Principal Component Analysis

Ugenteraan Manogaran

E-mail : m.ugenteraan15@gmail.com

24 February 2019

Contents

1.0 Data Set

Suppose there is a data set with zero mean that consists of m number of data points. Each of the data points are an n-dimensional vector.

- Each of the data points can be represented as $\mathbf{x}^{(i)}$. The superscript (i) is used to denote the *i-th* data point. Hence, i $\in 1, ..., m$.
- Since each of the data points are an n-vector,

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ . \\ . \\ . \\ x_n^{(i)} \end{bmatrix}$$

- The data set can be represented in a single n x m matrix $\mathbf{X}$ as shown below. $\mathbf{X}$ is called a data matrix.

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & . & . & . & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & . & . & . & x_2^{(m)} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ x_n^{(1)} & x_n^{(2)} & . & . & . & x_n^{(m)} \end{bmatrix}$$

Note : *Each of the data points could have been represented as 1* x *n matrices* $\mathbf{x}^{(i)^{\mathbf{T}}}$ *as well. That would have resulted* $\mathbf{X}$ *to be a m* x *n matrix* $\mathbf{X}^{\mathbf{T}}$ *as shown below.*

$$\mathbf{x}^{(i)^{\mathbf{T}}} = \begin{bmatrix} x_1^{(i)} & x_2^{(i)} & . & . & . & x_n^{(i)} \end{bmatrix}, \quad \mathbf{X}^{\mathbf{T}} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & . & . & . & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & . & . & . & x_n^{(2)} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ x_1^{(m)} & x_2^{(m)} & . & . & . & x_n^{(m)} \end{bmatrix}$$

However, in this note, we will be denoting each of the data points as n-vectors $\mathbf{x}^{(i)}$ and the data set as the n x m matrix $\mathbf{X}$.

2.0 Problem Statement

In many cases, the dimensionality of the data set is too high to be visualized or analyzed with some particular technique or method. Hence, the need to reduce the dimensionality of the data points in some "optimal" way is required.

For simplicity sake, we will be performing the dimensionality reduction linearly. In other words, we want to project our data points to a linear subspace of the *original vector space* [1]. The linear subspace will be a q-dimensional vector space where $q < n$.

The optimal linear subspace is the one that can reconstruct the data points back from the q-dimensional space to the n-dimensional space with as little *reconstruction error* [2] as possible. Now, the question is how do we find this optimal linear subspace?

Note : *The projection of the data points from the n-dimensional space to the q-dimensional space are the reconstructed data points.*

3.0 Principal Component Analysis

Principal Component Analysis (PCA) is an algorithm that reduces the dimensionality of a data set to a lower-dimensional linear subspace by linear projection in such a way that the reconstruction error made by the linear projection is as low as possible.

3.1 Reconstruction Error

Suppose the lower-dimensional subspace (q-dimensional) is spanned by a set of orthonormal n-vectors $\{\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, ..., \mathbf{p}^{(q)}\}$. Then, the projected data points $\mathbf{z}^{(i)}$ from the original space to the lower-dimensional subspace **i.e.** *the reconstructed data points* are given by :

$$\mathbf{z}^{(i)} = \sum_{j=1}^{q} (\mathbf{x}^{(i)} \bullet \mathbf{p}_j)\mathbf{p}_j \tag{1}$$

Since the reconstruction error, $\mathbf{E}_R$, is the mean squared distance between the original and the reconstructed data points,

[1] The n-dimensional space that is used to represent the original data points.

