Complex Engineering Project

Abdullah Farooq 2020023
FCS
GIKI
RehamYar Khan, Pakistan
u2020023@giki.edu.pk

Ammar Ahmad 2020071
FCS
GIKI
Faisalabad, Pakistan
u2020071@giki.edu.pk

Bilal Malik 2020102
FCS
GIKI
Karachi, Pakistan
u2020102@giki.edu.pk

***Abstract*—Processing multi-dimensional imagery using PCA and performing error analysis on it.**

Keywords—analysis, PCA, MSE, components

I. Introduction

The inclusion of more features in any data set might lead to worsening performance issues. Increasing the number of features will not always improve classification accuracy, which is also known as the curse of dimensionality. Dimensionality reduction is applied to improve classification accuracy by selecting the optimal set of lower dimensionality features. Hence, Principal Component analysis (PCA) is essential for data science, machine learning, data visualization, statistics, and other quantitative fields for accurate result. There are two techniques to make dimensionality reduction:

- Feature Selection
- Feature Extraction

It is essential to know about vector, matrix, and transpose matrix, eigenvalues, eigenvectors, and others to understand the concept of dimensionality reduction.

II. Methodology

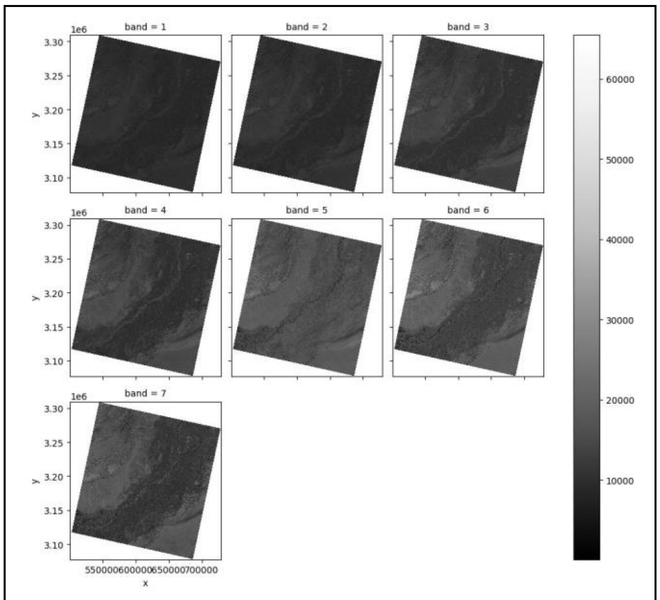
A. Landsat Multi-band Raster Data

A Landsat multi-band image of Rahimyar Khan (LC09_L2SP_151040_20221125_20221127_02_T1) was acquired first. Each band in a Landsat scene is often stored in an individual .tif file. The bands were then brought into a xarrray DataFrame, using Python.

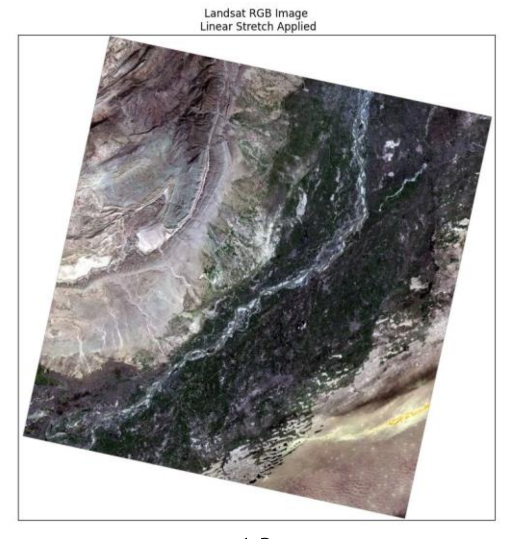
B. Plotting Bands

A function called open_clean_bands is used, in a loop, that opens a single .tif file and returns an xarray object. The result is a single array containing all bands in xarray. The bands are then concatenated using a python function and plotted using imshow(), (Fig 1.1). A 3 band color composite image is now plotted using the earthpy library function ep.plot_rgb(), (Fig 1.2).

w



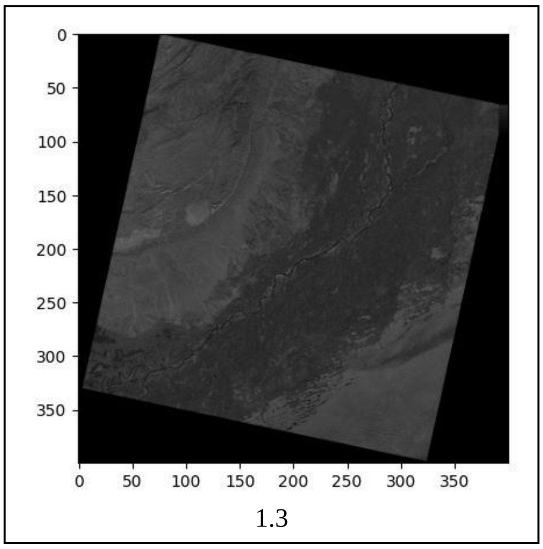
1.1



1.2

C. *Cropping Image*

The original image has been cropped so that PCA can be applied on it to reduce processing time and save RAM. This is done using the cv() library and converting the iamge into and np.array(). The original image was a 7741x7591 matrix, which has now been reduced to a square matrix of 400x400. (Fig 1.3)

