PCA and projection pursuits on high dimensional data reduction

Y Permanasasi, A N Falah, I Marethei, and B N Ruchjana

Department of Mathematics, Universitas Padjadjaran, Jalan Raya Bandung-Sumedang, km 21 Jatinangor, Sumedang 45363, Indonesia

*budi.nurani@unpad.ac.id

Abstract. The flood of multimedia data has affected various fields of research, including multimedia retrieval, database management, data mining, machine learning, social media analysis, image processing and so on. Many digital multimedia data collections are made in various fields and application contexts not only in advertising or journalism, but also often used in the context of cultural heritage applications, web search, or geographical information. Digital multimedia data is created in various formats and media modalities and available worldwide every day. The higher the quantity of data, the higher the complexity, diversity, dimensions and mutli modality, so that multimedia data grows exponentially. To reduce the impact of high-dimensional data, data reduction is carried out. Data reduction plays an important role in classification. Therefore the selection of data reduction techniques must be done carefully. Principal Component Transform is widely used in data reduction. This analysis attempts to eliminate correlations between bands and further determines the optimal linear combination of original data dimensions for variations in the value of data pixels. The mathematical principle of Principal Component depends on the decomposition of eigenvalues from the covariance matrix of the data matrix. When multimedia data is smaller or covered by data noise, Principal Component is no longer efficient at reducing. To overcome this, Projection Pursuit was introduced using sphering matrices and invariant translations in the transformations. In Projection Pursuits, only data that is considered interesting will be transformed. Optimization of reduced data selection is based on the maximum value of the projection index generated. The projection index used is Skewness and Kurtosis because it is considered in accordance with the data used, the CASI image of Bogor city.

1. Introduction

A lot of multimedia data can be accessed by everyone who needs it with various kinds of problems. As is not commonly known, technological advances have two opposing sides with their own advantages and disadvantages. The large amount of data available, of course, makes it easier for researchers to do research, while on the other hand, this large amount of data certainly requires sufficient time assistance even if it is assisted by a computer. This high-dimensional data also requires accurate and efficient data analysis techniques in addition to the dimensional curse, which is a lot of redundant or empty data.

Intuitively, what needs to be done to overcome this is to reduce the data dimension. In general, there are three different ways to perform data reduction, namely [1]: 1) Transformation, namely mapping feature dimensions by eliminating redundant and noise data or irrelevant data into smaller dimensions. 2) Selection, namely by selecting the features that are considered the most relevant. 3) Encoding, which encodes the features into a more efficient processing algorithm. This paper will focus on feature transformation.
This transformation or mapping from high-dimensional data to low-dimensional data aims to maintain the relevant features in the mapping results so that they can be used properly in the next recognition process. Principal Component Transform is a transformation that is often used in reducing dimensions [2]. Several studies have shown that PCT gives poor results in the pattern recognition process because the PCT projection always maintains its feature variations regardless of the factors that influence its appearance [3, 4]. The Projection Pursuit was introduced to overcome this.

Projection Pursuit (PP) is a multivariate exploratory data analysis to project a high-dimensional data set into a low-dimensional data space while retaining the desired information introduced by Friedman & Tuckey in 1974 [5]. The PP method is particularly suitable for the analysis of data sets with a large number of variables such as in high-dimensional spaces, because it involves a dimensional reduction step. The low-dimensional subspace is selected by maximizing a certain Projection Index. In this case, the Skewness and Kurtosis Projection Index will be selected because what become the problem in this paper is how to reduce data where the object to be recognized is much smaller than its background or commonly referred to as anomalies or outliers. PI skewness and Kurtosis are considered suitable for this case, because skewness and kurtosis are used in statistics to recognize outliers objects which are usually categorized as outliers or anomalies.

The purpose of this study is to examine the theoretical background of the stages of the Projection Pursuit process in implementing a high-dimensional data reduction mechanism using the Skewness Projection Index, as well as the Principal Component which utilizes an eigenvector matrix and the cumulative proportion of the contribution of eigenvalues to reduce data. High-dimensional data that will be used in the case study is the CASI image data for the city of Bogor with 12 bands obtained from the Indonesian National Aeronautics and Space Institution (LAPAN).

2. Method

2.1. Principal Component Transform

Data reduction method with Principal Component, is carried out based on the covariance matrix of the original images, with a set of feature images, namely X containing D features (D band), and each feature having N pixels (N observations), where the data is obtained from conversion of an analog image to a digital image, so that each pixel has a non-negative integer value, which is between 0 and 255, which indicates the gray level or often referred to as the gray level of the pixel concerned. From the existing D characteristics, the best d traits will be selected, namely the one that has the highest feature contribution [6].

