Principal Component Analysis

Rajath Kumar M.P.

July 21, 2015

Contents

1 Preface

This paper distills the knowledge on Principal Component Analysis gained after reading about it for a long time now. The concept has been explained in depth mathematically by taking an example.

2 Introduction

Principal Component Analysis is the oldest and the most widely used statistical multivariate technique which finds a pattern in the data under consideration. It can be further extended towards dimensionality reduction by extracting important features which are necessary and ignoring the ones which aren't. It find's its application in face recognition, compression, neuroscience etc

3 One Dimensional Dataset

3.1 Mean

Consider a random one dimensional dataset X with n number of data.

$$X = 5, 2, 3, 9, 1, 4$$

The mean of X is given by the formula

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Mean gives the center point of the dataset or the average of the entire dataset. Thus the mean, $\bar{X}$ for the considered example dataset will be equal to 4.

Consider another random one dimensional dataset Y

$$Y = 5, 3, 2, 7, 2, 5$$

The mean, $\bar{Y}$ of this dataset is also 4. The mean of both the datasets considered are the same but it is clear that the data are different and also it varies differently. The math concept which defines the variation of data is standard deviation.

3.2 Standard Deviation

Standard Deviation is a measure that is used to study the amount of variation or dispersion of a set of data values with respect to it's mean. For the X dataset, standard deviation is given by the formula

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

$(n-1)$ is Bessel's correction and is applied to yield a better answer.

3.3 Variance

Variance is the square of standard deviation.

$$var = SD^2$$

or

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

3.4 Conclusion

The variance of X is 8 and that of Y is 4. Which infers that the data of X is more varied or spread across than the data of Y.

4 Two Dimensional Dataset

Consider a two dimensional dataset, (X, Y)

Table 1: Two Dimensional Dataset

X	Y
2	1
3	4
5	8
7	7
2	1
1	3

By applying the same concepts as applied in 3, the following results are obtained.
Mean of X, $\bar{X} = 3.3335$ Mean of Y, $\bar{Y} = 4.0$ Variance of X, $X_{var} = 5.0667$ Variance of Y, $Y_{var} = 8.8$

The above discuss about the dataset itself and not about it's variation or relation with another dataset. The two datasets are compared or analyzed with a parameter called as *Covariance*.

4.1 Covariance

Covariance is similar to variance but between two parameters and not with itself. Hence defined as

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

Rewriting the above formula for $cov(X, X)$

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

or

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

which is same as the variance of X.

For the above considered example the covariance is, $cov(X,Y) = 5.6$
The same can be represented in a matrix. This matrix called as the variance-covariance matrix.

4.2   Variance-Covariance Matrix

The matrix is depicted as

$$\begin{pmatrix} cov(X,X) & cov(X,Y) \\ cov(Y,X) & cov(Y,Y) \end{pmatrix}$$

Here,
$cov(X,X) = X_{var}$
$cov(Y,Y) = Y_{var}$
$cov(X,Y) = cov(Y,X)$
Thus, The variance-covariance matrix can be re-written as

$$\begin{pmatrix} X_{var} & cov(X,Y) \\ cov(X,Y) & Y_{var} \end{pmatrix}$$

Substituting the values as calculated before we get the variance-covariance matrix for the example dataset (X,Y) as

$$\begin{pmatrix} 5.0667 & 5.6 \\ 5.6 & 8.8 \end{pmatrix}$$

To extract the principal information content from the datasets or to know which dataset has the most part of information, eigen values and eigen vectors are calculated.

4.3   Eigen Values and Eigen Vectors

Consider a set of vectors $\vec{v1}$, $\vec{v2}$, $\vec{v3}$ ... which contains a special vector $\vec{v_e}$ such that when a transformation is applied the vector does not go any changes but is just scaled.

$$T(\vec{v_e}) = \lambda \times \vec{v_e}$$

3

Here λ is the scaling factor and is defined as the eigen value. The vector $\vec{v_e}$ which remains unchanged is the eigen vector. For the above example (X,Y), Eigen values are

$$\begin{pmatrix} 1.03041503 \\ 12.83625163 \end{pmatrix}$$

and the eigen vectors are

$$\begin{pmatrix} -0.81124219 & 0.58471028 \\ 0.58471028 & 0.81124219 \end{pmatrix}$$

The two eigen vectors obtained are always perpendicular to one another.

4.4   Principal Component

Principal Component is the eigen vector with the highest eigen value of the data under consideration. It can also be understood as the eigen vector which has been scaled more than the other eigen vectors. This forms the basis of PCA.

