

A Comparative Exploration of PCA Variants for Clustering Analysis

1st Leo Ramos 

Kauel Inc., Houston, TX 77098, USA.

*Numerical Analysis and Data Science Research Group,
Yachay Tech University,
Urcuquí, 100119, Ecuador.
leo.ramos@kauel.com*


2nd Francklin Rivas-Echeverría 

Kauel Inc., Houston, TX 77098, USA.

*Pontificia Universidad Católica del Ecuador Sede Ibarra,
Ibarra, 100112, Ibarra, Ecuador.
francklin.rivas@kauel.com*

3rd Isidro R. Amaro 

*School of Mathematical and Computational Sciences,
Numerical Analysis and Data Science Research Group,
Yachay Tech University,
Urcuquí, 100119, Ecuador.
iamaro@yachaytech.edu.ec*

4th Franklin Camacho 

*School of Mathematical and Computational Sciences,
Numerical Analysis and Data Science Research Group,
Yachay Tech University,
Urcuquí, 100119, Ecuador.
fcamacho@yachaytech.edu.ec*

Abstract—In this study, we compare the performance of Principal Component Analysis (PCA), Sparse PCA (SPCA), Robust PCA (RPCA), and Weighted PCA (WPCA) on a high-dimensional dataset of economic indicators from G20 countries. We evaluate their effectiveness in retaining variance and enhancing the performance of K-means clustering. Our comparative analysis employs metrics including effectiveness of variance retention, mean variance of distance sample-centroid, mean distance among centroids, and the rand index for cluster similarity. Our analysis indicates that PCA exhibits a greater effectiveness compared to SPCA but is outperformed by RPCA and significantly by WPCA, which shows the highest variance retention among the four methods. In terms of clustering, SPCA coupled with K-means achieves the best balance between cluster compactness and separation, as indicated by a low mean variance of distance sample-centroid and a relatively high mean distance among centroids. RPCA, while exhibiting extremely compact clusters, demonstrates the least inter-cluster separation. The rand index comparisons reveal that while PCA, SPCA, and WPCA share similar clustering structures, RPCA distinguishes itself by detecting unique patterns, contributing to a broader perspective in the analysis of the high-dimensional datasets. The study provides insightful findings that emphasize the role of appropriate dimensionality reduction method selection in enhancing the effectiveness of unsupervised learning tasks.

Keywords—principal component analysis, clustering, K-means, classification, data analysis, machine learning

I. INTRODUCTION

In the realm of unsupervised learning, clustering is a fundamental technique aimed at grouping similar data points within an unlabeled dataset, facilitating pattern recognition, and data exploration [1]. Nevertheless, the high dimensionality of modern datasets can hinder clustering algorithms from efficiently identifying meaningful patterns [2]. Addressing this challenge, dimensionality reduction techniques offer a promising solution

by transforming the original data into a lower-dimensional space.

Dimensionality reduction is a common technique in machine learning and data science used to reduce the number of features in a dataset while preserving as much information as possible [3]. This can be beneficial for a variety of tasks, such as improving the performance of machine learning algorithms, making data visualization easier, and reducing the computational cost of data analysis [4].

One of the most well-known dimensionality reduction techniques is principal component analysis (PCA) [5]. PCA works by finding a set of orthogonal linear combinations of the original variables that account for as much of the variance in the data as possible [6]. This can be a useful way to summarize the main features of a dataset and identify relationships between variables.

However, PCA has some limitations. For example, it is not robust to outliers, and it can be sensitive to the scale of the variables [7]. Therefore, over the years, variants of PCA have been proposed to address these limitations. While PCA is the most widely known technique [8], the existence of alternative variants opens up new possibilities for improving clustering performance.

This research aims to analyze and compare PCA with three of its variants: sparse PCA (SPCA), robust PCA (RPCA), and weighted PCA (WPCA), seeking to identify the most suitable dimensionality reduction approach for optimizing clustering results. By understanding the strengths and limitations of each variant, we aim to provide valuable insights for researchers and practitioners in data analysis and machine learning tasks.

To achieve this, we utilize a dataset comprising economic indicators of G20 member countries, a diverse group of nations with significant economic, commercial, and population impact

worldwide. Following the dimensionality reduction step using PCA and its variants, we apply the K-means clustering algorithm to form the clusters.

To evaluate the performance of the dimensionality reduction techniques, we employ metrics such as effectiveness. Additionally, to assess the clustering results, we consider metrics such as mean variance of distance sample-centroid, mean distance among centroids, and rand index.

Through this comprehensive analysis, we aim to advance the state-of-the-art in dimensionality reduction and clustering, providing valuable insights for researchers and practitioners in data analysis and machine learning tasks. By understanding the strengths and limitations of each PCA variant, this research will aid in making informed decisions when dealing with high-dimensional datasets, with the ultimate goal of enhancing the utility of dimensionality reduction techniques in real-world data analysis.

