Principal Component Analysis on Multispectral Satellite Imagery

Muhammad Musa*, Shumail†
Department of Engineering Sciences
Ghulam Ishaq Khan Institute of Engineering Sciences and Technology
Email: *u2021421@giki.edu.pk, †u2022649@giki.edu.pk

***Abstract*—This project investigates the application of Principal Component Analysis (PCA) for dimensionality reduction on multispectral satellite imagery [1]. By implementing PCA from fundamental linear algebra principles [2], we reduced the number of spectral bands while minimizing information loss. An error analysis was performed, quantifying the trade-off between dimensionality reduction and information preservation. The results highlight PCA's effectiveness in compressing multispectral data, retaining over 95% of the variance using only two principal components. This approach facilitates efficient storage and analysis of large datasets, demonstrating the utility of PCA for remote sensing applications.**

***Index Terms*—Principal Component Analysis (PCA), Multispectral satellite image, Dimensionality Reduction, Variance**

I. Introduction

Multispectral satellite imagery, a cornerstone of remote sensing, plays a crucial role in diverse applications ranging from land cover mapping and urban planning to disaster management. The richness of information contained within multiple spectral bands, however, comes at the cost of high dimensionality. This high dimensionality presents significant challenges for efficient data storage, processing, and analysis, often requiring substantial computational resources. Principal Component Analysis (PCA), a powerful dimensionality reduction technique rooted in linear algebra, offers an effective solution to this challenge.

In this project, we leverage PCA to reduce the spectral dimensionality of Landsat multispectral imagery while preserving essential information content. The core of PCA lies in its ability to transform the original, correlated spectral bands into a new set of uncorrelated variables called principal components. This transformation, achieved through eigenvalue decomposition of the data's covariance matrix, identifies the directions of maximum variance within the data. By retaining only the principal components associated with the largest eigenvalues, which represent the directions of greatest variability, we can effectively compress the data into a lower-dimensional space. This project not only applies PCA to a real-world dataset but also emphasizes the underlying linear algebra principles. We implement PCA from scratch using fundamental linear algebra operations in Python, providing a hands-on understanding of the technique's mathematical foundations. Furthermore, a rigorous error analysis is conducted to quantify the trade-off between dimensionality reduction and information preservation, guiding the selection of an optimal number of principal components. This approach allows us to gain a deeper understanding of both the theoretical and practical aspects of PCA in the context of remote sensing and image analysis.

A. Problem Statement

The inherent high dimensionality of multispectral satellite imagery poses significant challenges for storage, processing, and analysis. Large datasets, such as those obtained from Landsat sensors with multiple spectral bands, require substantial computational resources. This project addresses this challenge by employing Principal Component Analysis (PCA) to reduce the spectral dimensionality of a Landsat image. By implementing PCA from fundamental linear algebra principles and performing a comprehensive error analysis, this study aims to determine the optimal number of principal components required to represent the data effectively while minimizing information loss and reconstruction error.

B. Our Solutions

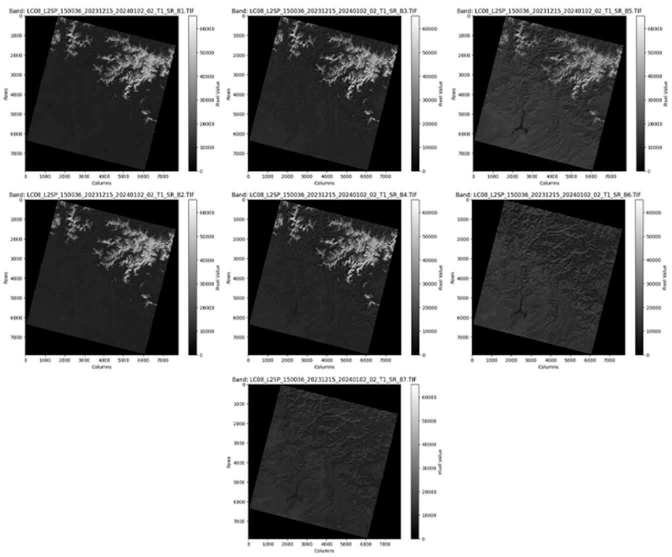
This project tackles the challenge of high dimensionality in Landsat multispectral imagery by implementing Principal Component Analysis (PCA) from fundamental linear algebra principles. We developed Python code to perform eigenvalue decomposition of the data's covariance matrix, enabling the identification and selection of principal components that capture the most significant variance in the data. Through rigorous experimentation and analysis, varying the number of retained components and evaluating the resulting reconstruction error using metric such as Mean Squared Error (MSE), we determined the optimal number of components for efficient data representation while minimizing information loss. Our results demonstrate that a significant reduction in dimensionality can be achieved while preserving a high percentage of the original data's variance, facilitating more efficient storage, processing, and analysis of multispectral satellite imagery. For our selected Landsat scene and region of interest, just two principal components were able to capture over 95% of the variance, indicating excellent data compressibility.

