Dimensionality Reduction with PCA

To reduce the dimensions of the original dataset using PCA, we multiply the desired number of components or loadings from the eigenvector matrix, A, by the design matrix X. Suppose the design matrix (or the original dataset) has m rows (or observations) and p columns (or features), if we want to reduce the dimensions of the original dataset to two dimensions, we will multiply the original dataset X by the first two columns of the eigenvector matrix, $A_{reduced}$. The result will be a reduced matrix of m rows and 2 columns.

If X is a $m \times p$ matrix and $A_{reduced}$ is a $p \times 2$ matrix,

$$T_{reduced} = X_{m \times p} \times A_{p \times 2}$$

Observe that the result $T_{reduced}$ is a $m \times 2$ matrix. Hence, T is a 2-D representation of the original dataset X as shown in Figure 26-2.

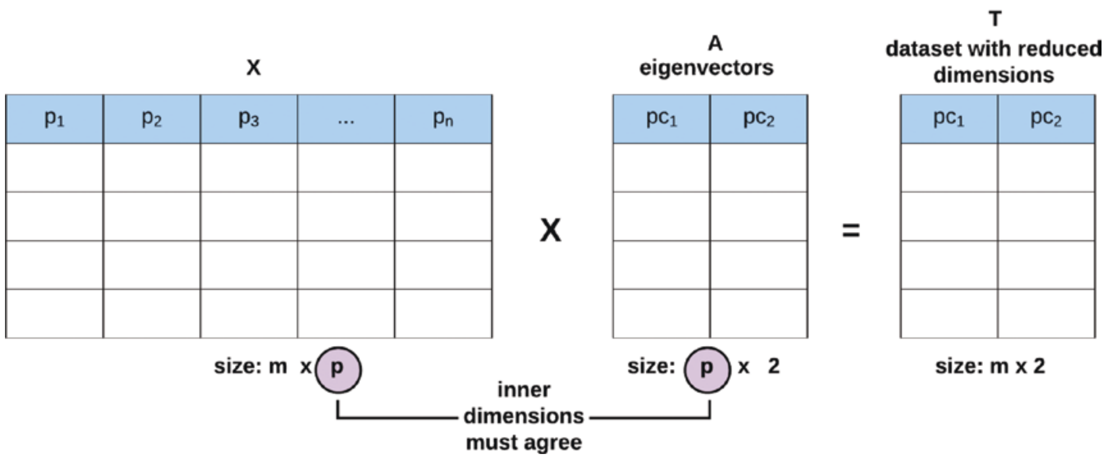

Figure 26-2. *Reducing the dimension of the original dataset*

In plotting the reduced dataset, the principal components are ranked in order of importance with the first principal component more prominent than the second and so on. Figure 26-3 illustrates a plot of the first two principal components.

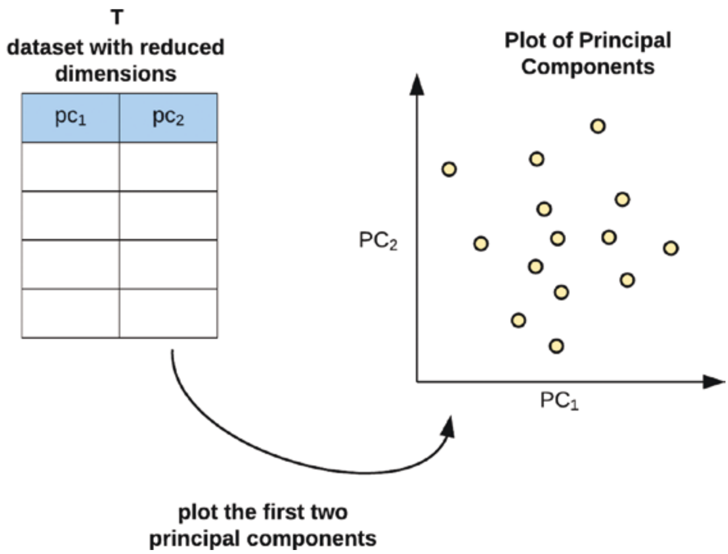

Figure 26-3. *Visualize the principal components*

Key Considerations for Performing PCA

It is vital to perform mean normalization and feature scaling on the variables of features of the original dataset before implementing PCA. This is because unscaled features can have stretched and narrow distance n-dimensional space, and this has a huge consequence when finding the principal components that explain the variance of the dataset (see Figure 26-4).

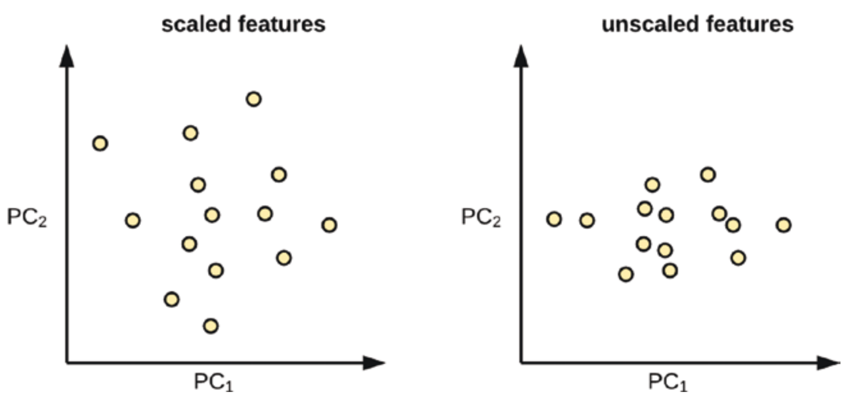

Figure 26-4. *Right: An illustration of PCA with scaled features. Left: An illustration of PCA with unscaled features.*