Applied Mathematics and Computational Intelligence
Volume 14, No. 3, 2025 [37-46]

Outlier Map of Classical and Robust Principal Component Analysis

Noor Wahida Jamil[1*], Shamshuritawati Sharif[2]

[1]Faculty of Computer and Mathematical Sciences, UiTM Jasin, Malacca, Malaysia
[1,2]School of Quantitative Sciences, UUM Sintok, Kedah, Malaysia

[*] Corresponding author : wahidajamil@uitm.edu.my

Received: 24 February 2025
Revised: 27 April 2025
Accepted: 30 May 2025

ABSTRACT

Classical Principal Component Analysis (CPCA) is widely used for dimensionality reduction, but it is highly sensitive to the presence of outliers, leading to distorted covariance estimates and unreliable principal components. To address this, Robust PCA (ROBPCA) integrates robust covariance estimation and projection pursuit to minimize the effects of outliers. Although CPCA and ROBPCA are often utilized for high-dimensional data $(p > n)$*, it is equally effective in low-dimensional settings, particularly when handling large datasets. This research illustrates the benefit of ROBPCA over CPCA by analyzing a large-scale gene expression dataset with 22 features and 47231 observations* $(p < n)$ *to demonstrate its efficiency in identifying and classifying outliers using outlier maps. Findings reveal that CPCA misidentifies outliers, leading to inflated variance structures and poor principal component estimation, whereas ROBPCA successfully isolates outliers, preserving data integrity and enhancing interpretability. This research emphasizes how ROBPCA improves data reliability and offers a reliable method for identifying outliers.*

Keywords: Outliers, Classical Principal Component Analysis (CPCA), Robust PCA (ROBPCA).

1 INTRODUCTION

Data preprocessing is important in ensuring that analytical models capture meaningful patterns while minimizing distortions. Both, high-dimensional and low-dimensional datasets are prone to noise, missing values and outliers, which can obscure structural relationships. Proper preprocessing techniques improve data representation, interpretability and stability in subsequent analyses. Without these steps, dimensionality reduction methods like CPCA may yield misleading or biased results.

Data dimensionality pertains to the number of variables or features in a dataset. High-dimensional data, where the number of features exceeds the number of observations $(p > n)$, poses challenges for traditional statistical methods [1]. The challenges include the curse of dimensionality, model overfitting, and high generalization errors [2]-[3]. The curse of dimensionality occurs when data becomes sparse and meaningful patterns are difficult to detect as dimensionality increases [4], while model overfitting occurs when models built on high-dimensional data fit noise rather than

underlying patterns [5]. Additionally, generalization error refers to the difference between a model's performance on training and test data [6]. A model with low generalization error performs well not only on the data it was trained on, but also on the new dataset. Conversely, high generalization error indicates that the model has overfitted to the training data [7].

To address these issues, dimensionality reduction techniques like CPCA help reduce the number of features while preserving the most important information [8]. This process enhances computational efficiency, reduces overfitting, assists machine learning models to perform better on unseen data, and helps mitigate generalization errors [9]-[10].

CPCA is popular due to its simplicity and low computational cost; however, a clear understanding of its theoretical foundations and limitations is required for effective implementation [11]. Additionally, CPCA plays an important role in dimensionality reduction; nevertheless, it focuses on data visualization, noise reduction, feature extraction and data preprocessing, in low-dimensional cases. CPCA is widely used in diverse fields, including machine learning, bioinformatics, image processing and medicine [12]-[15].

Despite the aforementioned advantages, its results are sensitive to outliers because CPCA relies on covariance matrices, which are significantly influenced by outliers [16]. When data contains extreme values, the accuracy of CPCA may be compromised. Researchers have proposed various robust alternatives to mitigate this issue, including Robust Principal Component Analysis (ROBPCA). ROBPCA was introduced by combining projection pursuit with robust scatter matrix estimation, yielding improved estimation in both contaminated and noncontaminated datasets [17]. Additionally, due to the presence of outliers, a diagnostic plot for outlier classification was also proposed in [17], complementing the ROBPCA in improving the reliability of CPCA. This approach has improved dimension reduction, data preprocessing, and classification results [14].