II. MATERIALS AND METHODS

A. Data set description

The data set used in this study was collected from Trading Economics¹, a website that aggregates millions of economic indicators from countries worldwide. This platform provides free access to indicators, historical data, charts, and forecasts.

For this research, our focus was on the G20 member countries, comprising the world's largest advanced and emerging economies. These countries collectively account for approximately two-thirds of the global population, 85% of the global gross domestic product, and over 75% of global trade. A total of 24 countries were analyzed, as G20 events often include guest members each year. The list of the 24 countries analyzed is presented in Table I.

TABLE I: G20 member countries. *guest member

Argentina	Australia
Brazil	Canada
China	European Union
France	Germany
India	Indonesia
Italy	Japan
Mexico	Netherlands*
Russia	Saudi Arabia
Singapore*	South Africa
South Korea	Spain*
Switzerland*	Turkey
United Kingdom	United States

We collected 14 economic, social, and trade indicators that were available from these countries. The data were gathered from the third quartile of 2022. A detailed description of the indicators is available in Table II.

The selection of this data set was strategic for our objective. Given the heterogeneity inherent in data from G20 member and guest countries, we are presented with diverse scaling, centrality, and distribution characteristics. This allows us to rigorously test the robustness and efficacy of the different

PCA variants. Likewise, possible noise in the data serve as an effective and realistic ground for comparison. Moreover, the inclusion of guest members provides atypical data points, further testing the versatility and adaptability of the K-means method under different dimensional reductions.

TABLE II: Description of the economic indicators collected for the G20 member countries.

Abbreviation	Description	Unit
gdp	Gross domestic product	USD Billions
gdp-agr	Gross domestic product annual growth rate	Percentage
gd-gdp	Government debt to gross domestic product	Percentage
ed	External debt	USD Millions
i	Imports	USD Millions
e	Exports	USD Millions
ir	Inflation rate	Percentage
cpi	Consumer price index	Points
fi	Food inflation	Percentage
pp	Producer prices	Points
ca	Current account	USD Millions

B. Dimensionality reduction techniques studied

1) *Principal component analysis (PCA)*: is a widely used statistical procedure that aims to transform a set of correlated variables into a smaller set of uncorrelated variables, known as principal components [7], [9]. The goal of PCA is to capture the maximum amount of variation in the data using a smaller number of variables [9]. PCA can be used for dimensionality reduction, data compression, and data visualization [10].

PCA works by finding the eigenvectors and eigenvalues of the covariance matrix of the data [11]. The eigenvectors represent the principal components, while the eigenvalues represent the amount of variance explained by each principal component. The principal components are ordered by the amount of variance they explain, with the first principal component explaining the most variance [12], [13]. Given a data matrix X with n observations and p variables, the covariance matrix C of X is calculated as:

$$C = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X})$$

where \bar{X} is the mean of X . The eigenvectors and eigenvalues of the covariance matrix C are calculated as:

$$Cv_i = \lambda_i v_i$$

where v_i is the i th eigenvector and λ_i is the i th eigenvalue. The principal components are calculated as:

$$PC_i = Xv_i$$

where PC_i is the i th principal component. The amount of variance explained by each principal component is calculated as:

$$\frac{\lambda_i}{\sum_{j=1}^p \lambda_j}$$

where λ_i is the i th eigenvalue.

¹<https://tradingeconomics.com/>

2) *Sparse principal component analysis (SPCA)*: is a dimensionality reduction technique that is designed to be more robust to noise and outliers than traditional PCA [14]. Sparse PCA works by adding a sparsity constraint to the PCA problem, which forces the principal components to be sparse [15]. This makes sparse PCA more resistant to noise and outliers, because the noise and outliers are less likely to be sparse [16].

Sparse PCA can be formulated as a minimization problem, where the goal is to minimize the reconstruction error subject to a sparsity constraint [17]. The minimization formulation of Sparse PCA can be expressed as:

$$\min_{v_i} \left\{ \frac{2n}{1} \|X - Xv_iv_i^T\|_F^2 + \lambda \|v_i\|_1 \right\}$$

subject to the constraint that $\|v_i\|_2 = 1$, where v_i is the i th principal component, $\|\cdot\|_F$ represents the Frobenius norm, and $\|\cdot\|_1$ represents the L_1 norm. The first term in the objective function represents the reconstruction error, while the second term represents the sparsity constraint. The parameter λ controls the trade-off between reconstruction error and sparsity.

3) *Robust principal component analysis (RPCA)*: is a modification of PCA that works well with respect to grossly corrupted observations [18]. The goal of RPCA is to separate low-rank trends from sparse outliers within a data matrix, that is, to approximate the data matrix as the sum of a low-rank matrix and a sparse matrix [18], [19]. Given a data matrix M , RPCA aims to decompose it into two matrices, L and S , such that:

$$M = L + S$$

where L is a low-rank matrix, and S is a sparse matrix. The low-rank matrix L captures the underlying structure of the data, while the sparse matrix S contains the outliers or noise in the data [20], [21].

