Effectiveness of GPCA in Reducing Data Dimensions and its Application to Human Development Dimension Indicators Data

Fahrezal Zubedi[1,2], I Made Sumertajaya[1*], Khairil Anwar Notodiputro[1], and Utami Dyah Syafitri[1]

[1]Department of Statistics, IPB University, Bogor, Indonesia
[2]Statistics Study Program, Universitas Negeri Gorontalo, Gorontalo, Indonesia
*Corresponding author: imsjaya@apps.ipb.ac.id

Received: 13 August 2024
Revised: 26 August 2024
Accepted: 3 September 2024

ABSTRACT – Analysis of human development growth at the regency/city level is challenging because the data is high-dimensional, indicators are correlated, and the regencies/cities are correlated. In this study, we propose a Generalized Principal Component Analysis to analyze human development growth by reducing the dimensions of regency/city and indicator. Thus, human development growth at the regency/city level is analyzed using the GPCA results in Biplot to describe each regency/city and its indicators. This study aims to evaluate GPCA in reducing the dimensionality of data whose observations are correlated, and indicators are correlated through simulation and empirical study; to analyze the growth of human development at the regency/city level based on the results of GPCA-Biplot. This research shows that GPCA works well in reducing data dimensions from correlated observations and correlated variables. Based on the results of the GPCA-Biplot visualization, the growth of human development in the Nduga regency from 2019 to 2022 showed significant fluctuations. Although some indicators show progress, especially in 2021, significant challenges remain. In the same way, the growth of human development in each regency/city can be analyzed. Thus, government policy focuses on real problems in the field.

Keywords– Biplot, GPCA, Human Development Growth, Procrustes

I. INTRODUCTION

Dimensionality reduction is a technique for reducing data dimensions so that the low-dimensional data formed can retain meaningful information from the original data. It is very efficient in the visualization and statistical analysis of high-dimensional data [1]. Principal Component Analysis (PCA) is one method for reducing data dimensions. PCA projects data in a d-dimensional space into a k-dimensional subspace, with k smaller than d. The resulting new set of dimensions is called Principal Components (PC) [2]. Furthermore, PCA is developed into Kernel PCA. KPCA is an extension of the standard PCA to nonlinear problems using the kernel method [3]. Then, PCA is developed into Generalized Principal Component Analysis (GPCA). The GPCA method was first proposed to obtain image compression and representation results. GPCA is used to reduce the dimensions of image data with the results of better visual quality and faster computation. This study shows the superiority of GPCA over PCA methods in image compression [4]. GPCA is often called Generalized Low-Rank Approximation of Matrices (GLRAM)[5] [6].

Various studies on GPCA have been conducted, including those on GPCA being applied to face identification. The idea is to learn low-rank approximation from raw-intensity face images such that the squared distance between all faces with the same identity should be smaller than those with different identities. In the research, GPCA was assisted by top-push constrained feature learning (TFL) in face identification [7]. Other research related to GPCA is that several data dimensionality reduction methods (including GPCA) have mathematically and experimentally evaluated the validity of reducing dimensions for calculating similarity in image pattern recognition. Image pattern recognition identifies instances of a particular object and distinguishes differences between images [8]. GPCA is also used to identify genes with overlapping patterns so that functions and interactions of genes can be identified. The experiment's results on gene expression pattern images show the effectiveness of GPCA in compressing biology images [9]. GPCA is also used to group multi-view data into several clusters based on the similarity between the data. The result of the study is that GPCA has superior performance in multi-view clustering compared to other clustering methods [10].

In the GPCA concept, data dimensions are reduced from a collection of observations (rows) correlated with each other and a collection of variables (columns) correlated with each other simultaneously. In addition, GPCA can simultaneously reduce a dataset's dimensions, provided that the dimensions of the dataset are the same [11]. Various studies on the GPCA method have been successfully applied to image data. In this research, GPCA is used on non-image data. Therefore, this research evaluates the effectiveness of the GPCA method in reducing the dimensions of non-image data from correlated observations and correlated variables using simulated data. On the empirical side, this study analyses the growth of social problems from year to year.

One of the topics that can be applied to the GPCA method is the analysis of human development growth from year to year. Measuring human development is called the Human Development Index (HDI). Indonesia's HDI increased from 71.92 in 2019 to 71.94 in 2020, then 72.29 in 2021. In the following year, the HDI increased to 72.91. Even though Indonesia's

HDI continues to increase, HDI growth in the last four years has slowed. Apart from that, gaps in human development achievements between the regencies/cities still exist [12–15]. This may be due to the policy focus not being in line with the real problems in the field. Therefore, it is necessary to comprehensively analyze the data on human development dimension indicators at the regency/city level from 2019 to 2022.

Data on human development dimension indicators at the regency/city level have properties dimensions of the indicators, and regency/city are high, between indicators are correlated, and between regencies/cities are correlated. Therefore, analyzing human development growth from year to year on high-dimensional data is challenging. The generalized Principal Component Analysis (GPCA) method is needed to obtain lower dimensions of indicator and regency/city, uncorrelated indicators, and uncorrelated regencies/cities. Thus, the growth of human development is analyzed based on data from dimension reduction using GPCA. With GPCA, the position of regencies/cities and indicators can be visualized yearly in a plot based on the Biplot concept. This way, the visualization plot will provide a specific description of each regency/city and its indicator so that the government's policy focus aligns with real problems in the field. Based on the problems above, this study aims to evaluate the GPCA method in reducing high-dimensional data from correlated observations and correlated indicators in empirical data and simulation data; analyze the growth of human development at the regency/city level from 2019 to 2022 based on visualization from GPCA-Biplot.

II. LITERATURE REVIEW

A. Generalized Principal Component Analysis

GPCA reduces the observations and variables dimension of the data. This means that GPCA not only reduces the dimensions of the variable but also considers the dimension of the observation, which is different from PCA, which only focuses on reducing the dimensions of the variables. GPCA aims to calculate two matrices, $\mathbf{L} \in \mathbb{R}^{r\,x\,l_1}$ and $\mathbf{R} \in \mathbb{R}^{c\,x\,l_2}$, with orthonormal columns to maximize the variance $(\mathbf{L}, \mathbf{R})$. An iterative procedure is used to obtain the optimal $\mathbf{L}$ and $\mathbf{R}$ matrices. To calculate $\mathbf{L}$, first calculate $\mathbf{R}$, which is obtained from the eigenvectors of the matrix $\mathbf{M_R}$. Calculate $\mathbf{L}$, which is obtained from the eigenvectors of the matrix $\mathbf{M_L}$ to obtain the matrix $\mathbf{R}$. This procedure is repeated until it converges. The solution depends on the initial choice, $\mathbf{L_0}$, for $\mathbf{L}$. Experiments show that choosing $\mathbf{L_0} = \mathbf{I} \in \mathbb{R}^{r\,x\,r}$, where $\mathbf{I}$ is the identity matrix, obtains excellent results. Given $\mathbf{L}$, $\mathbf{R}$, and $\{\mathbf{A_i}\}_{i=1}^{n}$, the projection of $\mathbf{A_i}$ by $\mathbf{L}$ and $\mathbf{R}$ can be calculated by $\mathbf{D_i = L^t A_i R}$. Conversely, given $\mathbf{L}, \mathbf{R}$ and $\{\mathbf{D_i}\}_{i=1}^{n}$, the estimated of original data $\{\mathbf{A_i}\}_{i=1}^{n}$ is obtained by the formula $\widehat{\mathbf{A}}_i \approx \mathbf{L\,D_i\,R^t}$. Checking the GPCA Algorithm uses the RMSE value [5]. More specifically, RMSE(i) and RMSE$(i-1)$ are the RMSE values in the i-th iteration and (i-1)-th iteration of the GPCA algorithm. Then, the convergence of the GPCA algorithm can be determined by checking RMSE$(i-1)$ − RMSE$(i) < \eta$. For some small thresholds, $\eta > 0$. The RMSE value in each GPCA iteration describes GPCA's performance in representing the original data in a lower dimensional space and shows the effectiveness of the dimension reduction process. Determining the number of principal components is based on the cumulative proportion of total variance. The selection of principal components from r observations in the eigenvector matrix $\mathbf{L}$ and principal components from c variables in the eigenvector matrix $\mathbf{R}$. This is based on the concept of PCA [16]. Like PCA, the standardization process is carried out because GPCA searches for principal components based on data variance. If features with high variance dominate, the resulting principal components will be more influenced by those features [17]. The GPCA algorithm in Table 1 has been added with a column standardization process [11].

