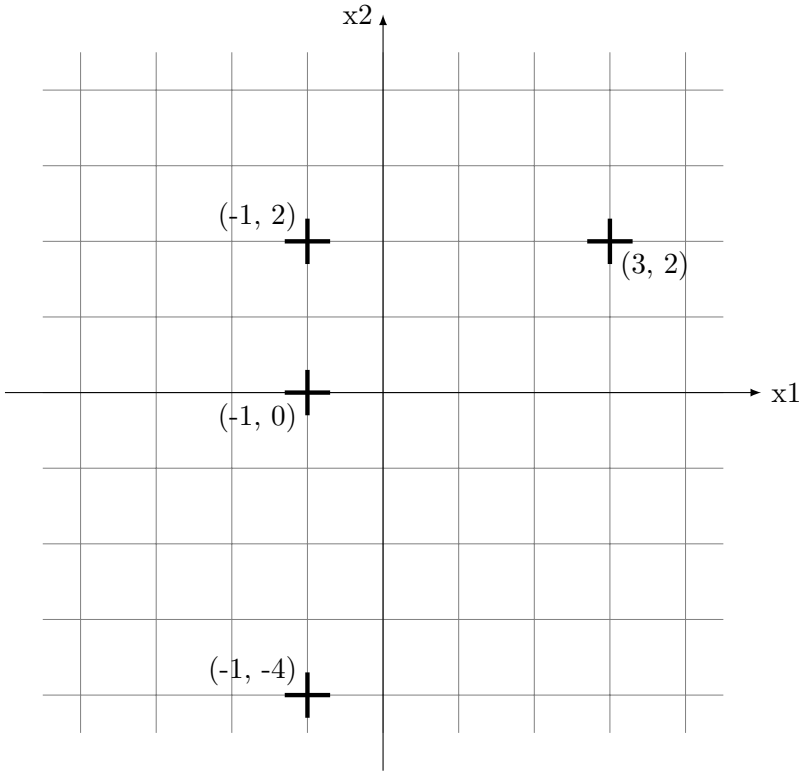
PRINCIPAL COMPONENT ANALYSIS

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 x1 and x2. The variances of these variables and their covariance are useful quantities to describe the structure of the dataset.

$$\begin{aligned}
Var(x1) &= \frac{1}{N}\sum_{i=1}^{N}(x1_i - \bar{x1})^2 \\
&= \frac{1}{4}\sum_{i=1}^{4}(x1_i - 0)^2 \\
&= \frac{1}{4}|x1|^2 \\
&= \frac{3^2 + (-1)^2 + (-1)^2 + (-1)^2}{4} \\
&= 3 \\
Var(x2) &= \frac{1}{4}\sum_{i=1}^{4}(x2_i - 0)^2 \\
&= \frac{1}{4}|x2|^2 \\
&= \frac{2^2 + 0^2 + 2^2 + (-4)^2}{4} \\
&= 6 \\
Cov(x1, x2) &= \frac{1}{N}\sum_{i=1}^{N}(x1_i - \bar{x1})(x2_i - \bar{x2}) \\
&= \frac{1}{4}\sum_{i=1}^{4}(x1_i - 0)(x2_i - 0) \\
&= \frac{1}{4}\vec{x1}.\vec{x2} \\
&= \frac{3 * 2 + (-1) * 0 + (-1) * 2 + (-1) * (-4)}{4} \\
&= 2
\end{aligned}$$

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.

$$\begin{aligned}\sum &= \begin{pmatrix} Var(x1) & Cov(x1,x2) \\ Cov(x1,x2) & Var(x2) \end{pmatrix} \\ &= \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}\end{aligned}$$

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.

$$\begin{aligned}\sum \phi &= \lambda\phi \\ (\sum -\lambda I)\phi &= 0\end{aligned}$$