In the example the second eigen vector becomes the principal component as it has a higher eigen value of 12.83625 thus we write a feature matrix for the data set containing the eigen vectors in order of higest to lowest importance. Thus,

$$\begin{pmatrix} 0.58471028 & -0.81124219 \\ 0.81124219 & 0.58471028 \end{pmatrix}$$

This matrix is defined as feature matrix denoted as $featuremat$.

4.5   Final Transformed Data

The final transformed dataset is denoted as $Findata$. The transpose of $feauturemat$ is taken so that the highest eigen vector comes to the first row and the next highest to the second row and so on. Now $featuremat$ becomes

$$\begin{pmatrix} 0.58471028 & 0.81124219 \\ -0.81124219 & 0.58471028 \end{pmatrix}$$

To make the dataset (X,Y) zero mean, we subtract the dataset with their respective mean as shown in $Table$ 3 i.e. $X - \bar{X}$ and i.e. $Y - \bar{Y}$

The zero mean dataset obtained is represented in a matrix and denoted as $datam$

$$\begin{pmatrix} -1.33333 & -3 \\ -0.33333 & 0 \\ 1.66667 & 4 \\ 3.66667 & 3 \\ -1.33333 & -3 \\ -2.33333 & -1 \end{pmatrix}$$

Table 2: Dataset (X,Y)		
X	Y
2	1
3	4
5	8
7	7
2	1
1	3

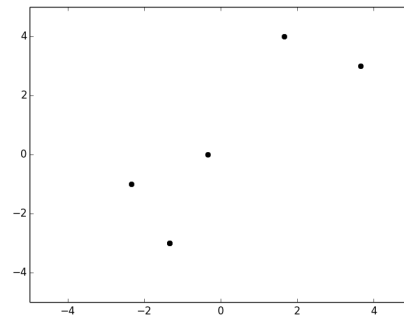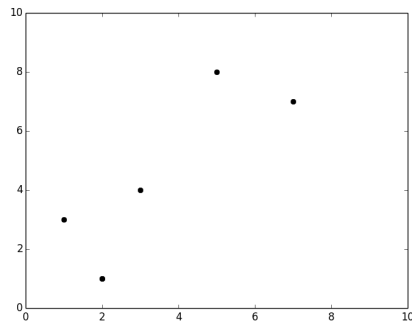| Table 3: Dataset(X,Y) - Mean | | |
|:---:|:---:|
| X | Y |
| -1.33333 | -3 |
| -0.33333 | 0 |
| 1.66667 | 4 |
| 3.66667 | 3 |
| -1.33333 | -3 |
| -2.33333 | -1 |

Figure 1: Plot of Dataset(X,Y)

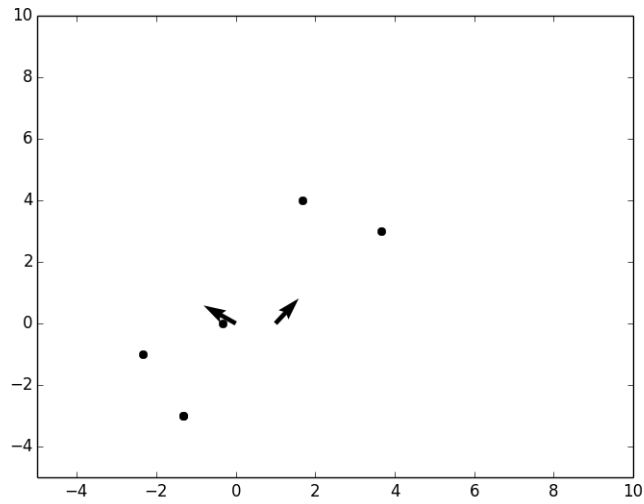Figure 2: Plot of Dataset(X,Y) - Mean

Figure 3: Plot of Dataset with it's Eigen Vectors

The transpose of the *datam* is taken only if the first column contains the first dataset and second in the second dataset so when transpose is taken first row will have the first dataset, second row the second dataset and so on. Hence taking transpose of *datam*

$$\begin{pmatrix} -1.33333 & -0.33333 & 1.66667 & 3.66667 & -1.33333 & -2.33333 \\ -3 & 0 & 4 & 3 & -3 & -1 \end{pmatrix}$$

The final transformed data is defined as,

$$Findata = featuremat \times datam$$

Thus we get the final transformed data. In this case $Findata$ is

$$\begin{pmatrix} -3.2133 & -0.1949 & 4.219 & 4.5777 & -3.2133 & -2.1756 \\ -0.6725 & 0.2704 & 0.9868 & -1.2204 & -0.6725 & 1.3082 \end{pmatrix}$$
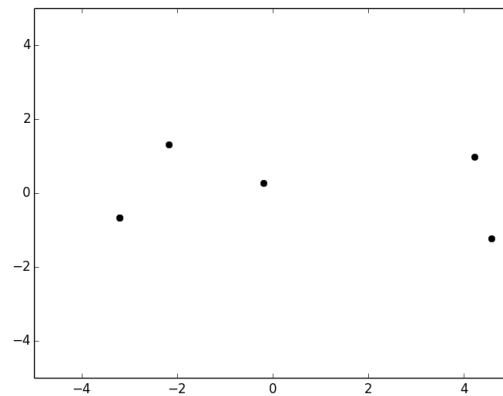
This is final data obtained.

