FNAS Journal of Mathematical Modeling and Numerical Simulation
Print ISSN: 3027-1282
www.fnasjournals.com
Volume 3; Issue 1; March 2026; Page No. 94-100.
DOI: https://doi.org/10.63561/jmns.v3i1.1212

Comparative Study of Low and High Sparsity of a Sparse Principal Component Analysis Model on Some Health Indices

***[1]Kalu, N.O., [2]Inamete., E.N.H., & [3]Biu, E.O.**
[1]Abia State Polytechnic, Aba, Abia State, Nigeria
[2]Federal Polytechnic of Oil and Gas, Bonny, Rivers State, Nigeria
[3]University of Port-Harcourt, Port-Harcourt, Nigeria

***Corresponding author email**: noka4nk@gmail.com

Abstract
This research investigates comparative study of low and high sparsity of a sparse principal component analysis model on some health indices which define the prevalence rate of some common diseases in Nigeria. Two cases of multivariate data sets were considered. The Kaiser-Meyer Olin (KMO) and Bartlett's tests were used to test the adequacy of the two cases of multivariate data sets to determine if they are fit for principal component analysis. The Pearson correlation coefficient was used to determine the relationship between the multiple dimensional data set. Then, investigate the impact of the sparse principal component analysis (SPCA) at the levels of sparse: (i.e. low sparsity at 15% and high sparsity at 75%). Some diseases considered are Malaria, TB, Diabetes, Diarrhea, Anemia, Overweight, HIV and Stunting Growth etc. The main objective was to explore the potential of this sparse model in identifying the key variable and then compare the structure and pattern of the sparsity levels, using SPCA. In the analysis, it was observed that the Kaiser-Meyer Olin (KMO) and Bartlett's tests result decreases as the percentage sparse increases. The KMO and Bartlett's test results are 0.561 (249.655) for 15%, and 0.465 (99.727) for 75%. It indicated that 15% sparse in the data will not impact the accuracy and robustness of statistical analyses, which will sufficiently account for the variation in the data. These research, reveal the impact of the levels based on adequacy, relationship between the values, structures and patterns of the variables at low sparse of 15% and high sparse of 75%. Furthermore, this research revealed that the SPCA results at different levels of the sparse; uncovering that the high sparse, has less adequacy on the data set analysis. This study was able to identify a smaller subset of the variables (diseases) and then identified that 15% sparse in the data sets will not impact the accuracy and robustness of statistical analyses, it will sufficiently account for the variance (variation) in the data.

Keywords:Principal Component Analysis,, Prevalence rate of the infections, Statistical Technique, Health, Sparsity

Introduction

Sparse Principal Component Analysis (SPCA) is a powerful statistical technique that helps reduce the dimensionality of high-dimensional datasets while retaining essential information. SPCA has found numerous applications in various fields, including genetics, imaging processing, and neuroimaging. The primary motivation behind SPCA is to obtain a more parsimonious representation of the data that is easier to interpret and analyze Chun et al(2016). Several techniques have been introduced for sparse principal component analysis, including penalized maximum likelihood estimation and non-negative sparse coding. These sparsity-inducing techniques have significantly impacted the field of data analysis and made possible the identification of essential features during data exploration and analysis with high dimensional data.

SPCA has been observed to be an improved variation of PCA since it has the ability to exploit the natural sparsity of data while extracting the principal component. Ama (2021) further stress that SPCA method can use a limited number of components, especially where number of non-zero elements is low. In fact, one of the challenges in our health

Cite this article as:
Kalu, N.O., Inamete., E.N.H., & Biu, E.O. (2026). Comparative study of low and high sparsity of a sparse principal component analysis model on some health indices. *FNAS Journal of Mathematical Modeling and Numerical Simulation, 3*(1), 94-100. https://doi.org/10.63561/jmns.v3i1.1212

system is how to analyze and give a reasonable view of the infection rate in our country. These infections include diarrhea, hepatitis A and E, typhoid fever, malaria, yellow fever, dengue fever, rabies, meningitis, and Lassa fever. Over the years, several researchers have tried to come up with different data and analysis about this infection and their sources.

