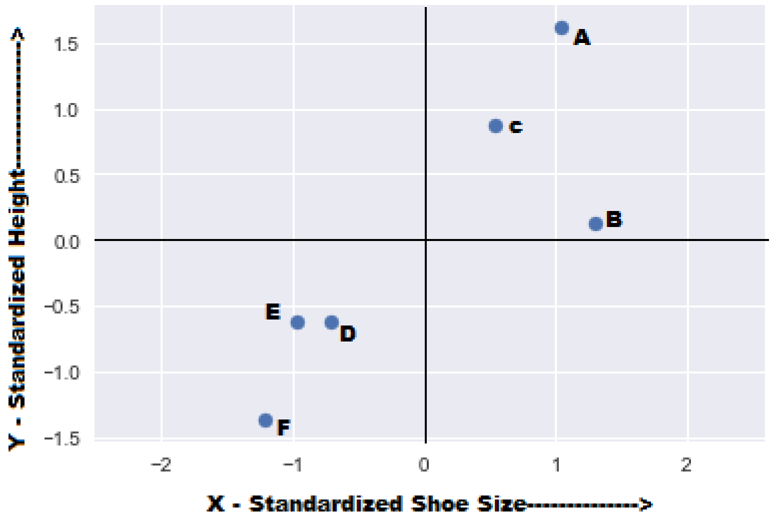
The Relation Between SVD, PCA and Eigenvectors

Note: If you are not familiar with the basic notion of eigenvectors and eigenvalues of a matrix, it is recommended to go through the following videos before reading this article:

- The intuition behind eigenvectors and eigenvalues: 3Blue1Brown
- Finding eigenvectors and eigenvalues - example: Khan Academy
- Eigendecomposition of a matrix: Wikipedia

You have seen that SVD is used to compute the principal components. Though SVD is used in most implementations (including python), there's an alternative way to compute the principal components - using the notion of eigenvectors and the covariance matrix.

Let's consider an example dataset - you have shoe sizes and heights of six people as shown below. We'll assume that the variables are scaled and centered, i.e. the mean of both variables is zero. You want to do a PCA on this (you can see the principal components already, can't you?).



Now, you intuitively know that principal components somehow capture the correlation among the original columns. For e.g., in the plot above, the two variables are correlated, i.e. x and y increase together in the direction of the line y=x, and the first principal component falls somewhere around that line. Let’s explore the notion of correlations (or rather its unscaled version ‘covariance’) in detail.

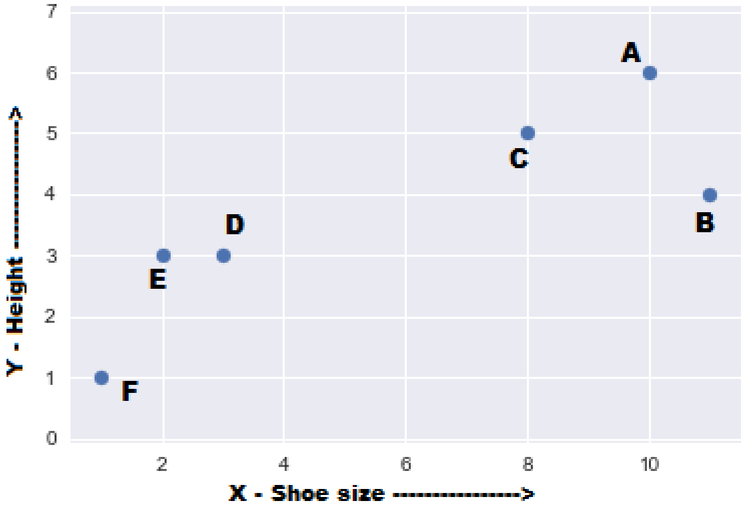
You know that the variance of any column having N data points is given by:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

.

However, variance can only be used to explain the spread of data in the direction of the axes, i.e. shoe size and height. For e.g. we could calculate the variance $\sigma(x,x)$in the X direction (i.e. the variance of the column ‘Shoe size’) and the variance $\sigma(y,y)$ in the Y direction (the variance of the column ‘Height’).