Figure 4: Plot of Final Transformed Data

4.6   Dimensionality Reduction

To reduce a dimension that is to make a two dimensional data into one dimensional data, The highest eigen vector in the feature matrix is considered and the other eigen vector is ignored and steps are continued by considering only one eigen vector thus the $Findata$ obtained when only the highest eigen vector is considered,

$$\begin{pmatrix} -3.2133 & -0.1949 & 4.219 & 4.5777 & -3.2133 & -2.1756 \end{pmatrix}$$

This is otherwise called as data reduction or compression.

4.7 Conclusion

By comparing the plots 3 and 4 we can conclude that upon principal component analysis the data set taken into consideration will change its axis into the eigen vector with the highest eigen value this can be observed from the plots. Hence highlighting the highest information content along the axes. Further upon dimensionality reduction the two dimensional dataset changes to one dimensional dataset as only one eigen vector was taken and when plotted the data will be on just one axes or on the eigen vector with the highest eigen value.

5 Multi Dimensional Dataset

For Multi dimensional data say $(A, B, C, ...)$. Everything is similar to two dimensional dataset but while considering eigenvectors to obtain the final transformed data one has to consider the vectors according to the problem statement or based on the desired reduction in data.

6 Reconstruction

To obtain back the original data, $datam$ from the final transformed data the equation

$$Findata = featuremat \times datam$$

can be rewritten as

$$datam = featuremat^{-1} \times Findata$$

$feauturemat$ contains the eigen vectors. If these vectors are unit vectors then the inverse of the matrix is nothing but the transpose of the matrix. But if it is not then the above statement is not true. But in most of the cases the output of most of the math libraries are unit eigen vectors. Now the equation becomes

$$datam = featuremat^{T} \times Findata$$

But $datam$ is zero mean data i.e. mean subtracted data. Hence,

$$datam = Originaldata - Mean$$

Substituting this in the previous equation

$$Originaldata - Mean = featuremat^{-1} \times Findata$$

Thus $Originaldata$ is expressed as

$$Originaldata = (featuremat^{-1} \times Findata) + Mean$$

7

7   References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[2] Rasmus Bro and Age K Smilde. Principal component analysis. *Analytical Methods*, 6(9):2812–2831, 2014.

[3] Andreas Daffertshofer, Claudine J C Lamoth, Onno G Meijer, and Peter J Beek. Pca in studying coordination and variability: a tutorial. *Clin Biomech (Bristol, Avon)*, 19(4):415–28, May 2004.

[4] James N. Damon and J. S. Marron. Backwards principal component analysis and principal nested relations. *Journal of Mathematical Imaging and Vision*, 50(1-2):107–114, 2014.

[5] Tauseef Gulrez and Abdullah Al-Odienat. A new perspective on principal component analysis using inverse covariance. *Int. Arab J. Inf. Technol.*, 12(1):104–109, 2015.

[6] Lassi Hyvärinen. Principal component analysis. In *Mathematical Modeling for Industrial Processes*, pages 82–104. Springer, 1970.

[7] Abdulrahman Oladipupo Ibraheem. Correlation of data reconstruction error and shrinkages in pair-wise distances under principal component analysis (PCA). *CoRR*, abs/1412.6752, 2014.

[8] Nojun Kwak. Principal component analysis by l_{p} -norm maximization. *IEEE T. Cybernetics*, 44(5):594–609, 2014.

[9] Zhihui Lai, Yong Xu, Qingcai Chen, Jian Yang, and David Zhang. Multilinear sparse principal component analysis. *IEEE Trans. Neural Netw. Learning Syst.*, 25(10):1942–1950, 2014.

[10] Shankar Narasimhan and Nirav Bhatt. Deconstructing principal component analysis using a data reconciliation perspective. *CoRR*, abs/1505.00314, 2015.

[11] Alvin C Rencher. Principal component analysis. *Methods of Multivariate Analysis, Second Edition*, pages 380–407, 2002.

[12] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52, 1987.