


Research Paper

Information Technology



# A Study on Principal Component Analysis For Lossless Data Compression

**\* Er. Prasannajit Dash \*\* Prof. (Dr.) Maya Nayak**

**\* Assistant Professor, Dept. of Information Technology, Orissa Engg. College, Navajyoti-bihar, Bhubaneswar,Odisha.**

**\*\* Head of Dept.(IT), Dept. of Information Technology, Orissa Engg. College,Navajyotibi-har, Bhubaneswar,Odisha.**

ABSTRACT

*Principal components analysis (PCA) is one of a family of techniques for taking high-dimensional data, and using the dependencies between the variables to represent it in a more tractable, lower-dimensional form, without losing too much information. PCA is one of the simplest and most robust ways of doing such dimensionality reduction. It is also one of the best, and has been rediscovered many times in many fields, so it is also known as the Karhunen-Loeve transformation, the Hotelling transformation, the method of empirical orthogonal functions, and singular value decomposition.*

**Keywords : mean, standard deviation, variance, covariance, eigenvector, eigenvalues**

INTRODUCTION

Principal component analysis is the way of finding patterns in data. These data is analyzed in such a way that the similarities and the differences are highlighted. 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. The other main advantage of PCA is that once it is found these patterns in the data, and the data is in compressed form, ie. by reducing the number of dimensions, without much loss of information, then this technique can be used in image compression. The entire subject of Principal Component Analysis is based around the idea that the data set is very big for more accuracy and to analyze that the data set in terms of the relationships between the individual points in that data set. The implementation of PCA is based on the entire subject of statistics that can be done on a big set of data, and what they do for the user about the data itself.

STANDARD DEVIATION

To understand standard deviation, we need a bigger data set. Statisticians are usually concerned with taking a sample of a population. To use election polls as an example, the population is all the people in the country, whereas a sample is a subset of the population that the statisticians measure. Here is an example data set :

X = [ 1 2 4 6 12 15 25 45 68 67 65 98 ]

It could be simply used the symbol X to refer to this entire set of numbers. If there is a need to refer to an individual number in this data set, the subscripts will be used on the symbol X to indicate a specific number. Eg. X<sub>3</sub> refers to the 3rd number in X, namely the number 4. It should be noted that X<sub>1</sub> is the first number in the sequence. Also, the symbol n will be used to refer to the number of elements in the set X There are a number of things that can be calculated about a data set i.e. the mean of the sample.

The mean is calculated by the formula:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \dots\dots\dots(Eq.1)$$

The symbol i.e.  $\bar{X}$  (said “X bar”) to indicate the mean of the set X . Unfortunately, the mean doesn’t tell us a lot about the data except for a sort of middle point. The Standard Deviation (SD) of a data set is a measure of how spread out the data is. The definition of the SD is: “The average distance from the mean of the data set to a point”. The way to calculate it is to compute the squares of the distance from each data point to the mean of the set, add them all up, divide by(n-1) , and take the positive square root. As a formula:

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n-1)}} \dots\dots\dots(Eq.2)$$

where s is the usual symbol for standard deviation.

TABLE-1  
THE ABOVE EXAMPLE DATA SET

	X	(X - $\bar{X}$ )	(X - $\bar{X}$ ) <sup>2</sup>
	1	-33	1089
	2	-32	1024
	4	-30	900
	6	-28	784
Standard Deviation	12	-22	484
	15	-19	361
	25	-9	81
	45	11	121
	68	34	1156
	67	33	1089
	65	31	961
	98	64	4096
Total	408		12146
Divided by (n-1)			1104.182
Square Root			33.2292

## VARIANCE

Variance is another measure of the spread of data in a data set. In fact it is almost identical to the standard deviation. The formula is this:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n-1)} \dots\dots\dots (\text{Eq.3})$$

It is to be noticed that this is simply the standard deviation squared, in both the symbol ( $S^2$ ) and the formula (there is no square root in the formula for variance).  $S^2$  is the usual symbol for variance of a sample. Both these measurements are measures of the spread of the data. Standard deviation is the most common measure, but variance is also used to provide a solid platform from which the covariance can launch from.