But these individual variances do not explain the *combined variance (or correlation)*, i.e. the direction in which *both* variables move together. The plot above clearly shows that, on average, if the x-value of a data point increases, then the y-value also increases, resulting in a positive correlation. This correlation can be captured by extending the notion of variance to what is called **covariance**. The covariance between x and y is denoted by $\sigma(x,y)$.



For 2-D data, we obtain $\sigma(x,x)$,$\sigma(y,y)$, $\sigma(x,y)$, and $\sigma(y,x)$as follows:

$$\sigma(x,y) = \sigma(y,x) = \frac{1}{N}\sum(x_i - \bar{X})(y_i - \bar{Y})$$

Since we have assumed that the data is centered (i.e. means are 0), then:

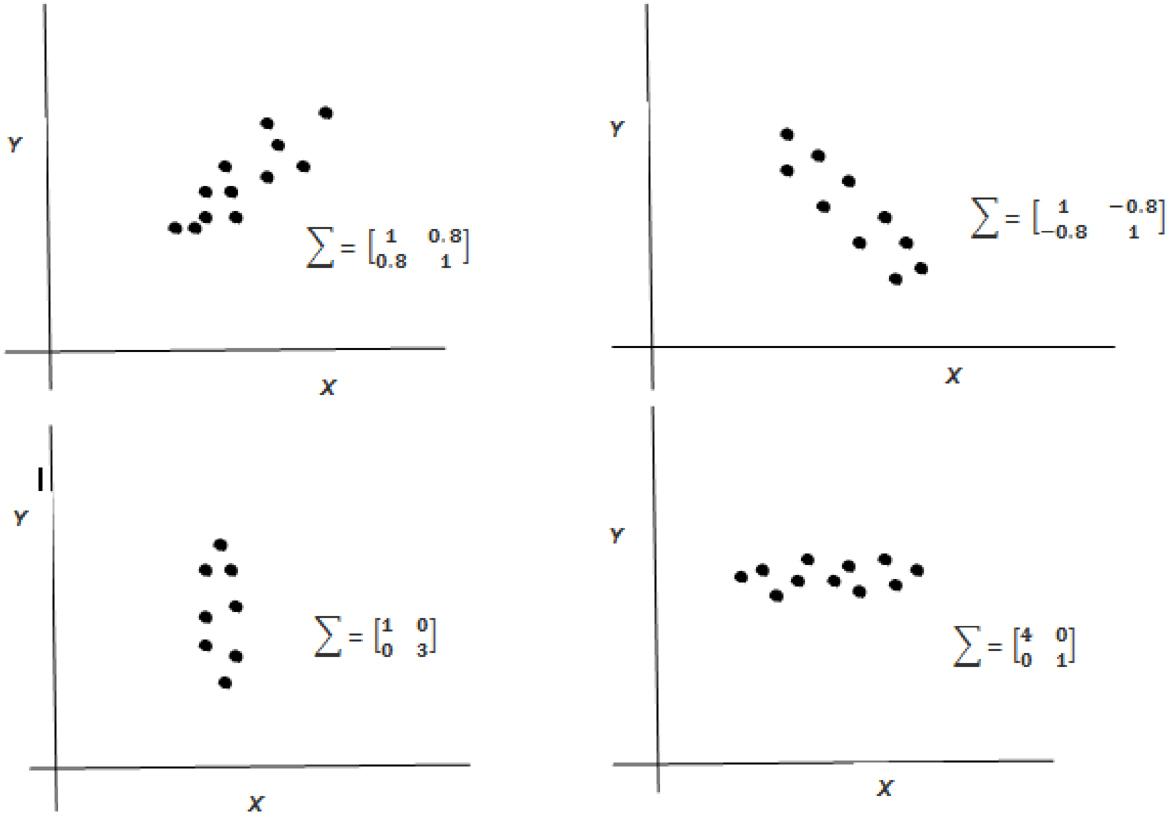
$$\sigma(x,y) = \sigma(y,x) = 1/n \sum \quad x_i y_i$$

These four values can be summarized in a matrix called the **covariance matrix** (note again that this only works if the data are centered):

$$\Sigma = \begin{bmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{bmatrix}$$

If x is correlated with y, then y is also correlated with x. In other words, $\sigma(x,y)$= $\sigma(y,x)$, and thus the covariance matrix is always a **symmetric matrix** with individual variances on its diagonal and the off-diagonal entries as covariances.

