*Proceedings of the 6th International Conference on Statistics: Theory and Applications (ICSTA'24)*
*Barcelona, Spain – August 19 - 21, 2024*
*Paper No. 173*
*DOI: 10.11159/icsta24.173*

# Noise Reduced Common PCA for High-Dimensional, Low-Sample Size Multi-View Data

**Hiroki Hasegawa[1], Homura Kawamura[1], Ryota Shin[2], Kazuyoshi Yata[3], Yukihiko Okada[2,4], Jun Kunimatsu[5]**

[1]Master's Program in Service Engineering / University of Tsukuba
1 Chome-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Tsukuba, Japan
First.s2420513@u.tsukuba.ac.jp; Second.s2320503@u.tsukuba.ac.jp

[2] Institute of Systems and Information Engineering / University of Tsukuba
1 Chome-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Tsukuba, Japan
Third.shin.ryota.gn@alumni.tsukuba.ac.jp

[3] Institute of Pure and Applied Sciences / University of Tsukuba
1 Chome-1-1 Tennodai, Tsukuba, Ibaraki 305-8571, Tsukuba, Japan
Fourth.yata@math.tsukuba.ac.jp

[4] Center for Artificial Intelligence Research / University of Tsukuba
1 Chome-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Tsukuba, Japan
Fifth. okayu@sk.tsukuba.ac.jp

[5] Institute of Medicine / University of Tsukuba
1 Chome-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Tsukuba, Japan
Sixth. jkunimatsu@md.tsukuba.ac.jp

***Abstract*** **-** High-Dimensional Low-Sample Size (HDLSS) data pose significant challenges in fields like medicine and neuroscience. Traditional principal component analysis (PCA) often fails under these conditions, leading to unstable eigenvalue estimation. This study introduces Noise Reduced-Common Principal Component Analysis (NR-CPCA), a method that combines Common Principal Component Analysis (CPCA) with a noise reduction technique to enhance eigenvalue stability and reliability in HDLSS data. By comparing eigenvalue estimations from NR-CPCA and traditional CPCA across various dimensions (1000, 2000, 3000) and sample sizes (10 to 120), we demonstrate that NR-CPCA mitigates noise effects more effectively, ensuring stable principal component selection. Simulation results confirm that NR-CPCA reduces variability in eigenvalue estimation, making it a valuable tool for dimensionality reduction in multi-view data. Despite limitations in simulation-based validation, NR-CPCA shows promise for real-world applications in data-intensive fields. Future research should focus on refining this method and applying it to diverse datasets to fully realize its potential. NR-CPCA provides an important advancement for researchers dealing with HDLSS data, promoting more accurate analysis and contributing to progress in data science, biology, and neuroscience.

***Keywords*****:** Big Data Analytics, High-Dimensional Data Analysis, Low-sample Size data, Common Principal Component Analysis, Noise Reduction Techniques, Dimensionality reduction

## 1. Introduction

High-Dimensional Low-Sample Size (HDLSS) data can be found in various fields such as medicine (Rahnenführer et al. [1]) and biology and neuroscience (Liu & Vinck [2]). Dimension reduction is effective for data analysis, and principal component analysis (PCA) is particularly widely used. However, traditional PCA is based on the assumption that the number of samples, $n$, is greater than the number of variables, $d$ (i.e., $n > d$). When this assumption is violated, particularly when $n \ll d$, the mathematical guarantees of the analysis are lost (Yata & Aoshima [3]). Specifically, the estimation of eigenvalues tends to become unstable, and due to the differing geometric properties of the data, the mathematical assumptions of PCA often break down (Yata & Aoshima [4]). There is a method known as common principal component analysis (CPCA) that allows for the fair selection of principal components across multiple databases (Flury [5]). CPCA is a powerful technique for achieving a unified understanding of data structures by finding common principal components among different datasets.