[2] Reconstruction error is the mean squared distance between the original and the reconstructed data points.

$$
\begin{aligned}
\mathbf{E}_R &= \frac{1}{m}\sum_{i=1}^{m} \parallel \mathbf{x}^{(i)} - \mathbf{z}^{(i)} \parallel^2 \\
&= \frac{1}{m}\sum_{i=1}^{m} \parallel \mathbf{x}^{(i)} - \sum_{j=1}^{q}(\mathbf{x}^{(i)} \bullet \mathbf{p}_j)\mathbf{p}_j \parallel^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{q} \parallel \frac{1}{q}\mathbf{x}^{(i)} - (\mathbf{x}^{(i)} \bullet \mathbf{p}_j)\mathbf{p}_j \parallel^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{q}(\frac{1}{q}\mathbf{x}^{(i)} - (\mathbf{x}^{(i)} \bullet \mathbf{p}_j)\mathbf{p}_j) \bullet (\frac{1}{q}\mathbf{x}^{(i)} - (\mathbf{x}^{(i)} \bullet \mathbf{p}_j)\mathbf{p}_j) \\
&= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{q}\frac{1}{q}(\mathbf{x}^{(i)} \bullet \mathbf{x}^{(i)}) - \mathbf{x}^{(i)} \bullet (\mathbf{x}^{(i)} \bullet \mathbf{p}_j)\mathbf{p}_j - (\mathbf{x}^{(i)} \bullet \mathbf{p}_j)\mathbf{p}_j \bullet \mathbf{x}^{(i)} \\
&\qquad\qquad + (\mathbf{x}^{(i)} \bullet \mathbf{p}_j)\mathbf{p}_j \bullet (\mathbf{x}^{(i)} \bullet \mathbf{p}_j)\mathbf{p}_j \\
&= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{q}\frac{1}{q} \parallel \mathbf{x}^{(i)} \parallel^2 -2(\mathbf{x}^{(i)} \bullet \mathbf{p}_j)^2 + (\mathbf{x}^{(i)} \bullet \mathbf{p}_j)^2\mathbf{p}_j \bullet \mathbf{p}_j
\end{aligned}
$$

Since $\mathbf{p}_j \bullet \mathbf{p}_j = 1$ due to the orthogonality,

$$
\begin{aligned}
&= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{q}\frac{1}{q} \parallel \mathbf{x}^{(i)} \parallel^2 -(\mathbf{x}^{(i)} \bullet \mathbf{p}_j)^2 \\
&= \frac{1}{m}(\sum_{i=1}^{m}\sum_{j=1}^{q}\frac{1}{q} \parallel \mathbf{x}^{(i)} \parallel^2) - \frac{1}{m}(\sum_{i=1}^{m}\sum_{j=1}^{q}(\mathbf{x}^{(i)} \bullet \mathbf{p}_j)^2) \\
&= \frac{1}{m}(\sum_{i=1}^{m} \parallel \mathbf{x}^{(i)} \parallel^2) - \frac{1}{m}(\sum_{i=1}^{m}\sum_{j=1}^{q}(\mathbf{x}^{(i)} \bullet \mathbf{p}_j)^2)
\end{aligned}
$$

Notice that :

$$
\begin{aligned}
\frac{1}{m}(\sum_{i=1}^{m}\sum_{j=1}^{q}(\mathbf{x}^{(i)} \bullet \mathbf{p}_j)^2) &= \frac{1}{m}(\sum_{i=1}^{m}(\mathbf{x}^{(i)} \bullet \mathbf{p}_1)^2 + (\mathbf{x}^{(i)} \bullet \mathbf{p}_2)^2 + ... + (\mathbf{x}^{(i)} \bullet \mathbf{p}_q)^2) \\
&= \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}^{(i)} \bullet \mathbf{p}_1)^2 + \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}^{(i)} \bullet \mathbf{p}_2)^2 + ... + \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}^{(i)} \bullet \mathbf{p}_q)^2 \\
&= var(\mathbf{p}_1) + var(\mathbf{p}_2) + ... + var(\mathbf{p}_q) \\
&= \sigma^2_{\mathbf{p}_1} + \sigma^2_{\mathbf{p}_2} + ... + \sigma^2_{\mathbf{p}_q} \\
&= \sum_{j=1}^{q}\sigma^2_{\mathbf{p}_j}
\end{aligned}
$$

Therefore,

$$\mathbf{E}_R = \frac{1}{m}\sum_{i=1}^{m} \parallel \mathbf{x}^{(i)} \parallel^2 - \sum_{j=1}^{q} \sigma^2_{\mathbf{P}_j} \tag{2}$$

From equation **(2)** above, it can be seen that minimizing $\mathbf{E}_R$ is equivalent to maximizing the variance of the reconstructed data points.

