspondence analysis and classification. In: Greenacre M, Blasius J, eds. Multiple Correspondence Analysis and Related Methods. Boca Raton, FL: Chapman & Hall. 371-392.

DOI 10.5455/ijlr.20170415115235

34. Saporta G and Niang N. 2009. Principal component analysis: application to statistical process control. In: Govaert G, ed. Data Analysis. London: John Wiley & Sons. 1-23.
35. Stewart GW. 1993. On the early history of the singular value decomposition. SIAM Rev, 35: 551-566.
36. Stone JV. 2004. Independent Component Analysis: A Tutorial Introduction. Cambridge: MIT Press.
37. Stone M. 1974. Cross validatory choice and assessment of statistical prediction. *J R Stat Soc [Ser A]*. 36: 111-133.
38. Strang G. 2003. Introduction to Linear Algebra. Cambridge, MA: Wellesley-Cambridge Press.
39. Thurstone LL. 1947. Multiple Factor Analysis. Chicago, IL: University of Chicago Press.
40. Will T. 1999. "Introduction to the Singular Value Decomposition" Davidson College. www.davidson.edu/academic/math/will/svd/index.html.
41. Wold S. 1995. PLS for multivariate linear modeling. In: van de Waterbeemd H, ed. Chemometric Methods in Molecular Design. Weinheim: Wiley-VCH Verlag. 195-217.

DOI 10.5455/ijlr.20170415115235