Table 1 Algorithm GPCA

Algorithm 1. Generalized Principal Component Analysis

Input : $\mathbf{A_1}, \mathbf{A_2}, \dots, \mathbf{A_n} \in \mathbb{R}^{r\,x\,c}$, Original dataset

Output: $\mathbf{D_1}, \mathbf{D_2}, \dots, \mathbf{D_n} \in \mathbb{R}^{l_1\,x\,l_2}$, Low-dimensional data that represents a original data

$\widehat{\mathbf{A}}_1, \widehat{\mathbf{A}}_2, \dots\ \widehat{\mathbf{A}}_n \in \mathbb{R}^{r\,x\,c}$, Estimated original dataset

RMSE

1 Standardize data using standardization columns
2 Initialize the matrix $\mathbf{L_0}$ with the identity matrix $\mathbf{L_0} = (\mathbf{I}, 0)^{\mathrm{T}}$
3 $i = 0$, RMSE$(i) = \infty$
4 Form the matrix $\mathbf{M_R}$ with the formula $\mathbf{M_R} = \sum_{j=1}^{n} \mathbf{A_j^t L_i L_i^t A_j}$
5 $i = i + 1$
6 Calculate l_1 eigenvectors $\{\boldsymbol{\beta}_j^{\boldsymbol{R}}\}_{j=1}^{l_1}$ of $\mathbf{M_R}$ corresponding to the cumulative proportion of variance (≤ 90%) so that $\mathbf{R_i} = \left[\boldsymbol{\beta}_1^{\boldsymbol{R}}, \dots, \boldsymbol{\beta}_{l_1}^{\boldsymbol{R}}\right]$
7 Form the matrix $\mathbf{M_L}$ with the formula $\mathbf{M_L} = \sum_{j=1}^{n} \mathbf{A_j\,R_i R_i^t\,A_j^t}$
8 Calculate l_2 eigenvectors $\{\boldsymbol{\beta}_j^{\boldsymbol{L}}\}_{j=1}^{l_2}$ dari $\mathbf{M_L}$ orresponding to the cumulative proportion of variance (≤ 90%) so that $\mathbf{L_i} = \left[\boldsymbol{\beta}_1^{\boldsymbol{L}}, \dots, \boldsymbol{\beta}_{l_2}^{\boldsymbol{L}}\right]$
9 Calculate RMSE$(i) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} ||\mathbf{A_j} - \mathbf{L_i\,L_i^t A_j R_i R_i^t}||_F^2}$
10 Repeat steps 4 to 9 until (RMSE$(i-1)$ − RMSE$(i) \leq$ 0,001)
11 Obtain the optimal $\mathbf{L}$ and $\mathbf{R}$ matrices from the last iteration
12 Form low-dimensional data with the formula
$\mathbf{D_j = L^t A_j R}$ for each j from i to n
13 Form an estimate of the dataset (standardization scale)
$\widehat{\mathbf{A}}_\mathbf{i} = \mathbf{L\,D_j\,R^t}$ for each j from i to n
14 Transform the data units to the initial scale

GPCA can analyze the position of regencies/cities, and the position of indicators based on the Biplot concept. Biplot is a graphical visualization method used in multivariate data analysis to display information about data in two dimensions. The main purpose of the biplot is to provide a comprehensive description of the data structure and relationships between indicators [18]. Mathematically, the process of finding the coordinates of the origin points in the GPCA biplot display is the same as the concept of PCA Biplot [19]. The following is an explanation for creating a Biplot in GPCA.

Suppose $\widehat{\mathbf{A}}$ is a data matrix containing r observations and c indicators, then with GPCA, it can be written as:

$$\widehat{\mathbf{A}} = \mathbf{L}\,\mathbf{D}\,\mathbf{R^t} \quad (1)$$

the matrix column $\mathbf{L}$ contains the eigenvectors from the multiplication matrix $\mathbf{A\,R\,R^tA^t}$, and the column matrix $\mathbf{R}$ contains the eigenvectors from the multiplication matrix $\mathbf{A^t\,L\,L^t\,A}$. After decomposition using the Singular Value Decomposition technique, matrix $\widehat{\mathbf{A}}$ can be factored in the form:

$$\widehat{\mathbf{A}} = \mathbf{X}\;\mathbf{Y^t} \quad (2)$$

where matrix $\mathbf{X} = \mathbf{L}\,\mathbf{D}^{\alpha}$ and matrix $\mathbf{Y^t} = \mathbf{D}^{1-\alpha}\mathbf{R^t}$, where $\alpha = 0$. To plot the position of regencies/cities and the position of indicators in GPCA, use the principal component score matrix as follows:

$$\mathbf{X} = \mathbf{L}\,\mathbf{D}^{\alpha} \text{ and } \mathbf{Y} = (\mathbf{Y^t})^{\mathbf{t}} \quad (3)$$

matrix $\mathbf{X}$ contains the principal component scores, which are the coordinates of the r observations. matrix $\mathbf{Y}$ contains the principal component scores, which are the coordinates of the c indicators.

B. Procrustes Analysis

Procrustes analysis measures the similarity of each original data ($\mathbf{A}$) with the estimated original data ($\widehat{\mathbf{A}}$). The basic principle of Procrustes analysis is that the original data ($\mathbf{A}$) is taken as specified, and the estimated original data ($\widehat{\mathbf{A}}$) is transformed so that the two data are as close as possible. This transformation involves processes like translation (shifting the data in a specific direction), rotation (turning the data around a point), and dilation (scaling the data up or down) [20].

Translation in Procrustes analysis is shifting all points in the original data ($\mathbf{A}$) and estimated original data matrix ($\widehat{\mathbf{A}}$) by a fixed distance and in the same direction so that both data configurations have the same centroid. The minimum distance between original data matrix ($\mathbf{A}$) and estimated original data matrix ($\widehat{\mathbf{A}}$) after the translation process is carried out as follows [21]:

$$D_T(\mathbf{A},\widehat{\mathbf{A}}) = D(\mathbf{A}_{\mathrm{T}},\widehat{\mathbf{A}}_{\mathbf{T}})$$
$$D_T(\mathbf{A},\widehat{\mathbf{A}}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left[(a_{ij} - \bar{a}_j) - (\hat{a}_{ij} - \bar{\hat{a}}_j)\right]^2 \quad (4)$$

$\mathbf{A}_{\mathrm{T}}$ is the translation result matrix from data $\mathbf{A}$ and $\widehat{\mathbf{A}}_{\mathbf{T}}$ is the translation result matrix from data $\widehat{\mathbf{A}}$.

Rotation is the movement of all points at a fixed angle without changing the distance of each point to its centroid. Rotation transformation is done by multiplying the matrix $\widehat{\mathbf{A}}_{\mathbf{T}}$ with an orthogonal matrix $\mathbf{Q}$, which minimizes the distance between data. To obtain the minimum $D_{TR}(\mathbf{A}_{\mathbf{T}},\widehat{\mathbf{A}}_{\mathbf{T}})$ value, an orthogonal matrix $\mathbf{Q} = \mathbf{VU^t}$ must be chosen from the decomposition of the singular values of $(\mathbf{A}_{\mathbf{T}})^{\mathbf{t}}\widehat{\mathbf{A}}_{\mathbf{T}} = \mathbf{ULV^t}$ so that the optimal distance after the rotation process is [22]

$$D_{TR}(\mathbf{A},\widehat{\mathbf{A}}) = \mathrm{trace}\left(\widehat{\mathbf{A}}_{\mathbf{T}}{}^{\mathbf{t}}\,\widehat{\mathbf{A}}_{\mathbf{T}}\right) + \mathrm{trace}(\mathbf{A}_{\mathbf{T}}{}^{\mathbf{t}}\,\mathbf{A}) - 2\,\mathrm{trace}(\mathbf{L}) \quad (5)$$

Dilation is data scaling by increasing or decreasing the distance of each point in the configuration to its centroid. The dilation transformation is carried out by multiplying the matrix $\widehat{\mathbf{A}}_{\mathbf{T}}\mathbf{Q}$ by a scalar k so that the configuration after the dilation transformation will be $k\widehat{\mathbf{A}}_{\mathbf{T}}\mathbf{Q}$. To obtain the minimum $D_{TRD}(\mathbf{A},\widehat{\mathbf{A}})$ value, k can be chosen as follows [21]:

$$k = \frac{\mathrm{trace}\,(\mathbf{A}_{\mathbf{T}}{}^{\mathbf{t}}\widehat{\mathbf{A}}_{\mathbf{T}}\widehat{\mathbf{A}})}{\mathrm{trace}(\widehat{\mathbf{A}}_{\mathbf{T}}{}^{\mathbf{t}}\,\widehat{\mathbf{A}}_{\mathbf{T}})} \quad (6)$$

thus, obtained:

$$D_{TRD}(\mathbf{A},\widehat{\mathbf{A}}) = \mathrm{trace}\,(\mathbf{A}_{\mathbf{T}}^{\mathbf{t}}\mathbf{A}_{\mathbf{T}}) - \frac{\mathrm{trace}^2(\mathbf{A}_{\mathbf{T}}^{\mathbf{t}}\widehat{\mathbf{A}}_{\mathbf{T}}\mathbf{Q})}{\mathrm{trace}(\widehat{\mathbf{A}}_{\mathbf{T}}^{\mathbf{t}}\widehat{\mathbf{A}}_{\mathbf{T}})} \quad (7)$$

The R-square value (R^2) is used to measure the similarity of data configurations in Procrustes analysis. The percentage of the two configurations that can be considered the same is shown by R^2. As the value gets closer to 100%, the similarity of the data configuration gets higher. The formula for calculating R^2 is as follows [21]:

$$R^2 = 1 - \frac{D_{TRD}(\mathbf{A},\widehat{\mathbf{A}})}{\mathrm{trace}\,(\mathbf{AA^t})} \quad (8)$$