## COVARIANCE

The standard deviation and the variance are purely one-dimensional. Data sets like this could be: heights of all the people in the room or marks for the last COMP101 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 for that 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 one 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 two dimensions. If calculate the covariance between one dimension and itself, the result should be variance. So, if you had a 3-dimensional data set (x,y,z) then it is measured as the covariance between the X and Y dimensions, the X and Z dimensions, and the Y and Z dimensions. Measuring the covariance between X and X, or Y and Y, or Z and Z would give you the variance of the X,Y and Z dimensions respectively. Similarly the covariance could be calculated for a 4-dimensional data set as discussed for 3-dimensional data set. The formula for covariance is very similar to the formula for variance.

The formula for variance could also be written like this:

$$\text{var}(x) = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{(n-1)} \dots\dots\dots (\text{Eq.4})$$

where it has been simply expanded the square term to show both parts. So given that knowledge,

$$\text{cov}(x,y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)} \dots\dots\dots (\text{Eq.5})$$

Let for example some 2-dimensional data is collected by asking a bunch of students how many hours in total that they spent studying COSC241, and the mark that they received. So the first is the H dimension, the hours studied, and the second is the M dimension, the mark received.

**TABLE-2 2-DIMENSIONAL DATA SET AND COVARIANCE CALCULATION**

	Hours(H)	Mark(M)
Data	9	39
	15	56
	25	93
	14	61
	10	50
	18	75
	0	32
	16	85
	5	42
	19	70
	16	66
	20	80
Totals	167	749
Averages	13.92	62.42

**TABLE-3 COVARIANCE**

	H	M	$(H_i - \bar{H})$	$(M_i - \bar{M})^C$	$(H_i - \bar{H}) \times (M_i - \bar{M})$
	9	39	-4.92	-23.42	115.23
	15	56	1.08	-6.42	-6.93
	25	93	11.08	30.58	338.83
	14	61	0.08	-1.42	-0.11
	10	50	-3.92	-12.42	48.69
	18	75	4.08	12.58	51.33
	0	32	-13.92	-30.42	423.45
	16	85	2.08	22.58	46.97
	5	42	-8.92	-20.42	182.15
	19	70	5.08	7.58	38.51
	16	66	2.08	3.58	7.45
	20	80	6.08	17.58	106.89
Total					1149.89
Average					104.54

If the value is positive, as it is here, then that indicates that both dimensions increase together, meaning that, in general, as the number of hours of study increased, so did the final mark. If the value is negative, then as one dimension increases, the other decreases. If it had been ended up with a negative covariance here, then that would have said the opposite, that as the number of hours of study increased as well as the final mark decreased. In the last case, if the covariance is zero, it indicates that the two dimensions are independent of each other. So, the definition for the covariance matrix for a set of data with n dimensions is:

$$C^{n \times n} = (C_{ij}, C_{ij} = \text{cov}(\text{Dim}_i, \text{Dim}_j)) \text{Eq(6)}$$

where  $C^{n \times n}$  is a matrix with n rows and n columns, and  $\text{Dim}_x$  is the x th dimension. The formula says is that if you have an n dimensional data set, then the matrix has n rows and n columns (so is square) and each entry in the matrix is the result of calculating the covariance between two separate dimensions. For example, the entry on row 2, column 3, is the covariance value calculated between the 2nd dimension and the 3rd dimension.

The covariance matrix for a three 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:

$$C = \begin{pmatrix} \text{cov}(x,x) & \text{cov}(x,y) & \text{cov}(x,z) \\ \text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\ \text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z) \end{pmatrix}$$

**EIGENVECTOR**

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

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

**Figure 1: Example of one non-eigenvector and one eigenvector**

The two matrices can be multiplied together, provided they are compatible sizes. Eigenvectors are a special case of this. Consider the two multiplications between a matrix and a vector in Figure 2.2. In the first, the resulting vector is not an integer multiple of the original vector, whereas in the second, the example is exactly 4 times the vector. So, the vector is a vector in 2 dimensional space. If you multiply this matrix on the left of a vector, the answer is another vector that is transformed from it's original position. It is the nature of the transformation that the eigenvectors arise from. Imagine a transformation matrix that, when multiplied on the left, got the reflected vectors in the line  $y=x$ . Then it can be seen that if there were a vector that lay on the line  $y=x$ , it's reflection is 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. The eigenvectors can only be found for square matrices. And, not every square matrix has eigenvectors. And, given an  $n \times n$  matrix that does have eigenvectors, there are  $n$  of them. Given a  $3 \times 3$  matrix, there are 3 eigenvectors.

