e-PGPathshala
Subject : Computer Science
Paper: Data Analytics
Module No 15: CS/DA/15 - Data Analysis Foundations – Principal Component Analysis (PCA)
Quadrant 1 – e-text

1.1 Introduction

Sometimes data are collected on a large number of variables from a single population. With a large number of variables, there would be too many pairwise correlations between the variables to consider. Graphical display of data may also not be of particular help in case the data set is very large. To interpret the data in a more meaningful form, it is therefore necessary to reduce the number of variables to a few, interpretable linear combinations of the data. Each linear combination will correspond to a principal component. Principal Component Analysis is a linear transformation method. PCA yields the directions (principal components) that maximize the variance of the data. This chapter gives a general view of PCA.

1.2 Learning Outcomes

- Learn how to compute principle components
- Understand the applications of PCA in analytics

1.3 Principal Component Analysis

PCA is a linear transformation method. PCA yields the directions (principal components) that maximize the variance of the data. In other words, PCA projects the entire dataset onto a different feature (sub) space. Often, the desired goal is to reduce the dimensions of a d-dimensional dataset by projecting it onto a (k)-dimensional subspace (where k<d) in order to increase the computational efficiency while retaining most of the information.

1.4 PCA Approach

I. Given N data vectors from k-dimensions, find c <= k orthogonal vectors that can be best used to represent data

II. The original data set is reduced to one consisting of N data vectors on c principal components (reduced dimensions)

III. Each data vector is a linear combination of the c principal component vectors

IV. Works for numeric data only

V. Used when the number of dimensions is large

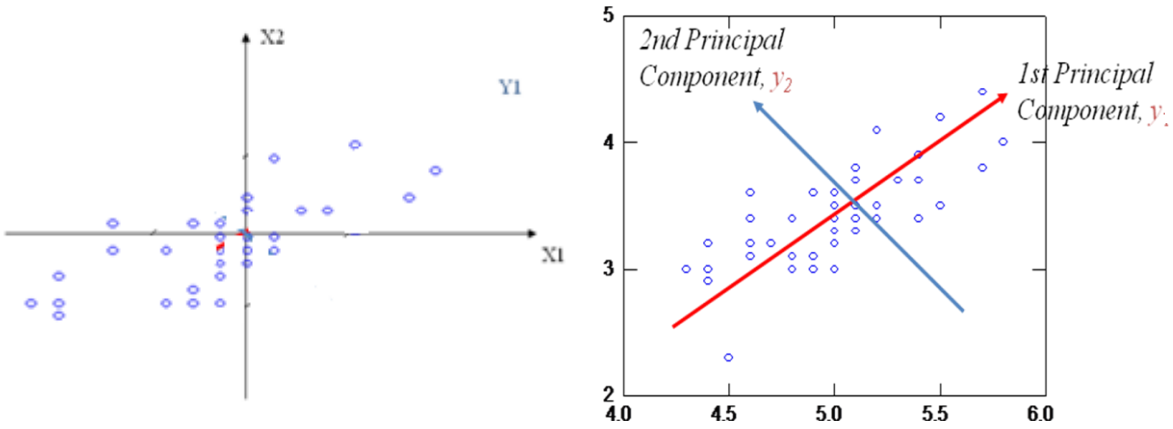


Figure 1 (a) Data with 2 variables X_1 and X_2 (b) Same data with 2 new variables Y_1 and Y2

From *k* original variables: $x_1,x_2,...,x_k$, (as shown in figure 1 (a) and (b)) produce *k* new variables: $y_1,y_2,...,y_k$:

$y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1k}x_k$

$y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2k}x_k$

...