The steps required to carry out the transformation with Principal Component Transform (PCT) on image data are as illustrated in the following flowchart [7]:

![Flowchart of principal component transform](image-url)
The first stage is data collection, because the CASI image is an analog image, it is necessary to convert the data to a digital image first, so that each pixel indicates a gray level or level of brightness (gray level) has a non-negative integer value, which is between 0 and 255. The second stage is to calculate the covariance of the matrix in order to calculate the eigenvalues in the third stage. The algebraic equation for the eigenvalue $\lambda$ corresponds to the eigenvector $v$ for the covariance matrix $S$ is as follows:

$$Sv = \lambda v$$

$$\iff Sv - \lambda v = 0$$

$$\iff (S - \lambda I)v = 0.$$  

Stage 4 performs the Principal Component transformation, that is calculating the eigenvectors of each original image band, determining the transformation matrix and then calculating the new image data in the fifth stage using the following formula [8]:

$$PCT = T * X^T \quad (1)$$

where:
- $PCT$: Transformation Result
- $T$: Transformation matrix
- $X$: Data Matrix.

The last is to select a new data set by selecting $d$ PCT features from $D$ new available image features ($d < D$), by looking at the contribution of each PCT feature image to the variance of the image as follows:

$$\text{Contribution of the } D\text{-th PCT feature} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_D}{L_D} \quad (2)$$

To choose the best $d$ PCT feature, then sort given $D: L_1, L_2, L_3, \ldots, L_D$ from largest to smallest or sort $D$ eigenvalues from largest to smallest, then select number $d$ from first largest to largest $d$, then the characteristic associated with the selected $d$ number is the desired best PCT feature.

2.2. Projection Pursuits

In principle, the flow of the Projection Pursuit (PP) data reduction method is the same as PCT, even in several papers it is stated that PCT is part of PP. The basic difference between the two methods is that PP uses the Projection Index for the selection of new data and the Sphering matrix as the transformation matrix. The Projection Index Skewness and Kurtosis as a measure for anomalous data exploration was considered in accordance with this study [9] [10]. The new image resulting from the transformation is formulated as follows:

$$PP = A * Y^T \quad (3)$$

where:
- $PP$: Transformation Result
- $A$: Transformation matrix
- $Y$: Data Matrix.

The transformation matrix used to find the transformation result is the Sphering $A$ matrix obtained by the following decomposition:

Assume that $\{\lambda_i\}_{i=1}^p$ is the eigenvalue for the covariance matrix $\Sigma$, and $\{v_i\}_{i=1}^p$ is the corresponding eigenvector. With $\Sigma$ definite non-negative, all eigenvalues are non-negative and there is a matrix $Q$ where $\Sigma$ can be decomposed as follows:

$$Q^T \Sigma Q = \Lambda. \quad (4)$$

Meanwhile, $Q = [v_1, v_2, \ldots, v_p]$ is a matrix consisting of eigenvectors $\{v_i\}_{i=1}^p$ and $\Lambda = \text{diag} \{\lambda_i\}_{i=1}^p$ is a diagonal matrix with $\{\lambda_i\}_{i=1}^p$ on the diagonal line.
Supposed $\Lambda^{1/2} = \text{diag}\left\{ \sqrt{\lambda_i} \right\}^{p}_{i=1}$, by multiplying it to the both sides of the equation (4) will yield:

$$\Lambda^{1/2} Q^T \sum Q \Lambda^{1/2} = I. \quad (5)$$

Then we get the sphering matrix $A$ as follows:

$$A = Q \Lambda^{1/2}. \quad (6)$$

While the data matrix used is the data centered matrix (invariant translation) [11]. If it is assumed that there are $N$ data, namely $\{x_{ki}\}_{k=1}^{N}$ each with the dimensions of $K$ and $X = [x_1, x_2, \ldots, x_N]$ is the matrix $K \times N$. With the assumption of data centering, the original matrix is transformed using the following transformation [12].

$$Y = X(I - (1/N) I^T). \quad (7)$$

While $I$ is the identity matrix $N \times N$, and $I = (I, I, \ldots I)^T$ is a column vector of $N$ dimension with all of its components being 1. If we look statistically, the process performed by equation (7) is as follows:

$$Y = X(I - (1/N) I^T) \quad (7)$$

$$Y = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NN} \end{pmatrix} \begin{pmatrix} 1/N & -1/N & \cdots & -1/N \\ 1/N & 1/N & \cdots & 1/N \\ \vdots & \vdots & \ddots & \vdots \\ 1/N & 1/N & \cdots & 1/N \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} 1/N & -1/N & \cdots & -1/N \\ 1/N & 1/N & \cdots & 1/N \\ \vdots & \vdots & \ddots & \vdots \\ 1/N & 1/N & \cdots & 1/N \end{pmatrix} \begin{pmatrix} x_1 - x \bar{x} \\ x_2 - x \bar{x} \\ \vdots \\ x_N - x \bar{x} \end{pmatrix}$$

the matrix equation above can be written as $Y = x_i - I/N \sum x_i = x_i - x \bar{x}$.

