Principal Component Analysis

Juan David Rodríguez Cuervo

December 17, 2017

The *Principal Component Analysis* is a statistical technique to help reduce complexity in multi-dimensional datasets. It aims at searching the "dimensions" that describe better the dataset. Thus, in reducing complexity, one might neglect the dimension vectors from which the information given is considerably small.

That being said, the concept of the analysis is to find vectors that are a linear combination of the original dimensions of the set in a way such that the information is maximized and where the vectors are orthogonal to the others. In a sense, it can be thought of as rotating the vector space in which the dataset relies where the principal axes give the most information from the set. From that point on, it is on the analyst to decide which dimension vectors to maintain and which to neglect to proceed with other operations.

In this example, the working is done in a 2-dimensional dataset. Let $\mathbf{M} = [\mathbf{X}, \mathbf{Y}]$ be the input dataset matrix of size $[N \times 2]$. Nonetheless, it is important to note that this process is better done with datasets having a higher number of dimensions. Therefore, we start defining a normalized matrix

$$\mathbf{M}' = [\mathbf{X} - \overline{x}, \mathbf{Y} - \overline{y}] \tag{1}$$

Here, $\overline{x}, \overline{y}$ are the arithmetic means of the corresponding columns.

As the general objective is to find the vectors from which the most information can be obtained from the set, the covariance values play a fundamental role in the analysis. As so, a covariance matrix is then obtained as

$$\mathbf{C} = cov\,\mathbf{M}' = \begin{bmatrix} cov(\mathbf{X}, \mathbf{X}) & cov(\mathbf{X}, \mathbf{Y}) \\ cov(\mathbf{Y}, \mathbf{X}) & cov(\mathbf{Y}, \mathbf{Y}) \end{bmatrix} \tag{2}$$

From (2) the eigenvectors are found. For this, it is applied the base equation of

$$\begin{aligned} \mathbf{C}v &= \lambda \mathbf{v} \\ \mathbf{C}v - \lambda \mathbf{v} &= 0 \\ (\mathbf{C}v - \lambda \mathbf{I})v &= 0 \end{aligned} \tag{3}$$

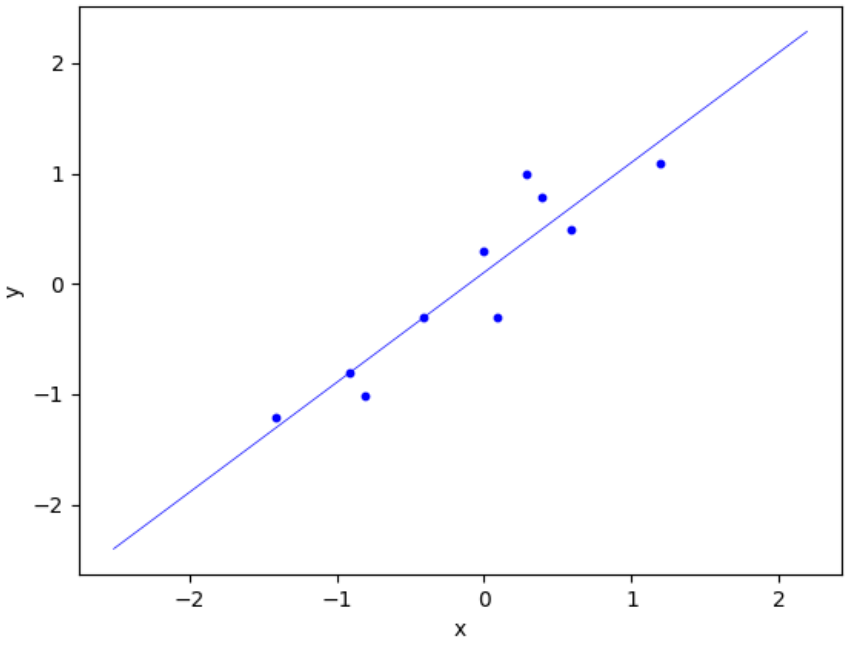
From 3 we obtain the eigenvalues (λ values), and then the eigenvectors (v). The calculations of the eigenvectors are beyond the scope of this document. Thus, if you want to investigate in detail the process, there are many resources explaining it.