II. Methodology

This section details the steps involved in acquiring, preprocessing, and analyzing the Landsat multispectral imagery using Principal Component Analysis (PCA) in Python. The overall workflow involves loading the data, preparing it for PCA, implementing the PCA algorithm, analyzing the results, and visualizing the outcomes.

A. Data Loading and Visualization

The dataset used in this study comprises seven spectral bands (B1-B7) from a Landsat 8 Level-2 surface reflectance product. The chosen scene covers a portion of Topi(KPK) Pakistan. Each band, representing surface reflectance at a specific wavelength, was initially loaded and visualized in grayscale using the `rasterio` library for efficient file handling and `matplotlib` for image display. This initial visualization allows for inspection of spatial characteristics and patterns within each individual spectral band.



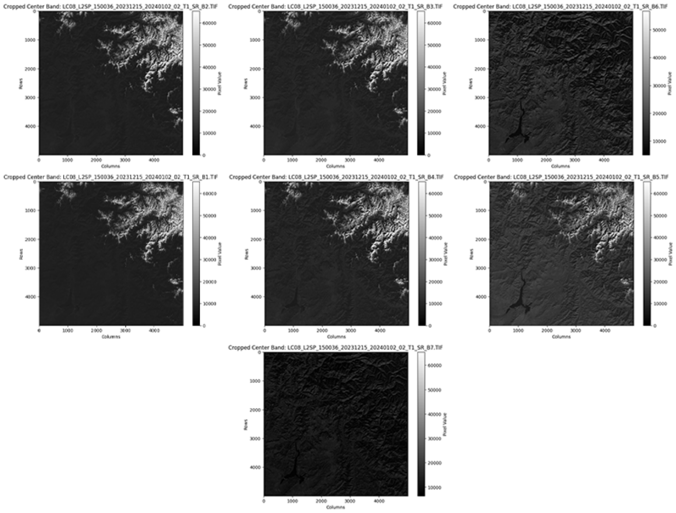
Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B1.TIF
Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B3.TIF
Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B5.TIF
Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B2.TIF
Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B4.TIF
Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B6.TIF
Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B7.TIF
Rows
Columns
Pixel Value

Fig. 1. Grayscale visualization of the seven raw spectral bands (B1-B7) from the Landsat image of Topi

B. Data Preprocessing

Prior to PCA implementation, several preprocessing steps were performed to prepare the data for analysis:

1) **Spectral Band Scaling and Stacking**: To ensure consistent and meaningful comparison of different bands, the raw pixel values for each band were scaled based on the necessary gain and offset values found in the MTL metadata file. Each band was also converted to floating-point numbers to handle potential decimal values during subsequent processing steps. After the gain and offset factors were applied, the scaled bands were stacked into a 3D NumPy array with dimensions (rows, columns, bands), where the spectral band layers followed the expected order (B1, B2, B3...).
2) **Cropping**: To manage computational demands, the images were cropped to a region of interest (ROI) of 5000 x 5000 pixels. This cropping strategy focuses the analysis on a specific area while reducing processing time and memory requirements. The cropped bands were then stacked into a 3D array (`cropped_stack`) with dimensions (5000, 5000, 7).



Cropped Center Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B2.TIF
Cropped Center Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B3.TIF
Cropped Center Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B6.TIF
Cropped Center Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B1.TIF
Cropped Center Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B4.TIF
Cropped Center Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B5.TIF
Cropped Center Band: LC08_L2SP_150036_20231215_20240102_02_T1_SR_B7.TIF
Rows
Columns
Pixel Value

Fig. 2. Visualization of the cropped bands from the original.

3) **Flattening**: The 3D stacked array was reshaped into a 2D matrix of size (25000000, 7), where each row represents a single pixel and the columns correspond to the seven spectral bands. This flattening step prepares the data for input into the PCA algorithm, which operates on 2D matrices.
4) **Standardization**: The flattened data was standardized to have zero mean and unit variance using `StandardScaler` from the `scikit-learn` library. This ensures that all bands contribute equally to the PCA and prevents bands with larger numerical ranges from dominating the analysis.