World Health Organization (WHO) 2022, reported that in 2020 there were over 245 million cases of malaria in the year 2020 and estimated number of deaths stand as 619,000. The problem of modeling and analyzing multiple dimensional mediator variables is a pressing issue in statistical research. By developing a multiple dimensional mediator model using SPCA, this research seeks to overcome the limitations of existing models and provide a more accurate, efficient, and interpretable tool for mediation analysis. The proposed research will contribute to the field of statistical analysis by advancing our understanding of high-dimensional mediation relationships and providing practical solutions for applications in social sciences and healthcare.

Materials and Methods

Nature and Source of Data

The data for this research is a secondary data that was collected from the record of the online data base of world bank. This list is made up of the yearly records of Malaria, Diarrhea, Tuberculosis, Syphilis, Typhoid, Diabetes, HIV and Anemia patients.

Method of Data Analysis

The main method of analysis in this research work is Sparse Principal Component Analysis and Percentage Prevalence Analysis which will compare multiple independent factors of the data set made up of the yearly records of Malaria, Diarrhea, Tuberculosis, Syphilis Typhoid, Diabetes, HIV and Anemia patients. The prevalence rate analysis of incidence will also be used to analyze the sparse percentage differences considering the high and low percentage sparse of 75% and 15% . The principal component consists theoretically of four assumptions.

1. Independence Assumption
2. Normality Assumption
3. Homogeneity of variance assumption
4. Linearity assumption
    1) Normality: This assumption assumes that all the random variable in the model are normally distributed, $Y_{ij} \sim N(\mu_{y,} \delta^2_y)$ $X_{ij} \sim N(\mu_{x,} \delta^2_x)$.
    2) Independence Assumption: This assumption that all the random variables in the model are independent.
    3) Linearity assumption: PCA assumes that the relationship between the observed variables and the underlying latent variables (principal components) is linear. This means that the changes in the observed variables are linearly related to changes in the principal components.
    4) Homogeneity: This assumption in principal component analysis (PCA) states that the variance-covariance matrix of the observed variables should be homogeneous. In other words, the variables should have similar variances and covariances with each other

Analysis with sparse principal components

In this section, we will indicate the analysis with the sparse PC of the mediators. Let

$M_t^{(k)} = M_i^{(T)}WK = \sum_j^p = M_{ijwk}$

for k=1, … ,q, fi

$$M_i^{(k)} = \alpha_0 + \alpha_k X_{i+} e_i^{(k)} \quad (1)$$

$$Yi = b0 = {}_{ck}x_i = bk\ M_i^{(k)} + \eta_i^{(k)} \quad (2)$$

where $e_i^{(k)}$ and $\eta_i^{(k)}$ are independent random errors normally distributed with mean zero.

Cite this article as:
Kalu, N.O., Inamete., E.N.H., & Biu, E.O. (2026). Comparative study of low and high sparsity of a sparse principal component analysis model on some health indices. *FNAS Journal of Mathematical Modeling and Numerical Simulation, 3*(1), 94-100. https://doi.org/10.63561/jmns.v3i1.1212

One advantage of the PCA mediation analysis is that the transformed mediator PCs are conditionally independent, and fitting the LSEM with multiple mediators is equivalent to using marginal LSEMs for each individual mediator. By specifying the loading vector, the orthogonality constraint is not explicitly imposed. To achieve the conditional independence, we include a regression projection step to remove the conditional linear dependence between the transformed mediators, analogous to the procedure proposed in Zou et al. (2006).

