Principal Component Analysis

Jap Purohit :1940109

Nihar Patel :1940119

Mohit Prajapati :1940171

Raj Gariwala :1940118

Purvam Sheth :1940151

Introduction:

Background

Principal components are the key to PCA; they represent what's underneath the hood of your data.

Principal Component Analysis or PCA is a dimensionality reduction technique. Many non-linear dimensionality reduction techniques exist out there (Sammons Mapping) but linear methods are more mature, if more limited.

1. Principal Component Analysis
2. Multidimensional Scaling
   PCA minimizes low-dimensional reconstruction error, but another sensible objective is to maximize the scatter of the projection, under the rationale that doing so would yield the most informative projection (this choice is sometimes called classical scaling).

There are other methods too for linear dimensionality reduction. We have included Multidimensional scaling as it is connected with PCA. This project focuses on obtaining the principal components of a matrix. The Principal Component Analysis helps find the component of the matrix that contribute most significantly. It is used for compression of data without losing any significant information. This project uses many concepts of linear algebra thus increasing our grasp and understanding of the concepts. It also helps us in seeing how these concepts are used in the real life and how it changes our lives in ways we can't imagine. Our group has tried to reduce the dimensions of a medical data set thus focusing on the components that really matter and even combining different components to form one singular variable.

Motivation and Overview

With the ongoing situation in the world medical reports and precision of medical reports have become increasingly important. One minor mistake can cost a person's life. For this project we decide to work on a dataset of a lung cancer victims. There are several critical variables which determine the life of the patient. In the third stage there are simply too many variables to be considered. Keeping in mind all these variables is simply not possible. This is where Principal Component Analysis comes in. It helps in reducing the dimensionality of the data set while keeping in mind that there is no significant loss of information. These cancer data set proves to be the perfect set for performing PCA. While each and every variable is equally important there are some which play a bigger role in the outcome of the health of a patient. The relation between different factors is based on a form of matrix. It analyses the data and tells us what can be clubbed together or in a crucial time even rejected for then.

Our dataset contains 5100 x 36 data points to work on. With each row representing a different patient and each column representing a different factor contributing to the lung cancer.

Keywords

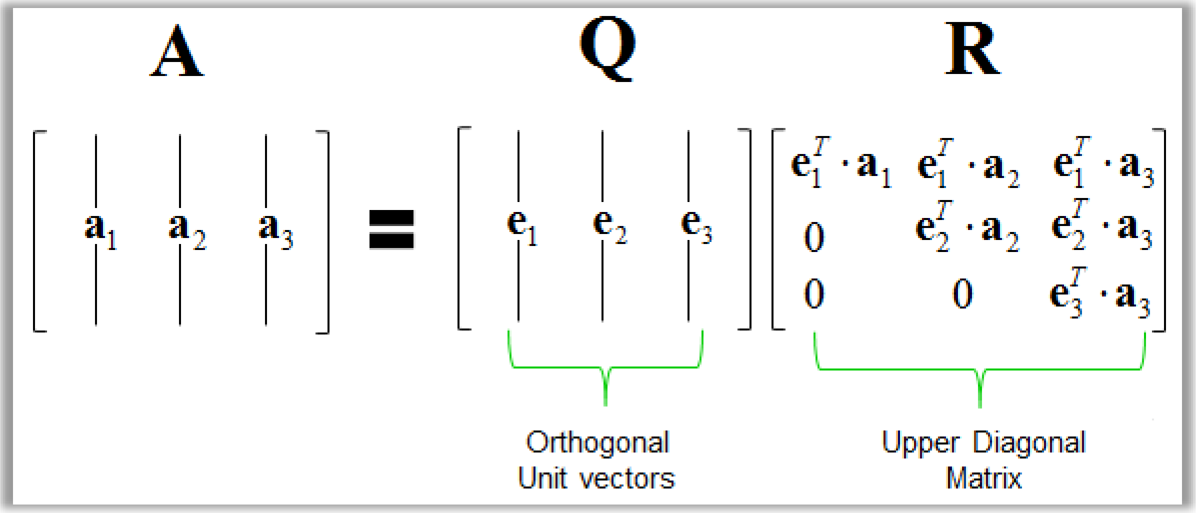
- PCA (Principal Component Analysis) and data reduction
- Householder Reflection and QR Decomposition of matrix