III. METHODOLOGY

A. Data

This research uses simulation data and empirical data. The simulation data used is 4 (four) data matrices ($\mathbf{A_j} \in \mathbb{R}^{500\,x\,100}, j = 1,2,3,4$). Each data matrix is designed as follows, consisting of 25 groups of observations and 5 groups of variables. Observations in the same group are correlated, while observations between groups are not correlated. The same applies to variables. In addition, the simulation data fulfills the condition that the matrices are correlated. The steps for generating simulation data are as follows:

1) Generating a covariance matrix $\mathbf{S} \in \mathbb{R}^{20\,\mathrm{x}\,20}$ that has symmetry and positive definite properties

$$\mathbf{S}_{20\,\mathrm{X}\,20} = \begin{pmatrix} 1 & 0{,}8 & \dots & 0{,}8 \\ 0{,}8 & 1 & \dots & 0{,}8 \\ \vdots & \vdots & \ddots & 0{,}8 \\ 0{,}8 & 0{,}8 & \cdots & 1 \end{pmatrix}$$

2) Doing Cholesky decomposition on matrix **S** aims to obtain the lower triangular matrix **P** and the upper triangular matrix $\mathbf{P^t}$. The matrix **P** has the property that observations are correlated. The matrix $\mathbf{P^t}$ has the property that variables are correlated. The Cholesky decomposition can only be applied to symmetric and positive definite matrix. a matrix is said to be positive-definite if all its eigenvalues are positive. If a matrix does not satisfy this condition, Cholesky decomposition cannot be performed [23].
3) Generating a matrix $\mathbf{Z} \in \mathbb{R}^{500 \text{ X } 20}$ using a partition approach. Each partition matrix is generated by a multivariate normal distribution with a mean vector (μ) and a covariance matrix (diagonal). The mean vector between columns is different.
4) Transforming each partition matrix from matrix $\mathbf{Z} \in \mathbb{R}^{500 \text{ x } 20}$ to matrix $\mathbf{A} \in \mathbb{R}^{500 \text{ x } 20}$ using the results of Cholesky decomposition by:
   Partition 1 $\quad \mathbf{A} \in \mathbb{R}^{20 \text{ X } 20} = \mathbf{P}_{20 \text{ x } 20} \text{ x } \mathbf{Z}_{20 \text{ x } 20} \text{ x } \mathbf{P}^{\mathbf{t}}_{20 \text{ x } 20}$
   ⋮
   Partition 25 $\quad \mathbf{A} \in \mathbb{R}^{20 \text{ x } 20} = \mathbf{P}_{20 \text{ x } 20} \text{ x } \mathbf{Z}_{20 \text{ x } 20} \text{ x } \mathbf{P}^{\mathbf{t}}_{20 \text{ x } 20}$
   Next, the transformation results of the partition matrices are combined (arranged downwards) to obtain a matrix $\mathbf{A} \in \mathbb{R}^{500 \text{ x } 20}$, which forms 25 groups of observations where observations in the same group have correlation properties and observations in different groups are uncorrelated.
5) Steps 3) to 4) are repeated four times. Then, it is combined (arranged on the right side) at matrix $\mathbf{A} \in \mathbb{R}^{500 \text{ x } 20}$ to obtain five groups of variables, where variables in the same group have correlation properties and variables in different groups are uncorrelated. So, matrix $\mathbf{A_1} \in \mathbb{R}^{500 \text{ X } 100}$ will be formed.
6) Steps 3 to 5) are repeated three times to obtain $\mathbf{X_i} \in \mathbb{R}^{500 \text{ x } 100}, \text{i} = 1,2,3$. Then, for the case of correlated matrices $\mathbf{A_j} \in \mathbb{R}^{500\,x\,100}, j = 1,2,3,4)$, as follows:
   $\mathbf{A_1} + \mathbf{X_1} = \mathbf{A_2}$
   $\mathbf{A_2} + \mathbf{X_2} = \mathbf{A_3}$
   $\mathbf{A_3} + \mathbf{X_3} = \mathbf{A_4}$

The empirical data used is indicators of human development dimensions at Indonesia's regency/city level from 2019 to 2022. The data used comes from the Statistic Indonesia (BPS) publication at https://www.bps.go.id /indicator/. This data has 20 indicators with a total of 514 observations. It is known that between indicators are correlated, and between regencies/cities are correlated. Table 2 shows the indicators used in this research [12–15].

Table 2 Indicators of Human Development Dimensions

Dimensions	Indicators (Unit)	Code
Long life and health life	Households with clean drinking water sources (%)	X_1
	Households that have access to adequate drinking water (%)	X_2
	Households that do not have defecation facilities (%)	X_3
	Morbidity (%)	X_4
Knowledge	School enrollment rates 7-12 years (%)	X_5
	School enrollment rates 13-15 years (%)	X_6
	School enrollment rates 16-18 years (%)	X_7
	Gross participation rates level ES/MI (%)	X_8
	Gross participation rates level JHS/MTs (%)	X_9
	Gross participation rates level SHS/VHS/MA (%)	X_{10}
	Net participation rates level ES/MI (%)	X_{11}
	Net participation rates level JHS/MTS (%)	X_{12}
	Net participation rates level SHS/VHS/MA (%)	X_{13}
The decent standard of living	Percentage of formal workers (%)	X_{14}
	Percentage of poor people (%)	X_{15}
	Open unemployment rate (%)	X_{16}
	Average wages of workers, and employees per month (rupiah)	X_{17}
	Gross regional domestic product per capita based on current prices (thousand rupiah)	X_{18}
	Percentage of informal workers (%)	X_{19}
	Gini ratio	X_{20}

B. *Data Analysis Procedures*

GPCA is applied to simulated data and empirical data. The simulation data analysis is used to study the effectiveness of GPCA in reducing the dimensionality of data whose observations correlated, and variables are correlated. The simulation data analysis procedure is as follows:

1) Preparing simulation data in the form of a matrix consisting of 4 data matrices ($\mathbf{A_j} \in \mathbb{R}^{500\,x\,100}, j = 1,2,3,4$)
2) Applying GPCA steps based on Table 1 to reduce the dimensions of observations and dimensions of variables
3) Generating RMSE and dimension from low-dimensional data
4) Estimating original data $\{\widehat{\mathbf{A}}_i\}_{i=1}^{4}$ based on the reduced data matrix $\{\mathbf{D}_i\}_{i=1}^{4}$ and the optimal **L** and **R** matrices, with the formula $\widehat{\mathbf{A}}_i = \mathbf{L}\,\mathbf{D}_j\,\mathbf{R^t}$ for each j from i to n

5) Measuring the similarity (R^2) of each original data with the estimated original data using Procrustes analysis
6) Steps 1) to 5) are repeated 100 times
7) Presenting the final RMSE value of each repetition in Boxplot form
8) Presenting the value of R^2 of each repetition in Boxplot form
9) Evaluating the GPCA method based on simulation results

The steps of the empirical data analysis procedure using GPCA are as follows:

1) Doing data description and data exploration. Data description is concise statistics such as mean, median, minimum, maximum, and standard deviation. Data exploration is used to identify relationships between indicators, which can provide valuable insights into correlations and dependencies between indicators.
2) Applying GPCA steps based on Table 1 to reduce the dimensions of regencies/cities and dimensions of indicators
3) Measuring the similarity (R^2) of each original data with the estimated original data using Procrustes analysis
4) Evaluating the performance of GPCA using RMSE and R-square value
5) Creating a biplot visualization from the result of GPCA
6) Interpreting of GPCA results on empirical data.

IV. RESULTS AND DISCUSSIONS

A. Simulation Study

The simulation dataset has dimensions of $500\ x\ 100$. Each data is divided into 25 groups of observations and 5 groups of variables. This means that each group of observation consists of 20 observations, and each group of variable consists of 20 variables. In this structure, the observations in the same group are correlated, indicating a similar relationship or pattern between the observations in one group. However, the observations in different groups are not correlated, meaning there is no similar relationship or pattern between the observations in different groups. Like the observations, the variables in the same group are also correlated, indicating a similar pattern or relationship between the indicators. However, the indicators in different groups are not correlated, meaning there is no similar relationship or pattern between the indicators in different groups.

Table 3 GPCA Result

Repetition	Dimension of matrix $\mathbf{A}_j$	Dimension of matrix L	Dimension of matrix R	Dimension of matrix $\mathbf{D}_j$
1	500 x 100	500 x 25	100 x 5	25 x 5
2	500 x 100	500 x 25	100 x 5	25 x 5
3	500 x 100	500 x 25	100 x 5	25 x 5
⋮	⋮	⋮	⋮	⋮
100	500 x 100	500 x 25	100 x 5	25 x 5

Table 3 shows the data dimensions of **L**, **R**, and $\mathbf{D_i}$ from the Generalized Principal Component Analysis (GPCA) results, which were repeated 100 times. Matrix **L** obtained in each replication has dimensions of 500 × 25. This matrix maps the original data from a 500-dimensional space to a 25-dimensional space, representing the observation structure. Matrix R obtained in each replication has dimensions of 100 × 5. This matrix maps data from a 100-dimensional space to a 5-dimensional space, representing the variable structure. Matrix **D** is the result of dimension reduction data with dimensions of 25×5. This matrix **D** represents data that has been reduced from the initial dimensions of 500×100 to a lower dimension of 25×5. This result is in line with the data structure that has been designed. The original data was designed with 25 observation groups and 5 variable groups. The study results show that GPCA works well in terms of its ability to reduce the dimensionality of data from correlated observations and correlated variables.