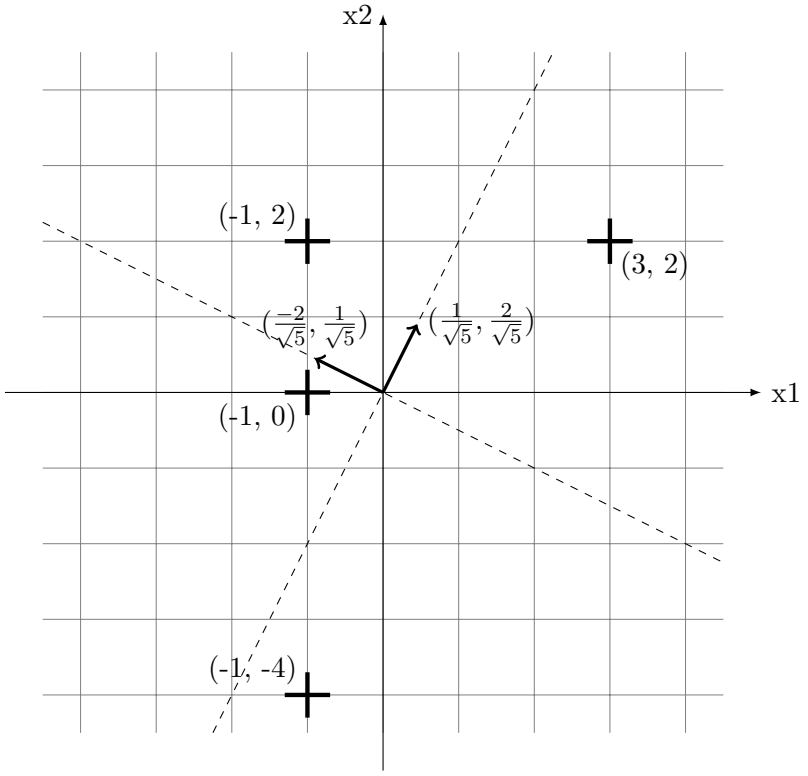
$(\sum -\lambda I)$ must be singular if eigenvectors are non-zero. Hence,

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

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.

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

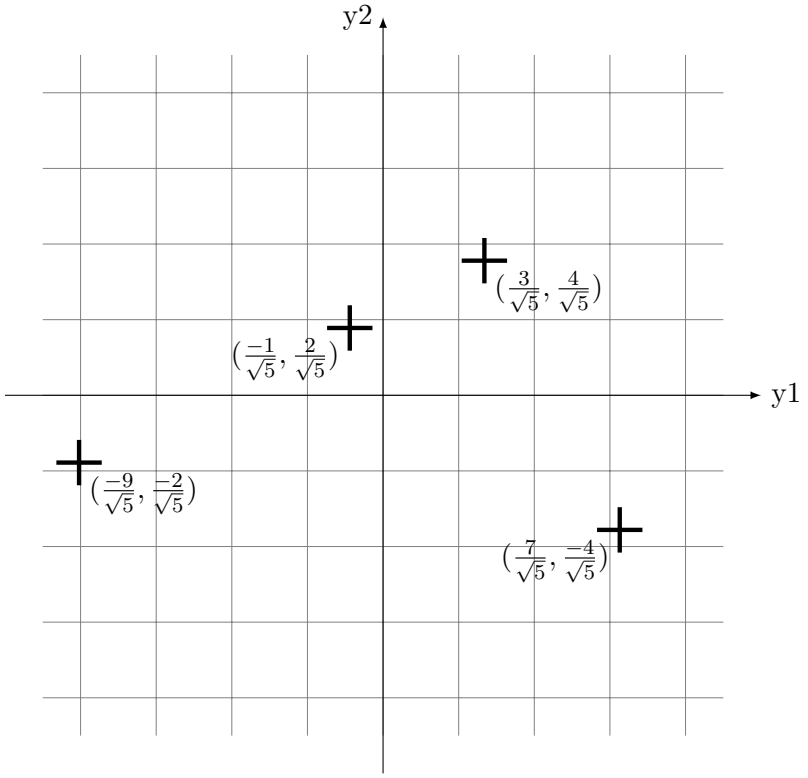
See the following visualization of eigenvectors.



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

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

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



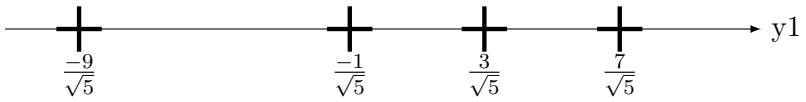
Visually, the points on the transformed graph arise due to a change of axes from the original axes to the axes defined by the eigenvectors of $\sum$. To see why this choice of axes was made, the variance covariance matrix of the transformed data is calculated.

$$\begin{aligned}
\sum_{new} &= \frac{1}{N} Y Y^T \\
&= \frac{1}{4}\frac{1}{\sqrt{5}} \begin{pmatrix} 7 & -1 & 3 & -9 \\ -4 & 2 & 4 & -2 \end{pmatrix} \frac{1}{\sqrt{5}} \begin{pmatrix} 7 & -4 \\ -1 & 2 \\ 3 & 4 \\ -9 & -2 \end{pmatrix} \\
&= \frac{1}{20} \begin{pmatrix} 140 & 0 \\ 0 & 40 \end{pmatrix} \\
&= \begin{pmatrix} 7 & 0 \\ 0 & 2 \end{pmatrix}
\end{aligned}$$

This results in the variance covariance of transformed data being diagonalized. Also notice that the variance values i.e. the diagonal terms are equal to the corresponding eigenvalues. PCA yields a set of a new set of variables/axes from the original set of variables/axes which have the following properties.