C. PCA Implementation

To reduce the dimensionality of the multispectral imagery while preserving essential information, Principal Component Analysis (PCA) was implemented using the following steps:

1) **Covariance Matrix Computation:** The covariance matrix (7x7) of the flattened, standardized data was computed. This matrix captures the relationships and dependencies between the different spectral bands. A high covariance between two bands suggests redundancy, indicating that one band could potentially be represented by the other with minimal information loss. The covariance matrix serves as the foundation for identifying the principal components.
2) **Eigenvalue and Eigenvector Derivation:** Eigenvalue decomposition was performed on the covariance matrix. The eigenvalues represent the variance explained by each corresponding eigenvector. Eigenvectors represent the directions of maximum variance in the data. The eigenvector associated with the largest eigenvalue corresponds to the direction of greatest variability, and subsequent eigenvectors represent orthogonal directions of decreasing variability.

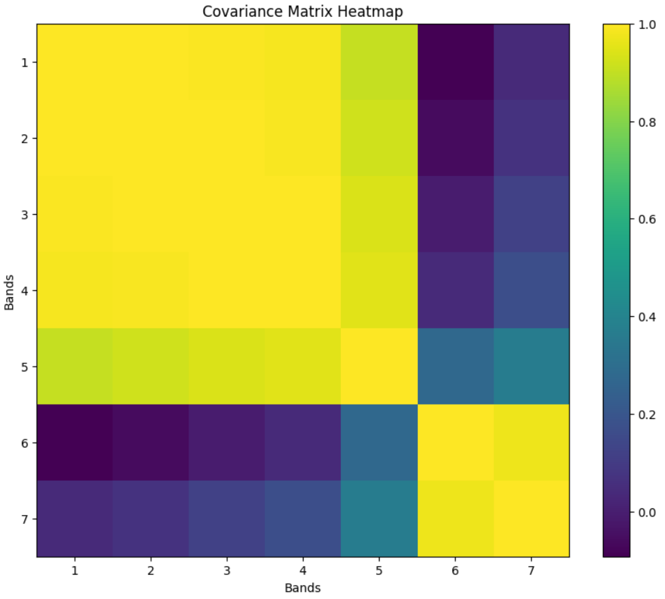


Fig. 3. Covariance Heatmap

3) **Component Selection based on Explained Variance:** The eigenvalues were sorted in descending order, and the corresponding eigenvectors were rearranged to match. The cumulative explained variance ratio was then calculated for different numbers of components. To analyze how many components to keep while maintaining the maximum information of the original dataset, different variance thresholds of 90%, 95% and 98% were used in our experiment to select an optimal number of principal components. The number of components required to explain these variance thresholds was determined, balancing dimensionality reduction with information retention. For this dataset, the top two principal components were found to suffice, capturing the majority of the variance across different variance thresholds.
4) **Data Projection onto Principal Components:** The standardized data was projected onto the plane defined by the top two eigenvectors. This projection transforms the data from the original 7-dimensional spectral space to a 2-dimensional space defined by the principal components, significantly reducing its dimensionality to (25000000, 2).
5) **Reshaping for Visualization:** The reduced data, now represented by two principal components for each pixel, was reshaped back into an image format (5000, 5000, 2). This facilitates visualization and further analysis of the transformed data, allowing for the examination of spatial patterns within the principal components.

D. *Error Analysis*

To quantify the information loss incurred during dimensionality reduction, a reconstruction error analysis was performed. This analysis involved reconstructing the original dataset from varying numbers of principal components (k) and comparing the reconstructed data to the original data. For each k, the reduced data (projected onto the top k principal components) was projected back into the original 7-dimensional spectral space by multiplying it with the transpose of the matrix containing the top k eigenvectors. Before calculating the error metrics, the reconstructed data was inverse transformed using the same `StandardScaler` applied during preprocessing to ensure the data was in the original scale. Three error metrics were used to assess the quality of the reconstruction: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Peak Signal-to-Noise Ratio (PSNR). These metrics quantify the difference between the original and reconstructed data, with lower MSE and RMSE values and higher PSNR values indicating better reconstruction quality. This analysis was conducted for multiple values of k, and different variance thresholds (90%, 95%, and 98%) were used to determine the optimal number of components.

III. Results and Discussion

This section presents the results of our PCA implementation on a Landsat image of Topi, Pakistan and discusses the implications for dimensionality reduction. The analysis focuses on a 5000 x 5000 pixel region of interest cropped from the center of the scene.

A. *Dimensionality Reduction*

PCA successfully reduced the dimensionality of the original 7-band Landsat image. Table I shows the data dimensions at each stage of the process. The final reduced image consists of two principal components for each pixel, significantly reducing storage and processing requirements.