However, traditional CPCA also faces problems when the sample size is small, leading to increased data variability and decreased stability of eigenvectors (Takane & Hunter [6]). On the other hand, methods such as Sparse PCA (Johnstone & Lu [7]) and the Noise Reduction (NR) Method (Yata & Aoshima [3]) have been developed for HDLSS data in single datasets. These methods aim to effectively reduce dimensionality and provide mathematical guarantees for the analysis. In this study, we propose extending CPCA using the NR method proposed by Yata & Aoshima [3] to create NR-CPCA, which enables mathematically guaranteed analysis in dimension reduction for HDLSS data. This research sets the following research question (RQ): Can NR-CPCA mitigate the impact of noise in general principal component analysis? Based on the above research question, we will verify the effectiveness of NR-CPCA. The contribution of this study lies in proposing for the first time a dimension reduction method for HDLSS multi-view data.

## 2. Method

In this study, we evaluate the consistency between estimated eigenvalues and actual eigenvalues, based on Yata & Aoshima [3], to verify the effectiveness of the proposed method. Specifically, by comparing the set eigenvalues, we assess the extent to which the estimated eigenvalues from CPCA and NR-CPCA are influenced by noise. CPCA involves calculating the covariance matrix, then computing the weighted average before determining the eigenvalues. In contrast, NR-CPCA takes into account the characteristics of high dimensional data by computing the dual covariance matrix, followed by the weighted average, before calculating the eigenvalues. Moreover, the eigenvalues and eigenvectors in NR-CPCA are computed based on NR method (Yata & Aoshima [3]), rather than using the eigenvalues and eigenvectors calculated by conventional PCA. Here, the estimated eigenvalues $\tilde{\lambda}_i$ of $S_D = (X-\bar{X})^T(X-\bar{X})$ and their corresponding estimated eigenvectors $\tilde{h}_i$ are given as (1) and (2).

$$\tilde{\lambda}_i = \hat{\lambda}_i - \frac{\mathrm{tr}(S_D) - \sum_{s=1}^{i} \hat{\lambda}_s}{(n_1 + n_2 + \cdots + n_k - 1) - i} \tag{1}$$

Let the number of dimensions of $X_i \in \mathbb{R}^{d \times n_i} (i = 1, 2, \cdots, k)$ be $d$, and the number of samples be $n_i (d \gg n_i)$. $X$ is the matrix represented by $X = (X_1\, X_2\, \cdots\, X_k) \in \mathbb{R}^{d \times (n_1+n_2+\cdots+n_k)} (d \gg \sum n_i)$, and $\bar{X}$ is the mean value of each row of $X$. $\tilde{\lambda}_i$ are the estimated eigenvalues, $\hat{\lambda}_i$ are the eigenvalues of $S_D$, $\mathrm{tr}(S_D)$ is the trace of the dual covariance matrix $S_D$.

$$\tilde{h}_i = \frac{X - \bar{X}}{\sqrt{(n_1 + n_2 + \cdots + n_k - 1)\tilde{\lambda}_i}} \hat{u}_i \tag{2}$$

$\tilde{h}_i$ are the estimated eigenvectors, $\hat{u}_i$ are the eigenvectors of $S_D$. The estimated eigenvalues $\tilde{\lambda}_i$ and eigenvectors $\tilde{h}_i$ are calculated up to $i = \min(n_1 - 2, n_2 - 2, \cdots, n_k - 2, d)$. The algorithm is summarized in Figure 1.

Next, we will explain the simulation settings. First, we preset the eigenvalues following Yata & Aoshima [3]. For the first set of data, the preset eigenvalues are $\lambda_1^{(1)} = d^{\frac{4}{5}}, \lambda_2^{(1)} = d^{\frac{3}{5}}, \lambda_3^{(1)} = d^{\frac{2}{5}}, \lambda_4^{(1)} = \cdots = \lambda_d^{(1)} = 1$. For the second set of data, the preset eigenvalues are$\lambda_1^{(2)} = d^{\frac{3}{4}}, \lambda_2^{(2)} = d^{\frac{1}{2}}, \lambda_3^{(2)} = d^{\frac{1}{4}}, \lambda_4^{(2)} = \cdots = \lambda_d^{(2)} = 1$.

