**PRINCIPAL COMPONENT ANALYSIS (PCA)**

Principal component analysis (PCA) is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

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 analysing data.

**PROCESS:**

**Step 1:** Consider the sample dataset and analyse the dataset.

**Step 2:** Finding the mean of all the dimensions.

**Step 3:** Subtracting the mean from each of the data dimensions.

This will give the average across each dimension. This will produce a dataset with mean Zero. This we are calling it as zero-mean data.

**Step 4:** Calculating the Covariance Matrix from the zero-mean data.

Covariance matrix can be found out by multiplying the zero-mean data and the transpose of the zero-mean transpose data and then dividing this whole matrix by the (m-1) where ‘m’ is the number of the samples. So if we consider the zero-mean data as ‘A’, then the Covariance of n dimension matrix can be found by (A\*AT ) / (m-1). Covariance matrix gives the n\*n matrix where n is the number of dimensions.

http://www0.gsb.columbia.edu/premba/analytical/images/s7/8290310392.gif

*x* = the independent variable  
 *y* = the dependent variable  
 *n* = number of data points in the sample  
 http://www0.gsb.columbia.edu/premba/analytical/images/s7/3661302035.gif = the mean of the independent variable *x*  
http://www0.gsb.columbia.edu/premba/analytical/images/s7/7251432727.gif = the mean of the dependent variable *y*

**Step 5:** Finding the Eigenvectors and Eigenvalues of the Covariance Matrix.

So first we find the Eigen values and this Eigen values are used to find the Eigen Vectors. Eigen vectors are unit vectors and they are perpendicular to each other. We make them unit vectors to make all the vectors same in length i.e to 1. Eigen vectors will provide us with the information about the patterns in the data.

**Step 6:** Choosing the Principal Components and Reduced Vector.

Eigenvector with the highest Eigen value is the Principal component of the data set. So we arrange the eigen values in the decreasing order. Principal component gives the significant relationship between the data dimensions. We can find the percentage of the variance of data present in the direction of Eigen vector of that particular Eigen value. This percentage is calculated by dividing the particular Eigen value by the sum of all the Eigen values and multiplied by 100. We can ignore the components of lesser significance. We may lose some information but if the eigen values are small, we don’t lose much information. So we are forming the ‘Reduced Vector’ which is constructed by considering the eigen vectors that we want to keep from a list of eigen vectors or considering the eigen vectors which of high significance i.e we are not considering all the vectors. So leaving the lesser significant components will give the dataset with less dimensions than the original. So here we are reducing the dimensions without much information loss. So what we are doing is if we have ‘n’ dimensions in our data and we will get ‘n’ eigen values and eigen vectors and choosing only the first ‘m’ eigen vectors of higher significance (i.e higher eigen values) will produce a final dataset which has only ‘m’ dimensions.

**Step 7:** Deriving the new Dataset.

After choosing the components (eigenvectors) that we wish to keep in our data, called as Reduced Vector, we need to transpose the vector. We need to transpose the Zero-Mean Data also.

Final Data = (Reduced Vector) T \* (Zero-Mean Data) T

So this will give us the original data in terms of vectors we choose.

**Tool we used to analyse the dataset for PCA:** We are using **XLSTAT** tool which is an MS EXCEL add-on to analyse the dataset.

**Dataset we used(5 records as sample):**

|  |  |  |  |
| --- | --- | --- | --- |
| **Days** | **Temp** | **Attendance** | **Month** |
| d1 | 20 | 21 | 1 |
| d2 | 19 | 8 | 1 |
| d3 | 21 | 13 | 2 |
| d4 | 23 | 15 | 2 |
| d5 | 26 | 51 | 3 |

**Table 1:** The above tables illustrates 5 sample records from the original dataset we are using.

The dataset we are using is Akron Zoo dataset and we are considering 1074 records for the PCA analysis. This dataset will determaine the Temperature(F), Attendance of the people on a particular day and in which Month. We are making use of this dataset to represent the relation between different attributes or dimensions.