The ‘‘robust’’ part of this analysis involves splitting the original data matrix into a low-rank matrix and a sparse matrix before performing PCA [21], [22]. RPCA has many real-life applications, particularly when the data under study can naturally be modeled as a low-rank plus a sparse contribution. The decomposition of the data matrix into low-rank and sparse matrices can be achieved by different approaches, including an idealized version of RPCA [23], which aims to recover a low-rank matrix from highly corrupted measurements.

4) *Weighted principal component analysis (WPCA)*: is an extension of traditional PCA that allows the user to weight the different features of the data [24]. This can be useful when some features are more important than others.

In WPCA, the goal is to find a set of principal components that capture the maximum amount of variance in the data while considering the weights assigned to each variable [25]. This is achieved by finding the eigenvectors of the weighted covariance matrix of the data, and then projecting the data onto the eigenvectors with the largest eigenvalues [24]. The mathe-

matical formulation of Weighted PCA involves modifying the covariance matrix calculation to incorporate the weights.

Let’s consider a data matrix X with n observations and p variables. The weighted covariance matrix C_w is calculated as:

$$C_w = \frac{1}{n-1} (X - \bar{X})^T W (X - \bar{X})$$

where \bar{X} is the mean of X , and W is a diagonal matrix containing the weights assigned to each variable. The eigenvectors and eigenvalues of the weighted covariance matrix C_w are then computed as:

$$C_w v_i = \lambda_i v_i$$

where v_i is the i th eigenvector and λ_i is the i th eigenvalue. The principal components are obtained by projecting the data onto the eigenvectors:

$$PC_i = X v_i$$

We present a summary of the key characteristics of these dimensionality reduction techniques in Table III.

C. K-means clustering

K-means is a popular unsupervised machine learning algorithm used for partitioning data into distinct non-overlapping subgroups or clusters [26]. The algorithm works by iteratively assigning data points to the nearest cluster center and updating the cluster centers based on the mean of the assigned data points [27], [28]. Below are the steps involved in the K-means clustering algorithm:

- 1) Choose the number of clusters (k) to be formed.
- 2) Initialize the cluster centers randomly.
- 3) Assign each data point to the nearest cluster center based on the Euclidean distance.
- 4) Recalculate the cluster centers as the mean of the assigned data points.
- 5) Repeat steps 3 and 4 until convergence is achieved.

K-means clustering is widely used in various domains and applications, including customer segmentation, anomaly detection, market analysis, and more [29], [30]. It is chosen as a clustering algorithm due to its simplicity, efficiency, scalability, versatility, and interpretability. The algorithm is easy to implement and can handle large datasets with high dimensionality [31]. It produces clusters that are easy to interpret and understand, providing insights into the underlying structure of the data.

D. Performance metrics

1) *Effectiveness*: quantifies the amount of retained or lost variance through the Frobenius norm comparison of two dimensionality reduction methods [32]. For instance, to assess the effectiveness of the RPCA method relative to PCA, we employ the following formulation:

$$e = \frac{\|L\|_F^2 - \|D\|_F^2}{\|X\|_F^2}$$

TABLE III: Summary of features of the studied dimensionality reduction techniques

Feature	PCA	SPCA	RPCA	WPCA
Number of components	Unconstrained	Constrained to be sparse	Constrained by rank and sparsity	Constrained by weights
Robustness to noise and outliers	Sensitive	More robust	More robust	More robust
Ability to weight features	No	Yes	Yes	Yes

where L represents the low-rank matrix generated by the RPCA method, D represents the matrix formed by the principal components generated by PCA, and X represents the original data matrix.

The range of effectiveness (e) varies between -1 and 1. A positive e value ($e > 0$), indicates that the first method under comparison retains more variance than the second. Conversely, a negative e value ($e < 0$), signifies that the second method retains more variance than the first [32], [33].

2) *Mean variance of distance sample-centroid*: is a metric that measures the average distance between each sample and the centroid of its assigned cluster [34], [35]. A smaller variance indicates better clustering, as it suggests that the samples within each cluster are closer to their respective centroids.

3) *Mean distance among centroids*: is a metric that measures the average distance between all the centroids of the clusters [35]. A larger distance is desired, as it indicates that the clusters are well-separated and distinct. This metric is based on the distance between cluster centers [36], which is a fundamental concept in clustering

4) *Rand index*: is a widely used metric for evaluating clustering algorithms. This metric measures the similarity between two clusterings [35], [37]. It is a measure of the agreement between the true labels and the predicted labels. The rand index ranges from 0 to 1, with 0 indicating complete dissimilarity between the groupings and 1 indicating that they are the same.

E. Workflow and implementation details

Before being used by the algorithms, the data was analyzed and preprocessed. The first step was to verify whether the data is suitable for applying multivariate statistical analysis methods. For this, we calculated the determinant of the correlation matrix. Let X be our data matrix, and we obtained the following:

$$\det(\text{corr}(X)) = 0.0066$$

This value is close to zero, indicating that we can proceed and apply the different multivariate statistical analysis techniques. Then, the data was scaled using `MinMaxScaler` to normalize it and ensure the proper performance of the dimensionality reduction and clustering techniques addressed.