**EIGENVALUES**

Looking for a transformation of the data matrix, say  $X(n \times p)$  such that

$$Y = \delta^T X = \delta_1 X_1 + \delta_2 X_2 + \dots + \delta_p X_p \quad \text{Eq.(7)}$$

$$\text{Where } \delta = (\delta_1, \delta_2, \dots, \delta_p)^T \text{ is a}$$

column vector of weights with

$$\delta_1^2 + \delta_2^2 + \dots + \delta_p^2 = 1.$$

Maximize the variance of the projection of the observations on the  $Y$  variables by finding  $d$  so that the variance is  $\text{Var}(\delta^T X) = \delta^T \text{Var}(X) \delta$  is maximal. The matrix  $C = \text{Var}(X)$  is the covariance matrix of the  $X_i$  variables. The direction of  $d$  is given by the eigenvector  $g_1$  corresponding to the largest eigenvalue of matrix  $C$ . The second vector that is orthogonal (uncorrelated) to the first is the one that has the second highest variance which comes to be the eigenvector corresponding to the second eigenvalue and so on.

New variables  $Y_i$  that are linear combination of the original variables ( $x_i$ ):

$$Y_i = a_{i1}x_1 + a_{i2}x_2 + \dots + a_{ip}x_p; \quad i=1..p \quad \text{Eq.(8)}$$

The new variables  $Y_i$  are derived in decreasing order of importance, so they are called 'principal components'.

The eigenvalues  $\lambda_i$  are found by solving the equation  $\det(C - \lambda I) = 0$  where  $C$  is the covariance matrix.

Let us take two variables with covariance  $C > 0$ , hence

$$\begin{pmatrix} 1-\lambda & c \\ c & 1-\lambda \end{pmatrix} \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix} \text{ so } C - \lambda I = \begin{pmatrix} 1-\lambda & c \\ c & 1-\lambda \end{pmatrix} \text{ as } \det(C - \lambda I) = (1-\lambda)^2 - c^2 \dots \dots \dots \text{Eq.9}$$

Solving this it has been found that  $I_1 = (1+C)$  and  $I_2 = (1-C)$

**APPLYING PCA METOHDS****STEP 1: GET REAL-TIME DATA SET**

The actual data set is taken from Tata Motors's sensx data

listed in National stock exchange for last three months from 3<sup>rd</sup> November 2013. The survey has been done for daily Open Price, High Price, Low Price and Close Price for sensx data relative to each day transaction. This is the historical data being taken from website of National Stock Exchange. This data is of lossless in nature i.e. text data the is  $61 \times 4$  matrix. It means there are 61 rows and four columns where data set is 4 dimensional in nature for the Open Price, High Price, Low Price and Close Price as the date is the index.

Suppose that the data to be reduced consists of tuples or data vectors described by 'n' attributes or dimensions. PCA also called (Karhunen- Loeve or K-L method), searches for  $k$  n-dimensional orthogonal vectors that can best be used to represent the data, where  $k \leq n$ . The original data are thus projected on to a much similar space, resolving in dimensionality reduction. Unlike attribute subset selection, which reduces the attribute set size by retaining a subset of the initial set of attributes, PCA combines the essence of attributes by creating an alternative, smaller set of variables. The initial data can then be projected onto this smaller set.

The basic procedure is that the input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domain will not dominate with smaller domains.

PCA computes  $K$  orthogonal vectors that provide a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others. These vectors are referred to as the principal components. The input data are a linear combination the principal components

Sources: [http://www.nseindia.com/live\\_market/dynaContent/live\\_watch/get\\_quote/GetQuote.jsp?symbol=TATAMOTORS](http://www.nseindia.com/live_market/dynaContent/live_watch/get_quote/GetQuote.jsp?symbol=TATAMOTORS)

