Exploring Sentinel-2 Imagery through Principal Component Analysis (Group 17)

.Sermad Mehdi 2021578 *Faculty of Computer Science and Engineering*  
*Ghulam Ishaq Khan Institute of Engineering Sciences and Technology*   
Topi, Khyber Pakhtunkhwa, Pakistan

Syed Basim Mehmood 2021613 *Faculty of Computer Science and Engineering*  
*Ghulam Ishaq Khan Institute of Engineering Sciences and Technology*   
Topi, Khyber Pakhtunkhwa, Pakistan

Syed Ibrahim Hamza 2021623 *Faculty of Computer Science and Engineering*  
*Ghulam Ishaq Khan Institute of Engineering Sciences and Technology*   
Topi, Khyber Pakhtunkhwa, Pakistan

*Abstract*—This paper presents a comprehensive analysis of Principal Component Analysis (PCA) applied to Sentinel-2 satellite imagery. The primary objective was to explore the application of PCA as a dimensionality reduction technique on remote sensing data, specifically focusing on Sentinel-2 multi-spectral images. Leveraging Python and the rasterio library, we initiated the project by preprocessing the Sentinel-2 images, stacking bands, and selecting a targeted region for analysis.

The methodology encompassed the implementation of PCA on the cropped image area, involving data reshaping, standardization, and the derivation of eigenvalues and eigenvectors. By systematically selecting principal components, we examined their impact on image reconstruction and information preservation.

Key findings include the successful application of PCA for dimensionality reduction in remote sensing imagery, evidenced through reconstructed images and individual band visualizations. The analysis of reconstruction error provided insights into the trade-off between reduced dimensions and information loss.

This study demonstrates the efficacy of PCA in distilling critical information from Sentinel-2 imagery while reducing dimensionality. The findings underscore the significance of PCA as a valuable tool for remote sensing data analysis, paving the way for enhanced information extraction and analysis from multi-spectral satellite imagery.

Keywords— Sentinel-2 images, Principal Component Analysis (PCA), Image Preprocessing, Information Loss, Image Reconstruction, Remote Sensing, Data Analysis, Image Processing

# INTRODUCTION

The fusion of Sentinel-2 satellite data and the principles of linear algebra, particularly Principal Component Analysis (PCA), marks a groundbreaking endeavor in modern remote sensing. Sentinel-2 imagery, comprising multi-spectral data, poses a challenge of high dimensionality, demanding sophisticated techniques for information distillation and interpretation.

At the crux of this project lies the application of PCA, a fundamental linear algebra concept, instrumental in decomposing the covariance matrix of multi-dimensional data into eigenvalues and their corresponding eigenvectors. This mathematical technique facilitates dimensionality reduction by selecting the principal components, thereby enabling a concise representation of the original data while preserving its intrinsic variance.

The significance of leveraging PCA on Sentinel-2 imagery extends beyond data reduction; it encapsulates the ability to unveil latent structures and distinguish significant spectral signatures within the vast data ensemble. By translating the complex relationships among spectral bands into eigenvalues and eigenvectors, PCA transcends mere reductionism, offering a pathway to comprehend and extract crucial information critical for diverse applications.

This report embarks on a rigorous exploration of PCA's application to Sentinel-2 imagery, intricately entwining linear algebra concepts with remote sensing methodologies. It delineates the theoretical underpinnings, methodological implementations, and the profound implications of PCA in the realm of multi-spectral remote sensing data analysis.

# CEP TASK AND METHODOLOGY

## A. Preprocessing Sentinel-2 Images

The preprocessing of Sentinel-2 images involved several pivotal steps to ensure data readiness for subsequent analysis:

1. **Data Collection:** The project initiated by sourcing Sentinel-2 satellite images stored in a designated folder. These images encompassed multiple spectral bands capturing diverse information about Earth's surface.
2. **Band Extraction:** Python, in conjunction with the rasterio library, facilitated the extraction of individual bands from the Sentinel-2 image collection. Each band represented specific spectral information, spanning various wavelengths across the electromagnetic spectrum.
3. **Stacking Bands:** Following extraction, the individual bands were stacked into a unified, multi-band image matrix. This stacking process amalgamated the spectral information, creating a comprehensive representation of the satellite data.
4. **Area Selection:** An interactive interface was implemented to enable the selection of a specific area of interest within the multi-band image. This interactive selection process allowed for the focused analysis of a targeted region, considering both analytical requirements and computational efficiency. Figure 1 shows the uncropped Sentinel 2 image in greyscale, we can select the area by cropping it and applying PCA on this.