It is known that 25 principal components are formed from the **L** matrix and 5 principal components from the **R** matrix. Twenty-five principal components represent the observation structure, and five principal components represent the variable structure. Furthermore, it is determined which observations are included from PC 1 to PC 25 and which variables are included from PC 1 to PC 5 using the loading value. Loading with the largest value means having a major role in the PC. Based on the results, the observations that enter PC 1 to PC 25 follow the designed data structure, and the variables that enter PC1 to PC5 follow the designed data structure. This process is repeated 100 times. Each repetition produces the same results. GPCA has an excellent ability to detect and separate observation and variable group structures in data.

Based on Figure 1, the median RMSE value of around 1.90 indicates that the average error produced by GPCA in reducing the data dimensionality is relatively low. The range of RMSE values of around 1.80 to 2.05 indicates that the results of dimensionality reduction by GPCA are consistent, with small variations in errors between repetitions.

Based on Figure 2, the results of the Procrustes analysis for 100 replications show that the original data estimates ($\widehat{\mathbf{A}}_1$, $\widehat{\mathbf{A}}_2$, $\widehat{\mathbf{A}}_3$, $\widehat{\mathbf{A}}_4$) have very high similarity to the original data ($\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$, $\mathbf{A_4}$), with R-squared values generally above 0.984. The high median R-squared value and narrow interquartile range indicate that the GPCA method is consistent and reliable in producing estimates similarity to the original data. The few outliers that appear in pairs $\mathbf{A_3}$ & $\widehat{\mathbf{A}}_3$ and $\mathbf{A_4}$ & $\widehat{\mathbf{A}}_4$

indicate some cases where the similarity value is slightly lower, but this is not significant compared to the overall excellent results.

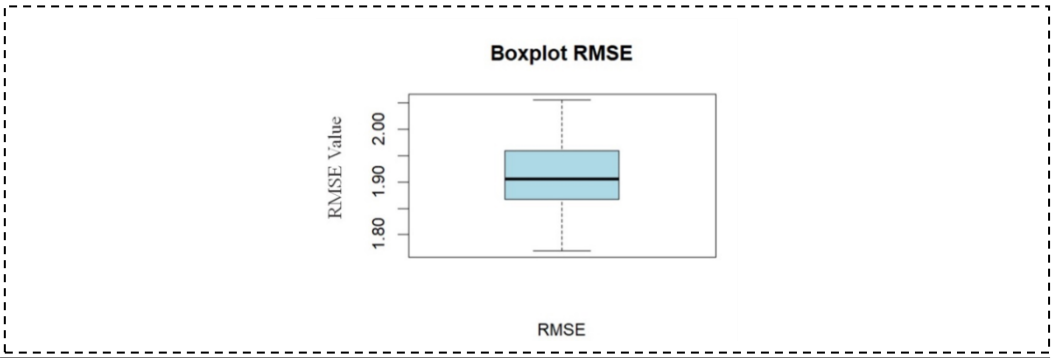

Figure 1 Boxplot RMSE

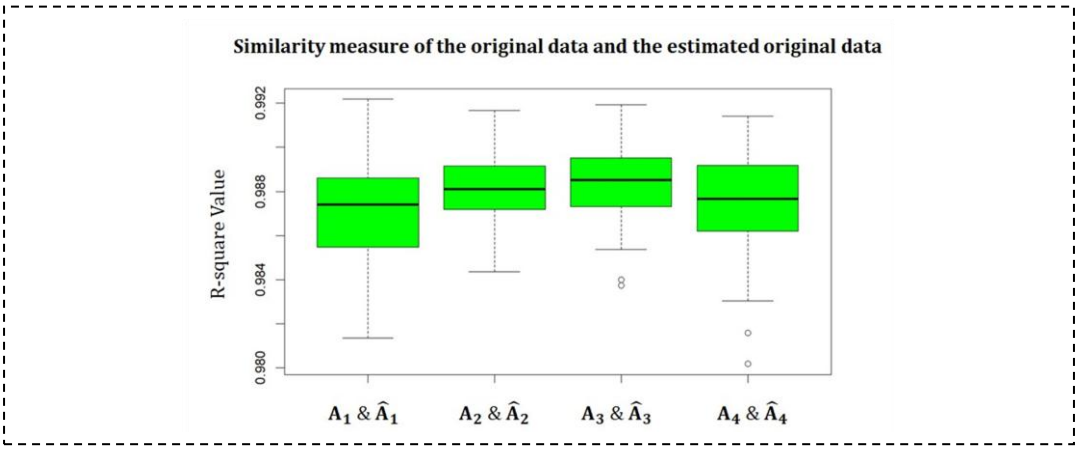

Figure 2 Boxplot of R-Square Value

B. Empirical Study

The data description of the human development dimension indicators is shown in Tables 4 and 5. X_{17} and X_{18} have different units from other indicators. Based on Tables 4 and 5, each indicator has very different maximum and minimum values. Based on this statement, this data may have outliers. The mean and median of each indicator have different values. Based on this statement, outliers in the data are also possible. The median value higher than the mean value indicates a possible negative skew in the data distribution. For example, in indicator X_2 in 2020. a high standard deviation indicates that the dataset values are spread further from the average. Conversely, a small standard deviation indicates that the data is more concentrated around the average. For example, in indicator X_{20} in 2021, a standard deviation value of 15.21 indicates that most of the dataset values are 62.63 ± 15.21.

Table 4 Descriptive Statistics of Human Development Dimension Indicators Data in 2019 and 2020

Code	2019					2020				
	Mean	Median	Minimum	Maximum	Standard Deviation	Mean	Median	Minimum	Maximum	Standard Deviation
X_1	70.30	73.33	0.29	100	21.60	70.35	73.97	0.29	99.84	20.58
X_2	73.49	82.71	1.06	100	25.27	67.57	74.76	2.40	100	28.03
X_3	9.82	6.73	0	93.36	10.9	9.61	5.94	0	90.87	11.54
X_4	16.40	15.21	0.77	51.09	6.81	15.67	14.41	0.06	51.05	7.1
X_5	98.33	99.53	52.21	100	4.91	98.38	99.48	52.22	100	4.741
X_6	94.94	96.30	30.44	100	6.22	94.97	96.41	32.43	100	6.16
X_7	74.36	74.50	28.20	97.72	10	73.90	74.39	22.55	98.17	10.4
X_8	107.82	108.51	53.55	126.17	7.06	107.19	107.75	58.82	126.78	6.75
X_9	90.46	91.08	14.54	134.97	12.04	91.69	92.36	35.71	125.32	10.20
X_{10}	85.10	85.68	0.71	154	17.77	85.83	86.35	12.69	131.83	15.55
X_{11}	96.22	98.23	43.14	100	6.14	96.59	98.42	48.88	99.88	5.69
X_{12}	76.90	78.05	12.08	98.92	10.77	77.77	78.96	12.61	97.45	10.21
X_{13}	62.16	62.63	1.13	91.26	12.37	62.84	63.20	6.46	98.89	12.19
X_{14}	39.39	37.64	0	74.46	15.5	35.32	33.60	0	71.77	14.4

Code	2019					2020				
	Mean	Median	Minimum	Maximum	Standard Deviation	Mean	Median	Minimum	Maximum	Standard Deviation
X_{15}	12.12	9.88	1.68	66.21	8.12	11.95	10.09	2.02	41.76	7.47
X_{16}	4.40	3.99	0	12.37	2.28	5.55	4.90	0.21	15.92	2.73
X_{17}	2579261.12	2409472.50	639908	5787902	749871.03	2527590.45	2359009	1398405	5926620	720009
X_{18}	31403.47	12072.76	207.47	699838.12	70597.82	30871.10	12093.39	226.97	700985.69	69010.69
X_{19}	0.32	0.32	0.19	0.48	0.04	0.32	0.32	0.19	0.47	0.05
X_{20}	59.43	61.94	7.59	100	16.23	63.53	65.71	8.47	100	15.32

Table 5 Descriptive Statistics of Human Development Dimension Indicators Data in 2021 and 2022

