# PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is non-parametric approach to transform a set of observations - of possibly correlated variables - into another set of values defined by a calculated set of new orthogonal uncorrelated variables called principal components.

1   Example

Consider the following set of four observations in two dimesions,
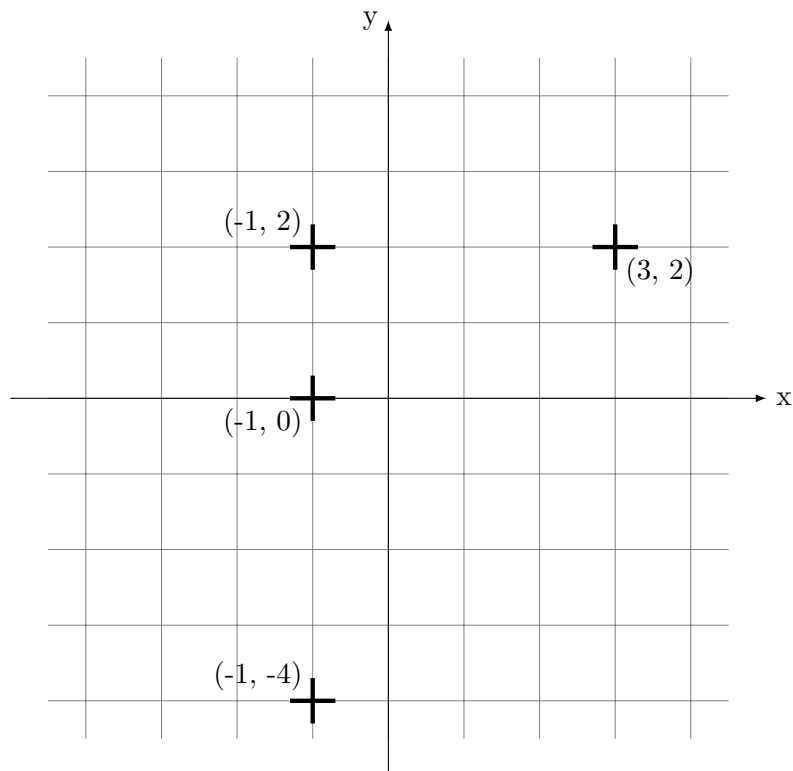
To see how PCA alters the representation of this tiny dataset, the dataset is first written in matrix form as follows,

$$X = \begin{pmatrix} 3 & -1 & -1 & -1 \\ 2 & 0 & 2 & -4 \end{pmatrix}$$

The x-axis intercepts are represented by the first row of this matrix while y-axis intercepts are represented by the second row. This matrix can be thought of as a representation composed of two random variables x and y. The variances of these variables and their covariance are useful quantities to describe the structure of the dataset.

$$Var(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$
$$= \frac{1}{4} \sum_{i=1}^{4} (x_i - 0)^2$$
$$= \frac{1}{4} |x|^2$$
$$= \frac{3^2 + (-1)^2 + (-1)^2 + (-1)^2}{4}$$
$$= 3$$
$$Var(y) = \frac{1}{4} \sum_{i=1}^{4} (y_i - 0)^2$$
$$= \frac{1}{4} |y|^2$$
$$= \frac{2^2 + 0^2 + 2^2 + (-4)^2}{4}$$
$$= 6$$
$$Cov(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$
$$= \frac{1}{4} \sum_{i=1}^{4} (x_i - 0)(y_i - 0)$$
$$= \frac{1}{4} \vec{x} . \vec{y}$$
$$= \frac{3 * 2 + (-1) * 0 + (-1) * 2 + (-1) * (-4)}{4}$$
$$= 2$$

Note that the mean of both the vectors is zero in the dataset. This is a necessary precondition for PCA, the reason for this constraint will become

apparent in later sections. Next, these variances and covariance are arranged in a matrix as follows.

$$\sum = \begin{pmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{pmatrix}$$
$$= \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}$$

Due to zero mean constraint, the following also holds true,

$$\sum = \frac{1}{N} X X^T$$

Now eigenvalues and eigenvectors of $\sum$ are calculated.

$$\sum \phi = \lambda \phi$$
$$(\sum - \lambda I) \phi = 0$$

$(\sum - \lambda I)$ must be singular if eigenvectors are non-zero. Hence,

$$\begin{vmatrix} 3 - \lambda & 2 \\ 2 & 6 - \lambda \end{vmatrix} = 0$$
$$(3 - \lambda)(6 - \lambda) = 4$$