ROBPCA has appeared as an alternative approach that aims to minimize the effects of outliers, providing a more stable and accurate framework for data analysis. Although CPCA and ROBPCA are often utilized for high-dimensional data ($p > n$), it is equally effective in low-dimensional settings, particularly when handling large datasets. The primary research question of this study is: How does ROBPCA perform compared to CPCA in handling outliers and identifying reliable principal components? To investigate this question, this research utilized a publicly available gene expression dataset focused on understanding the regulatory role of the KDM3A histone demethylase enzyme in estrogen receptor (ER) signalling within breast cancer cells. A full description of the dataset and the methodologies involved will be discussed thoroughly in the next section.

2 METHODS

2.1 Classical PCA

Classical PCA (CPCA) is a statistical approach that reconstructs the original set of variables X_j into a smaller number of uncorrelated and orthogonal variables $\tilde{P}_j$, called principal components. Basically, $\tilde{P}_j$ represents the linear combinations of the mean-centred variables $\tilde{X}_j = X_j - \bar{X}$. It should be emphasized that these components align with the eigenvectors of the sample covariance matrix $S =$

$1/(n-1)\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'$ of the data. For each component vector $\tilde{P}_j$, the corresponding eigenvalue $\tilde{l}_j$ of S indicates the amount of the data's variability that is explained by $\tilde{P}_j$ trough the relation $\widetilde{l_j}$= Var $(\tilde{P}_j)$. Usually, these component vectors are arranged in descending order of the eigenvalues. Therefore, the first k principal components describe most of the data's variability.

In CPCA process, after k is selected, the p-dimensional data points can be projected onto the subspace formed by the k component vectors, and their coordinates can be calculated with respect to these $\tilde{P}_j$; This produces the score vector for observation i, $\tilde{t}_i$

$$\tilde{t}_i = \tilde{P}'(x_i - \bar{x}) \tag{1}$$

For each $i = 1, \ldots, n$, which naturally have a zero mean and $\tilde{P}'$ is the transpose of the component vector $\tilde{P}_j$. In the original coordinate system, the projected data point is calculated as the fitted value.

$$\hat{x}_i = \bar{x} + \tilde{P}\,\tilde{t}_i \tag{2}$$

Note that the *(p × k)* component matrix $\tilde{P}$ comprises the components in columns. The *(k × k)* diagonal matrix $\tilde{L} = \left(\tilde{l}_j\right)_j$ will be used to represent the eigenvalues (in descending order). Many criteria exist to select the suitable number of components k. A common graphical method is based on the scree plot, where eigenvalues representing the variance explained by each component are plotted in decreasing order. The plot is examined for a point where the curve flattens, known as the "elbow," which marks the last significant component to retain [20]. Components after this point explain little additional variance and are typically excluded, as they contribute minimally to the analysis. A more structured approach considers the total variation accounted by the first k components, and necessitates, approximately,

$$\left(\sum_{j=1}^{k}\tilde{l}_j\right)/\left(\sum_{j=1}^{p}\tilde{l}_j\right) \geq 80\% \tag{3}$$

CPCA is unreliable with the presence of outliers because the sample covariance matrix S and mean vector are easily affected by outliers. Furthermore, from the score plot, the groups of outliers can be distinguished, but several undesirable effects are foreseen. The plot has superimposed the 97.5% tolerance ellipse, defined by the set of vectors whose squared Mahalanobis distance is the 0.975 quantile of the $\chi^2_{\cdot}$– distribution with k degrees of freedom:

$$E_{0.975} = \left\{t \in R^k; MD(t) = \sqrt{\chi^2_{k,0.975}}\right\},$$

with

$$MD(t) = \sqrt{\tilde{t}'\,\tilde{L}^{-1}\,\tilde{t}} \tag{4}$$

Typically, the inner region of the ellipse, where the tolerance ellipse represents a 97.5% confidence region for the data center μ, illustrates the covariance structure among variables, when data points are normally distributed with k-dimensions. Additionally, the coordinate axes (the components) aligned with the primary axes of the ellipse due to the uncorrelated nature of the scores. Data points are classified as regular if they fall inside the ellipse, because the different variances of the

components caused them to be allocated not too far from the estimated center. Contradict to this, when data points fall in the remaining 0.025 tail probability outside the ellipse, it represents potential outliers.

