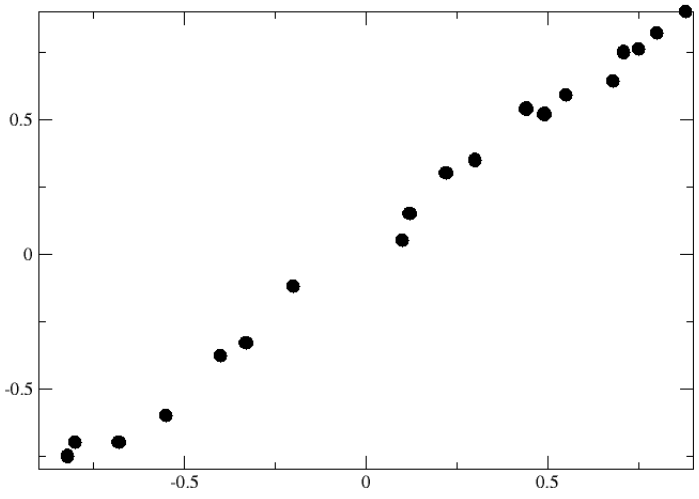
1 Principle Component Analysis

a. Geometry Intuition

Imagine we have a two dimensional plane with axis x_1 and x_2 perpendicular to each other. On this plane we have a data set(x_{1i}, x_{2i}), as shown in the graph below. We notice most of the data lie along the line 45 degree angle between the x_1 and x_2 axis. If we do a coordinate transformation by rotating the $x1$ and $x2$ axis by 45 degree counterclockwise, we get new axis z_1 and z_2. Then we see our data mainly lies along z_1 axis. So if we eliminate coordinate z_2, we are still able to keep most information in the data. We reduce a two dimensional data to one dimension. The z_1 axis here is called the principal component.



The 45 degree rotation can be written as

$$z_1 = \frac{\sqrt{2}}{2}(x_1 + x_2)$$
$$z_2 = \frac{\sqrt{2}}{2}(x_1 - x_2)$$

The above equation can be written in matrix form

$$\begin{pmatrix} z_1 & z_2 \end{pmatrix} = \begin{pmatrix} z_1 & z_2 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}$$

We call $W_1 = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})^T$, and $W_2 = (\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})^T$
In general, we can write the transform with W

$$\begin{pmatrix} z_1 & z_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \end{pmatrix}$$

Writing Z and X in row vectors instead of column vector looks a little weird. The reason we do this is when we have multiple samples, we will increase number of rows to accommodate more samples.

Miltiple Sample representation and Algorithm Review

If there are n samples, the transformation can be written in matrix notation

$$\begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ ... & ... \\ z_{n1} & z_{n2} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ ... & ... \\ x_{m1} & x_{m2} \end{pmatrix} \begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \end{pmatrix}$$

In order to reduce to dimension, the goal is to find the transformation matrix W so that $(z_{11}, z_{21}, ...z_{n1})^T$ have the maximum variance. The dimension which has to maximum variance is the principle component. If we assume the data we have is processed and have mean at 0. Then the variance of Z is

$$\begin{aligned} Z^T Z = & \begin{pmatrix} W_1^T \\ W_2^T \end{pmatrix} \\ & \begin{pmatrix} x_{11} & x_{21} & ... & x_{m1} \\ x_{21} & x_{22} & ... & x_{m2} \end{pmatrix} \\ & \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ ... & ... \\ x_{m1} & x_{m2} \end{pmatrix} \begin{pmatrix} W_1 & W_2 \end{pmatrix} \\ & = W^T X^T X W. \end{aligned}$$

Here we use some intuition to analyze the variance and leave the rigorous proof in next. By intuition we imagine W_1, W_2 ... W_d be the eigenvectors of $X^T X$, namely

$$X^T X W_k = \lambda_k W_k$$

Then

$$W_k^T X^T X W_k = \lambda_k W_k^T W_k = \lambda_k$$

Therefore, in order to maximize $W_k^T X^T X W_k$, W_k has to be the corresponding eigenvector of the maximum eigenvalue λ_{max}.

c. Rigorous Proof

Suppose we have a data set (X,y) where x is the feature variable. It is an mxn

matrix where m is the data size, and n is the dimension of the features

$$\begin{pmatrix} x_{11} & x_{12} & ... & x_{1n} \\ x_{21} & x_{22} & ... & x_{2n} \\ ... & ... & ... & ... \\ x_{m1} & x_{m2} & ... & x_{mn} \end{pmatrix}$$

We call the feature vector associated with the ith data x(i), we consider a coordinate transformation:

$$\begin{pmatrix} z_{11} & z_{12} & ... & z_{1n} \\ z_{21} & z_{22} & ... & z_{2n} \\ ... & ... & ... & ... \\ z_{m1} & z_{m2} & ... & z_{mn} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1n} \\ x_{21} & x_{22} & ... & x_{2n} \\ ... & ... & ... & \\ x_{m1} & x_{m2} & ... & x_{mn} \end{pmatrix} \begin{pmatrix} w_{11} & w_{21} & ... & w_{n1} \\ w_{12} & w_{22} & ... & w_{n2} \\ ... & ... & ... & ... \\ w_{1n} & w_{22} & ... & w_{nn} \end{pmatrix}$$

The goal is to reduce the dimension of the feature to d, still have a good representation of the data. When Z has only d (d¡n) dimension, then

$$\begin{pmatrix} w_{11} & w_{21} & ... & w_{d1} \\ w_{12} & w_{22} & ... & w_{d2} \\ ... & ... & ... & ... \\ w_{1n} & w_{22} & ... & w_{dn} \end{pmatrix}$$

The problem is how to choose d dimensions out of n. We define the error function as

$$\sum_{i}^{m} ||x_i - z_i||_2^2 = ||X - WWTx||_2^2 \text{This is a little involved}$$

$$W = argmax||X - XWWT||22 = argmaxtr(WTX^TXW)$$

Let w_1, w_2,, w_n be the column vectors of matrix W, then

$$\begin{aligned} & tr(WTX^TXW) \\ &= w_1^TX^TXw_1 + w_2^TX^TXw_2 + \ + w_d^TX^TXw_d \\ &= \sum w_i^Tw_{ii}(\lambda_i \text{ is the ith eigenvalues of } X^TX) \end{aligned}$$

If w_i are the eigenvector corresponding to λ_i Then the maximum value of the trace is achieved when we take w_1 to w_d as the eigenvectors associated with the first d maximum eigenvalues.