Subsequently, we will conduct numerical simulations to validate the accuracy of the proposed method. Independent pseudo-random normal distributions will be generated using covariance matrices based on predetermined eigenvalues. This approach allows us to create high dimensional data with small sample sizes. The dimensionality is set at 1000, 2000, and 3000, and the sample size varies from 10 to 120. For each set of conditions, 10 simulations will be performed, and the average of these 10 simulations will be adopted as the estimated eigenvalue. This simulation will be executed using both CPCA and NR-CPCA. Through this method, we aim to comprehensively evaluate the utility of common principal components in HDLSS data and examine the extent to which they are affected by noise.

## 3. Results and Discussion

In Figure 2, we compare the estimated eigenvalues calculated using NR-CPCA and CPCA across various dimensions and sample sizes. Specifically, it visualizes the ratio of the estimated eigenvalues to the true eigenvalues. The left column displays the first, second, third, and fourth principal components for dimensions $d = 1000$, $2000$, and $3000$, respectively. The horizontal axis of each graph represents the sample size, ranging from $n = 10$ to $n = 120$. The red lines show the results for NR-CPCA, while the black lines show the results for CPCA. The closer the ratio is to 1, the closer the estimated eigenvalues are to the true eigenvalues. Furthermore, by examining the variation in these values, we can assess the stability of the estimated eigenvalues as the sample size increases.

```
Algorithm 1 NR-CPCA
Require: $\{X_i\}_{i=1}^{k}$ {Set of $k$ data matrices, $X_i \in \mathbb{R}^{d \times n_i}$}
$X_i \leftarrow \frac{1}{\sqrt{n_i - 1}} X_i$
$X \leftarrow (X_1 X_2 \cdots X_k) \in \mathbb{R}^{d \times (n_1 + n_2 + \cdots + n_k)}$
$\bar{X} \leftarrow$ row means of $X$
$S_D \leftarrow (X - \bar{X})^T (X - \bar{X})$
Compute eigenvalues and eigenvectors of $S_D$:
$\hat{\lambda}_i \leftarrow$ eigenvalues of $S_D$
$\hat{u}_i \leftarrow$ eigenvectors of $S_D$
Determine the constants: $r \leftarrow \min(n_1 - 2, n_2 - 2, \cdots, n_k - 2, d)$
for $i \leftarrow 1$ to $r$ do
    $\tilde{\lambda}_i \leftarrow \hat{\lambda}_i - \frac{\mathrm{tr}(S_D) - \sum_{s=1}^{i} \hat{\lambda}_s}{(n_1 + n_2 + \cdots + n_k - 1) - i}$
    $\tilde{h}_i \leftarrow \frac{X - \bar{X}}{\sqrt{(n_1 + n_2 + \cdots + n_k - 1)\tilde{\lambda}_i}} \hat{u}_i$
end for
return $list(values = \tilde{\lambda}_i, vectors = \tilde{h}_i)$
```