Code	2021					2022				
	Mean	Median	Minimum	Maximum	Standard Deviation	Mean	Median	Minimum	Maximum	Standard Deviation
X_1	70.16	73.08	0.30	99.79	20.16	69.52	72.57	0.33	100	20.18
X_2	84.94	89.74	0.87	100	15.87	86.42	91.32	1.33	100	14.44
X_3	8.22	4.75	0	100	10.65	8.32	5.03	0	87.33	9.80
X_4	12.38	10.45	0.13	53.30	7.35	19.42	13.98	0.13	97.31	18.27
X_5	98.26	99.35	51.61	100	4.84	98.08	99.44	34.51	100	5.97
X_6	94.83	96.49	31.98	100	6.54	93.83	96.45	23.41	99.98	9.33
X_7	75.08	75.34	27.71	99.98	10.46	72.81	74.78	11.14	99.98	15.56
X_8	106.61	107.16	59.58	129.54	6.90	106.58	107.11	57.06	126.99	7.08
X_9	92.62	92.45	36.67	135.50	9.97	91.87	91.30	37.06	125.35	11.03
X_{10}	87.71	87.60	9.40	143.34	16.62	87.49	87.48	15.12	143.48	14.99
X_{11}	96.41	98.41	0.87	99.98	7.05	96.61	98.22	51.96	99.99	5.95
X_{12}	78.15	79.44	14.58	97.70	10.15	95.76	79.60	16.05	99.53	10.33
X_{13}	63.02	63.22	8.61	99.39	12.16	63.31	63.70	9.99	99.4	11.84
X_{14}	36.17	34.49	0	75.55	14.41	36.05	33.67	0	72.04	15.07
X_{15}	12.27	10.46	2.38	41.66	7.45	11.68	9.82	2.28	42.03	7.27
X_{16}	5.06	4.57	0	13.37	2.63	4.62	4.31	0.12	11.82	2.32
X_{17}	2411408.11	2296867	1269597	5617088	615182.27	2618555.04	2452670	1382402	6989775	727767.398
X_{18}	33142.38	13112.36	242.95	728386.10	73114.01	37350.28	14866	256	794936	80968.988
X_{19}	0.32	0.32	0.20	0.52	0.05	0.32	0.31	0.18	0.54	0.05
X_{20}	62.63	64.90	9.05	100	15.21	62.81	65.34	8.75	100	15.99

Based on the correlation heatmap from each year, as shown in Figure 3, several indicators have a strong positive correlation, especially in the same group. Several indicators show a significant negative correlation, indicating that the other indicator tends to decrease when one indicator increases. For example, in the 2019 data, indicators X_5, X_6, X_8, X_{11} and X_{12} strongly correlate. Indicators X_3, X_{15} and X_{20} negatively correlate with indicators X_1, X_{14} and X_{16}.

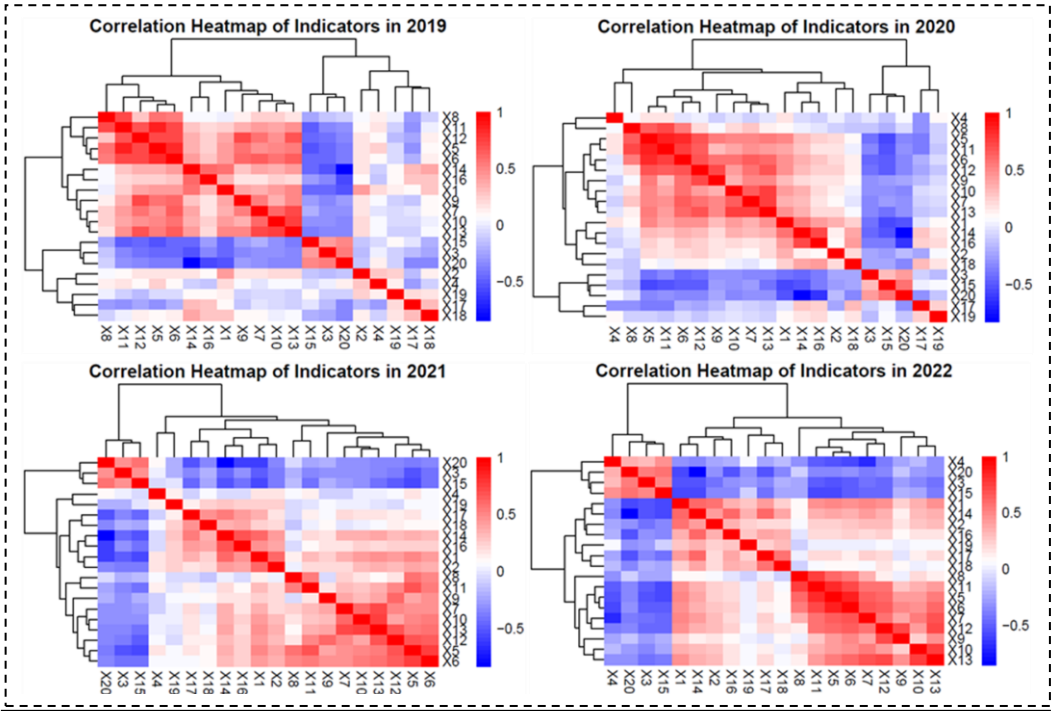

Figure 3 Heatmap of Correlations Between Observations

Based on Figure 4, each heatmap shows that most observations have a very high correlation, as indicated by the dominant red color in all years. High correlations indicate regional patterns or uniform policies that affect human development dimension indicators in many regencies/cities. Although high correlations indicate similarities, it is essential to conduct further analysis.

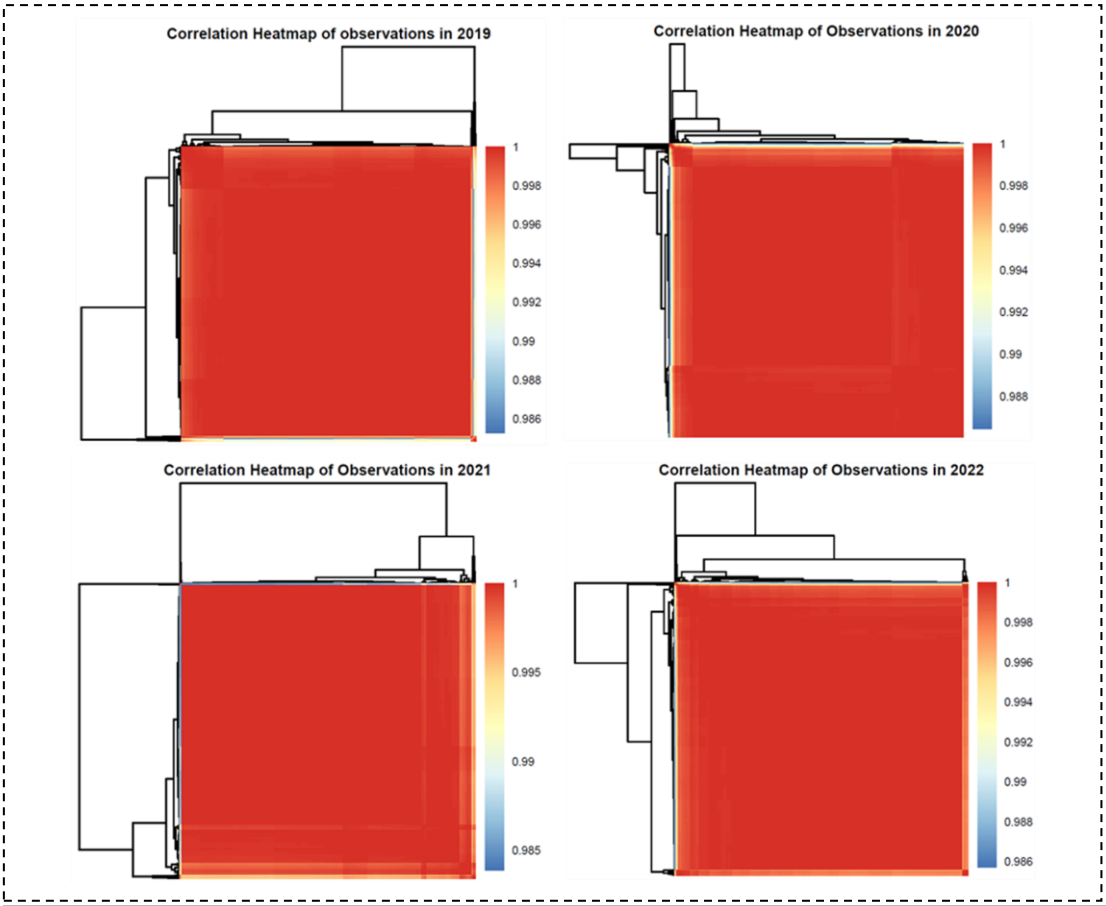

Figure 4 Heatmap of Correlations between Observations

The column standardization process is carried out before starting the GPCA process. This process transforms the data with an average value of 0 and a standard deviation of 1. The GPCA process stops at the sixth iteration. The GPCA results show that from the matrix **L**, the number of principal components is 30, with a cumulative proportion of variance of 89.35%. From the matrix **R**, the number of principal components is 6, with a cumulative proportion of variance of 88.46%. Table 6 shows that the indicators of the Human Development dimension consisting of 20 indicators can be reduced to 6 PC. This result is obtained from the loading value in the principal component equation formed from **R**.