Some steps and code snippets that are to be in analyzing the sparse principal components in Matlab:

1. Load the data: Start by loading the data into Matlab using the appropriate function. For example, if you have a CSV file, The `csvread` function can be use to load it into Matlab (MathWorks, 2021a).
2. Normalization of the data: It is essential to normalize the data to ensure that all variables have the same scale. The data can be normalize by using the `normalize` function in Matlab or by subtracting the mean of each column and dividing by the standard deviation (MathWorks, 2021b).
3. Perform dimension reduction: Dimension reduction is the process of reducing the number of variables in a dataset while retaining the most important information. The most common method for dimension reduction is principal component analysis (PCA). The PCA can be performed using the `pca` function in Matlab (MathWorks, 2021c).
4. Select the number of principal components: After performing PCA, a set of principal components will be generated. Decide how many of these components to keep. One way to do this is to look at the percentage of variance explained by each component and select the components that explain the most variance (Jolliffe & Cadima, 2016).
5. Apply sparsity constraints: To perform sparse principal component analysis, apply sparsity constraints on the principal components. This can be achieved by adding an L1 penalty term to the PCA objective function. The `lasso` function or the `lassoglm` function in Matlab can be us to apply sparsity constraints (MathWorks, 2021d).
6. Analyze the results: After performing sparse principal component analysis, analyze the results by looking at the coefficients of the principal components. The coefficients that are close to zero represent the variables that have no influence on that component. To visualize the results, plot the principal components graph.

Sample codes to start are:

```
% Load data
data = csvread('data.csv');
% Normalize data
data_norm = normalize(data);
% Perform PCA
[coeff,score,latent] = pca(data_norm);
% Select the number of principal components
num_components = 2;
% Apply sparsity constraints
sparse_coeff = lassoglm(data_norm, score(:,1:num_components), 'lasso');
% Analyze the results
plot(sparse_coeff); % plot the coefficients
scatter(score(:,1), score(:,2)); % plot the principal components
```

Let $\eta_i^{(k\,.\,1,\ldots,k-1)}$ denote the residual after adjusting for $\eta_i^{(1)}, \ldots, \mathrm{M}_i^{(k-1)}$ when controlling $\mathrm{X_i}$ (for $i = 1, \ldots, n$). That is,

$$\mathrm{M}_i^{(k\,.\,1,\ldots,k-1)} = \mathrm{M}_i^{(k)} - \mathrm{M}_i^{(k\,.\,1,\ldots,k-1)\mathrm{T}} \Pi_{k\,.\,1,\ldots,k-1,}$$

Where $\mathrm{M}_i^{(1,\ldots,k-1)} = \mathrm{M}_i^{(1)}, \ldots., \mathrm{M}_i^{(k-1)})^{\mathrm{T}}, and\ \Pi_{k\,.\,1,\ldots,k-1,}$ is the estimate coefficient model.

$$\mathrm{M}_i^{(1)} = \pi_0 + \pi_1 x_1 + \mathrm{M}_i^{(1,\ldots,k-1)T} \Pi_{k\,.\,1,\ldots,k-1,} + \mathrm{t}_i^{(k)}$$

Cite this article as:
Kalu, N.O., Inamete., E.N.H., & Biu, E.O. (2026). Comparative study of low and high sparsity of a sparse principal component analysis model on some health indices. *FNAS Journal of Mathematical Modeling and Numerical Simulation, 3*(1), 94-100. https://doi.org/10.63561/jmns.v3i1.1212

Where $t_i^{(k)}$ is normally distributed model error with mean zero. The new mediators are $M^{(1,\ldots,k-1)}$uncorrelated given the treatment *X*. Thus, we can use a model to estimate the indirect effect of each individual mediation pathway.
We summarize the steps of mediation analysis with sparse principal components in Algorithm 1. To perform inference on model parameters, we propose a boot strap procedure.

Results
The percentage of sparse (or zero) observations in the data sets is computed as number of sparse observations divides by the overall observation multiplied by 100.

Table 1: Testing for adequacy of the percentage sparse: with low sparsity of 15% sparse observations

.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.561
Bartlett's Test of Sphericity	Approx. Chi-Square	249.655
	Df	78
	Sig.	.000

The results in Table 2 showed that the **KMO/BARTLETT'S** test measurement of the sample adequacy.

Table 2: Summary of the Prevalence Rate Correlation Coefficients with low sparsity of 15% sparse observations.