Note: *The first term in the equation is the variance of the original data points. For* $\mathbf{E}_R$ *to be zero, the variance of the reconstructed points have to be equal to the variance of the original data points. This can only happen when the q-dimensional subspace pass through all the data points in the n-dimensional space. Also note that the variance of the projected data points will never be bigger than the variance of the original data points.*

3.2 Covariance Matrix

A covariance matrix of a data matrix is a symmetric matrix where the main diagonal's entries are the variance of each variable while the off-diagonal's entries are the covariance between every pair of variables.

As been mentioned in **section 1**, the data set will be represented by the n x m matrix $\mathbf{X}$ as below.

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & . & . & . & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & . & . & . & x_2^{(m)} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ x_n^{(1)} & x_n^{(2)} & . & . & . & x_n^{(m)} \end{bmatrix}$$

Note : *The mean of the data set is zero.*

The covariance matrix of $\mathbf{X}$, $\mathbf{S_X}$ is calculated as such.

$$\mathbf{S_X} = \frac{1}{m}\mathbf{X}\mathbf{X^T} \tag{3}$$

Proof :

$$\mathbf{XX^T} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & . & . & . & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & . & . & . & x_2^{(m)} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ x_n^{(1)} & x_n^{(2)} & . & . & . & x_n^{(m)} \end{bmatrix} \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & . & . & . & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & . & . & . & x_n^{(2)} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ x_1^{(m)} & x_2^{(m)} & . & . & . & x_n^{(m)} \end{bmatrix}$$

Since,

$$cov(\mathbf{x}_j, \mathbf{x}_k) = \frac{1}{m}\sum_{i=1}^{m}[(x_j^{(i)})(x_k^{(i)})] \qquad \text{and} \qquad cov(\mathbf{x}_j, \mathbf{x}_j) = var(\mathbf{x}_j) \qquad ,$$

$$\mathbf{XX^T} = m \begin{bmatrix} var(\mathbf{x}_1) & cov(\mathbf{x}_1, \mathbf{x}_2) & . & . & . & cov(\mathbf{x}_1, \mathbf{x}_n) \\ cov(\mathbf{x}_2, \mathbf{x}_1) & var(\mathbf{x}_2) & . & . & . & cov(\mathbf{x}_2, \mathbf{x}_n) \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ cov(\mathbf{x}_n, \mathbf{x}_1) & cov(\mathbf{x}_n, \mathbf{x}_2) & . & . & . & var(\mathbf{x}_n) \end{bmatrix}$$

Since $cov(\mathbf{x}_j, \mathbf{x}_k) = cov(\mathbf{x}_k, \mathbf{x}_j)$, $\mathbf{XX^T}$ *is an* n x n symmetric matrix where the main diagonal's entries are the variance of each variable and the off-diagonal's entries are the covariance of each pair of variables.

i.e.

$$\begin{aligned} [\mathbf{XX^T}]_{j,k} &= \sum_{i=1}^{m}[(x_j^{(i)})(x_k^{(i)})] \\ &= m[cov(\mathbf{x}_j, \mathbf{x}_k)] \end{aligned}$$

Therefore,

$$\mathbf{S_X} = \frac{1}{m}\mathbf{XX^T} \qquad or \qquad \mathbf{S_X} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}^{(i)}\mathbf{x}^{(i)\mathbf{T}} \tag{4}$$

3.3 Maximizing Variances in a Covariance Matrix

How do we maximize the variances **i.e.** *the diagonal entries,* in a covariance matrix?