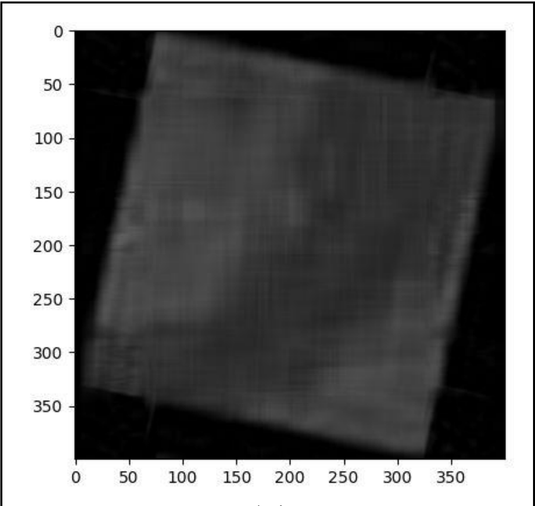

1.3

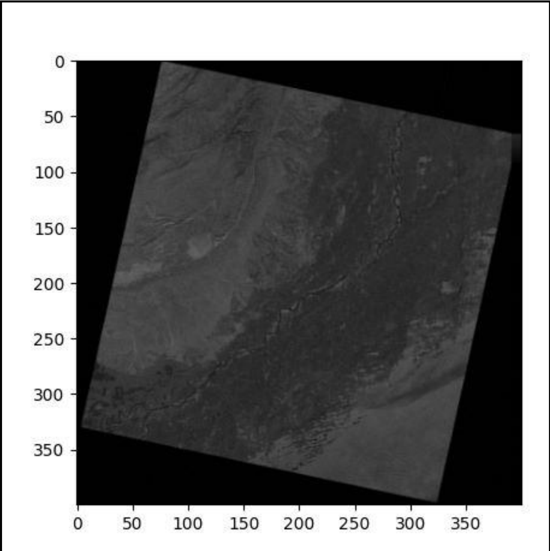
D. *PCA*

In exploratory data analysis, the Principal Component Analysis (PCA) method is used to reduce the dimensionality of the data set to 2D or 3D in order to create predictive models. We use PCA to reduce the number of dimensions of a dataset (Dimensional Reduction), to visualize the data of higher dimensionality and to capture the original variance in the data as much as we can. In summary, PCA can be defined as the transformation of a high number of variables. present in a dataset, into a smaller number of uncorrelated variables, also called PCs (principal components).

Steps which are involved in PCA include standardizing the PCA, calculating the covariance matric. Eigenvalues and the respective eigenvectors are found for the covariance matric and the data (vectors) is plotted on the scaled data using pythons libraries.

The image is split into its RBG components and normalized by dividing it by 255, so that a value between 0 and 1 can be calculated. The components are projected into a smaller subspace now, where the eigenvectors form the axes. A covariance matrix, which stores the eigenvalues, is calculated for each of the components. To decide which eigenvector(s), we can discard without losing much information in the subspace construction and checking the corresponding eigenvalues. The eigenvectors with the highest values are the ones that include more information about the distribution of our data. The components are concatenated after PCA is performed on them individually. The inversetransform() function is used to reconstruct the band so that it can be plotted using the MATLAB library. The number of eigenvectors selected is provided to the system, a lower number would result in a blurry image (Fig 1.4) while a higher number would result in a sharper image (Fig 1.5).


1.4

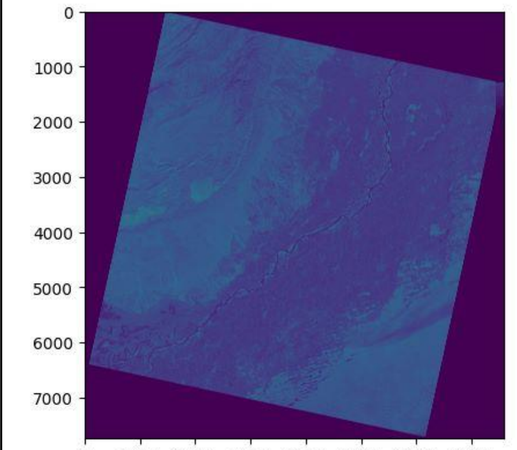

1.5

E. *Error Analysis*

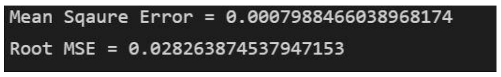
Error analysis on the PCA is done using Mean Square Error. In Statistics, Mean Squared Error is defined as Mean or Average of the square of the difference between actual and estimated values. Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD). The calculations for the mean squared error are similar to the variance. To find the MSE, take the observed value, subtract the predicted value, and square that difference. Repeat that for all observations. Then, sum all of those squared values and divide by the number of observations. Squaring the differences eliminates negative values for the differences and ensures that the mean squared error is always greater than or equal to zero. It is almost always a positive value.

Only a perfect model with no error produces an MSE of zero. And that doesn't occur in practice.

The original image has three components, RGB. To conduct error analysis on it, the image was converted to a greyscale, which means one component (Fig 1.6). MSE was conducted on it with 10 PCA (Fig 1.6) and 100 PCA (Fig 1.7) applied.

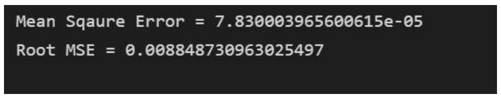


1.6



```
Mean Sqaure Error = 0.0007988466038968174
Root MSE = 0.028263874537947153
```

1.7



```
Mean Sqaure Error = 7.830003965600615e-05
Root MSE = 0.008848730963025497
```

1.8

III. Task Distribution

- 2020023
  Conducted research on the syntax and libraries used to perform PCA and wrote the program. Debugged the code. Provided clippings for the report
- 2020071
  Learned the syntax to help deign the code for the program. Studied basic python libraries used and read manuals on them, which would help in the program.
- 2020102
  Conducted research on the mathematical operations of PCA to help design the code for the PCA function itself. Compiled research into a report.