Once this preprocessing was completed, each of the studied techniques was applied. We first reduced dimensionality and then applied clustering. Subsequently, we evaluated the results

obtained using the metrics described earlier. The workflow followed is illustrated in Fig. 1.

The implementation was carried out in Google Colab under the Python programming language. PCA and SPCA were utilized through the `sklearn` library, a widely used machine learning framework. For RPCA and WPCA, we obtained the implementations from two external repositories^{2,3} that contained the respective algorithms.

III. RESULTS AND DISCUSSION

The results from Table IV reflect the effectiveness of different dimensionality reduction methods, each in relation to PCA, SPCA, RPCA (Robust PCA), and WPCA.

In the comparison between PCA and SPCA, an effectiveness score of 0.1327 implies that PCA is more effective at retaining variance than its sparse counterpart, SPCA. However, when contrasted with RPCA, PCA shows a slightly lower effectiveness score of -0.0582. This score suggests that RPCA retains slightly more variance than PCA, demonstrating the potential benefits of using RPCA in datasets with complex structures.

The analysis becomes more pronounced when PCA is compared to WPCA, where the effectiveness score is found to be -0.5620. This large negative value underscores the superior capacity of WPCA in retaining variance relative to PCA. The potential advantage of weighted PCA might be attributed to its ability to account for variances differently across different components, which proves beneficial in the presence of heteroscedastic data.

Furthermore, when comparing SPCA and RPCA, an effectiveness score of -0.1909 is found, denoting that RPCA holds a slight edge in retaining more variance than SPCA. This could be a result of RPCA's ability to robustly handle outliers, compared to the sparse nature of SPCA.

The comparison of SPCA to WPCA yields an effectiveness score of -0.6947, suggesting that WPCA substantially outperforms SPCA in terms of variance retention. The ability of WPCA to assign differential weights to components likely explains this performance differential.

Finally, when comparing RPCA to WPCA, the effectiveness score is -0.5038, indicating that WPCA maintains a greater amount of variance compared to RPCA. This further underscores the robustness of WPCA against other methods.

²<https://github.com/dganguli/robust-pca>

³<https://pypi.org/project/wpca/>

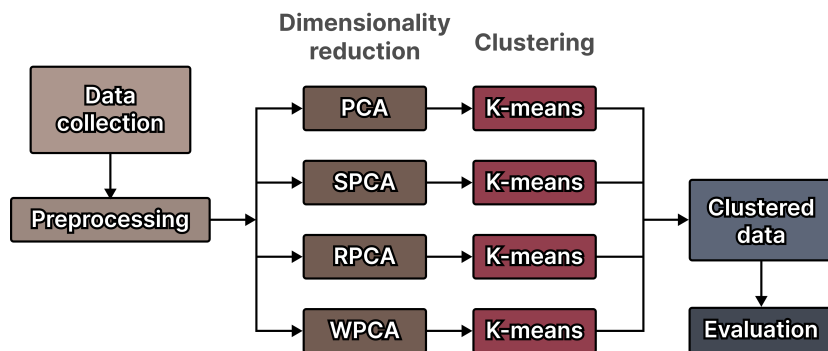


Fig. 1: Schematic illustration of the workflow followed in this research.

TABLE IV: Results of effectiveness of dimensionality reduction methods.

Method	Effectiveness
PCA relative to SPCA	0.1327
PCA relative to RPCA	-0.0582
PCA relative to WPCA	-0.5620
SPCA relative to RPCA	-0.1909
SPCA relative to WPCA	-0.6947
RPCA relative to WPCA	-0.5038

Table V presents the results of clustering, specifically looking at the mean variance of distance sample-centroid and the mean distance among centroids.

For the combination of PCA and K-means clustering, the mean variance of distance sample-centroid is 0.0347, and the mean distance among centroids is 3.3733. The variance value suggests a reasonable compactness within clusters, while the larger mean distance implies good differentiation among clusters.

The SPCA with K-means results show a mean variance of distance sample-centroid of 0.0036, and a mean distance among centroids of 3.0526. These results indicate excellent compactness of clusters, as evidenced by the significantly lower mean variance compared to PCA + K-means. The mean distance among centroids, while slightly less than the PCA + K-means results, still indicates a respectable separation of clusters.

The combination of RPCA with K-means exhibits the smallest mean variance of distance sample-centroid, at 0.000843, suggesting extremely tight clusters. However, the mean distance among centroids is dramatically lower than the previous methods, at 0.3852. This indicates less differentiation among clusters, which may signify a higher likelihood of misclassification between clusters.

Finally, the WPCA with K-means results provide a mean variance of distance sample-centroid of 0.0193, and a mean distance among centroids of 1.1493. The variance is notably lower than PCA + K-means, suggesting better clustering. However, the mean distance among centroids is also less, indicating less separation among clusters compared to PCA and SPCA with K-means.