The selection of the best $k$ features is carried out using a genetic algorithm by optimizing the Projection Index for Skewness and Kurtosis.

### 3. Result and Discussion

The experimental data used Bogor imagery obtained from LAPAN. Data using experiments recorded in these 12 Bands is called CASI data. It has technical specifications as follows: Dim: 2678 x 1831 x 12 [BIL], Size: [Unsigned Int] 117,682,032 bytes. File Type: ENVI Standard Sensor Type: Unknown Byte Order: Host (Intel) Projection: UTM, Zone 48 South Datum: WGS 84, X Pixel: 1.00000 Meters, Wavelength: None Upper Left Corner: 1,1 Description: [Fri Jun 01 21:17:51 2001].

This data is an attitude data (roll, pitch and deviation) recorded by the aircraft and position data by GPS, corrected differential. Both data are corrected by combining each part to another (bonding point), then recombinig each of these bonding points into the Bakosurtanal digital topographical map, scale: 1 : 25000, the last photo is taken (Control Point). The last coordinate represents the Reality Numbers of
projection system DGN-95 UTM. The following Table 1 describes the spectral data used in the CASI data experiment.

Table 1. CASI spectral ranges

| Bands | Row data |
|-------|----------|
| 1     | 445.7nm+/-.5.7nm (rows 266-271) DNSRU:1000.000000 |
| 2     | 503.0nm+/-.5.8nm (rows 235-240) DNSRU:1000.000000 |
| 3     | 532.9nm+/-.5.8nm (rows 219-224) DNSRU:1000.000000 |
| 4     | 571.3nm+/-.4.9nm (rows 199-203) DNSRU:1000.000000 |
| 5     | 602.4nm+/-.5.8nm (rows 182-187) DNSRU:1000.000000 |
| 6     | 651.6nm+/-.5.8nm (rows 156-161) DNSRU:1000.000000 |
| 7     | 674.3nm+/-.5.9nm (rows 144-149) DNSRU:1000.000000 |
| 8     | 710.5nm+/-.5.9nm (rows 125-130) DNSRU:1000.000000 |
| 9     | 741.0nm+/-.5.9nm (rows 109-114) DNSRU:1000.000000 |
| 10    | 775.4nm+/-.5.9nm (rows 91-96) DNSRU:1000.000000 |
| 11    | 800.3nm+/-.5.9nm (rows 78-83) DNSRU:1000.000000 |
| 12    | 844.3nm+/-.5.9nm (rows 55-60) DNSRU:1000.000000 |

The data consists of 12 bands, after cropping to 200 rows and 300 columns in size to save memory. The trial scenarios carried out in this study include:

*Experiment 1*: Selecting the best 3 bands from 12 bands based on Principal Component Transform. The transformation uses the original data whose histogram is shown in Figure 2 below:

![Figure 2. Original data distribution histogram](image)

The three best bands are selected based on the 3 largest eigenvalues obtained, namely band 1, band 2, and band 3 with the percentage of eigenvalues obtained as follows:

Table 2. The results of data reduction using PCT

| Selected Band | Eigenvalues | Cumulative Eigen Proportion |
|---------------|-------------|----------------------------|
| 1 2 3         | 5,1997      | 97%                        |

The value of the cumulative proportion in Table 2 produced by PCT is the percentage of eigenvalues as a variant with a large enough value. This happened because in the PCT the difference in the value of the variants of each band was quite striking so that even if one band was chosen, the percentage was still large enough to believe that one band could represent all twelve bands.

Principal Component collects the largest variant value. In terms of data being images, the largest variant could be background images that are not required in the study. CASI data Table 1 shows an aerial photograph of an area with several features displayed such as grass, roads, trees, water, and buildings. The image is taken from the air so that the sky background may be recognized as the largest variant.
Experiment 2: Select the best 3 bands from 12 bands based on the Projection Pursuit; Selection of the best 3 bands based on the projection index of the largest skewness and kurtosis using a genetic algorithm. The data to be transformed first is translated using equation (7) to produce a histogram as shown in Figure 3 below:

![Figure 3. Histogram data after invariant translation](image)

The x-axis shows the degree of gray scale, while the y-axis is the frequency. In Figure 2, the degree of grayness of the original data is 0-255 with the highest frequency around 1200, whereas after being translated the degrees of gray are uniformed (standardized) to the mean so that the scale becomes -200 to 200 with the highest frequency around 1600. This shows that the data is standardized on a smaller scale.