2.2 Robust PCA

A brief description of the Robust PCA (ROBPCA) algorithm is provided here, with more details available in Robust PCA and classification in biosciences [21]. Minimum Covariance Determinant (MCD) is popular because of the development of the fast algorithm to perform its calculations and the strong resistance to outliers [22].

In defining the MCD estimator, subsets of size h from the complete dataset of size n are considered. The number h determines the robustness of the estimator and should be no less than $[(n + p + 1)/2]$. Then, the MCD estimator identifies the h-subset with the covariance matrix S has the smallest determinant. The MCD location estimate, $\hat{\mu}_{MCD}$ is given by the average of this optimal h-subset, while the MCD scatter estimate, $\hat{\Sigma}_{MCD}$ is given by its covariance matrix, multiplied by a consistency factor. Following the raw MCD estimation, a step of reweighting can be added to considerably improve the efficiency of the finite-sample estimator considerably. A weight of 1 is received by each data point $\boldsymbol{x_i}$ if its robust distance is defined as

$$RD(\boldsymbol{x_i}) = \sqrt{(x_i - \hat{\mu}_{MCD})' \hat{\Sigma}_{MCD}^{-1} (x_i - \hat{\mu}_{MCD})} \tag{5}$$

is lesser than $c = \sqrt{\chi^2_{p,0.975}}$ and assigned a weight of zero. The reweighted MCD estimator is then equal to the average of data points and covariance matrix S. The first k eigenvectors of the MCD covariance matrix, ordered by decreasing eigenvalues, yield robust components.

In technical terms, the MCD estimator has a breakdown value of $(n–h+1)/n$, meaning that at least n−h+1 outliers are needed to render the estimates ineffective. Due to this, it is understood that the MCD estimator can resist $n–h$ outliers. Typically, h is set to approximately $0.75n$, allowing up to 25% of the data to be outliers.

When dealing with high-dimensional regressors ($p > n$), the covariance matrix for $h < p$ observations will always be zero, which cannot be minimized, hence the MCD method cannot be applied [21]. In such cases, ROBPCA proceeds differently. First, the x-data are preprocessed by reducing the data space to the affine subspace spanned by the n observations. This reduction is not difficult to achieve through a singular value decomposition (SVD) of the data matrix X, allowing the data to be represented using up to $n - 1 = rank(\tilde{X})$ variables, preserving all relevant information.

In the next step of the ROBPCA algorithm, a measure of "outlyingness" is calculated for each data point by projecting high-dimensional data points onto several univariate directions, $\mathbf{v}$. For each direction, a robust center and scale of the projected data points $\boldsymbol{x_i'v}$ are calculated, namely the univariate MCD estimator for location m_{MCD} and s_{MCD}. Then, the standardized distance of each data point from this center is measured, and the maximum distance across all directions is selected for each point. This yields the outlyingness

$$outl(\boldsymbol{x_i}) = \max_{v} \frac{|\boldsymbol{x_i'v} - m_{MCD}(\boldsymbol{x_i'v})|}{s_{MCD}} \quad (6)$$

In ROBPCA, the h data points with the minimum outlyingness are retained, and from the covariance matrix Σ_1 of this h-subset, the number of principal components to retain, k, is selected. In the final stage of ROBPCA, data points are projected onto the k-dimensional subspace defined by the largest k eigenvectors of Σ_1 and of calculating their shape and center using the reweighted MCD estimator. The resulting eigenvectors form the robust principal components and the MCD location estimate providing a robust data center. The outcome of the ROBPCA method is thus a robust estimate of the data center, $\hat{\boldsymbol{\mu}}$, a set of robust components, P, and eigenvalues, l_j (for j=1,..., k) and, similar to (1) robust scores

$$t_i = P'(x_i - \hat{\mu}) \quad (7)$$

2.3 Outlier Map

The results of the CPCA analysis can also be visualized in a diagnostic plot, often called an "outlier map." This map identifies and categorizes outliers based on their locations relative to the primary data. In CPCA terms, an outlier is any observation that either lies far from the subspace defined by the k principal components or is located distant from the majority of data in the subspace [20]. The orthogonal distance (OD) and the score distance (SD) are used to evaluate the outlyingness. The OD measures how far an observation is from its projection in the k-dimensional CPCA subspace.