As stated in **section 3.1**, the total variance of the reconstructed data points is

$$\frac{1}{m}(\sum_{i=1}^{m}\sum_{j=1}^{q}(\mathbf{x}^{(i)} \bullet \mathbf{p}_j)^2)$$

The above expression can be rewritten as

$$\begin{aligned}
\frac{1}{m}(\sum_{i=1}^{m}\sum_{j=1}^{q}(\mathbf{x}^{(i)} \bullet \mathbf{p}_j)^2) &= \frac{1}{m}(\sum_{i=1}^{m}\sum_{j=1}^{q}(\mathbf{x}^{(i)^\mathbf{T}}\mathbf{p}_j)^2) \\
&= \frac{1}{m}(\sum_{i=1}^{m}\sum_{j=1}^{q}(\mathbf{x}^{(i)^\mathbf{T}}\mathbf{p}_j)^\mathbf{T}(\mathbf{x}^{(i)^\mathbf{T}}\mathbf{p}_j)) \\
&= \frac{1}{m}(\sum_{i=1}^{m}\sum_{j=1}^{q}(\mathbf{p}_j{}^\mathbf{T}\mathbf{x}^{(i)})(\mathbf{x}^{(i)^\mathbf{T}}\mathbf{p}_j)) \\
&= \sum_{j=1}^{q}(\mathbf{p}_j{}^\mathbf{T}\frac{1}{m}\sum_{i=1}^{m}[\mathbf{x}^{(i)}\mathbf{x}^{(i)^\mathbf{T}}]\mathbf{p}_j)
\end{aligned}$$

Referring to equation **(4)**, we have

$$= \sum_{j=1}^{q}(\mathbf{p}_j{}^\mathbf{T}\mathbf{S}_\mathbf{X}\mathbf{p}_j)$$

Hence, the total variance of the reconstructed data points can be expressed as

$$\sum_{j=1}^{q}(\mathbf{p}_j{}^\mathbf{T}\mathbf{S}_\mathbf{X}\mathbf{p}_j) \tag{5}$$

To maximize the total variance of the reconstructed data points, we need to perform partial derivative on expression **(5)** with respect to each $\mathbf{p}_j$. However, to prevent $\| \mathbf{p}_j \| \rightarrow \infty$, the maximization has to be constrained. Hence, we shall make use of the condition $\mathbf{p}_j^\mathbf{T}\mathbf{p}_j = 1$. A Lagrange multiplier that is denoted by λ_j shall be introduced to enforce the constraint and then make an unconstrained maximization of

$$\sum_{j=1}^{q}(\mathbf{p}_j{}^\mathbf{T}\mathbf{S}_\mathbf{X}\mathbf{p}_j + \lambda_j(1 - \mathbf{p}_j^\mathbf{T}\mathbf{p}_j)) \tag{6}$$

By setting the partial derivative with respect to each $\mathbf{p}_j$ equal to zero, we get

$$2\mathbf{S}_\mathbf{X}\mathbf{p}_1 - 2\lambda_1\mathbf{p}_1 \ + \ldots \ + \ 2\mathbf{S}_\mathbf{X}\mathbf{p}_q - 2\lambda_q\mathbf{p}_q = 0$$

Therefore,

$$\mathbf{S_X p}_j = \lambda_j \mathbf{p}_j \tag{7}$$

It can be seen that each $\mathbf{p}_j$ is an eigenvector of $\mathbf{S_X}$ and λ_j is the eigenvalue of $\mathbf{S_X}$ that corresponds to $\mathbf{p}_j$. Because $\mathbf{p}_j^\mathbf{T}\mathbf{p}_j = 1$, we can multiply both side by $\mathbf{p}_j^\mathbf{T}$ on the left and obtain

$$\mathbf{p}_j^\mathbf{T}\mathbf{S_X p}_j = \lambda_j \tag{8}$$

The eigenvector that corresponds to the largest eigenvalue is known as the first principal component and the second largest by the second principal component and so on. By referring to expression **(5)** it can be concluded that

$$var(\mathbf{p}_j) = \lambda_j \tag{9}$$

Even though we've explicitly defined that $q < n$, PCA still holds for $q = n$, in which case there is no dimensionality reduction but simply a rotation of the coordinate axes to align with the principal components **i.e.** *the eigenvectors of the covariance matrix of the data matrix*. This is also known as matrix diagonalization (*Refer to Appendix 4.1*). Hence, the total variance in the original data points is

$$\sum_{j=1}^{n} \lambda_j \tag{10}$$

and the total variance of the reconstructed data points is

$$\sum_{j=1}^{q} \lambda_j \tag{11}$$

Therefore, the reconstruction error $\mathbf{E_R}$ is

$$\mathbf{E_R} = \sum_{j=1}^{n} \lambda_j - \sum_{j=1}^{q} \lambda_j = \sum_{j=q+1}^{n} \lambda_j \tag{12}$$