Reliability level	Correlation Coefficients	Correlation Coefficients Remark
0.00 to 0.19	Slightly	24
0.20 to 0.39	Fair	15
0.40 to 0.59	Moderate	18
0.60 to 0.79	Substantial	11
0.80 to1.00	Prefect	8

Table 3: First to Fifth Components of the Prevalence Rate of the diseases with low sparsity of 15% sparse observations.

- Variable	PC1	PC2	PC3	PC4	PC5
- TB	-0.334	-0.118	0.277	0.095	0.019
- HIV15-49	0.373	-0.030	-0.009	-0.057	-0.078
- HIVM15-49	0.174	-0.022	-0.055	-0.773	0.386
- HIVF15-49	0.378	0.013	0.082	0.015	0.060
- ANEMIAWR 15-49	0.202	-0.097	0.620	0.189	-0.135
- ANEMIA PW	0.358	0.119	-0.286	0.119	0.009
- ANEMIA PW 15-49	0.169	0.415	-0.207	0.512	0.445
- ANEMIA C 0-5	0.225	-0.221	-0.480	0.024	-0.569
- TB 100 P	0.277	-0.008	0.264	-0.048	-0.350
- MALARIA Gen	0.313	0.150	0.295	0.006	0.076
- HIVC 15-49	0.373	-0.047	0.103	-0.034	0.142
- DIARR C 0-5	-0.105	0.578	0.070	-0.246	-0.209
- DIABETES 20-49	-0.062	0.618	0.042	-0.115	-0.330

Cite this article as:
Kalu, N.O., Inamete., E.N.H., & Biu, E.O. (2026). Comparative study of low and high sparsity of a sparse principal component analysis model on some health indices. *FNAS Journal of Mathematical Modeling and Numerical Simulation, 3*(1), 94-100. https://doi.org/10.63561/jmns.v3i1.1212

- From Table which present the first to fifth components of the prevalence rate of the actual data the result which the PCI, PC2, PC3, PC4 and PC5 15% sparse.

High sparsity

Table 4: KMO and Bartlett's test with High sparsity of 75% sparse observations.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.465
	Approx. Chi-Square	99.727
Bartlett's Test of Sphericity	Df	10
	Sig.	.000

75% Principal Component Analysis: stunting growth, overweight, Diabetes, Diarrhea with Feeding, Diarrhea with ORS

Table 5: Summary of the Prevalence Rate Correlation Coefficients with High sparsity of 75% sparse observations.

Reliability level	**Correlation Coefficients**	**Correlation Remark**	**Coefficients**
0.00 to 0.19	Slightly	5	
0.20 to 0.39	Fair	2	
0.40 to 0.59	Moderate	1	
0.60 to 0.79	Substantial	0	
0.80 to1.00	Prefect	2	

The principal components model of the prevalence rate of the diseases with High sparsity of 75% sparse observations.

Table 6: First to Fifth Components of the Prevalence Rate of the diseases with High sparsity of 75% sparse observations.

Variable	PC1	PC2	PC3	PC4	PC5
Stunting Growth	0.025	-0.001	0.108	-0.617	0.779
overweight	0.020	-0.002	0.136	-0.767	-0.626
Diabetes	0.162	0.287	-0.930	-0.160	-0.002
Diarrhea Feeding	0.654	-0.747	-0.117	0.002	-0.004
Diarrhea ORS	0.738	0.599	0.301	0.075	-0.00

Discussion

The results in Table 1 showed that the **KMO/BARTLETT'S** degree of freedom is 78 with the approximated chi-square value of 249.655. The result of the Kaiser-Meyer oikin (KMO) and Bartlett's test measurement of the sample adequacy is 0.561 which can be approximated to 0.6. This is an indication that the data is adequate, since the critical value is approximately 0.6 which is the minimum value. Furthermore, the p-value is less than 0.05. This implies that there is no identity matrix in the sample which can stop the analysis on the principal components, although the sample is not strongly adequate but a sparse principal component can be applied.