$$OD_i = ||\boldsymbol{x_i} - \hat{\boldsymbol{x}}_i|| \quad (8)$$

Eigenvalues provide information of the score's covariance structure, while SD measures the distance within the CPCA subspace and is defined as

$$SD_i = \sqrt{t_i' L^{-1} t_i} = \sqrt{\sum_{j=1}^{k} \left(t_{ij}^2 / l_j\right)} \quad (9)$$

When employing CPCA, the score distance aligns precisely with Mahalanobis distance, making it a crucial measure for detecting outliers. The OD and SD together classify observations into four distinct categories, as illustrated in Figure 1a**.**

Regular observations exhibit both small OD and SD values, indicating that they lie within the principal component subspace and conform to the general data structure. In contrast, good leverage points are data points with large SD but small OD. These observations lie near the principal component subspace but remain somewhat distinct from the main data cluster. Observations 1 and 4 in Figure 1a fall in this category, suggesting that their influence is limited, while they differ from most data points. Since only a small amount of information is lost when they are replaced by their estimated values in the CPCA subspace, their presence does not significantly distort the PCA results.

Another category, orthogonal outliers, consists of observations with large OD but small SD. These data points lie outside the principal component subspace but appear similar to regular observations when projected onto it. Observation 5 in Figure 1a is an example of this type. Because orthogonal

outliers do not deviate significantly in the principal component space, they can be mistaken for regular observations. However, their actual distance from the subspace suggests their projected values should not replace them, as this would conceal their outlier nature.

The fourth category, bad leverage points, includes observations with both large OD and large SD. These data points are far from both the subspace and the main data cluster, often influencing CPCA results by shifting eigenvectors toward them. Observations 2 and 3 in Figure 1a belong to this category and are considered the most problematic outliers, as they can distort the principal component structure and affect the overall analysis.

Figure 1b, provides an outlier map, where each observation's orthogonal distance (OD_i) is plotted against its score distance (SD_i), allowing the classification of observations. In this figure, the boundary lines are drawn to differentiate between observations with small and large OD and SD values, helping to categorize them based on their deviation from the main data structure. The cut-off for SD is derived from the chi-squared (χ^2) distribution with k degrees of freedom, as normally distributed data lead to normally distributed scores, whose squared Mahalanobis distances follow a chi-squared (χ^2) distribution. Hence the cut-off value $c = \sqrt{X^2_{p,0.975}}$ is used, while the cut-off for OD follows an approximation introduced in theorems on quadratic forms applied in the study of analysis of variance [24], where the squared OD values can be approximated by a scaled chi-squared (χ^2) distribution with $g1$ degrees of freedom, $OD^2 \sim g_2 x^2_{g_1}$. Estimates for $g1$ and $g2$ are obtained using the Wilson–Hilferty transformation, which transforms chi-squared distributed data into a form that approximates normality (see Hubert et al., 2004, http://www.wis.kuleuven.ac.be/stat/robust.html).

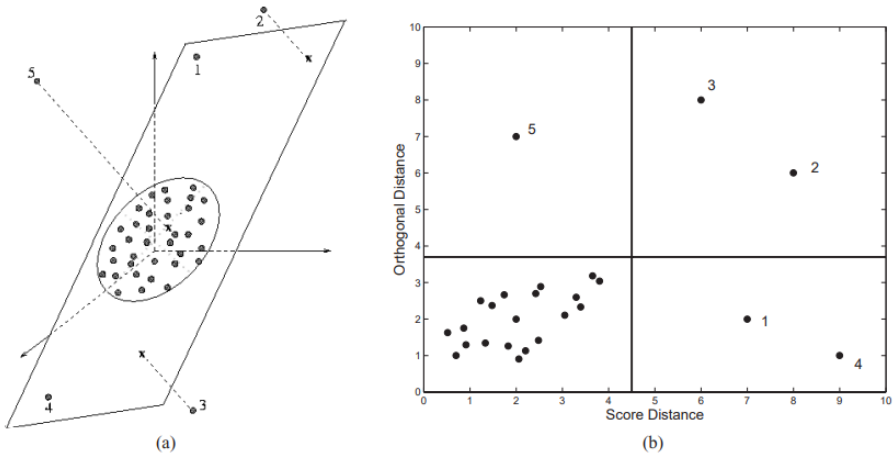


Figure 1: (a) Different types of outliers when a three-dimensional dataset is projected on a robust two-dimensional CPCA-subspace. (b) The corresponding outlier map [21].