The following plots illustrates an example of the covariance matrix for a 2-D setting. In clockwise direction - the first plot shows positive covariance between x and y, the second one shows a negative covariance, the third and the fourth plots shows almost no covariance. The diagonal entries in each plot represent the variances of x and y respectively.



Now, if the original data matrix is A = **n x p,** i.e. n data points and p variables, then it can be shown that the **covariance matrix of the dataset A** will be a **p x p** matrix given by

$$\Sigma = \frac{1}{N} A^T A$$

Where N is number of data points.

Let’s look at a simple example. Consider the following data matrix with N=5 and p=2. Here, the two variables x and y are the scaled ‘shoe size’ and ‘height’ columns —

$$\text{A} = \begin{pmatrix} -2 & -4 \\ -1 & -2 \\ 0 & 0 \\ 1 & 2 \\ 2 & 4 \end{pmatrix}$$

$$A^T = \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ -4 & -2 & 0 & 2 & 4 \end{pmatrix}$$

$$\Sigma = \frac{1}{N} A^T A$$

$$\Sigma = \frac{1}{N} A^T A = \frac{1}{5} \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ -4 & -2 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} -2 & -4 \\ -1 & -2 \\ 0 & 0 \\ 1 & 2 \\ 2 & 4 \end{pmatrix}$$

$$\Sigma = \frac{1}{5} \begin{pmatrix} 10 & 20 \\ 20 & 40 \end{pmatrix} = \begin{pmatrix} 2 & 4 \\ 4 & 8 \end{pmatrix}$$

You can see that the diagonal entries are $\sum \quad x^2/5$and $\sum \quad y^2/5$ while the off-diagonal entries are $\sum \quad x_i y_i/5$.

So what does the covariance matrix have to do with principal components (and eigenvectors)? It turns out that the **eigenvectors of the covariance matrix are the principal components** of the dataset A.

It can be proved that the largest **eigenvector of a covariance matrix** always points to the **direction of the largest variance** in the data. Note that by largest eigenvector, we mean the vector having the largest eigenvalue. The second largest eigenvector is always **orthogonal** to the largest eigenvector and points to the direction of the second largest spread in the data, and so on.

Thus, doing a *PCA of dataset A* essentially boils down to computing the *eigenvectors of the covariance matrix of A*.

More formally, the eigenvectors of any square matrix Σ can be represented as (this is known as the eigen decomposition of the matrix):

$$\Sigma = VLV^T,$$

where V is a matrix of the eigenvectors (i.e. each column being an eigenvector) and L is a diagonal matrix with the corresponding eigenvalues (in decreasing order of magnitude). In our case, this square matrix is the covariance matrix.

Let’s now see how SVD and and the new eigendecomposition approach relate to each other. In SVD, you decompose the original data matrix A as:

$$A = USV^T,$$

Where the matrix V (p x k) represents the relation between themes (PCs) and the original variables, i.e. each of the k columns of V contains a principal component vector. In other words, each column vector of V is a principal component vector, which is an eigenvector of Σ .

Since we have already established that the covariance matrix Σ is nothing but A^TA/n, Σ can be rewritten as

$$\Sigma = 1/N\,A^TA = 1/N\,(USV^T)^{\mathrm{T}}\,(USV^T) = 1/NVS^TU^T\,USV_{\blacksquare}^{T}$$
$$\Sigma = 1/NVS_{\blacksquare}^{2}V_{\blacksquare}^{T}$$
$$\Sigma = VLV^T$$

where $L = S^2/N$, i.e. the diagonal matrix S from SVD basically contains the square root of the eigenvalues l_i:

$$S = \sqrt{Eigenvalues\ of\ covariance\ matrix}$$

To summarize, one can use either SVD or eigenvalue decomposition to find the principal components. So why would most packaged use SVD rather than the eigenvalue method? It turns out understanding the reasons for it involve a bit more complex linear algebra (one being that in some datasets, doing an SVD is possible though computing AA^T leads numeric instability; read more below). Please go through the additional reading material if you want to read about that in detail.

Additional Reading

- Relationship between SVD and PCA
- Why SVD for PCA