A grey surface with white text

Description automatically generated with medium confidence

**Figure 1**

Loading, Stacking, and Cropping:

1. **Image Loading:** Leveraging the capabilities of the rasterio library, the code facilitated the loading of individual band files, providing access to their pixel-level data, Image shown in Figure 1.
2. **Band Stacking:** Stacking involved the consolidation of individual bands into a multi-dimensional array, effectively encapsulating the entire spectral range in a single comprehensive structure.
3. **Cropping for Analysis:** The interactive interface empowered users to delineate and select a region of interest within the stacked image. This approach facilitated a more targeted and focused analysis by isolating specific regions for in-depth examination, we can crop image shown in figure 1. Figure 2 shows the portion of image which is cropped from figure 1 for applying PCA on it.

**A close-up of a river

Description automatically generated**

**Figure 2**

Considerations and Challenges:

1. **Data Integrity:** Ensuring alignment and consistency among the multi-band images were paramount to prevent discrepancies during subsequent analyses.
2. **Computational Efficiency:** The selection of the area of interest considered computational resources, aiming to balance analytical depth without compromising computational feasibility.
3. **Quality Assurance:** Maintaining image quality during the cropping process was critical to ensure that the selected area represented the data's true characteristics..

## B. Principal Component Analysis (PCA)

PCA Methodology on Cropped Image:

1. **Data Reshaping:** The cropped image underwent reshaping to prepare it for PCA. The multi-dimensional cropped image was flattened to a two-dimensional array, aligning rows as observations and columns as features (pixels or spectral bands).
2. **Standardization And Scaling:** Standardization of the data was executed through Standard deviation and scaling. This process involved subtracting the mean of each feature (pixel) across observations and scaling by their respective standard deviations.
3. **Covariance Calculation:** The covariance matrix was computed from the standardized data. This matrix elucidated the relationships and variances among different spectral bands or pixels in the cropped image.
4. **Eigenvalues and Eigenvectors:** The eigenvalues and eigenvectors of the covariance matrix were derived through matrix decomposition. Eigenvalues represented the variance along principal components, while eigenvectors denoted the directions or axes of maximum variance.

Functions and Procedures in the Code:

1. **PCA Function:** The code encapsulated PCA within a function, orchestrating the step-by-step process outlined above, thereby encapsulating the mathematical operations required for PCA implementation.
2. **Basic Operations:** The code implements the basic and fundamental operations like cropping of Sentinel 2 image for applying PCA upon it. It further visualises all the bands in Sentinel 2 image and concatenates them to construct the image for applying PCA.
3. **Threshold Application:** A threshold value was introduced to retain only the most significant components and discard insignificant ones. This thresholding process involved setting a limit on the eigenvalues or components to retain, thus further aiding in dimensionality reduction and noise elimination.

Selection of Components:

1. **Number of Components:** The selection of the number of principal components involved a trade-off between retaining critical information and reducing dimensionality. This decision was guided by assessing the cumulative explained variance or considering application-specific requirements.

## C. Results and Analysis

A close-up of a grey surface

Description automatically generated

Figure

Results from PCA:

1. **PCA Applied Image**: Figure 3 shows the cropped image on which we applied PCA, we can clearly observe the difference from Figure 2, PCA applied image less details in the image and lightened iamge.
2. **Reconstructed Images:** The application of PCA yielded reconstructed images from the selected principal components. These images showcased the transformed representation of the original cropped image. The variations in the number of components used for reconstruction provided insights into the fidelity of the reconstructed images compared to the original. Here shows the Reconstruction Formula.

Figure 5 shows the Reconstructed image after applying PCA. Image successfully reconstructed back.

1. **Individual Bands Visualization:** Visualizations of individual bands derived from the principal components facilitated an understanding of each band's contribution to the overall image. This analysis highlighted the spectral information retained by different principal components. Figure 4 shows the 11 bands of cropped imaged used for PCA.

A group of images of band

Description automatically generated

Figure

Impact of Varying Components:

1. **Information Retention:** Observing the reconstructed images with varying numbers of principal components illustrated the trade-off between reduced dimensions and information preservation. Higher component numbers retained more information but risked overfitting, while lower numbers resulted in information loss but offered a more concise representation.
2. **Visual Interpretation:** The visual impact of varying components on reconstructed images and individual bands provided a qualitative understanding of the data transformation, enabling the identification of crucial spectral information preserved across different principal components.