In Table 4 which presents the KMO And Bartlett's Test of The Data For high sparsity with 75% Sparse Observations. The degree of freedom is 10 showing the approximate Chi-square of 99.727. The result of the Kaiser-Meyer Olkin (KMO) test shows a sampling adequacy of 0.465. This is < 0.6 which indicate that the data is not adequate enough since the value is less than 0.6; that is the minimum value of adequacy. The criterion for adequacy acceptance is 0.6 which is the minimum acceptance value. The P-value also show (0.000) < 0.05. This implies that there is no identity

Cite this article as:
Kalu, N.O., Inamete., E.N.H., & Biu, E.O. (2026). Comparative study of low and high sparsity of a sparse principal component analysis model on some health indices. *FNAS Journal of Mathematical Modeling and Numerical Simulation, 3*(1), 94-100. https://doi.org/10.63561/jmns.v3i1.1212

matrix in the data set which can stop the analysis of the principal components.

The correlation matrix of the variable with 75% sparse indicates that diabetes and stunting growth with 0.168, diabetes and weight 0.88 (0.353) diarrhea with feeding and overweight 0.189 (0.206), diarrhea with ORS and overweight 0.195 (0.199) and diarrhea with feeding and diabetes 0.112 (0.314) shows a slightly correlation relationship. Diarrhea with feeding and stunting growth 0.282 (0.108) and Diarrhea with ORS and diarrhea with feeding 0.462 (0.17) shows moderate relationship while overweight and stunting growth 0.983 (0.00) show perfect relationship. The correlation results were computed like the study of Omprakash and Gokila (2018), where they describe principal components as multivariate procedure that investigates a data slab in which clarifications are designated by numerous inter-correlated measurable reliant variables.

From Table 3 which present the first to fifth components of the prevalence rate with 15% sparse observations, the result which the PCI, PC2, PC3, PC4 and PC5. In PC1 the Tuberculosis show a negative strong influence on the data with PC, of the tuberculosis being -0.334. In this, it shows that the Tuberculosis has a weak contribution in the data building. HIV on its own has a PCI of 0.373 for general contribution and 0.378 for only women. This is an indication that HIV which recorded a common reduction rate has a positive influence on the data set. It is also an indication that information of the data HIV maintains a Justifiable influence on the data set. Other common diseases such as Anemia for women, men and children show positive contribution. Malaria also shows a First principal component with positive influence in the data set influence in the data set. Diarrhea and diabetes show a negative influence with values -0105 and -0.062 as its PCI respectively. For the PC2 result, infectious diseases show negative impact in the data set. In this, Tuberculosis with 0.118 has the most negative impact. At the PC2 infections, anemia for children record shows negative influence on the data set, which is an indication that their impact is reducing the data set.

Diabetes with 0.618 show a strong positive influence on the data set and diarrhea with 0.578 also show positive impact on the data set. Other infections with positive influence include Anemia, Malaria and HIV record for female. Analyzing the table 6 which is the PC1 to PC3 model of the prevalence rate with high sparsity of 15% sparse observations, the model analysis shows that in PC1 T.B with the value of 0.431 has the highest contribution in the data. It also shows that Diarrhea with the value 0.271 also contributed positively. Making it the variable with a good positive influence in the PC1 model. Also, diabetes with the value 0.190 also show a positive contribution in the model. HIV with the value -0.531 have the lowest negative significance influence on the model. It indicates an absolute value of 0.531. Malaria has a value of -0.465 which makes its absolute value to be 0.465. Anemia also shows a value of -0.454. In PC2, four infectious diseases have positive values of 0.693, 0.584, 0.325 and 0.090 which are values from diabetes, diarrhea, malaria and anemia respectively. Diabetes has the most significant contribution in the PC2 model. Observing the PC3 model, four infectious diseases out of the six variables also indicate positive contribution in the PC3 model with T.B 0.660 showing the most significance contribution.