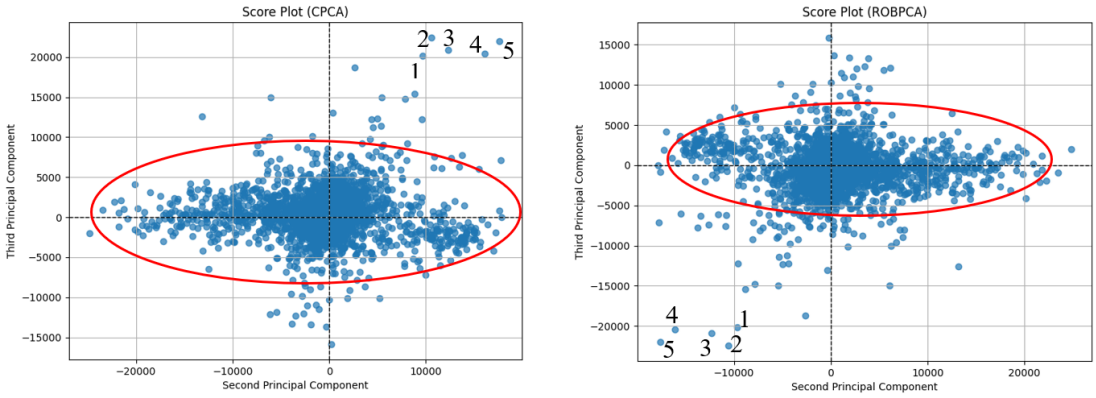
3 RESULTS

Normalizing data by dividing each row by its maximum value is not robust, as it hinges on the largest value measured [21]. To avoid this issue, this study only caters for missing values in the preprocessing part and analyzes the raw data without normalization.

The dataset used in this study was obtained from the Gene Expression Omnibus (GEO) repository, specifically GSE68918 published in Nucleic Acids Research [27]. It comprises gene expression profiles of the human breast cancer cell line MCF-7, where an RNA interference (RNAi) screen identified the histone demethylase enzyme KDM3A as a key regulator of estrogen (E2) signalling. This dataset has 22 features ($p = 22$) and 47231 observations ($n = 47231$), including a group of outliers (1-5) in Figure 2a. The first six eigenvalues describe nearly 100% of the total variation, so six principal components ($k = 6$) were selected for analysis.

Figure 2a illustrates distinguishable outlier groups, but several undesirable effects are foreseen. Although CPCA scores have a zero mean, biased mean estimation shifts regular data points away from zero. This moved the data center toward the outlying group, causing the origin to fall outside the regular data cluster. Furthermore, outliers 1-5 remain within the 97.5% tolerance ellipse, as Mahalanobis distance fails to detect them, leading to inflated variance and misclassification.

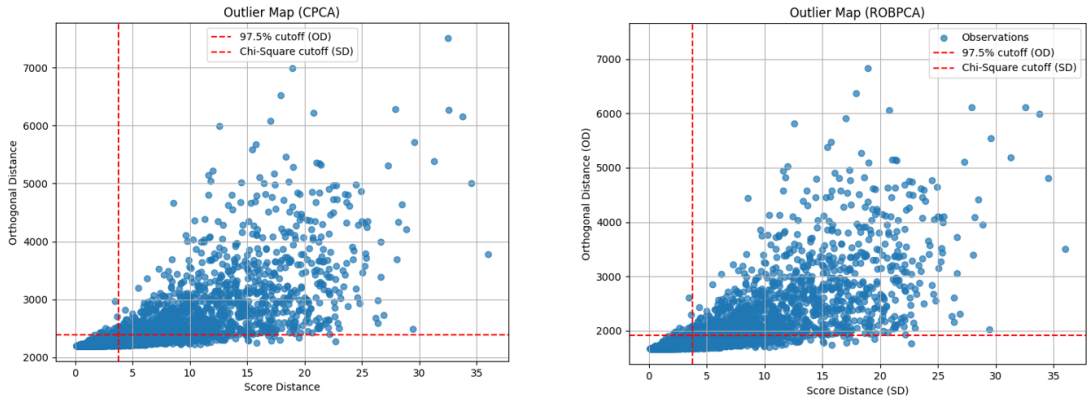
Figure 2b displays the score of the KDM3A data as determined by ROBPCA. It is now evident that the center has been accurately estimated, and positioned among the regular observations. The 97.5% tolerance ellipse effectively encloses these points while excluding all 5 outliers.