This gives eigenvalues $\lambda = 2, 7$ and eignevectors $\phi = (\frac{-2}{\sqrt{5}}, \frac{1}{\sqrt{5}}), (\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})$ respectively. Now the eigenvectors are arranged as rows in a matrix in order of their corresponding eigenvalue.

$$P = \begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix}$$
$$= \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$$

Now the transformed dataset Y is calculated simply as product of P and X.

$$P = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 3 & -1 & -1 & -1 \\ 2 & 0 & 2 & -4 \end{pmatrix}$$
$$= \frac{1}{\sqrt{5}} \begin{pmatrix} 7 & -1 & 3 & -9 \\ -4 & 2 & 4 & -2 \end{pmatrix}$$

To visualize the result, Y is plotted on a graph,

3
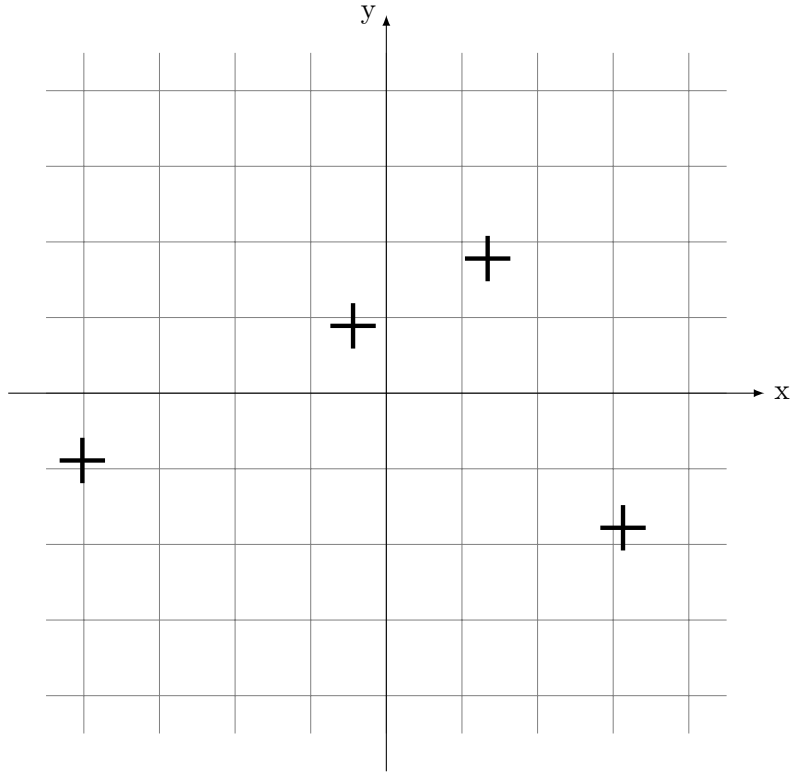
Visually, the points on the transformed graph arise due to a change of axes from the original axes to the axes defined by the eigenvectors of $\sum$. To see why this choice of axes was made, the variance covariance matrix of the transformed data is calculated.

$$\sum_y = \frac{1}{N}YY^T$$

$$= \frac{1}{4}\frac{1}{\sqrt{5}}\begin{pmatrix} 7 & -1 & 3 & -9 \\ -4 & 2 & 4 & -2 \end{pmatrix}\frac{1}{\sqrt{5}}\begin{pmatrix} 7 & -4 \\ -1 & 2 \\ 3 & 4 \\ -9 & -2 \end{pmatrix}$$

$$= \frac{1}{20}\begin{pmatrix} 140 & 0 \\ 0 & 40 \end{pmatrix}$$

$$= \begin{pmatrix} 7 & 0 \\ 0 & 2 \end{pmatrix}$$

Two notes - the variance covariance matrix of transformed data is diagonalized meaning there is no correlation between the new variables and the variance values i.e. the diagonal terms are equal to the eigenvalues.

This is the central point of PCA: it gives you a transformation of variables which are uncorrelated and you are free to choose a subset of these

4

variables depending on the tradeoff between smaller number of dimensions vs loss in variance. In practice, a large chunk of calculated variables in the bottom right part of the diagonal are near zero in their variance and thus are ignored resulting in dimensionality reduction.