From the Table 1 which compares the impact at different degree. The K.M.O shows a downward pattern as the percentage sparse increases. At 15% the sampling adequacy has the value of 0.561 which is approximately 0.6, while at 75% it has an adequacy rate of 0.465. This trend indicates that as the percentage sparse increases, the sampling adequacy reduces. Since the K.M.O rule states that, for a sample to be adequate enough, its adequacy rate must be at 0.6 or above for adequacy, it shows that above 15% sparse, the sample may not be adequate enough to explain the component in the data sets with sparse and produce a reliable result, even if the P-value may be showing that the sample is not an identity matrix. The Bartlett's test result which has the value of the approx. chi-square as at 15% sparse sample produce a value of 249.6, while at 75% sparse, the value is 99.727. At Bartlett's test result, they all show that the P-value is 0.000 which is less than 0.05. It shows that samples are identity matrix and that the percentage sparse of the sample does not affect the Bartlett's test.

Analyzing the correlation test at 15% sparse 24 units show slight relationship, 15 units indicates fair relationship, 18 units show moderate relationship, 11 unit indicates substantial relationship while 8 units observed perfect relationship. 75% sparse also show 5 slight relationship 2 fair relationship, 1 moderate and 2 perfect relationship at every level it was observed that the slight relationship are more in number while the other correlation levels are not regular. Analyzing the Eigen values at 90% proportion 15% sparse show the data set can be reduced to 8 variables out of the 13 variables, while at the 75% the variable can be reduced to 2 variables out of 6.

Cite this article as:
Kalu, N.O., Inamete., E.N.H., & Biu, E.O. (2026). Comparative study of low and high sparsity of a sparse principal component analysis model on some health indices. *FNAS Journal of Mathematical Modeling and Numerical Simulation, 3*(1), 94-100. https://doi.org/10.63561/jmns.v3i1.1212

Conclusion

This study was able to compare the impact of the SPCA at lower sparsity of 15% and high sparsity of 75% sparse; then identify a smaller subset of the variables (diseases) and then identified that 15% sparse in the data sets will not impact the accuracy and robustness of statistical analyses, although, it may not sufficiently account for the variance (variation) in the data. 75% sparse in the data sets will impact the accuracy and robustness of statistical analyses, it will sufficiently account for the variance (variation) in the data.

References

Ama, S. (2021). Sparse principal component analysis for dimension reduction in high-dimensional genomic data. *BMC Bioinformatics, 22(*1), 1-16.

Chun-Mei, L., Yu-Cheng, L., & Tien-Yu, L. (2016). Improved principal component analysis using sparse representation for hyperspectral image classification.

Brown, R. A. (2009). Dimensionality reduction of data. Wiley Interdisciplinary Reviews: *Computational Statistics, 1*(3), 261–269. https://doi.org/10.1002/wics.14

Brusch, J. L. (2009). Typhoid Fever. In StatPearls [Internet]. StatPearls Publishing.

Buzby, J. C., Roberts, T., & MacDonald, J. M. (2001). Bacterial foodborne disease: Medical costs and productivity losses. *Agricultural Economics Reports, 794*.

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Jolliffe, I. T., & Cadima, J. (2016). Principal component Analysis: A review and recent developments. philosophical transactions of the royal society. *A Mathematical, Physical and Engineering Science*, 374, Article 20150202. https://doi.org/10.1098/rsta.2015.020

MathWorks, (2021) 2021b MATLAB and Simulink. Eastern DayLight Time; Sep 28, 2021. www.businesswire.com

MathWorks, (2021) 2021a MATLAB and Simulink. Microwave Journal; March 16, 2021. www.businesswire.com

WHO. (2022). Tuberculosis. Retrieved from [insert URL]

World Health Organization. (2017). Global health estimates 2020: Deaths by cause, age, sex, by country and by region, 2000-2019. Retrieved from [insert URL]

World Health Organization. (2019). Malaria. Retrieved from https://databank.worldbank.org/databases

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics, 15*(2), 265-286. https://doi.org/10.1198/106186006X113430

Cite this article as:
Kalu, N.O., Inamete., E.N.H., & Biu, E.O. (2026). Comparative study of low and high sparsity of a sparse principal component analysis model on some health indices. *FNAS Journal of Mathematical Modeling and Numerical Simulation, 3*(1), 94-100. https://doi.org/10.63561/jmns.v3i1.1212