Based on these metrics, SPCA + K-means appears to offer the best balance of compactness within clusters (low variance of distance sample-centroid) and good separation among clusters (relatively high mean distance among centroids).

TABLE V: Results of clustering according to mean variance of distance sample-centroid and mean distance among centroids.

Method	Mean variance of distance sample-centroid	Mean distance among centroids
PCA + Kmeans	0.0347	3.3733
SPCA + Kmeans	0.0036	3.0526
RPCA + K-means	0.000843	0.3852
WPCA + K-means	0.0193	1.1493

Table VI details the cluster assignments for each country using four different dimensionality reduction techniques combined with K-means clustering.

At a glance, the PCA, SPCA, and WPCA methods seem to assign most countries to the same cluster, indicating that these techniques might be identifying similar structures within the data. However, the RPCA method appears to provide more varied cluster assignments, suggesting that it might be detecting different patterns or structures compared to the other methods.

Moreover, to complement these initial observations and provide a quantitative measure of the similarity between the cluster assignments from each technique, Fig. 2 presents the rand index (RI) comparisons for our dimensionality reduction techniques studied, each paired with the K-means clustering algorithm.

The results indicate a high degree of similarity in the clustering outcomes between PCA, SPCA, and WPCA, as evidenced by RI values of 1.0. This suggests these methods identify similar cluster structures within the data. However, when these methods are compared with RPCA, the RI drops to 0.51. This consistent decrease in RI suggests that RPCA is identifying different clustering structures compared to the other methods.

A possible explanation for this could be the robustness of RPCA to noisy data and outliers. This suggests that the other methods might not be capturing certain features of the dataset that RPCA is able to identify.

TABLE VI: Clusters assignment of each method.

Country	PCA+K-means	SPCA+K-means	RPCA+K-means	WPCA+K-means
Argentina	2	1	2	0
Australia	0	0	1	2
Brazil	0	0	1	2
Canada	0	0	1	2
China	1	2	0	1
European Union	1	2	2	1
France	0	0	2	2
Germany	0	0	1	2
India	0	0	2	2
Indonesia	0	0	1	2
Italy	0	0	2	2
Japan	0	0	1	2
Mexico	0	0	1	2
Netherlands	0	0	2	2
Russia	0	0	0	2
Saudi Arabia	0	0	1	2
Singapore	0	0	2	2
South Africa	0	0	0	2
South Korea	0	0	1	2
Spain	0	0	2	2
Switzerland	0	0	1	2
Turkey	2	1	2	0
United Kingdom	0	0	2	2
United States	1	2	0	1

In summary, these results suggest that while PCA, SPCA, and WPCA all provide similar clustering results when used in conjunction with K-means, RPCA tends to produce different results. This discrepancy might be due to the unique features of RPCA, particularly its robustness to outliers, which may cause it to identify different data structures.

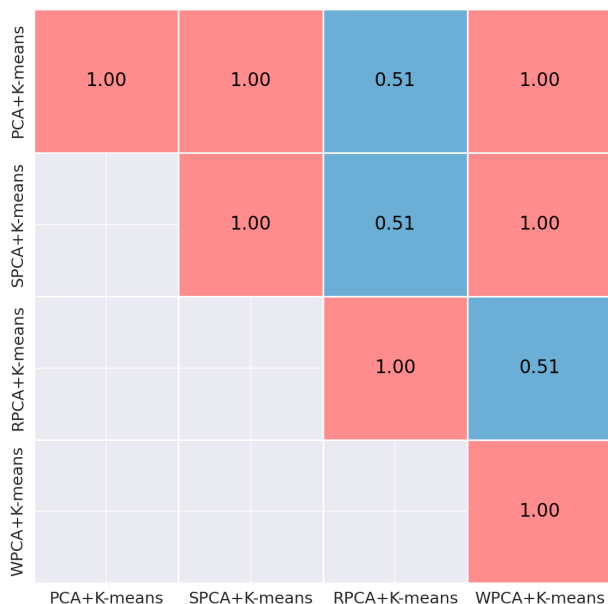


Fig. 2: Comparison of clustering methods using rand index

Finally Fig. 3 graphically shows the results of the clusters generated by the techniques under study. The results are shown as a scatter plot, with each point representing a country in the

original data set. The colors of the points represent the cluster that each data point was assigned to by K-Means.

The first observation is that the PCA+K-Means plot exhibits a clear separation of clusters, which is a testament to PCA being a linear dimensionality reduction technique. It identifies the principal components that account for the most variance in the data, thereby facilitating the clustering process for K-Means.

In the case of the SPCA+K-Means plot, the separation between the clusters is discernible, but slightly smaller than the PCA+K-Means plot. SPCA, as a sparse PCA technique, denoises the data which can add complexity to the clustering process performed by K-Means. However, it enhances the interpretability of the clusters.

Contrarily, the RPCA+K-Means plot exhibits the least separation between clusters. This could be attributed to the fact that RPCA, as a robust PCA technique, focuses on outlier removal. While this might add difficulty to the clustering process, it improves the robustness of the clusters against noise. This technique also demonstrates the most dispersion among the clusters, implying greater variability within each cluster.