Approach

Description and the Approach used

- **QR Decomposition of matrix**
  To compute with matrixes QR Decomposition is a stable method in comparison with other methods. Matrix **A**, in system of equation in matrix form **Ax=b** is converted to **A=QR** and which is further written as $\mathbf{Rx=Q^{-1}b}$, which is much easier to compute.
  In QR Decomposition, **Q** is the orthonormal matrix which means $\mathbf{Q^{-1}=Q^{T}}$, whereas **R** is the upper diagonal matrix. So, we can also rewrite the equation as $\mathbf{Rx=Q^{T}b}$.

  $\mathbf{Rx=Q^{T}b}$ is easier to solve because right hand side is just a vector. And R is an upper triangular. So, this can be easily solved by back-substitution.

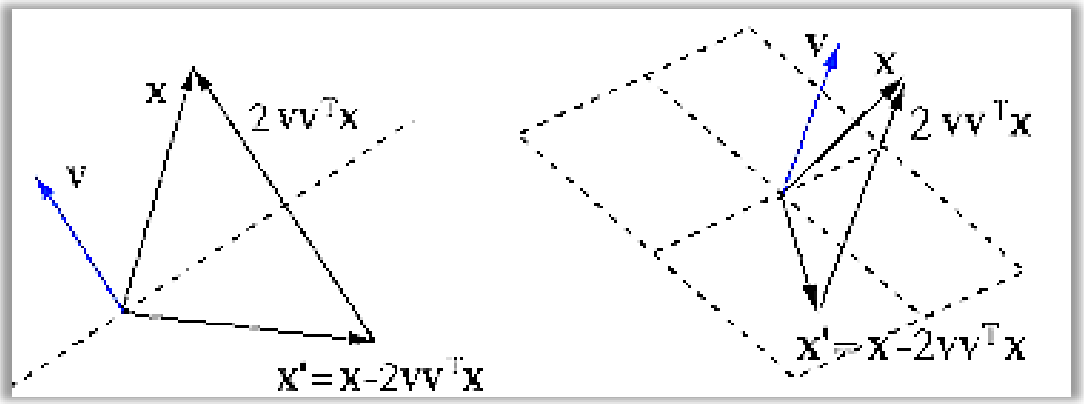


- **Householder Reflection**

  Householder reflection is one of the important aspects in finding QR decomposition.

  Householder reflection technique is used in finding reflection of any vector in any dimension which respect to any plane in the given subspace. $\mathbf{H} = (\boldsymbol{I} - 2(\mathbf{v}\mathbf{v}^{T}))$, this equation is known as householder reflection equation, where **v** is the orthonormal vector to the plane and **H** is the Householder reflector matrix.

$$
\begin{aligned}
\mathbf{x}' &= \mathbf{H}\mathbf{x} \\
&= \mathbf{x} - 2(\mathbf{x}\mathbf{v})\mathbf{v} \\
&= \mathbf{x} - 2\mathbf{v}(\mathbf{x}\mathbf{v}) \\
&= \boldsymbol{I}\mathbf{x} - 2\mathbf{v}(\mathbf{v}^{T}\mathbf{x}) \\
&= \boldsymbol{I}\mathbf{x} - 2(\mathbf{v}\mathbf{v}^{T})\mathbf{x} \\
&= (\boldsymbol{I} - 2(\mathbf{v}\mathbf{v}^{T}))\mathbf{x}
\end{aligned}
$$



Representation of vector and it's reflection with respect to a plane.

Conclusion

On an ending note this project was a wonderful opportunity for us to understand the concepts of linear algebra such as QR Decomposition, eigen values and vectors. Not only did we learn much more about them but also applied it in real life. This was a great hand on activity for us and we understood how the concepts are applied in real life. Besides this there are several other components of linear algebra that are yet to be discovered and applied by us. One can't cover everything in a single file. But apart from that it was a great experience working with a team, getting their feedbacks on how to solve a particular problem moving ahead and finally finishing the project.

Each and every group member has contributed equally on the outcome of this project.