**Covariance Matrix:**

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Temperature | Attendance | Month |
| Temperature | **405.01293** | 11916.782 | 14.737 |
| Attendance | 11916.782 | **761940.209** | 52.880 |
| Month | 14.737 | 52.880 | **11.60531** |

**Table 2:** The above table gives the covariance matrix for the records we are considering.

After the dataset is obtained, we need to find the mean of all the dimensions and then subtracting the mean from each of the data dimensions gives the zero-mean data. The above mentioned covariance matrix is obtained by multiplying the zero-mean data and the transpose of the zero-mean data and then dividing this whole matrix by the (m-1) where ‘m’ is the number of the samples(we are considering 1074 samples).

**Eigenvalues:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **F1** | **F2** | **F3** |
| **Eigenvalue** | 1.717 | 0.990 | 0.293 |
| **Variability (%)** | 57.230 | 32.992 | 9.778 |
| **Cumulative %** | 57.230 | 90.222 | 100.000 |

**Table 3:** The above table gives the eigenvalues for the 3 factors or dimensions we have considered.

The above figure determines the Eigen values, Variability (%), Cumulative (%) for each of the factors. Variability is obtained by dividing the corresponding eigen value by sum of all the factors eigen values and multiplied by 100. This variability will determine the percentage of the variance of data represented by corresponding factors. So we can see from the cumulative row that about 90.22% of the variance of data is represented by first 2 factors. So there is very much less loss of information. From the below screen plot, we can see Eigen values for the different factors.

**Figure 1:** The above screen plot represents the Eigen values for different factors. X axis represent the factors, Left Y-axis represent the Eigen values and Right Y-Axis represent the cumulative variability(%) with respect to factors.The bars represent the eigen value for each of the individual factor. The red color dots represent the cumulative variability(%). Red color dot on the factor 1 tells us that factor 1 represents about 57.23% of the variance of data. Red dot on factor 2 represent 90.22% of the variance of the data along with factor1. Similarly for other factors.

**Eigen Vectors:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | F1 | F2 | F3 |  |
| Temperature | 0.704 | -0.020 | 0.709 | Temperature |
| Attendance | 0.672 | -0.302 | -0.676 | Attendance |
| Month | 0.228 | 0.953 | -0.199 | Month |

**Table 4:** The above tabular form represents the Eigen vectors for the corresponding factors eigen values. Eigen vectors will provide us with the information about the patterns in the data. So we are considering only first 2 factors eigen values which correspond about 90.22% of variance of data.

**Graph Plots:**

**Figure 2:** The above Bi-Plotted graph represent the variance of the data in a 2D space considering factors F1(X-axis) and F2(Y-axis) axes without much loss of information.

The above graph plot from figure 3 represent the variance of the data in the 2D space considering F1 on X-axis and F2 on Y-axis. We can see that F1 and F2 represent about 90.22% of the variance of the data in a 2D space. We can see the red color lines in the above image which represent the 3 dimensions variation of the data in a 2D space. So here we are reducing the dimensions from 3 dimensions to 2 dimensions without much loss of information.

The direction of the month represents the increase in that direction from bottom to top(for example we considered 1- January(bottom), 12-December(top) etc). The direction of temperature represents the increase from left to right. Similarly for the attendance. So we can conclude from the above Bi-Plot graph that as the temperature is increasing, attendance is also increasing(consider the temperature vs attendance vs month axes(Summer time)). Similarly as the temperature decreases, attendance decreases from the top and bottom cluster data(Winter time).

**Figure 2:** The above observations graph plot represent the 3 dimensions in a 2 D space considering F1 as X-axis and F2 as Y-Axis.

The above figure describes the variation of data in 2D space considering F1 on the X-axis and F2 on the Y-axis. We can see the clusters of data from top to bottom. The cluster data on the top represent the months of December, November, October, September which tells us that when the temperature is less, the attendance is also less(Winter time). The middle cluster of data represent the months of summer where the temperature is high, so the attendance is also high. The bottom cluster of data represent the months of January to April(Winter and Spring) which tells us that temperature is less and the attendance is also less. Overall we can say that when the temperature is high, attendance to the Akron Zoo is also high.

So we can conclude that we are able to represent the variance of data in a lesser dimension space without much loss of information. This is how PCA reduces the dimensions.