Table 6 Indicators that Describe the Principal Components

Principal Components	Indicators
1	Households with clean drinking water sources (X_1), Households that have access to adequate drinking water (X_2) and Households that do not have defecation facilities (X_3).
2	School enrollment rates 7-12 years (X_5), School enrollment rates 13-15 years (X_6), School enrollment rates 16-18 years (X_7), Gross participation rates level SHS/VHS/MA (X_{10}), Net participation rates level ES/MI (X_{11}), Net participation rates level SHS/VHS/MA (X_{13})
3	Morbidity (X_4), Percentage of formal workers (X_{14}) and Percentage of poor people (X_{15}).
4	Average wages of Workers, and Employees per Month (X_{17}) and Gini ratio (X_{20})
5	Open unemployment rate (X_{16}), Gross regional domestic product per Capita Based on Current Prices (X_{18}) and Percentage of informal workers (X_{19})
6	Gross participation rates level ES/MI (X_8), Gross participation rates level JHS/MTs (X_9) and Net Participation rates Level JHS/MTS (X_{12})

Table 7 shows that 514 regencies/cities can be reduced to 30 principal components. The regencies/cities in each PC are obtained from the loading values in the principal component equation formed from **L**. After obtaining the optimal $\mathbf{L} \in \mathbb{R}^{514\,x\,30}$ and $\mathbf{R} \in \mathbb{R}^{20\,x\,6}$, the low-dimensional data formed has a dimension of $30x6$. Based on the results obtained, it can be concluded that the data of human development dimension indicators at the regency/city level consisting of 514

regencies/cities and 20 indicators can be reduced to 30 PC and 6 PC. This low-dimensional data describes the original data.

Table 7 Regency/City that Describe the Principal Components

Principal Components	Regency/City
1	Aceh Timur, Gayo Lues, Nias, Aceh Jaya, Pidie, Tapanuli Utara, Tapanuli Selatan, Biruen and others
2	Jakarta Timur, Semarang, Makassar, Jakarta Utara, Jakarta Selatan, Jakarta Barat, Bandung, Jakarta Pusat and others
⋮	...
30	Jayapura, Intan Jaya, Asmat, Puncak, Yalimo, Buru, Maluku Tengah, Sumba Tengah, Beru, Alor and others.

In the first iteration, the RMSE value was very high, namely 32.884, which shows that the initial representation of **L** and **R** was not optimal in capturing information from the original data. The significant decrease in RMSE in the second iteration to 12.379 indicates that the GPCA begins to capture the important structure of the data, reducing the reconstruction error drastically. a further decrease in the third iteration, with the RMSE being 3.425, indicates an improvement in the quality of data representation by **L** and **R**. However, the rate of decrease begins to slow down. In the fourth iteration, the RMSE value slightly decreases to 3.403, indicating that most of the variability in the original data has been explained by the model, and the further decrease in the reconstruction error becomes smaller. The fifth and sixth iterations show almost unchanged RMSE values, namely 3.393 and 3.392, indicating the GPCA convergence. The stability of the RMSE at the final iteration shows that the matrix **L** and **R** are optimal. Based on our findings in the data exploration section, the research data contains outliers. GPCA is not robust to image data containing noise or outlier [6,24]. If using the GPCA method which is robust to outliers, the resulting RMSE is lower than the RMSE results obtained in this study. PCA is also not robust to data containing outliers [25].

Based on Table 8, the R-square values obtained indicate that the Procrustes analysis produces estimates of the original data that are similarity to the original data, with more than 87% of the variability of the original data explained by the estimated data in each data pair. This shows that the GPCA method is very effective in maintaining the similarity between the original and estimated data.

Table 8 R-square Value

Data	R-square value
$\mathbf{A_1}$ & $\widehat{\mathbf{A}}_1$	0.894
$\mathbf{A_2}$ & $\widehat{\mathbf{A}}_2$	0.896
$\mathbf{A_3}$ & $\widehat{\mathbf{A}}_3$	0.877
$\mathbf{A_4}$ & $\widehat{\mathbf{A}}_4$	0.878

Biplot visually depicts how the dataset's indicators and regencies/cities relate to each other in the space reduced by GPCA. The visualization obtained from the GPCA-Biplot illustrates the characteristics of the data. In addition, the advantage of presenting with GPCA-Biplot is that it can determine the closeness between observations and the relationship between indicators. GPCA-biplot divides the regency/city into four quadrants. Each quadrant consists of a regency/city with characteristics that are close to each other regarding human development indicators. Conversely, regencies/cities in opposite quadrants will have characteristics that are opposite to each other. The interpretation of the GPCA-Biplot visualization is the same as the PCA-Biplot visualization [26].

Based on Figure 5, the regencies/cities in Quadrant 1 are characterized by indicators X_6, X_7, X_{14}, X_{17} and X_{19}. Therefore, regencies/cities in Quadrant 1 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 2 are characterized by indicators X_5, X_{12}, X_{15}, and X_{16}. Therefore, regencies/cities in Quadrant 2 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 3 are characterized by indicators X_3, X_4, X_{11}, and X_{20}. Therefore, regencies/cities in Quadrant 3 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 4 are characterized by indicators X_1, X_2, X_8, X_9, X_{10}, and X_{18}. Therefore, regencies/cities in Quadrant 4 have relatively higher values for their characteristic indicators. The total cumulative variance of information from the data that can be explained by GPCA-Biplot is 52.56%. The regency/city that are close in the biplot have similar characteristics based on the principal components obtained from GPCA. For example, regencies/cities such as Jayapura City, Nabire, and Marauke show similarities in the dimensions of human development.

Based on Figure 6, the regencies/cities in Quadrant 1 are characterized by indicators X_6, X_9, X_{13}, X_{14}, X_{16} and X_{19}. Therefore, regencies/cities in Quadrant 1 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 2 are characterized by indicators X_4, X_5, X_8, X_{11}, X_{12}, X_{15} and X_{20}. Therefore, regencies/cities in Quadrant 2 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 3 are characterized by indicators X_2, X_3, X_{10} and X_{18}. Therefore, regency/city in Quadrant 3 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 4 are characterized by indicators X_1, X_7 and X_{17}. Therefore, regency/city in Quadrant 4 have relatively higher values for their characteristic indicators. The total cumulative variance of information from the data that can be explained by GPCA-Biplot is 51.87%. The regency/city that are close in the biplot have similar characteristics based on the principal components obtained from GPCA. For example, regencies/cities such as Belitung, Bangka Tengah, and Bangka show similarities in the dimensions of human development.