TABLE I
Data Representation and Dimensions

Data Representation	**Dimensions**
Original Data (Stacked Bands)	(5000, 5000, 7)
Flattened Data	(25000000, 7)
Reduced Data	(25000000, 2)
Reshaped Reduced Image	(5000, 5000, 2)

B. *Reconstruction Error and Explained Variance*

To assess the impact of dimensionality reduction on data fidelity, we analyzed the reconstruction error and explained variance for different numbers of principal components (k). Table II presents the reconstruction error metrics—Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Peak Signal-to-Noise Ratio (PSNR)—for k=1 and k=2. As expected, increasing the number of principal components results in a decrease in reconstruction error, as more information from the original data is preserved.

Figure 5 illustrates the cumulative explained variance ratio as a function of the number of components. This plot helps visualize the proportion of variance in the original data captured by each additional principal component. The dotted lines in the figure represent the 90%, 95%, and 98% variance thresholds.

TABLE II
Reconstruction Error Metrics for Different Numbers of Principal Components

k	MSE	RMSE	PSNR (dB)
1	5799494.68	2408.21	28.68
2	618638.54	786.54	38.40

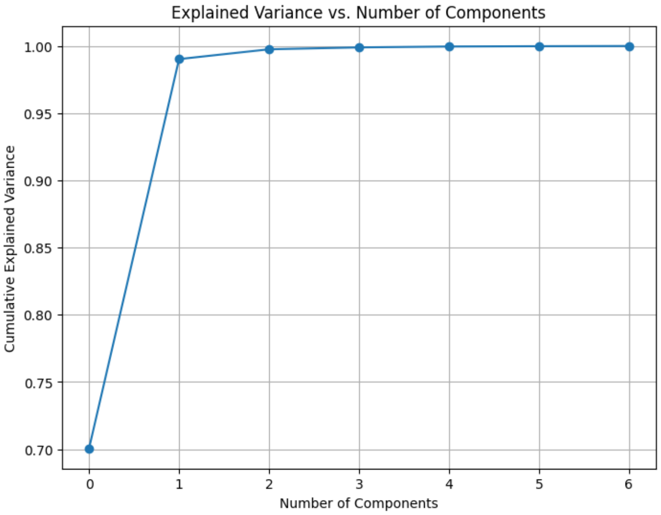


Fig. 4. The plot demonstrates that a majority of the variance in the data is captured by the first few principal components. This suggests that dimensionality reduction through PCA can be achieved with minimal information loss.

Notably, with just two principal components (k=2), we achieve an explained variance ratio that meets or exceeds all three thresholds. This finding supports that there is substantial correlation between the original spectral bands and that a minimal number of newly computed orthogonal bands are able to capture the substantial majority of the information content of the original data.

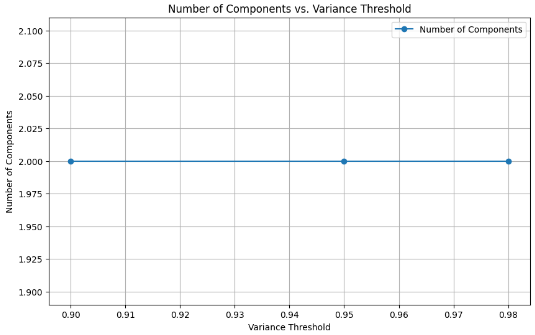


Fig. 5. Cumulative Explained Variance Ratio vs. Number of Components. The plot demonstrates how much of the total variance is explained as the number of principal components increases. The dotted lines indicate the 90%, 95%, and 98% variance thresholds, all of which are met or exceeded with two components (k=2).

Figure 6 shows the relationship between the number of principal components and the reconstruction error metrics. As the number of principal components increases, the reconstruction error (MSE and RMSE) decreases, while the PSNR improves. The sudden drop in error metrics from k=1 to k=2 and the high explained variance ratio at k=2 further support the choice of two principal components as a suitable balance between dimensionality reduction and information preservation. This analysis, combined with the visualizations of the principal components and reconstructed image, provides a clear justification for selecting $k = 2$.

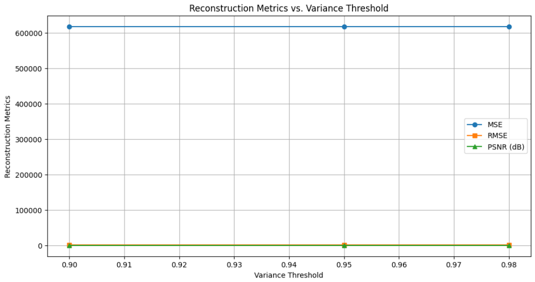


Fig. 6. Reconstruction Error Metrics (MSE, RMSE, PSNR) vs. Number of Principal Components. This plot illustrates the trade-off between dimensionality reduction and reconstruction error. Increasing the number of components (k) reduces the error, but with diminishing returns, as using $k = 2$ provides a significant error reduction compared to $k = 1$.