3.4 Goal of Principal Component Analysis

As mentioned in **section 3** and **section 3.1**, the goal of PCA is to linearly project the data points to a lower-dimensional subspace such that the reconstruction error is minimized, or equivalently, the total variance of the projected data points are maximized.

The basis that spans the q-dimensional subspace $\{\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, ..., \mathbf{p}^{(q)}\}$ will be represented as a q x n matrix $\mathbf{P}$ as below.

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}^{(1)^\mathbf{T}} \\ \mathbf{p}^{(2)^\mathbf{T}} \\ . \\ . \\ . \\ \mathbf{p}^{(q)^\mathbf{T}} \end{bmatrix} = \begin{bmatrix} p_1^{(1)} & p_2^{(1)} & . & . & . & p_n^{(1)} \\ p_1^{(2)} & p_2^{(2)} & . & . & . & p_n^{(2)} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ p_1^{(q)} & p_2^{(q)} & . & . & . & p_n^{(q)} \end{bmatrix}$$

The goal now is to find $\mathbf{P}$ such that

$$\mathbf{PX} = \mathbf{Y} \tag{13}$$

where the total variance of the data points in $\mathbf{Y}$ is maximized. **i.e.** *Covariance matrix of* $\mathbf{Y}$, $\mathbf{S_Y}$ *is diagonalized.* We begin by rewriting $\mathbf{S_Y}$.

$$\begin{aligned} \mathbf{S_Y} &= \frac{1}{m}\mathbf{YY^T} \\ &= \frac{1}{m}(\mathbf{PX})(\mathbf{PX})^T \\ &= \frac{1}{m}\mathbf{PXX^TP^T} \\ &= \mathbf{P}(\frac{1}{m}\mathbf{XX^T})\mathbf{P^T} \end{aligned}$$

Based on equation **(4)**,

$$= \mathbf{PS_XP^T}$$

By matrix diagonalization (*Refer to Appendix A*),

$$\mathbf{S_X} = \mathbf{QDQ^T} \tag{14}$$

By selecting the matrix $\mathbf{P}$ to be a matrix where each row is an eigenvector of $\mathbf{S_X}$ **i.e.** $\mathbf{P} = \mathbf{Q^T}$,

$$\begin{aligned} \mathbf{S_Y} &= \mathbf{PS_XP^T} \\ &= \mathbf{QQ^TDQ^TQ} \end{aligned}$$

Because, $\mathbf{Q}^{-1} = \mathbf{Q^T}$ *or* $\mathbf{QQ^T} = \mathbf{Q^TQ} = \mathbf{I}$,

$$\mathbf{S_Y} = \mathbf{D} \tag{15}$$

We see that our selection of $\mathbf{P}$ diagonalizes $\mathbf{S_Y}$. The rows of $\mathbf{P}$ are the principal components of $\mathbf{X}$. This was the goal for PCA. By selecting the number of principal components q to be lesser than n, the data matrix $\mathbf{X}$ is transformed to a lower dimensional vector space $\mathbf{Y}$.