- New variables are mutually uncorrelated.
- New variables are orthogonal to each other.

You are free to choose a subset of these new 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. In this example, if the new variable corresponding to lower variance is dropped, the dataset is transformed from 2-d to 1-d as shown below.



2 Mathematics

Assume a (MxN) matrix X as a set of N observations of M random variables such that each observation is a column in the matrix X. Furthermore, assume that each variable has zero mean meaning that each row of the matrix sums up to zero. Denote by $\sum_x$ the variance covariance matrix of the variables. Due to the zero mean condition,

$$\sum_x = XX^T$$

Let $\phi_i, \phi_2, ..., \phi_m$ be the m orthonormal eigenvectors of $\sum_x$. Let P be defined as the matrix $[\phi_1^T, \phi_2^T, ..., \phi_m^T]^T$. Consider the following transformation,

$$Y = PX$$

It can be shown that the resulting variables are also zero-mean based. So, their variance covariance matrix can be written as,

$$\begin{aligned}\sum_y &= YY^T \\ &= (PX)(PX)^T \\ &= PXX^TP^T \\ &= P\sum_x P^T\end{aligned}$$

$\sum_x$ is symmetric and all symmetric matrices are diagonalized by a matrix of their orthonormal eigenvectors. Or more precisely,

$$\sum_x = P^T D P$$

where D is a diagonal matrix. Upon substitution in the preceding equation,

$$\sum_y = P P^T D P P^T$$

Since P is orthonormal, its transpose is its inverse and thus $PP^T = I$ which reduces $\sum_y$ to D.

3 Kernel PCA

Let X be $[x_1, x_2, ..., x_N]$ where each x_i is an observation of M variables. Consider mapping each x_i to $f(x_i)$ where f is non-linear and may increase the dimensionality of each x_i. So,

$$f(X) = [f(x_1), f(x_2), ..., f(x_N)]$$

Now examine the variance covariance matrix of $f(X)$.

$$\begin{aligned}
\sum_f &= \frac{1}{N} f(X) f(X)^T \\
&= \frac{1}{N} [f(x_1), f(x_2), ..., f(x_N)] \begin{pmatrix} f(x_1^T) \\ f(x_2^T) \\ ... \\ f(x_N^T) \end{pmatrix} \\
&= \frac{1}{N} \sum_{i=1}^{N} f(x_i) f(x_i^T)
\end{aligned}$$

To find principal components, eigenvalues and eigenvectors of the variance covariance matrix must be found out. since there are N eigenvalue, eigenvector pairs, they will be subscripted by j.

$$\lambda_j \phi_j = \sum_f \phi_j$$

$$\lambda_j \phi_j = \frac{1}{N} \sum_{i=1}^{N} f(x_i) f(x_i^T) \phi_j$$

Since $f(x_i^T)\phi_j = f(\vec{x_i}).\vec{\phi_j}$ is just a scalar, the equation can be simplified to,

$$\phi_j = \sum_{i=1}^{N} \frac{f(\vec{x_i}).\vec{\phi_j}}{N\lambda_j} f(x_i)$$

Substituting $\frac{f(\vec{x_i}).\vec{\phi_j}}{N\lambda_j} = a_{ij}$, the equation becomes,

$$\phi_j = \sum_{i=1}^{N} a_{ij} f(x_i)$$

This equation reveas that the principal components are linear combinations of the observations.

$$\phi_j = a_{1j} f(x_1) + a_{2j} f(x_2) + ... + a_{Nj} f(x_N)$$

Taking dot products by $f(x_1), f(x_2)...f(x_N)$ successively,

$$\begin{aligned}
f(\vec{x_1}).\vec{\phi_j} &= a_{1j} f(\vec{x_1}).f(\vec{x_1}) + a_{2j} f(\vec{x_1}).f(\vec{x_2}) + ... + a_{Nj} f(\vec{x_1}).f(\vec{x_N}) \\
f(\vec{x_2}).\vec{\phi_j} &= a_{1j} f(\vec{x_2}).f(\vec{x_1}) + a_{2j} f(\vec{x_2}).f(\vec{x_2}) + ... + a_{Nj} f(\vec{x_2}).f(\vec{x_N}) \\
... &= ... \\
f(\vec{x_N}).\vec{\phi_j} &= a_{1j} f(\vec{x_N}).f(\vec{x_1}) + a_{2j} f(\vec{x_N}).f(\vec{x_2}) + ... + a_{Nj} f(\vec{x_N}).f(\vec{x_N})
\end{aligned}$$

