# MSc Data Mining and Machine Learning (2019)

# Lab 2 – PCA

|  |  |  |
| --- | --- | --- |
| Instructor: |  | Dr Peter Jancovic |
| Written by: |  | James J Nkhata |
|  |  |  |
| Experiment Date: |  | 07/02/2019 |
| Report Delivery Date: |  | 21/02/2019 |

# Part 2: PRINCIPLE COMPONENTS ANALYSIS

## Purpose

The aim of part 2 of the lab was to discover the structure of the modified dataset “Lab2Data-matlab” after applying Principle Component Analysis (PCA).

“Lab2Data-matlab” was the same dataset as “Lab2Data” without headers for better compatibility with MATLAB software.

## Procedure

### Loaded “Lab2Data-matlab” into MATLAB as a matrix X:

#### Typed the command ‘>>X=load(‘Lab2Data-Matlab’), to read the data from Lab2Data-matlab into matrix X.

### Computed the covariance of matrix X using the “cov” function in MATLAB:

#### Typed the command ‘>> C= cov (X)’, to store the results of covariance of matrix X to Matrix C.

### Applied eigenvector/eigenvalue decomposition to the covariance matrix C:

#### Typed the command ‘>> [U, D] = eig(C)’, to decompose the matrix C into matrices U and D.

## Experimental data

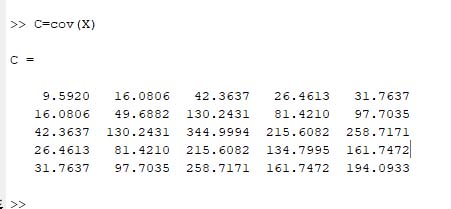


Figure 3: Resulting matrix C after computing the covariance of matrix X in MATLAB

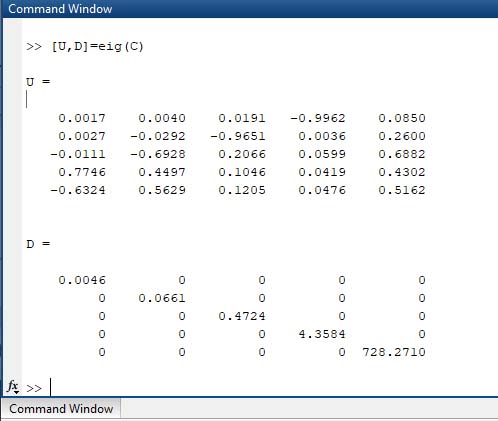


Figure 4: Resulting matrix U and D after applying eigenvector/eigenvalue decomposition to matric C

## Questions

The eigenvalues computed were

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Eigenvalues | 0.0046 | 0.0661 | 0.4724 | 4.3584 | 728.2710 |

The eigenvalues of 0.0046 in the new x-direction and that of 728.2710 in the new y-direction implied that there was high variance of data along the dimensions corresponding to the y-direction.

## Conclusion

This lab report demonstrated the results of applying Principle Components Analysis (PCA) on a multi-dimensional (5-dimensions) dataset to discover its structure.

Through the results of the eigenvalue decomposition of covariance matrix C (of dataset Lab2Data-matlab), the eigenvalues obtained corresponded to a high variance of data along the new y-direction (728.2710) while the new x-direction (0.0046) had a very small variance. These results indicated that significant components of the data were in the y-direction. This led to the belief that the data within the other dimensions was insignificant and possibly noise.

Therefore, the data contained in the 5-dimensions seemed redundant and could be reduced to 1-dimension retaining the data that corresponded to the eigenvalue of 728.2710.

### References

Jancovic, P (2019) **Data Mining and Machine Learning** [Online]. Available from: <https://canvas.bham.ac.uk/courses/34771> [Accessed 30 January 2019].