(a) (b)

Figure 2: Score plot and 97.5% tolerance ellipse of the KDM3A dataset acquired with (a) CPCA and (b) ROBPCA.

Additionally, Figure 3 compares CPCA and ROBPCA in detecting outliers using score distance (SD) and orthogonal distance (OD). CPCA which is sensitive to outliers, misclassifies more outliers, particularly those classified as bad leverage points, due to its inability to handle them effectively. In contrast, ROBPCA minimizes outlier influence, resulting in fewer misclassifications of such points and demonstrating its superior ability to manage outliers.



(a) (b)

Figure 3: The outlier map of the KDM3A dataset, based on two principal components obtained with (a) CPCA and (b) ROBPCA.

4 DISCUSSION

The results from the study demonstrate the advantages of ROBPCA in handling data containing outliers, particularly when compared to CPCA. A detailed analysis proved that while CPCA is significantly influenced by outliers, leading to inflated variance-covariance structures and inaccurate data representations, ROBPCA employs advanced techniques such as projection pursuit and robust covariance estimation to rectify this problem effectively. The use of outlier maps in the study further highlighted ROBPCA's ability to accurately identify and classify outliers based on score and orthogonal distances, resulting in a more precise data analysis framework. This distinction is particularly critical in datasets like the KDM3A case, where CPCA struggled to handle outliers, leading to distorted eigenvectors and inaccurate data center estimation. In contrast, ROBPCA delivered stable results by efficiently excluding the outliers from the main cluster. These findings highlight ROBPCA's potential to improve the reliability and interpretability of analysis, particularly when dealing with complex datasets that consist of outliers.

5 CONCLUSION

ROBPCA is demonstrated to be a superior alternative to CPCA for datasets affected by outliers, particularly for large, complex datasets. While both methods are commonly applied to high-dimensional data ($p > n$), they remain equally effective in low-dimensional settings, especially when handling large datasets. By employing robust covariance estimation, ROBPCA preserves the data structure and improves the accuracy of principal components. This makes ROBPCA particularly valuable in fields like bioinformatics, finance, and engineering, where outliers can significantly distort analytical results and compromise reliability.

ACKNOWLEDGEMENT

The authors sincerely thank the anonymous reviewers for their constructive comments and insightful suggestions. Financial support from Universiti Teknologi MARA (UiTM) and the Ministry of Higher Education Malaysia (MOHE) under the Skim Latihan Akademik IPTA 2.0 (SLAI) is gratefully acknowledged.

REFERENCES

[1] H. Momeni and A. Ebrahimkhanlou, "High-dimensional data analytics in structural health monitoring and non-destructive evaluation: a review paper," in *Smart Materials and Structures*, vol. 31, no. 4, p. 043001, 2022. https://doi.org/10.1088/1361-665x/ac50f4.

[2] F. Al-nagashi, N. Rahim, S. Shukor, and M. Hamid, "Mitigating overfitting in extreme learning machine classifier through dropout regularization", in *AMCI*, vol. 13, no. 1, pp. 26-35, 2024. https://doi.org/10.58915/amci.v13ino.1.561.

[3] M. Gahrooei, J. Whitehurst, Y. Ampatzidis, & P. Pardalos, "Dimensionality reduction techniques for high-dimensional data in precision agriculture", in *Modeling for Sustainable Management in Agriculture, Food and the Environment,* pp. 28-39, 2021. https://doi.org/10.1201/9780429197529-2.

[4] D. Peng, Z. Gui, and H. Wu, "Interpreting the curse of dimensionality from distance concentration and manifold effect," *Preprint at arXiv*, vol. 2401.00422v3, pp. 1-21, 2025.