$y_k = a_{k1}x_1 + a_{k2}x_2 + ... + a_{kk}x_k$

such that:

y_k's are uncorrelated (orthogonal) Principal Components

y_1 explains as much as possible of original variance in data set

y_2 explains as much as possible of remaining variance

etc.

and

$\{a_{11},a_{12},...,a_{1k}\}$ is 1st **Eigenvector** of correlation/covariance matrix, and **coefficients** of first principal component

$\{a_{21},a_{22},...,a_{2k}\}$ is 2nd **Eigenvector** of correlation/covariance matrix, and **coefficients** of 2nd principal component

…

$\{a_{k1},a_{k2},...,a_{kk}\}$ is *k*th **Eigenvector** of correlation/covariance matrix, and **coefficients** of *k*th principal component

In summary, PCA rotates multivariate dataset into a new configuration which is easier to interpret for two major reasons:

– simplify data

– look at relationships between variables

– look at patterns of units

1.5 Fundamentals

To get to PCA, we're going to quickly define some basic statistical ideas – *mean, standard deviation, variance* and *covariance* – so we can weave them together later. Their equations are closely related.

Mean is simply the average value of all *x*'s in the set X, which is found by dividing the sum of all data points by the number of data points, *n*.

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

Variance (s^2) is the measure of the data's spread.

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

Standard deviation, is simply the square root of the average square distance of data points to the mean. In the equation below, the numerator contains the sum of the differences between each data point and the mean, and the denominator is simply the number of data points (minus one), producing the average distance.

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

Variance is the spread, or the amount of difference that data expresses.

Covariance (*cov(X,Y)*) is the joint variability between two random variables X and Y, and covariance is always measured between 2 or more dimensions. If you calculate the covariance between one dimension and itself, you get the variance.

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

For both variance and standard deviation, squaring the differences between data points and the mean makes them positive, so that values above and below the mean don't cancel each other out.

Input to various regression techniques can be in the form of correlation or covariance matrix. Covariance Matrix is a **matrix** whose element in the i, j position is the **covariance** between the i th and j th elements of a random vector. A random vector is a random variable with multiple dimensions.

Correlation Matrix is a table showing **correlation** coefficients between sets of variables.

1.5.1 Covariance Matrix vs Correlation Matrix:

Covariance Matrix	Correlation Matrix
• Variables must be in same units • Emphasizes variables with most variance • Mean eigenvalue ≠1.0	• Variables are standardized (mean 0.0, SD 1.0) • Variables can be in different units • All variables have same impact on analysis • Mean eigenvalue = 1.0

1.6 PCA steps

1. Consider a Data
2. Subtract the mean - from each of the data dimensions
3. Calculate the covariance matrix
4. Calculate the eigenvalues and eigenvectors of the covariance matrix
5. Reduce dimensionality and form feature vector
    1. order the eigenvectors by eigenvalues, highest to lowest. This gives you the components in order of significance and *ignore* the components of lesser significance
    2. Feature Vector = (eig1 eig2 eig3 … eign)
6. Deriving the new data:
   FinalData = RowFeatureVector x RowZeroMeanData
    1. RowFeatureVector is the matrix with the eigenvectors in the rows, with the most significant eigenvector at the top
    2. RowZeroMeanData is the mean-adjusted data transposed, ie. the data items are in each column, with each row holding a separate dimension

1.6.1 PCA steps explained with an examples

Step 1: Get some data
Consider a data with just 2 dimensions, and its 2D plots of the data to show what the PCA analysis is doing at each step. The data used is found in Figure 2 (a), along with a plot of that data in Figure 2 (b).

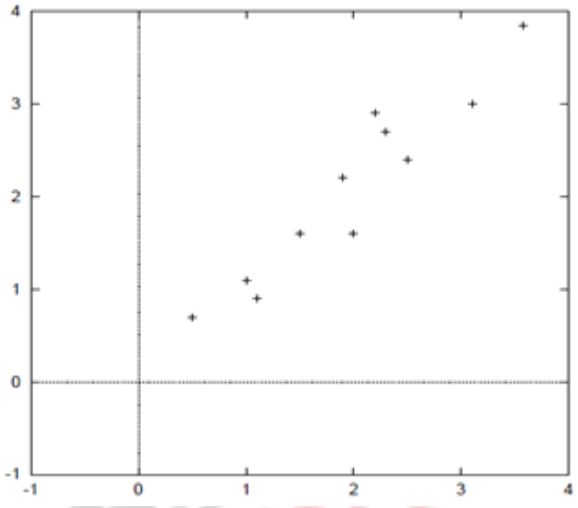
Step 2: Subtract the mean

For PCA to work properly, you have to subtract the mean from each of the data dimensions. The mean subtracted is the average across each dimension. So, all the x values have $\bar{x}$ (the mean of the < values of all the data points) subtracted, and all the y values have $\bar{y}$ subtracted from them. This produces a data set whose mean is zero as shown in figure 2 (c).