The result of the Projection Pursuit is an optimization of the skewness and kurtosis values. The largest skewness value obtained is the sum of the absolute skewness values of each selected band, because the skewness value $= 0$ indicates normally distributed data while the negative and positive values have the same magnitude, only distinguishing the opposite direction of transformation. The cumulative proportion in Table 3 is the cumulative value of the selected band compared to the total cumulative skewness value of the entire band obtained.

| Selected Band | Skewness | Cumulative Proportions |
|---------------|----------|------------------------|
| 3 10 12       | 1,4064   | 28%                    |

Likewise with the value of kurtosis, because the value of kurtosis $= 3$ indicates normal, $0 < k < 3$ is negative kurtosis and $k > 3$ is positive kurtosis.

| Selected Band | Kurtosis | Cumulative Proportions |
|---------------|----------|------------------------|
| 1 3 6         | 2,6523   | 32%                    |

The skewness and kurtosis values in Table 3 and Table 4 produced by PP tend to be the same, so that if 3 bands are taken, the cumulative percentage is still below 50%, this indicates that the transformed data is more identical in distribution. This happens because PP uses invariant translation data. Therefore, if we want a higher cumulative percentage (above 50%) then the band selection must be more than three.

4. Conclusion
The fundamental difference between PCT and Projection Pursuit transformations is in taking the transformation matrix and the image data to be transformed. In PCT, the transformation matrix is an eigenvector matrix (V) while in Projection Pursuit, the transformation matrix is a sphering matrix (A). The data to be transformed in PCT is original data, while PP uses data that has been done using invariant
translation. Invariant translation is the standardization of data on a smaller scale so that it ensures that all data that is transformed will not change its characteristics after being transformed and can save memory.

PCT can not inform the existence of anomalies because the PCT algorithm only collects the largest variance or features in its main components. When the biggest feature is a background image that is not included in the study, PCT will fail to recognize the desired object. In contrast to the Projection Pursuit, with the Skewness and Kurtosis Projection Index, it can indicate outlier data or anomalies, so it can be used as a detection of targets or small objects.

The value of the cumulative proportion of eigenvalues on PCT is greater than the value of the cumulative proportion of skewness and kurtosis values in PP. This happens because PP uses more identical data so that it has almost the same skewness and kurtosis values for each band, unlike PCT which has very different eigenvalues in each band.

Acknowledgments
The authors thank to Rector Universitas Padjadjaran who provided financial assistance for the dissemination of research through Academic Leadership Grant 2020 with contract number: 1427/UN6.3.1/LT/2020. Also thanks to the reviewers for the valuable review of this paper.

References
[1] L 2017 Learning in High-Dimensional Multimedia Data: The State of The Art University of Electronic Science and Technology of China
[2] Pei Y 2015 Linear Principal Component Discriminant Analysis IEEE International Conference System Man and Cybermatic
[3] Departemen Ilmu Komputer dan Elektronika FMIPA 2018 Reduksi Dimensi Menara Ilmu Pengenalan Pola UGM
[4] Rui Z, Feiping N, and Xuelong L 2017 Auto-weight two-dimensional PCA with Robbust Outlier IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)
[5] Christian P 1995 Projection Pursuits Exploration Data Analysis Computational Statistics and Data Analysis Vol 20 No 6, pp 669-687
[6] Camacho J 2019 Group-Wise Principal Component Analysis for Exploratory Instrusion Detection IEEE Access Vol 7 pp 113081-113093
[7] Asha T and Roopa H 2019 A Linear Model Based on Principal Componen Analysys for Disease Prediction IEEE Access Vol 7 pp 105314-105318
[8] Pei Y 2015 Linear Principal Component Discriminant Analysis IEEE International Conference on System, Man, and Cybermatic
[9] Rencher A C and Christensen W F 2012 Methods of Multivariate Analysis WILEY
[10] Xiurui G, Meng L, Li L, Ji L, and Sun K 2015 Momentum Principal Skewness Analysis IEEE Geoscience and Remote Sensing Letters Vol 12 No 11 pp 2262-2266
[11] Akira T 2018 Kernel-Induced Sampling Theorem for Translation-Invariant Reproducing Kernel Hilbert Sapce with Uniform Sampling ICASSP 2018
[12] Fan C, Li Y, Kang G, and Li Y 2018 Learning Translation-Invariant Representation for Image Recognition with Drop Transformation Network IEEE Vol 6 pp 73357-7336