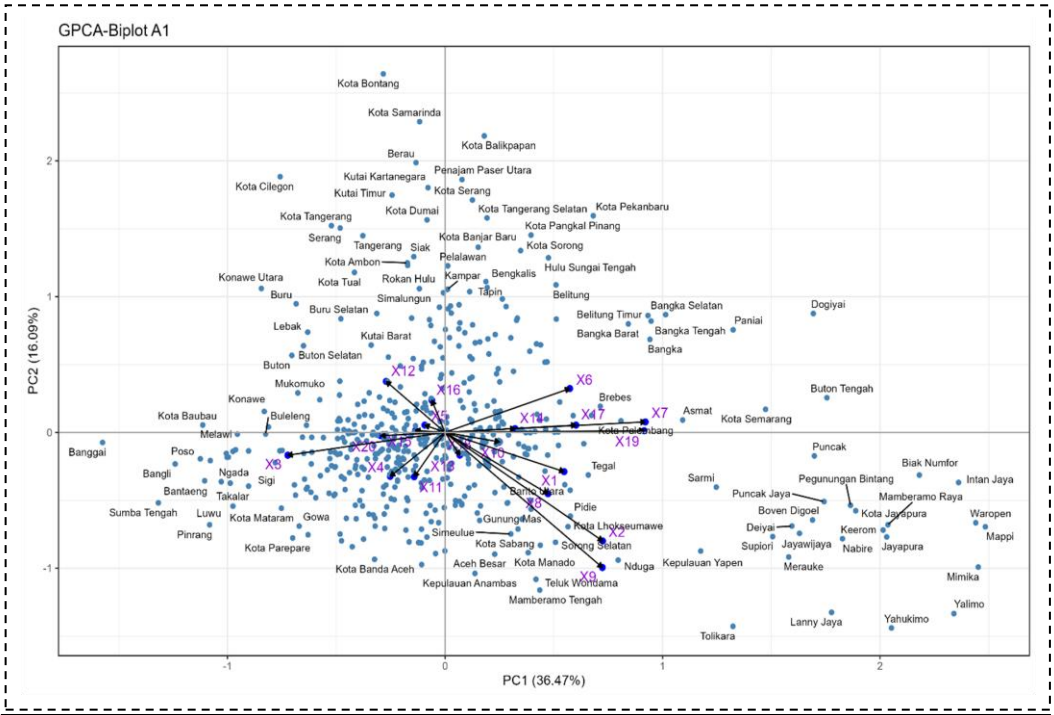

Figure 5 Visualization of GPCA Results in Biplot from 2019 Data

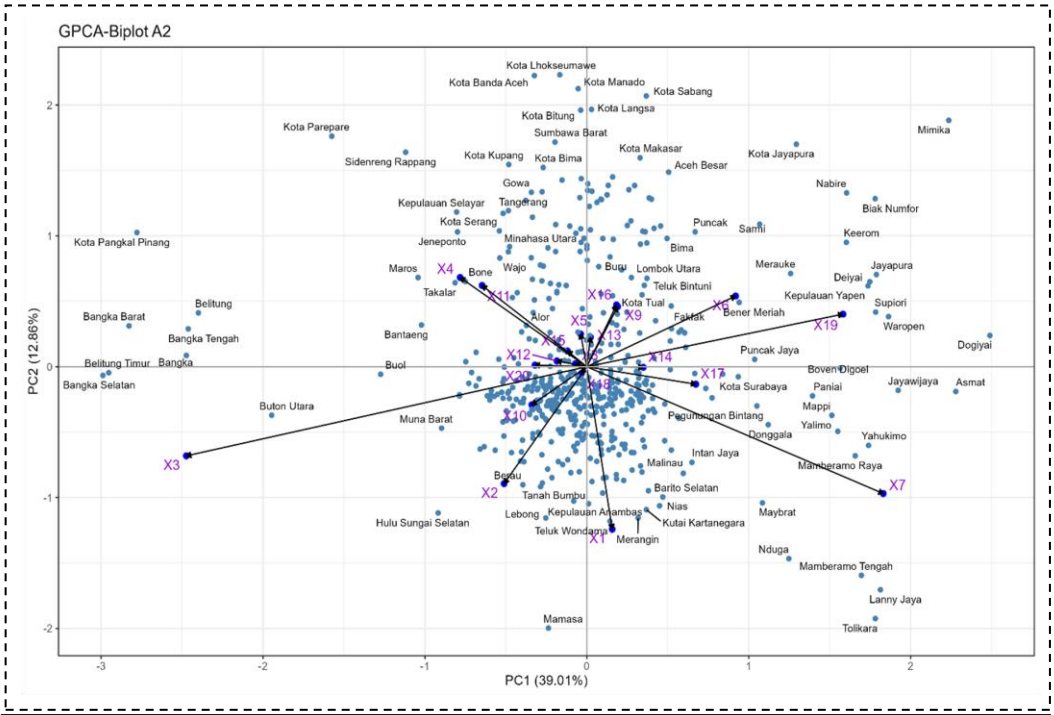

Figure 6 Visualization of GPCA Results in Biplot from 2020 Data

Based on Figure 7, the regencies/cities in Quadrant 1 are characterized by indicators X_3, X_6, X_7, X_9, X_{12} and X_{14}. Therefore, regencies/cities in Quadrant 1 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 2 are characterized by indicators X_1, X_2, X_{10}, X_{18} and X_{19}. Therefore, regencies/cities in Quadrant 2 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 3 are characterized by indicators X_4, X_{16}, X_{17} and X_{20}. Therefore, regencies/cities in Quadrant 3 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 4 are characterized by indicators X_5, X_8, X_{11}, X_{13} and X_{15}. Therefore, regencies/cities in Quadrant 4 have relatively higher values for their characteristic indicators. The total cumulative variance of information from the data that can be explained by GPCA-Biplot is 52.17%. The regencies/cities that are close in the biplot have similar characteristics based on the principal components obtained from GPCA. For example, regencies/cities such as Yalimo, Membramo Raya and Jayawijaya show similarities in the dimensions of human development.

Based on Figure 8, the regencies/cities in Quadrant 1 are characterized by indicators X_3, X_4, X_{10} and X_{20} Therefore, regencies/cities in Quadrant 1 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 2 are characterized by indicators X_1, X_2, X_8, X_{11} and X_{17}. Therefore, regencies/cities in Quadrant 2 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 3 are characterized by indicators X_5, X_6, X_7, X_9, X_{12}, X_{13}, X_{14} and X_{19} Therefore, regencies/cities in Quadrant 3 have relatively higher values for their characteristic indicators. The regencies/cities in Quadrant 4 are characterized by indicators X_{15} and X_{16}. Therefore, regencies/cities in Quadrant 4 have relatively higher values for their characteristic indicators. The total cumulative variance of information from the data that can be explained by GPCA-Biplot is 52.68%. The regencies/cities that are close in the biplot have similar characteristics based on the principal components obtained from GPCA. For example, regencies/cities such as Buol, Sigi and Banggai show similarities in the dimensions of human development.

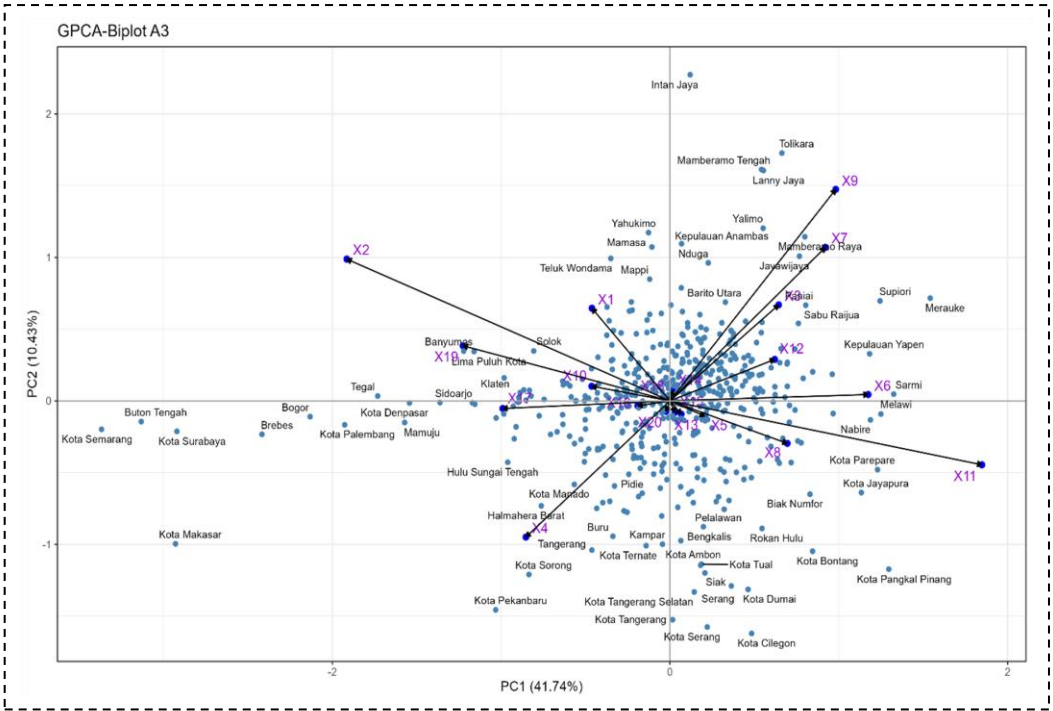

Figure 7 Visualization of GPCA Results in Biplot from 2021 Data

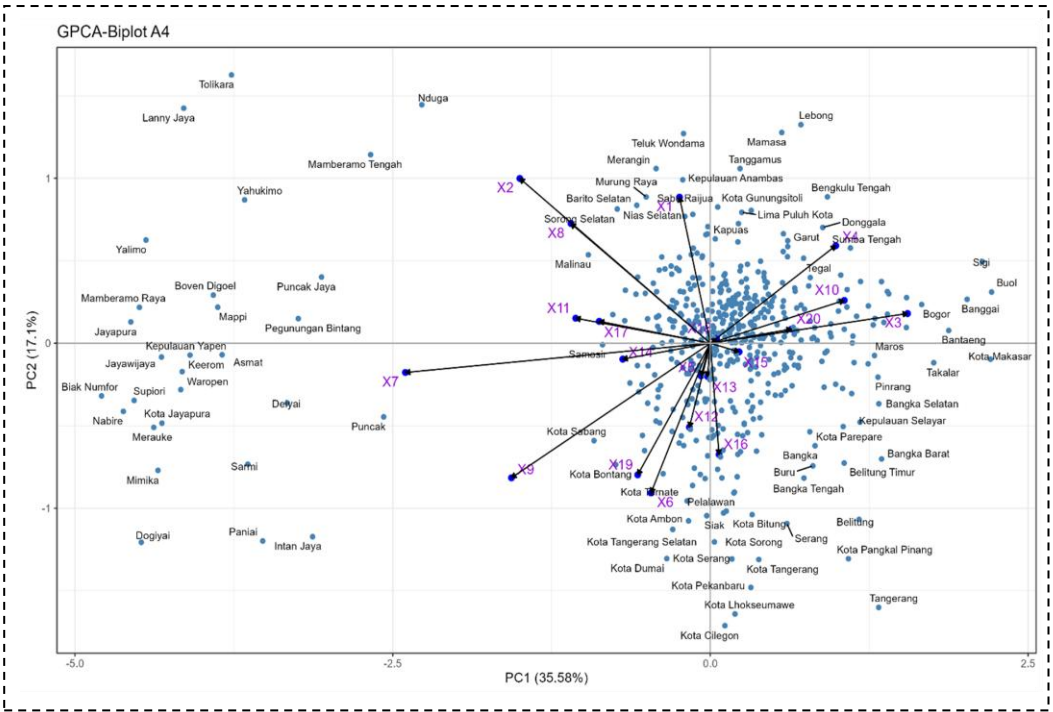

Figure 8 Visualization of GPCA Results in Biplot from 2022 Data

In 2019, Nduga Regency was in Quadrant 4. Indicators X_1, X_2, X_8 and X_9 are close to Nduga regency. This shows that the Nduga regency has adequate access to the indicators of X1 (households with clean drinking water sources) and X_2 (households that have access to adequate drinking water). Apart from that, indicators of X_8 (Gross participation rates level ES/MI) and X_9 (Gross participation rates level JHS/MTs) are high in the Nduga regency. The indicators far from Nduga regency are X_{12}, X_{15}, and X_{16}. This shows that the X_{12} (net participation rates level JHS/MTS) is low, X_{15} (percentage of poor people), and X_{16} (open unemployment rate) are high. The long distance from Nduga regency to indicators X_2, X_8, and X_9 in 2020 shows no significant increase in these indicators. In 2020, Nduga Regency is close to X_7. This shows an increase in X_7 (school enrollment rates 16-18 years). Indicators X_{12}, X_{15}, and X_{16} remain far from Nduga Regency in 2020, the same as in 2019. In 2021, Nduga is very close to the indicators X_7 and X_9. This shows a significant improvement in the X_7 (school enrollment rates 16-18 years) and X_9 (Gross participation rates level JHS/MTs) compared to 2020. Nduga Regency also showed an increase in the X_{12} (Net participation rates level JHS/MTs) and X_{14} (percentage of formal workers) in 2021. This shows that the indicators X_{12} and X_{14} have improved. Apart from that, other indicators are still very far from Nduga Regency, which means there is no improvement in these indicators. In 2022, several indicators, such as X_{12} and X_{14}, decline. However, there is an improvement in the indicators X_2 and X_8.

