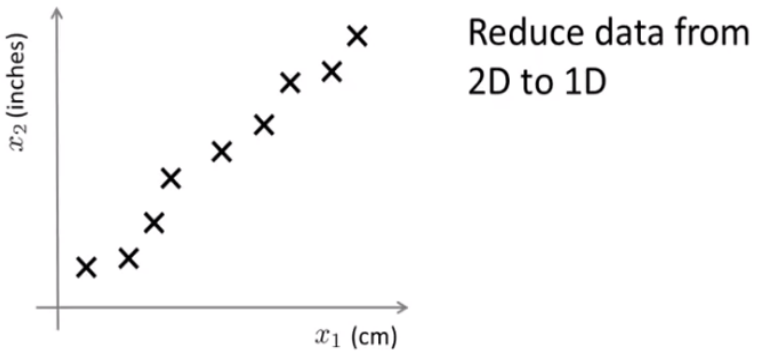
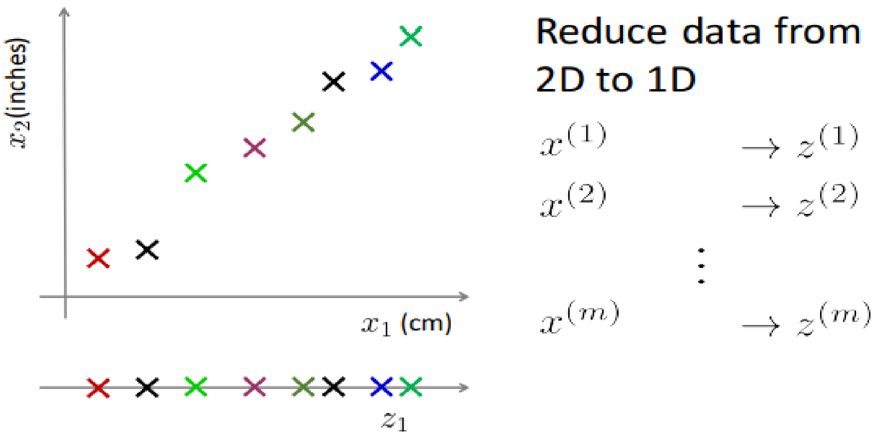
PCA (Principal Component Analysis)

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

The idea of PCA is simple — reduces the number of variables of a data set, while preserving as much information as possible.



If we draw a line between these points and project all points on that line, we will get a good approximation of the points, but only on the line we've found. It means we are going to have new points on the line, which will be just one dimension. We are going to have something like this



Eigenvectors and Eigenvalues

Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the **principal components** of the data.
Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components. So, the idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible

information in the first component, then maximum remaining information in the second and so on, until having something like shown in the scree plot below.

Let our data matrix X be the score of three students:

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

Technical Procedure of PCA

1. **Compute the mean of every dimension of the whole dataset.**

   The data from the above table can be represented in matrix A, where each column in the matrix shows scores on a test and each row shows the score of a student.

$$\mathbf{A} = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

   So, the mean of matrix A would be

$$\bar{\mathbf{A}} = [\ 66 \quad 60 \quad 60\ ]$$

2. **Compute the covariance matrix of the whole dataset.**

   So, we can compute the covariance of two variables X and Y using the following formula

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n} (Xi - \bar{x})(Yi - \bar{y})$$

   Using the above formula, we can find the covariance matrix of A. Also, the result would be a square matrix of d ×d dimensions.

$$\begin{array}{cc} & \begin{array}{ccc} Math & English & Art \end{array} \\ \begin{array}{c} Math \\ English \\ Art \end{array} & \begin{bmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{bmatrix} \end{array}$$

Covariance Matrix of A

Few points that can be noted here is:

- Shown in Blue along the diagonal, we see the variance of scores for each test. The art test has the biggest variance (720); and the English test, the smallest (360). So we can say that art test scores have more variability than English test scores.
- The covariance is displayed in black in the off-diagonal elements of the matrix A
- The covariance between English and art, however, is zero. This means there tends to be no predictable relationship between the movement of English and art scores.

3. **Compute Eigenvectors and corresponding Eigenvalues.**

Intuitively, an eigenvector is a vector whose direction remains unchanged when a linear transformation is applied to it.
Now, we can easily compute eigenvalue and eigenvectors from the covariance matrix that we have above.

Let A be a square matrix, ν a vector and λ a scalar that satisfies Aν = λν, then λ is called eigenvalue associated with eigenvector ν of A.

The eigenvalues of A are roots of the characteristic equation

$$\det(A - \lambda I) = 0$$

Calculating det(A-λI) first, I is an identity matrix :

$$\det\left(\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)$$

Simplifying the matrix first, we can calculate the determinant later,

$$\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

$$\begin{pmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{pmatrix}$$

Now that we have our simplified matrix, we can find the determinant of the same:

$$\det\begin{pmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{pmatrix}$$

$$-\lambda^3 + 1584\lambda^2 - 641520\lambda + 25660800$$

We now have the equation and we need to solve for λ, so as to get the eigenvalue of the matrix. So, equating the above equation to zero:

$$-\lambda^3 + 1584\lambda^2 - 641520\lambda + 25660800 = 0$$

After solving this equation for the value of λ, we get the following value

$$\lambda \approx 44.81966\ldots, \lambda \approx 629.11039\ldots, \lambda \approx 910.06995\ldots$$

Now, we can calculate the eigenvectors corresponding to the above eigenvalues.

$$A\nu = \lambda\nu,$$

$$Av_1 = \lambda_1 v_1$$

$$\begin{pmatrix} -3.75100\ldots \\ 4.28441\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} -0.50494\ldots \\ -0.67548\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} 1.05594\ldots \\ 0.69108\ldots \\ 1 \end{pmatrix}$$

$$v_1, \qquad v_2, \qquad v_3$$

4. **Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d × k dimensional matrix W.**

The eigenvectors only define the directions of the new axis, since they have all the same unit length 1. So, in order to decide which eigenvector(s) we want to drop for our lower-dimensional subspace, we have to take a look at the corresponding eigenvalues of the eigenvectors.

The eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data, and those are the ones we want to drop.

The common approach is to rank the eigenvectors from highest to lowest corresponding eigenvalue and choose the top k eigenvectors. So, after sorting the eigenvalues in decreasing order, we have

$$\begin{pmatrix} 910.06995 \\ 629.11039 \\ 44.81966 \end{pmatrix}$$

For our simple example, where we are reducing a 3-dimensional feature space to a 2-dimensional feature subspace, we are combining the two eigenvectors with the highest eigenvalues to construct our d×k dimensional eigenvector matrix W.

So, eigenvectors corresponding to two maximum eigenvalues are:

$$\mathbf{W} = \begin{bmatrix} 1.05594 & -0.50494 \\ 0.69108 & -0.67548 \\ 1 & 1 \end{bmatrix}$$

5. Transform the samples onto the new subspace

In the last step, we use the 2×3 dimensional matrix W that we just computed to transform our samples onto the new subspace via the equation $y = W' \times x$ where W' is the transpose of the matrix W.