The WPCA+K-Means plot, however, manifests the most clear-cut separation between clusters. This is due to WPCA's nature as a weighted PCA technique, assigning different weights to various features in the data. Although this can complicate the K-Means clustering, carefully chosen weights can enhance the interpretability of the clusters. Interestingly, this plot shows the least dispersion among the clusters, suggesting a more compact, tightly grouped cluster structure.

Considering the dispersion, WPCA+K-Means clustering illustrates the least dispersion, indicating tighter, more homogeneous clusters. This can be particularly advantageous when greater intra-cluster similarity is needed.

Conversely, RPCA+K-Means presents the most dispersion, reflecting higher variability within clusters, possibly due to its robustness to outliers. PCA+K-Means and SPCA+K-Means exhibit similar dispersion levels, both sitting between RPCA and WPCA. These differences in dispersion offer valuable insights into the inherent variability within each cluster produced by the different dimensionality reduction techniques.

IV. CONCLUSIONS AND FUTURE WORKS

In this work, we embarked on a comparative exploration of several dimensionality reduction techniques, namely, Principal Component Analysis (PCA), Sparse PCA (SPCA), Robust PCA (RPCA), and Weighted PCA (WPCA). Utilizing a dataset of economic indicators from G20 member countries, we employed these techniques to transform the high-dimensional data into a more manageable format conducive to clustering analysis.

We evaluated the effectiveness of each dimensionality reduction method based on their ability to retain the inherent variance in the data. Subsequently, the reduced-dimensionality data was subjected to K-means clustering, with the resulting clusters assessed through metrics such as mean variance of distance sample-centroid and mean distance among centroids.

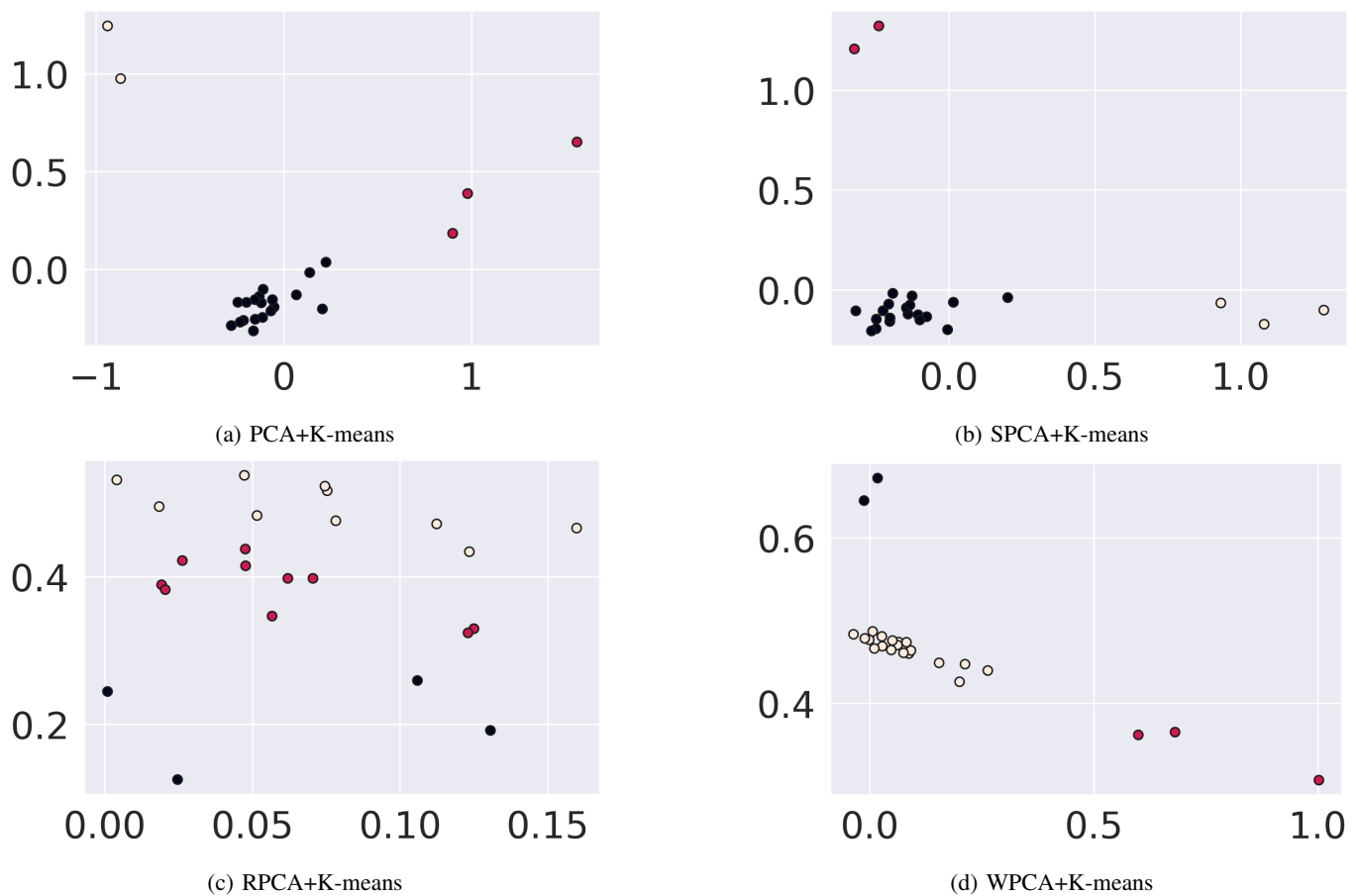


Fig. 3: Visualization of the clusters generated by each of the studied techniques.