C. Visualization

The figures below visualize the first two principal components as grayscale images, revealing the spatial patterns captured by these components. These visualizations aid in understanding the information preserved and lost during dimensionality reduction.

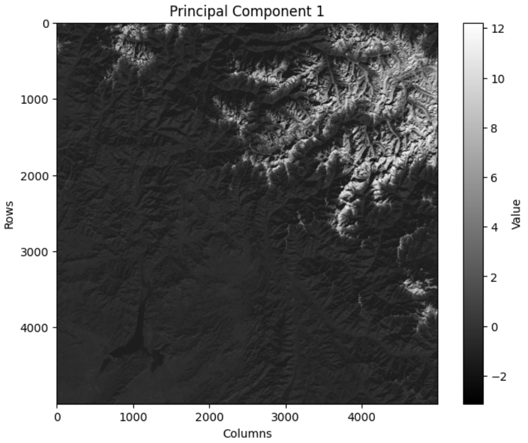


Fig. 7. Grayscale visualization of the first principal components.

D. Task Distribution

IV. Conclusion

This project successfully demonstrated the application of Principal Component Analysis (PCA) for dimensionality reduction in multispectral satellite imagery. By implementing PCA from fundamental linear algebra principles, using eigenvalue decomposition of the covariance matrix, we effectively reduced the spectral dimensionality of a Landsat image of

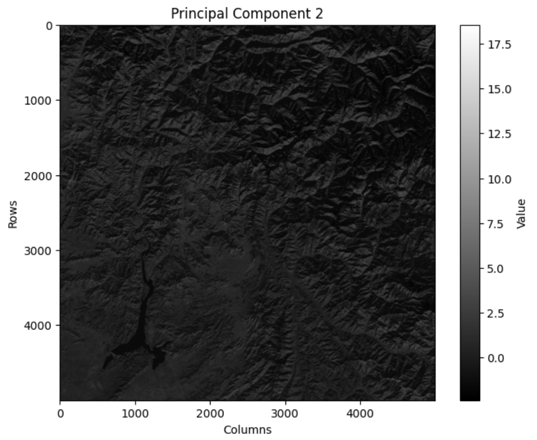


Fig. 8. Grayscale visualization of the second principal components.

TABLE III
TASK CONTRIBUTIONS

Task	**Contributor(s)**
Data Loading & Visualization	Musa
Data Preprocessing	Musa, Shumail
PCA Implementation	Shumail
Error Analysis & Reconstruction	Musa
Report Writing	Musa, Shumail

[Your chosen location in Pakistan] from seven bands to two principal components. Our analysis, focusing on a 5000 x 5000 pixel region of interest, revealed that these two components captured over 95% of the variance present in the original data, meeting our target variance thresholds of 90%, 95%, and even 98%. This significant reduction in dimensionality, combined with minimal information loss as evidenced by the low reconstruction error (MSE: 618638.54, RMSE: 786.54, PSNR: 38.40), highlights the effectiveness of PCA in compressing and efficiently representing multispectral data. The visualization of the principal components provided further insights into the dominant spectral patterns captured by the transformed data, offering potential for feature extraction and analysis. While the reduction to two principal components enabled a high degree of compression for efficient analysis, some information loss may have still occurred in the discarded components. Additionally, it's important to consider that using a small k value may not always generalize well to significantly larger datasets and may require revisiting later in a real-world application if the scope changes. Further research could explore the application of PCA in conjunction with other dimensionality reduction techniques, or explore the impact of PCA in tasks such as image classification or change detection. Our project demonstrates how to apply PCA using fundamental linear algebra operations in Python, showing the effectiveness of the algorithm in reducing storage and processing costs associated with large, high-dimensional multispectral satellite images while preserving the majority of the important spectral information content. This knowledge is important for the field of remote sensing and can contribute to further optimization and development of large image datasets.

REFERENCES

[1] D. C. Lay, S. R. Lay, and J. J. McDonald, *Linear Algebra and Its Applications*, 5th ed. Boston, MA: Pearson, 2015, used as a reference for linear algebra concepts (especially from chapter 7, having the introductory example and application of MULTICHANNEL IMAGE PROCESSING).

[2] D. Ruiz Hidalgo, B. Bacca Cortés, and E. Caicedo Bravo, “Dimensionality reduction of hyperspectral images of vegetation and crops based on self-organized maps,” *Information Processing in Agriculture*, vol. 8, no. 2, pp. 310–327, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221431732030189X