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

(a)



(b)

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

(c)

Figure 2. PCA Steps 1 and 2

Step 3: Calculate the covariance matrix

This is done in exactly the same way as was discussed in section 2.1.4. Since the data is 2 dimensional, the covariance matrix will be 2 x 2.as given below:

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

So, since the non-diagonal elements in this covariance matrix are positive, we should expect 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, we can calculate the eigenvectors and eigenvalues for this matrix. These are rather important, as they tell us useful information about our data. The eigenvectors and eigenvalues are given below:

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

It is important to notice that these eigenvectors are both unit eigenvectors ie. their lengths are both 1. So, by this process of taking the eigenvectors of the covariance matrix, we have been able to extract lines that characterise the data. The rest of the steps involve transforming the data so that it is expressed in terms of them lines.

Step 5: Choosing components and forming a feature vector

Here is where the notion of data compression and reduced dimensionality comes into it. If you look at the eigenvectors and eigenvalues from the previous step, you 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.

In our example, the eigenvector with the larges eigenvalue was the one that pointed down the middle of the data. It is the most significant relationship between the data dimensions. 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. To be precise, if you originally have n dimensions in your data, and so you calculate n eigenvectors and eigenvalues, and then you choose only the first p eigenvectors, then the final data set has only p dimensions. What needs to be done now is you need to form a *feature vector:*

$$FeatureVector = (eig_1\ eig_2\ eig_3\ ....\ eig_n)$$

Given our example set of data, and the fact that we have 2 eigenvectors, we have two choices. We can either form a feature vector with both of the eigenvectors or we can choose to leave out the smaller, less significant component and only have asingle column:

$$\begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix} \quad \textbf{(OR)} \quad \begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

Step 6: Deriving the new data set

This the final step in PCA, and is also the easiest. Once we have chosen the components (eigenvectors) that we wish to keep in our data and formed a feature vector, we simply take the transpose of the vector and multiply it on the left of the original data set, transposed.

$$FinalData = RowFeatureVector \times RowDataAdjust$$

where *RowFeatureVector* 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 *RowDataAdjust* is the mean-adjusted data *transposed*, ie. the data items are in each column, with each row holding a separate dimension. FinalData is the final data set, with data items in columns, and dimensions along rows.

Our original data set had two axes, x and y, so our data was in terms of them. It is possible to express data in terms of any two axes that you like. If these axes are

perpendicular, then the expression is the most efficient. This was why it was important that eigenvectors are always perpendicular to each other. We have changed our data from being in terms of the axes x and y, and now they are in terms of our 2 eigenvectors.

In the case of keeping both eigenvectors for the transformation, we get the data and the plot found in Figure 3. This plot is basically the original data, rotated so that the eigenvectors are the axes. This is understandable since we have lost no information in this decomposition.

Transformed Data=

x	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

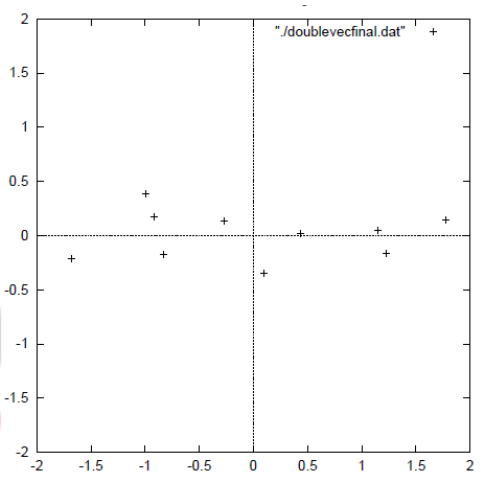


Figure 3. Transformed Data and New Plot

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.