The clustering assignments for each dimensionality reduction method were further compared and quantified using the rand index.

Our analysis revealed WPCA as the most effective technique for variance retention, a finding that could be attributed to its differential weighting of components. Despite this, the results also illuminated that a higher variance retention does not necessarily equate to superior clustering performance. In this regard, SPCA combined with K-means emerged as the most balanced approach, offering excellent compactness within clusters and commendable separation among clusters.

While there was substantial consistency in clustering assignments from PCA, SPCA, and WPCA, the clusters derived from RPCA were more diverse, hinting at its ability to uncover unique patterns within the data.

These insights underline the importance of methodological selection tailored to the data and analysis objectives, shedding light on the varied and context-specific strengths of each dimensionality reduction technique.

Overall, this work contributes to the evolving landscape of dimensionality reduction techniques and their application in data analysis, providing a foundational reference point for researchers and practitioners navigating high-dimensional datasets. Through continued refinement of these methods and

broadened understanding of their strengths and limitations, the power and utility of dimensionality reduction can be enhanced in diverse real-world contexts.

Looking ahead, future research could extend the present findings by applying these dimensionality reduction methods to different datasets, enabling a deeper understanding of their applicability and generalizability. Exploration of other clustering algorithms paired with these techniques could further enrich the discourse on the compatibility and performance of different method combinations. Furthermore, comprehensive studies into the computational efficiencies of these methods will be instrumental as data volumes continue to grow.

REFERENCES

- [1] S. Pitafi, T. Anwar, and Z. Sharif, "A taxonomy of machine learning clustering algorithms, challenges, and future realms," *Applied Sciences*, vol. 13, no. 6, 2023.
- [2] S. Aysha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020.
- [3] M. Paricherla, S. Babu, K. Phasinam, H. Pallathadka, A. S. Zamani, V. Narayan, S. K. Shukla, and H. S. Mohammed, "Towards development of machine learning framework for enhancing security in internet of things," *Security and Communication Networks*, vol. 2022, p. 1–5, 2022.
- [4] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: A review," *Complex & Intelligent Systems*, vol. 8, no. 3, p. 2663–2693, 2022.