Overall, human development in Nduga Regency from 2019 to 2022 shows significant fluctuations. Despite progress in some indicators, especially in 2021, major challenges remain, and the focus of development is shifting from one dimension to another without improvements in all areas. More consistent and integrated efforts are needed to ensure sustainable and equitable progress across all dimensions of human development in the Nduga Regency. In the same way, the growth of human development in another regency/city can be described.

V. CONCLUSIONS AND SUGGESTIONS

The application of GPCA in this study is different from other studies that discuss GPCA. This study applies GPCA to analyze human development growth by reducing the dimensions of observations and dimensions of variables in the data set. Based on the analysis and discussion above, it can be concluded that GPCA works well in terms of its ability to reduce the dimensions of data from correlated observations and correlated variables. This conclusion is based on the results obtained from simulation studies and empirical studies. Based on the simulation study results, the dimensions of the low-dimensional data are by the correlation structure of the designed data. The range of RMSE values around 1.80 to 2.05 indicates that the results of dimensionality reduction by GPCA are consistent, with small error variations between repetitions. The results of the Procrustes analysis show that the original data estimates are similarity to the original data. Based on empirical studies, GPCA stops at the sixth iteration with an RMSE value of 3.392. The results of the Procrustes analysis show that the original data estimates are similarity to the original data, with the R-square value of each data pair of more than 87%. The dimensions of the low-dimensional data are 30x6. Based on the results of the GPCA-Biplot visualization, the growth of human development in Nduga Regency from 2019 to 2022 showed significant fluctuations. Although some indicators show progress, especially in 2021, major challenges remain, and the focus of development is shifting from one dimension to another without any improvement in all areas. In the same way, the growth of human development in each regency/city can be analyzed based on the GPCA-Biplot visualization. Thus, government policy focuses on real problems in the field.

Based on our findings, the empirical data contains outliers. Thus, this study suggests reducing the dimension of data containing outliers, especially in the GPCA. Another research suggestion is how to reduce the dimension of data containing missing data, especially in the GPCA.

ACKNOWLEDGEMET

This research is funded by the Directorate General of Higher Education, Research, and Technology Ministry of Education, Culture, Research, and Technology in accordance research program implementation contract No: 027/E5/PG.02.00.PL/2024.

REFERENCES

[1] A. Genender-Feltheimer, "Visualizing high dimensional and big data," *Procedia Comput. Sci.*, vol. 140, pp. 112–121, 2018, doi: 10.1016/j.procs.2018.10.308.

[2] K. Keerthi Vasan and B. Surendiran, "Dimensionality reduction using Principal Component Analysis for network intrusion detection," *Perspect. Sci.*, vol. 8, pp. 510–512, 2016, doi: 10.1016/j.pisc.2016.05.010.

[3] L. C. Djoufack Nkengfack, D. Tchiotsop, R. Atangana, B. S. Tchinda, V. Louis-Door, and D. Wolf, "A comparison study of polynomial-based PCA, KPCA, LDA and GDA feature extraction methods for epileptic and eye states EEG signals detection using kernel machines," *Informatics Med. Unlocked*, vol. 26, pp. 1–16, 2021, doi: 10.1016/j.imu.2021.100721.

[4] J. Ye, R. Janardan, and Q. Li, "GPCA: an efficient dimension reduction scheme for image compression and retrieval," *Proc. tenth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 354–363, 2004, doi: 10.1145/1014052.1014092.

[5] J. Ye, "Generalized low rank approximations of matrices," *Mach. Learn.*, vol. 61, pp. 167–191, 2005, doi: 10.1007/s10994-005-3561-6.

[6] S. Jiarong, W. Yang, and X. Zheng, "Robust generalized low rank approximations of matrices," *PLoS One*, vol. 10, no. 9, pp. 1–23, 2015, doi: 10.1371/journal.pone.0138028.

[7] Y. Chen *et al.*, "Face identification with top-push constrained generalized low-rank approximation of matrices," *IEEE Access*, vol. 7, pp. 160998–161007, 2019, doi: 10.1109/ACCESS.2019.2947164.

[8] H. Itoh, A. Imiya, and T. Sakai, "Dimension reduction and construction of feature space for image pattern recognition," *J. Math.*

Imaging Vis., vol. 56, pp. 1–31, 2016, doi: 10.1007/s10851-015-0629-1.

[9] J. Ye, R. Janardan, and S. Kumar, "Biological Image Analysis via Matrix Approximation," in *Encyclopedia of Data Warehousing and Mining, Second Edition*, 2008, pp. 166–170. doi: 10.4018/9781605660103.ch027.

[10] Z. Li, Z. Hu, F. Nie, R. Wang, and X. Li, "Multi-view Clustering based on Generalized Low Rank Approximation," *Neurocomputing*, vol. 471, pp. 251–259, 2022, doi: 10.1016/j.neucom.2020.08.049.

[11] S. Ahmadi and M. Rezghi, "Generalized low-rank approximation of matrices based on multiple transformation pairs," *Pattern Recognit.*, vol. 108, pp. 1–16, 2020, doi: 10.1016/j.patcog.2020.107545.

[12] BPS, *Indeks pembangunan manusia 2019*. Jakarta: Badan Pusat Statistik, 2020.

[13] BPS, *Indeks Pembangunan Manusia 2020*. Jakarta: Badan Pusat Statistik, 2021.

[14] BPS, *Indeks Pembangunan Manusia 2021*. Jakarta: Badan Pusat Statistik, 2022.

[15] BPS, *Indeks Pembangunan Manusia 2022*. Jakarta: Badan Pusat Statistik, 2023.

[16] I. T. Jolliffe and J. Cadima, "Principal component analysis : a review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, pp. 1–16, 2016, [Online]. Available: https://doi.org/10.1098/rsta.2015.0202

[17] N. Salem and S. Hussein, "Data dimensional reduction and principal components analysis," *Procedia Comput. Sci.*, vol. 163, pp. 292–299, 2019, doi: 10.1016/j.procs.2019.12.111.

[18] R. Nariswari, T. S. Prakoso, N. Hafiz, and H. Pudjihastuti, "Biplot analysis: A study of the change of customer behaviour on e-commerce," *Procedia Comput. Sci.*, vol. 216, pp. 524–530, 2023, doi: 10.1016/j.procs.2022.12.165.

[19] S. Gardner, N. J. Le Roux, and C. Aldrich, "Process data visualisation with biplots," *Miner. Eng.*, vol. 18, pp. 955–968, 2005, doi: 10.1016/j.mineng.2004.12.010.

[20] J. C. Gower, "Procrustes Analysis," in *International Encyclopedia of the Social & Behavioral Sciences*, Second Edi., vol. 19, Elsevier, 2015, pp. 79–81. doi: 10.1016/B978-0-08-097086-8.43078-3.

[21] J. L. Kern, "On the Correspondence Between Procrustes Analysis and Bidimensional Regression," *J. Classif.*, vol. 34, pp. 35–48, 2017, doi: 10.1007/s00357-017-9224-z.

[22] T. Bakhtiar and S. Siswadi, "Orthogonal procrustes analysis : Its transformation arrangement and minimal distance," *Int. J. Appl. Math. Stat.*, vol. 20, no. M11, pp. 16–24, 2011.

[23] R. Cantó, M. J. Peláez, and A. M. Urbano, "Full rank Cholesky factorization for rank deficient matrices," *Appl. Math. Lett.*, vol. 40, pp. 17–22, 2015, doi: 10.1016/j.aml.2014.09.001.

[24] M. Amini Omam and F. Torkamani-Azar, "Noise adjusted version of generalized principal component analysis," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 24, pp. 50–60, 2016, doi: 10.3906/elk-1303-151.

[25] S. Chenouri, J. Liang, and C. G. Small, "Robust dimension reduction," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 7, pp. 63–69, 2015, doi: 10.1002/wics.1331.

[26] J. C. Gower, N. J. Le Roux, and S. Gardner-Lubbe, "Biplots: quantitative data," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 7, pp. 42–62, 2015, doi: 10.1002/wics.1338.