[5] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, p. 022022, 2019. https://doi.org/10.1088/1742-6596/1168/2/022022.

[6] Z. Zhang and X. Zhang, "A Review of Research on Generalization Error Analysis of Deep Learning Models," in *2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, Chengdu, China, 2023, pp. 906-912, doi: 10.1109/ICICML60161.2023.10424837.

[7] N. S. V. Rao, "Study of overfitting by machine learning methods using generalization equations," in *2023 26th International Conference on Information Fusion (FUSION)*, Charleston, SC, USA, 2023, pp. 1-8, doi: 10.23919/FUSION52260.2023.10224198.

[8] F. R. Mulla, and A. K. Gupta, "A Review Paper on Dimensionality Reduction Techniques," in *Journal of Pharmaceutical Negative Results*, vol. 13, no. 3, 2022. pp. 1263-1272, https://doi.org/10.47750/pnr.2022.13.s03.198.

[9] X. Song, Z. Yon, D. Gong, & X. Gao, "A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data", in *IEEE Transactions on Cybernetics*, vol. 52, no. 9, pp. 9573-9586, Sept. 2022, doi: 10.1109/TCYB.2021.3061152.

[10] M. A. Salam, A. M. Taher, M. R. Samy, and K. A. Mohamed, "The effect of different dimensionality reduction techniques on machine learning overfitting problem," in *International Journal of Advanced Computer Science and Applications,* vol. 12, no. 4, 2021.

https://doi.org/10.14569/ijacsa.2021.0120480.

[11] J. Palma and G. Pierdominici-Sottile, "On the uses of PCA to characterise molecular dynamics simulations of biological macromolecules: basics and tips for an effective use," in *ChemPhysChem*, vol. 24, no. 2, pp. e202200491, 2022. https://doi.org/10.1002/cphc.202200491.

[12] S. Aslam and T. F. Rabie, "Principal component analysis in image classification: A review," in *2023 Advances in Science and Engineering Technology International Conferences (ASET)*, 2023, pp. 1–7. https://doi.org/10.1109/aset56582.2023.10180847.

[13] B. M. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," in *Computer Science, Medicine,* vol. 2, no. 1, pp. 20–30, 2021, https://doi.org/10.30880/jscdm.2021.02.01.003.

[14] L. He, Y. Yang, and B. Zhang, "Robust PCA for high-dimensional data based on characteristic transformation," in *Australian & New Zealand Journal of Statistics,* vol. 65, no. 2, pp. 127-151, 2023. https://doi.org/10.1111/anzs.12385.

[15] G. T. Reddy, M. P. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and I. T. Srivastava, "in Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776-54788, 2020. https://doi.org/10.1109/access.2020.2980942.

[16] K. Liu and Y. Cao, "Robust principal component analysis: a construction error minimization perspective", *Preprint at arXiv*:2111.12132v1, pp. 1-13, 2021.

[17] M. Hubert, P. Rousseeuw, & K. Branden, "Robpca: a new approach to robust principal component analysis", in *Technometrics*, vol. 47, no. 1, p. 64-79, 2005. https://doi.org/10.1198/004017004000000563.

[20] R. B. Cattell, "The scree test for the number of factors," in *Multivariate Behavioral Research*, vol. 1, no. 2, pp. 245–276, 1966. https://doi.org/10.1207/s15327906mbr0102_10.

[21] M. Hubert and S. Engelen, "Robust PCA and classification in biosciences," in *Bioinformatics*, vol. 20, no. 11, pp. 1728–1736, 2004. https://doi.org/10.1093/bioinformatics/bth158.

[22] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," in *Technometrics,* vol. 41, no. 3, pp. 212–223, 1999. https://doi.org/10.1080/00401706.1999.10485670.

[24] G. E. P. Box, "Some theorems on quadratic forms applied in the study of analysis of variance problems: effect of inequality of variance in one-way classification," in *Annals of Mathematical Statistics*, vol. 25, no. 2, pp. 484–498. http://www.jstor.org/stable/223683.

[27] M. A. Wade, D. Jones, L. Wilson, J. Stockley, et al., "The histone demethylase enzyme KDM3A is a key estrogen receptor regulator in breast cancer," Nucleic Acids Research, vol. 43, no. 1, pp. 196–207, 2015.