- [5] M. Allaoui, M. L. Kherfi, and A. Cheriet, "Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study," in *Image and Signal Processing*, A. El Moataz, D. Mammass, A. Mansouri, and F. Nouboud, Eds. Cham: Springer International Publishing, 2020, pp. 317–325.
- [6] M. Greenacre, P. J. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, 2022.
- [7] F. Kherif and A. Latypova, "Chapter 12 - principal component analysis," in *Machine Learning*, A. Mechelli and S. Vieira, Eds. London: Academic Press, 2020, pp. 209–225.
- [8] M. Alkhayrat, M. Aljnidi, and K. Aljoumaa, "A comparative dimensionality reduction study in telecom customer segmentation using deep learning and pca," *Journal of Big Data*, vol. 7, no. 1, 2020.
- [9] C. Acal, A. M. Aguilera, and M. Escabias, "New modeling approaches based on varimax rotation of functional principal components," *Mathematics*, vol. 8, no. 11, 2020.
- [10] F. L. Gewers, G. R. Ferreira, H. F. D. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. D. F. Costa, "Principal component analysis: A natural approach to data exploration," *ACM Comput. Surv.*, vol. 54, no. 4, may 2021.
- [11] H. Pan, D. Badawi, I. Bassi, S. Ozev, and A. E. Cetin, "Detecting anomaly in chemical sensors via l1-kernel-based principal component analysis," *IEEE Sensors Letters*, vol. 6, no. 10, pp. 1–4, 2022.
- [12] N. Migenda, R. Möller, and W. Schenck, "Adaptive dimensionality reduction for neural network-based online principal component analysis," *PLOS ONE*, vol. 16, no. 3, 2021.
- [13] S. V. Narwane and S. D. Sawarkar, "Dimensionality reduction of unbalanced datasets: Principal component analysis," in *2021 Asian Conference on Innovation in Technology (ASIANCON)*, 2021, pp. 1–6.
- [14] D. Bertsimas and R. Cory-Wright, "Solving large-scale sparse pca to certifiable (near) optimality," *J. Mach. Learn. Res.*, vol. 23, no. 1, jan 2022.
- [15] J. Bian, D. Zhao, F. Nie, R. Wang, and X. Li, "Robust and sparse principal component analysis with adaptive loss minimization for feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.
- [16] N. B. Erichson, P. Zheng, K. Manohar, S. L. Brunton, J. N. Kutz, and A. Y. Aravkin, "Sparse principal component analysis via variable projection," *SIAM Journal on Applied Mathematics*, vol. 80, no. 2, pp. 977–1002, 2020.
- [17] H. Robert Frost, "Eigenvectors from eigenvalues sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 31, no. 2, p. 486–501, 2021.
- [18] A. Saeed, Z. Bangash, H. Farooq, and I. Akhtar, "Robust principal component analysis of vortex-induced vibrations using particle image velocimetry measurements," in *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*, 2021, pp. 774–781.
- [19] J. Pan, Y. Cai, Y. Xie, T. Lin, Y. Gao, and C. Cao, "Robust principal component analysis based on purity," in *2022 34th Chinese Control and Decision Conference (CCDC)*, 2022, pp. 2017–2023.
- [20] Y. Gao, T. Lin, Y. Zhang, S. Luo, and F. Nie, "Robust principal component analysis based on discriminant information," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 2, pp. 1991–2003, 2023.
- [21] I. Scherl, B. Strom, J. K. Shang, O. Williams, B. L. Polagye, and S. L. Brunton, "Robust principal component analysis for modal decomposition of corrupt fluid flows," *Phys. Rev. Fluids*, vol. 5, p. 054401, May 2020.
- [22] S. Wang, F. Nie, Z. Wang, R. Wang, and X. Li, "Robust principal component analysis via joint reconstruction and projection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022.
- [23] P. Kumar and A. Khani, "Evaluating special event transit demand: A robust principal component analysis approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 12, pp. 7370–7382, 2021.
- [24] L. Delchambre, "Weighted principal component analysis: a weighted covariance eigendecomposition approach," *Monthly Notices of the Royal Astronomical Society*, vol. 446, no. 4, pp. 3545–3555, 12 2014.
- [25] H. Yue and M. Tomoyasu, "Weighted principal component analysis and its applications to improve fdc performance," in *2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No.04CH37601)*, vol. 4, 2004, pp. 4262–4267 Vol.4.
- [26] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, 2020.
- [27] S. A. Abbas, A. Aslam, A. U. Rehman, W. A. Abbasi, S. Arif, and S. Z. H. Kazmi, "K-means and k-medoids: Cluster analysis on birth data collected in city muzaffarabad, kashmir," *IEEE Access*, vol. 8, pp. 151 847–151 855, 2020.
- [28] W. Chang, X. Ji, Y. Liu, Y. Xiao, B. Chen, H. Liu, and S. Zhou, "Analysis of university students' behavior based on a fusion k-means clustering algorithm," *Applied Sciences*, vol. 10, no. 18, 2020.
- [29] L. Li, J. Wang, and X. Li, "Efficiency analysis of machine learning intelligent investment based on k-means algorithm," *IEEE Access*, vol. 8, pp. 147 463–147 470, 2020.
- [30] H. Zare and S. Emadi, "Determination of customer satisfaction using improved k-means algorithm," *Soft Computing*, vol. 24, no. 22, p. 16947–16965, 2020.
- [31] A. Ashabi, S. B. Sahibuddin, and M. Salkhordeh Haghighi, "The systematic review of k-means clustering algorithm," in *Proceedings of the 2020 9th International Conference on Networks, Communication and Computing*, ser. ICNCC '20. New York, NY, USA: Association for Computing Machinery, 2021, p. 13–18.
- [32] M. Enríquez, S. Naranjo, I. Amaro, and F. Camacho, "Dimensionality reduction using pca and cur algorithm for data on covid-19 tests," *Advances in Intelligent Systems and Computing*, vol. 1326 AISC, pp. 121–134, 2021.
- [33] G. Hunt, "Cur: An interpretable alternative to principal components analysis," 2013, ph.D. thesis, Drew University.
- [34] S. Xia, D. Peng, D. Meng, C. Zhang, G. Wang, E. Giem, W. Wei, and Z. Chen, "Ball kk -means: Fast adaptive clustering with no bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 87–99, 2022.
- [35] J. Riofrío, C. Muñoz-Moncayo, I. R. Amaro, and I. Pineda, "Identifying similar groups of countries according to the impact of corona virus (covid-19) by a two-layer clustering method," in *Artificial Intelligence, Computer and Software Engineering Advances*, M. Botto-Tobar, H. Cruz, and A. Díaz Cadena, Eds. Cham: Springer International Publishing, 2021, pp. 34–48.
- [36] L. Cao, Z. Zhao, and D. Wang, *Clustering Algorithms*. Singapore: Springer Nature Singapore, 2023, pp. 97–122.
- [37] V. Robert, Y. Vasseur, and V. Brault, "Comparing high-dimensional partitions with the co-clustering adjusted rand index," *Journal of Classification*, vol. 38, no. 1, p. 158–186, 2020.