References

[1] *Nakos, G., and Joyner, D. (1998). Linear algebra with applications. PWS Publishing Company.*

[2] *Wong Y.P. (2015) Principal Component Analysis*

[3] *Bishop, C. M. (2006). Pattern recognition and machine learning. springer.*

[4] *Shlens, J. (2014). A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100.*

[5] *Principal Component Analysis. (2019). Retrieved from https://www.stat.cmu.edu/ cshalizi/uADA/12/lectures/ch18.pdf*

A Diagonalization

A vector $\mathbf{v}$ is an eigenvector of an r x r matrix $\mathbf{A}$ if for some scalar λ

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{16}$$

The scalar λ is called an eigenvalue of $\mathbf{A}$ corresponding to the eigenvector $\mathbf{v}$. Eigenvectors are not unique. Any scalar multiple of an eigenvector $\mathbf{v}$ is also an eigenvector of $\mathbf{A}$. Furthermore, all the scalar multiples of an eigenvector have the same eigenvalue. Note that the eigenvectors that corresponds to a set of distinct eigenvalues of $\mathbf{A}$ are linearly independent. If $\mathbf{A}$ has r linearly independent eigenvectors $\begin{bmatrix}\mathbf{e}_1 & \mathbf{e}_2 & \dots & \mathbf{e}_r\end{bmatrix}$, then

$$\begin{aligned}
\mathbf{A}\begin{bmatrix}\mathbf{e}_1 & \mathbf{e}_2 & \dots & \mathbf{e}_r\end{bmatrix} &= \begin{bmatrix}\mathbf{A}\mathbf{e}_1 & \mathbf{A}\mathbf{e}_2 & \dots & \mathbf{A}\mathbf{e}_r\end{bmatrix} \\
&= \begin{bmatrix}\lambda_1\mathbf{e}_1 & \lambda_2\mathbf{e}_2 & \dots & \lambda_2\mathbf{e}_r\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{e}_1 & \mathbf{e}_2 & \dots & \mathbf{e}_r\end{bmatrix}\begin{bmatrix} w & 0 & . & . & . & 0 \\ 0 & \lambda_2 & 0 & . & . & 0 \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & 0 \\ 0 & . & . & . & 0 & \lambda_r \end{bmatrix}
\end{aligned}$$

Let

$$\mathbf{E} = \begin{bmatrix}\mathbf{e}_1 & \mathbf{e}_2 & \dots & \mathbf{e}_r\end{bmatrix} \qquad \text{and} \qquad \mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & . & . & . & 0 \\ 0 & \lambda_2 & 0 & . & . & 0 \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & 0 \\ 0 & . & . & . & 0 & \lambda_r \end{bmatrix}$$

Therefore,

$$\mathbf{A}\mathbf{E} = \mathbf{E}\mathbf{D} \tag{17}$$

Because $\mathbf{E}$ is a square with linearly independent columns, it is invertible. Therefore,

$$\mathbf{E}^{-1}\mathbf{A}\mathbf{E} = \mathbf{D} \tag{18}$$

Matrix $\mathbf{A}$ is said to be similar to the diagonal matrix $\mathbf{D}$ and $\mathbf{A}$ is called diagonalizable.

If matrix $\mathbf{A}$ turns out to be a symmetric matrix **i.e.** $\boldsymbol{A} = \boldsymbol{A}^{\boldsymbol{T}}$ with real values, then by Spectral Theorem, $\mathbf{A}$ is orthogonally diagonalizable and has only real eigenvalues. This means that there exist an orthogonal r x r matrix $\mathbf{Q}$ and a diagonal matrix $\mathbf{D}$ such that

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}$$

Since $\mathbf{Q}$ is an orthogonal matrix, $\mathbf{Q}^{-1} = \mathbf{Q}^{T}$. Therefore,

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{T} \tag{19}$$

or equivalently,

$$\mathbf{Q}^{T}\mathbf{A}\mathbf{Q} = \mathbf{D} \tag{20}$$

As before, $\mathbf{D}$ is a diagonal matrix with eigenvalues on its main diagonal and $\mathbf{Q}$ contains a set of linearly independent eigenvectors. However, unlike before, the eigenvectors are orthonormal to each other.

Since a covariance matrix is a real symmetric matrix, it is therefore orthogonally diagonalizable. Diagonalization maximize the variances while the covariances are reduced to zero.

Note: *Since high covariance means that values in the two particular dimensions are highly related to each other, it is therefore indicates redundancy. Hence, reducing the covariances to zero further helps to filter out redundancy in the data.*