1.6 Interpretation of PCs

Consider Places Rated Almanac data (Boyer and Savageau) which rates 329 communities according to nine criteria:

- Climate and Terrain (C1)
- Housing (C2)
- Health Care & Environment (C3)
- Crime (C4)
- Transportation (C5)
- Education (C6)
- The Arts (C7)
- Recreation (C8)
- Economics (C9)

Step 1: Examine the eigenvalues to determine how many principal components should be considered.

- If you take all of these eigenvalues and add them up and you get the total variance of 0.5223.
- The proportion of variation explained by each eigenvalue is given in the third column. For example, 0.3775 divided by the 0.5223 equals 0.7227, or, about 72% of the variation is explained by this first eigenvalue.
- The cumulative percentage explained is obtained by adding the successive proportions of variation explained to obtain the running total. For instance, 0.7227 plus 0.0977 equals 0.8204, and so forth. Therefore, about 82% of the variation is explained by the first two eigenvalues together.
- Next we need to look at successive differences between the eigenvalues. Subtracting the second eigenvalue 0.051 from the first eigenvalue, 0.377 we get a difference of 0.326. The difference between the second and third eigenvalues is 0.0232; the next difference is 0.0049.
- A sharp drop from one eigenvalue to the next may serve as another indicator of how many eigenvalues to consider.
- The first three principal components explain 87% of the variation. This is an acceptably large percentage.

Table 1. Eigenvalues, and the proportion of variation explained by the principal components

Component	Eigenvalue	Proportion	Cumulative
1	0.3775	0.7227	0.7227
2	0.0511	0.0977	0.8204
3	0.0279	0.0535	0.8739
4	0.0230	0.0440	0.9178
5	0.0168	0.0321	0.9500
6	0.0120	0.0229	0.9728
7	0.0085	0.0162	0.9890
8	0.0039	0.0075	0.9966
9	0.0018	0.0034	1.0000
Total	0.5225		

Step 2: Next, compute the principal component scores:

- For example, the first principal component can be computed using the elements of the first eigenvector:

$$\begin{aligned} Y_1 = {} & 0.0351 \times (\text{climate}) + 0.0933 \times (\text{housing}) + 0.4078 \times (\text{health}) \\ & + 0.1004 \times (\text{crime}) + 0.1501 \times (\text{transportation}) + 0.0321 \times (\text{education}) \\ & 0.8743 \times (\text{arts}) + 0.1590 \times (\text{recreation}) + 0.0195 \times (\text{economy}) \end{aligned}$$

Step 3: To interpret each component, compute the correlations between the original data for each variable and each principal component.

First Principal Component Analysis - PCA1

The first principal component is a measure of the quality of Health and the Arts, and to some extent Housing, Transportation and Recreation. Health increases with increasing values in the Arts. If any of these variables goes up, so do the remaining ones. They are all positively related as they all have positive signs.

Second Principal Component Analysis - PCA2

The second principal component is a measure of the severity of crime, the quality of the economy, and the lack of quality in education. Crime and Economy increase with decreasing Education. Here we can see that cities with high levels of crime and good economies also tend to have poor educational systems.

	Principal Component		
Variable	**1**	**2**	**3**
Climate	0.190	0.017	0.207
Housing	**0.544**	0.020	0.204
Health	**0.782**	**-0.605**	0.144
Crime	0.365	0.294	**0.585**
Transportation	**0.585**	0.085	0.234
Education	0.394	-0.273	0.027
Arts	**0.985**	0.126	-0.111
Recreation	**0.520**	0.402	**0.519**
Economy	0.142	0.150	0.239

Third Principal Component Analysis - PCA3

The third principal component is a measure of the quality of the climate and poorness of the economy. Climate increases with decreasing Economy. The inclusion of economy within this component will add a bit of redundancy within our results. This component is primarily a measure of climate, and to a lesser extent the economy.

Summary

- PCA is a dimensionality reduction technique for very high dimensionality data.
- One benefit of PCA is that we can examine the variances associated with the principle components.
- Used for identifying key dimensions for analysis.