A close-up of a river

Description automatically generated

Figure

Reconstruction Error and Information loss Analysis:

1. **Analysis of graphs:** The reconstruction error, computed between the original cropped image and its reconstructed counterpart, served as a quantitative metric to evaluate the efficacy of dimensionality reduction through PCA. Higher errors indicated a greater discrepancy between original and reconstructed images. Figure 6 illustrates the relation between Reconstruction errors and the numbers of components.

Information loss graph in Figure 7, clearly illustrating as number of components increasing resulting in least information loss, as reconstruction is happening.

A graph with a dotted line

Description automatically generated

Figure

A graph with a green line

Description automatically generated

Figure

1. **Implications:** Analyzing the reconstruction error and information recovered by reconstruction shed light on the effectiveness of PCA in retaining essential information while reducing dimensions. Understanding this error was crucial in assessing the suitability of selected principal components for meaningful analysis.

# Task Distribution

## A. Initial Research and Pre-Work Meeting:

## **Combined Research:** All team members engaged in collective research efforts, exploring concepts related to Sentinel-2 imagery, PCA applications, and their intersection with remote sensing. This joint effort ensured a shared understanding of the project's scope, goals, and methodologies.

1. **Pre-Work Meeting:** A comprehensive meeting was conducted before commencing individual tasks. The team discussed the research findings, outlined task allocations, and established a clear roadmap for the project. This meeting served as a foundation for collaborative efforts and task delineation.

## B. Task Division

1. **Sermad and Basim:** Collaboratively managed Preprocessing and PCA. Sermad focused on sourcing images and initial data preparation, while Basim concentrated on implementing PCA, including eigenvalue/eigenvector calculations and component selection.
2. **Ibrahim:** Dedicated to Results, Analysis, and Report. Ibrahim conducted in-depth analyses, computed reconstruction errors, and interpreted the implications of PCA components on information loss. Additionally, he took charge of finalizing the report, ensuring coherence and integrating overall findings.

## C. Collaboration and Coordination:

1. **Constant Communication:** The team maintained regular communication channels, sharing insights and progress updates. This facilitated aligned efforts and ensured that individual contributions were in harmony with the project's objectives.
2. **Code and Report Finalization:** All members participated in the finalization stage, reviewing and refining the codebase and report collectively. Ibrahim led the report finalization, integrating inputs from all members for a comprehensive and cohesive document.

# Conclusion

Our exploration into applying PCA to Sentinel-2 images unveils significant insights crucial for sophisticated image analysis.

1. **Main Findings and Outcomes:** PCA's Transformative Power: The application of PCA to Sentinel-2 images unveiled the potential for dimensionality reduction while preserving essential spectral information. The reconstructed images and individual band visualizations demonstrated PCA's efficacy in distilling vast spectral data into concise, informative components.
2. **Information Loss and Component Variation:** The analysis of varying components showcased the delicate balance between dimensionality reduction and information retention. This exploration illuminated the trade-off, emphasizing the need for an optimal number of components for meaningful analysis.
3. **Significance in Project Context:** Relevance in Remote Sensing: The findings underscore the significance of PCA in remote sensing applications, especially in processing multi-spectral imagery. PCA's ability to distill complex data into principal components holds promise for efficient data representation and subsequent analysis.
4. **Insights for Decision-Making:** The results offer insights into the impact of component selection on information loss. Understanding this trade-off is pivotal in guiding decisions related to data compression and analysis methodologies.

# References

1. Lay, David C., Steven R. Lay, and Judi J. McDonald. "Linear Algebra and its applications 5th edition." *Pearson* (2016). *(references)*
2. Mudrova, M., and Aleš Procházka. "Principal component analysis in image processing." *Proceedings of the MATLAB technical computing conference, Prague*. 2005.
3. Wikipedia contributors. "Earth observation satellite." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 23 Sep. 2023. Web. 17 Dec. 2023.
4. Wikipedia contributors. "Multispectral imaging." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 17 Nov. 2023. Web. 17 Dec. 2023.
5. Wikipedia contributors. "Principal component analysis." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 15 Dec. 2023. Web. 17 Dec. 2023.
6. Copernicus Documentation, "Sentinel Applications Platform - Browser,"https://documentation.dataspace.copernicus.eu/Applications/Browser.html. 15 Dec. 2023.