NATIONAL STOCK EXCHANGE - TATA MOTORS SENSEX DATA FOR THREE				
Date	Open Price	High Price	Low Price	Close Price
3-Nov-13	386.00	392.60	386.00	391.00
1-Nov-13	384.00	389.00	383.20	384.60
31-Oct-13	380.00	386.40	379.20	381.15
30-Oct-13	383.00	384.15	377.20	379.60
29-Oct-13	377.70	384.20	375.00	381.65
28-Oct-13	377.15	381.80	374.10	371.60
25-Oct-13	378.10	379.40	372.60	376.25
24-Oct-13	377.10	384.70	374.40	379.30
23-Oct-13	381.15	382.35	368.10	374.65
22-Oct-13	380.40	385.10	376.10	380.00
21-Oct-13	380.00	391.50	378.65	380.05
18-Oct-13	376.55	381.80	374.50	379.65
17-Oct-13	385.45	385.45	370.60	373.40
15-Oct-13	390.85	393.00	382.30	388.85
14-Oct-13	386.10	393.30	385.70	390.45
11-Oct-13	375.00	389.60	368.55	385.30
10-Oct-13	354.00	373.85	352.10	371.65
9-Oct-13	350.80	356.35	350.00	354.20
8-Oct-13	352.20	359.40	349.80	350.55
7-Oct-13	350.00	352.80	343.50	347.90
4-Oct-13	345.00	364.80	344.40	350.05
3-Oct-13	335.70	348.35	334.10	345.95
1-Oct-13	332.90	339.80	330.30	335.65
30-Sep-13	338.10	339.50	330.00	332.50
27-Sep-13	343.00	346.20	338.35	340.10
26-Sep-13	342.30	347.30	341.00	343.55
25-Sep-13	339.10	347.45	339.10	344.10
24-Sep-13	334.50	342.25	334.15	337.10
23-Sep-13	337.65	341.35	331.10	333.85
20-Sep-13	344.90	354.85	332.40	338.35
19-Sep-13	344.80	351.55	341.40	349.15
18-Sep-13	341.00	341.00	332.95	336.85
17-Sep-13	329.00	337.65	328.25	335.05
16-Sep-13	339.00	340.80	324.85	331.35
13-Sep-13	332.00	338.30	330.10	334.05
12-Sep-13	341.90	342.00	329.50	333.45
11-Sep-13	348.50	350.55	336.20	340.45
10-Sep-13	320.10	353.35	320.10	349.60
6-Sep-13	318.30	320.00	312.20	317.80
5-Sep-13	316.05	324.35	314.65	318.10
4-Sep-13	302.40	313.75	300.25	311.70
3-Sep-13	302.00	311.35	295.65	297.35
2-Sep-13	296.00	302.00	292.05	299.70
30-Aug-13	308.05	308.70	293.50	299.25
29-Aug-13	299.50	309.00	295.30	306.00
28-Aug-13	290.05	299.50	286.30	297.95
27-Aug-13	295.00	295.50	287.75	290.05
26-Aug-13	301.50	304.30	293.50	298.10
23-Aug-13	295.00	304.60	294.10	300.95
22-Aug-13	284.00	294.90	284.00	291.95
21-Aug-13	290.80	294.60	278.55	283.35
20-Aug-13	297.00	298.40	285.60	287.95
19-Aug-13	307.20	308.75	297.40	301.70
16-Aug-13	317.50	323.40	311.35	313.95
14-Aug-13	298.25	324.40	298.20	319.10
13-Aug-13	280.90	294.00	276.05	290.85
12-Aug-13	282.00	285.40	278.25	281.30
8-Aug-13	285.50	293.30	275.30	278.80
7-Aug-13	284.10	286.75	271.80	278.90
6-Aug-13	284.45	288.90	280.40	287.55
5-Aug-13	290.25	302.00	279.20	284.75

**Figure 2: NSE Sensex datafor past 3 months**

**STEP 2: SUBTRACT THE MEAN**

For PCA to work properly, we need to subtract the mean from each of the data dimensions. The mean subtracted is the average across each dimension.

**STEP 3: CALCULATE COVARIANCE MATRIX**

This is done in exactly the same way as was discussed in earlier section. Since the data is 4 dimensional, the covariance matrix will be 4 X 4 as it produces the result as below :

PCA on the 61-by-4 data matrix X, and returns the principal component coefficients, also known as loadings. Rows of X correspond to observations, columns to variables. Covariance is 8-by-4 matrix, each column containing coefficients for one principal component. The columns are in order of decreasing component variance.

COEFF =

0.4958	0.6978	0.2774	0.4365
0.4921	-0.3409	0.6696	-0.4397
0.5104	0.2279	-0.6204	-0.5502
0.5015	-0.5873	-0.2998	0.5601
0.4958	0.6978	0.2774	0.4363
0.4921	-0.3409	0.6696	-0.4397
0.5104	0.2279	-0.6204	-0.5502

**STEP 4: CALCULATE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX**

Latent is a vector containing the eigenvalues of the covariance matrix i.e. COEFF.