In matrix form, these equations look like,

$$\begin{pmatrix} f(\vec{x_1}).\vec{\phi_j} \\ f(\vec{x_2}).\vec{\phi_j} \\ ... \\ f(\vec{x_N}).\vec{\phi_j} \end{pmatrix} = \begin{pmatrix} f(\vec{x_1}).f(\vec{x_1}) & f(\vec{x_1}).f(\vec{x_2}) & ... & f(\vec{x_1}).f(\vec{x_N}) \\ f(\vec{x_2}).f(\vec{x_1}) & f(\vec{x_2}).f(\vec{x_2}) & ... & f(\vec{x_2}).f(\vec{x_N}) \\ ... & & & \\ f(\vec{x_N}).f(\vec{x_1}) & f(\vec{x_N}).f(\vec{x_2}) & ... & f(\vec{x_N}).f(\vec{x_N}) \end{pmatrix} \begin{pmatrix} a_{1j} \\ a_{2j} \\ ... \\ a_{Nj} \end{pmatrix}$$

The first matrix on the right is a NxN matrix. Denote it by the kernel matrix: K. For every term on the left hand side matrix use the fact that $f(\vec{x_i}).\vec{\phi_j} = N\lambda_j a_{ij}$.

$$N\lambda_j \begin{pmatrix} a_{1j} \\ a_{2j} \\ ... \\ a_{Nj} \end{pmatrix} = K \begin{pmatrix} a_{1j} \\ a_{2j} \\ ... \\ a_{Nj} \end{pmatrix}$$

It is apparent that $N\lambda_j, [a_{1j}, a_{2j}, ..., a_{Nj}]^T \forall j \in [1, N]$are eigenvalue, eigenvector pairs for the matrix K. The PCA transform for $f(X)$ can be written as,

$$\begin{aligned} Y &= \begin{pmatrix} \phi_1^T \\ \phi_2^T \\ ... \\ \phi_N^T \end{pmatrix} f(X) \\ &= \begin{pmatrix} \phi_1^T \\ \phi_2^T \\ ... \\ \phi_N^T \end{pmatrix} [f(x_1), f(x_2), ..., f(x_N)] \\ &= \begin{pmatrix} \vec{\phi_1}.f(\vec{x_1}) & \vec{\phi_1}.f(\vec{x_2}) & ... & \vec{\phi_1}.f(\vec{x_N}) \\ \vec{\phi_2}.f(\vec{x_1}) & \vec{\phi_2}.f(\vec{x_2}) & ... & \vec{\phi_2}.f(\vec{x_N}) \\ ... & & & \\ \vec{\phi_N}.f(\vec{x_1}) & \vec{\phi_N}.f(\vec{x_2}) & ... & \vec{\phi_N}.f(\vec{x_N}) \end{pmatrix} \end{aligned}$$

Using the relation $f(\vec{x_i}).\vec{\phi_j} = N\lambda_j a_{ij}$,

$$\begin{aligned} Y &= \begin{pmatrix} N\lambda_1 a_{11} & N\lambda_1 a_{21} & ... & N\lambda_1 a_{N1} \\ N\lambda_2 a_{12} & N\lambda_2 a_{22} & ... & N\lambda_2 a_{N2} \\ ... & & & \\ N\lambda_N a_{1N} & N\lambda_N a_{2N} & ... & N\lambda_N a_{NN} \end{pmatrix} \\ &= N \begin{pmatrix} \lambda_1 a_{11} & \lambda_1 a_{21} & ... & \lambda_1 a_{N1} \\ \lambda_2 a_{12} & \lambda_2 a_{22} & ... & \lambda_2 a_{N2} \\ ... & & & \\ \lambda_N a_{1N} & \lambda_N a_{2N} & ... & \lambda_N a_{NN} \end{pmatrix} \end{aligned}$$

This is the kernel trick for PCA where it is not required to explicit map each point to higher dimensions whcich may be computationally expensive. PCA transformation in nonlinear higher dimensions can be calculated by performing the following steps in the order below.

- Choose a kernel $k(x_i, x_j) = f(\vec{x}_i).f(\vec{x}_j)$.
- Calculate kernel matrix $\mathrm{K} = \begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & ... & k(x_1,x_N) \\ k(x_2,x_1) & k(x_2,x_2) & ... & k(x_2,x_N) \\ ... & & & \\ k(x_N,x_1) & k(x_N,x_2) & ... & k(x_N,x_N) \end{pmatrix}$.
- Find the eigenvalue, eigenvalue pairs of K.
- Use eigenvalue, eigenvector pairs to calcuate the kernel PCA transform as shown above.
- Do dimensinality reduction as appropriate.

4 Assumptions and Limitations

- The new variables are linear combinations of the original ones. Kernel PCA helps alleviate this problem.
- PCA assumes that mean and variance are sufficient statistics to describe a random variable. Since only Gaussian distributions are described fully by their variance, the assumption that input dataset is Gaussian is inherent.
- The new variables are orthogonal which can be limiting but makes the solution to be readily expressible through linear algebra.