After all the process, a matrix of eigenvectors is obtained in the form $\mathbf{V} = [v_1, v_2]$, where v_1, v_2 are the corresponding eigenvectors of the covariance matrix. At last, the new matrix space is obtained by following the product of the initial matrix and the eigenvector matrix as in

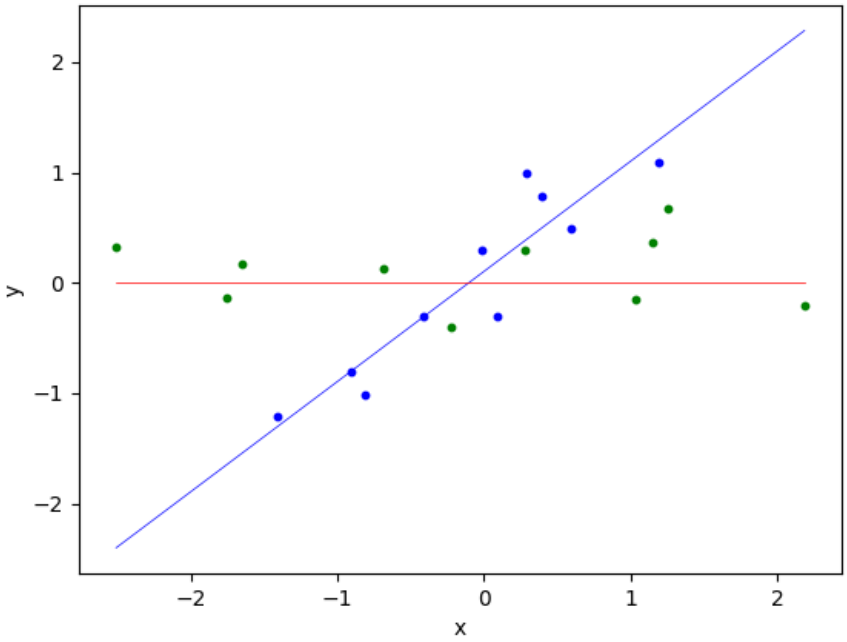
$$\mathbf{R} = \mathbf{M}' \times \mathbf{V} \tag{4}$$

As stated in the beginning, the dimensions can be then neglected as desired. From 4 only the first columns can be used for analysis or visualization. The following graphs with summarize the process visually.

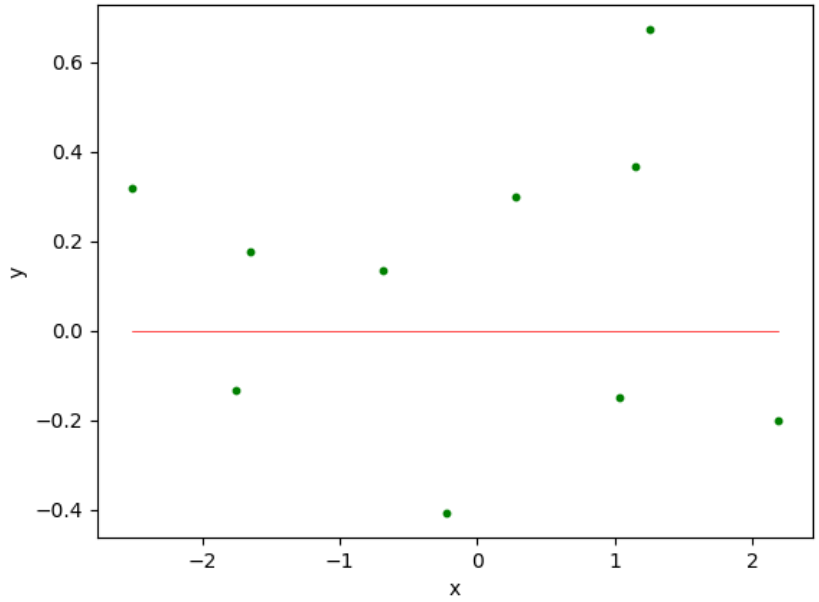
At first, the original data points as shown. In the following graph it is also plotted the line corresponding to the principal component or main vector dimension from which the space is going to be rotated.



Later on, it is shown the process in which the points are rotated in the space. The blue dots are the original points, whilst the green ones are the rotated data. As seen, the new x-axis is the principal component in the original space. This creates a new horizontal axis made from a linear combination of the original data dimensions.



The new space assures that the dispersion is maximum in the horizontal axis. This can be interpreted as the maximum amount of information is gotten from the horizontal axis.



Finally, let's assume the vertical axis is not providing enough information as desired, but it is obscuring the data in general. Thus, a strategy that may help to find other patterns, includes neglecting the vertical axis, and so, projecting the data points over the principal component.