4.8756 X 1.0e+03
0.0315 X 1.0e+03
0.0107 X 1.0e+03
0.0032 X 1.0e+03
4.8756 X 1.0e+03
0.0315 X 1.0e+03
0.0107 X 1.0e+03
0.0032 X 1.0e+03

**STEP 5: TO CALCULATE THE CUMULATIVE SUM OF THE VARIANCES**

0.9908
0.9972
0.9993
1.0000
0.9908
0.9993
0.9972
1.0000

**STEP 6: THE TARGET REDUCED DATA SET**

0.4958	0.6978	0.2774	0.4363
0.4921	-0.3409	0.6696	-0.4397
0.5104	0.2279	-0.6204	-0.5502
0.5015	-0.5873	-0.2998	0.5601

**STEP7:** biplot(coefs) creates a biplot of the coefficients in the matrix coefs. The biplot is 2-D if coefs has two columns or 3-D if it has three columns. coefs usually contains principal component coefficients created with princomp, pcacov, or factor loadings estimated with factoran. The axes in the biplot represent the principal components or latent factors (columns of coefs), and the observed variables (rows of coefs) are represented as vectors. A biplot allows you to visualize the magnitude and sign of each variable's contribution to the first two or three principal components, and how each observation is represented in terms of those components. Biplot imposes a sign convention, forcing the element with largest magnitude in each column of coefs to be positive. This flips some of the vectors in coefs to the opposite direction, but often makes the plot easier to read. Interpretation of the plot is unaffected, because changing the sign of a coefficient vector does not change its meaning.

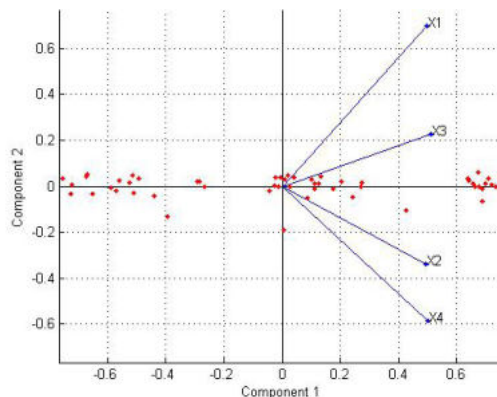
'Scores' : Scores in the matrix are the scores in the biplot. Scores usually contains principal component scores created with princomp or factor scores estimated with factoran. Each observation (row of scores) is represented as a point in the biplot. The scores are the data formed by transforming the original data into the space of the principal components. The values of the vector latent are the variance of the columns of SCORE. Hotelling's T2 is a measure of the multivariate distance of each observation from the center of the data set.

X1 = variable holds for Open Price

X2 = variable holds for High Price

X3 = variable holds for Low Price

X4 = variable holds for Close Price



**Figure 3: PCA's biplot for component1 & 2**

**CONCLUSIONS**

Finally this shows that almost 90% of the variance is accounted for by the first two principal components. PCA is useful for finding new, more informative, uncorrelated features; it reduces dimensionality by rejecting low variance features. PCA should be applied on data that have approximately the same scale in each variable.

## REFERENCES

- [1]Principal Component Analysis in Technology Original Research Article CIRP Annals-Manufacturing Technology, Volume 38,Issue 1,1989, Pages 107-109 G. Lorenz | [2] Fast computation of PCA bases of image subspace using its inner-product subspace, Review Article Applied Mathematics and Computation, Volume219, Issue 12,15 February 2013, Pages 6729-6732 E.S.Gopi, P.Palanisamy | [3] Incipient fault detection and diagnosis based on Kullback-Leiber divergence using Principal Component Analysis:Part1 Original Research Article Signal Processing,Volume 94, January 2014, Pages278-287 Jinane Harmouche, Claude Delpha, Demba Diallo | [4]Principal-componentanalysis in combination with case-based reasoning for detecting therapeutically correct and incorrect measurements in continuous glucose monitoring systems Original Research Article Biomedical Signal Processing and Control, Volume 8, Issue 6, November 2013, Pages603-614Yenny Leal, Magda Ruiz, Carol Lorenzo, Jorge Bondia, Luis Mujica, Josep Vehi |