Fig. 1: NR-CPCA Algorithm

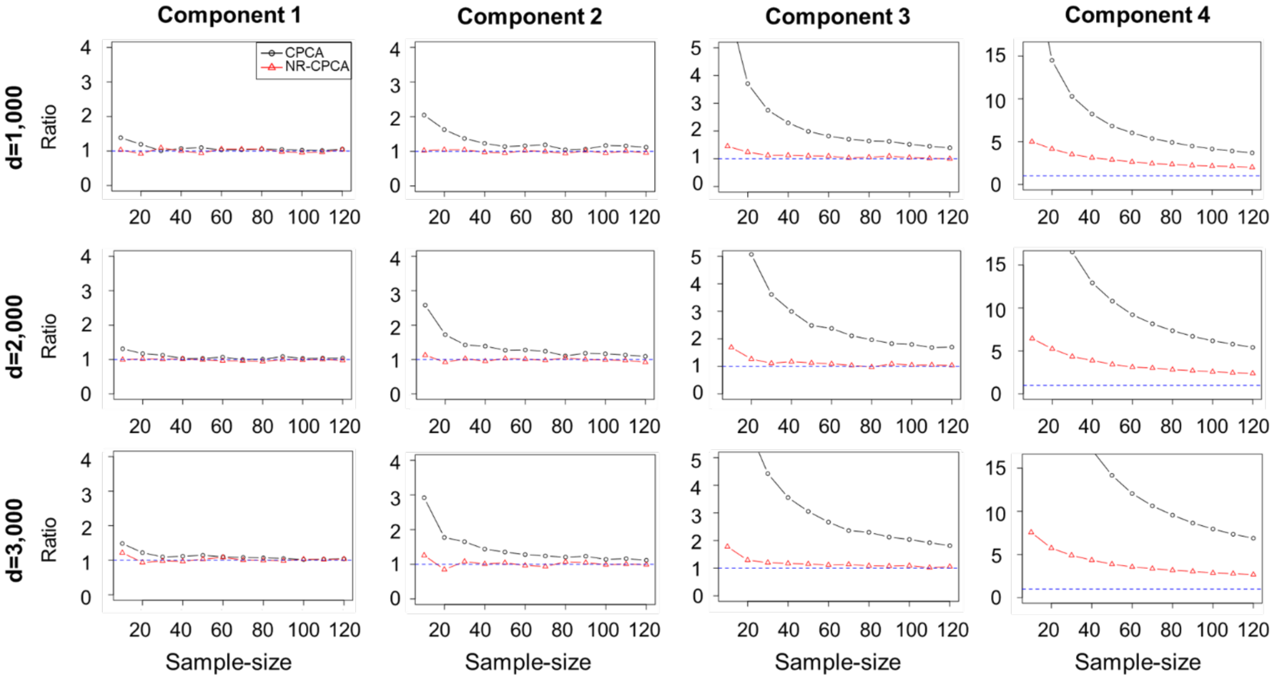

Fig. 2  Simulation Results

The comparison between NR-CPCA and CPCA reveals that NR-CPCA exhibits lower variability in the estimated eigenvalues compared to CPCA. This observation aligns with the findings of Yata and Aoshima [3], demonstrating that using the NR method relatively reduces the variability of eigenvalues. This trend is particularly notable from the first to the fourth principal components. Additionally, it has been observed that as the dimensionality of the data increases, the ratio between the set eigenvalues and the estimated eigenvalues increases for small sample sizes. Analysis reveals that the ratios computed by NR-CPCA are smaller than those computed by CPCA for all eigenvalues. This suggests that NR-CPCA provides more stable and reliable eigenvalue estimates. Moreover, as the sample size increases, NR-CPCA estimates the eigenvalues more accurately compared to CPCA. In conclusion, NR-CPCA not only reduces the variability in eigenvalue estimation but also demonstrates superior performance in providing more accurate eigenvalue estimates than CPCA.

The results of this study demonstrate the usefulness of NR-CPCA in estimating eigenvalues of HDLSS multi-view data. Consequently, applying NR-CPCA enables eigenvalue estimation with mathematical guarantees for noise reduction compared to traditional methods. In this study, we selected $w_i = 1$ ($i \in \{1,2,\cdots,k\}$) as the weight for the weighted average. Future research should explore optimizing these weights to potentially yield better estimates. Additionally, it will be crucial to evaluate and optimize parameters related to the degree of noise reduction and weighting methods. Furthermore, exploring how the results change by preparing eigenvalues in various settings will offer deeper insights into the method's reliability and applicability. Another important task for future research is to verify whether the analysis using real data can produce results equivalent to those obtained through simulations.

## 4. Conclusion

In this study, we propose the Noise Reduction Common Principal Component Analysis (NR-CPCA) method and verify its effectiveness. By integrating the noise reduction (NR) technique with Common Principal Component Analysis (CPCA), our proposed approach demonstrates greater resilience to noise in high-dimension, low-sample-size (HDLSS) data. Our research findings reveal that, irrespective of the sample size, NR-CPCA facilitates more stable and less variable eigenvalue estimation compared to traditional CPCA. This enhanced stability was consistently observed across various dimensions ($1000$, $2000$, $3000$) and a range of sample sizes (ranging from $10$ to $120$).

However, validation based solely on simulations cannot fully capture the complexity of real-world data. Future research needs to apply NR-CPCA to various real-world datasets to further demonstrate its utility. Additionally, to further improve NR-CPCA, it is essential to evaluate its performance across different data distributions and noise levels. The impact of parameters such as the degree of noise reduction and the weighting method on NR-CPCA performance also needs to be investigated. Furthermore, it is necessary to explore the generalizability of NR-CPCA. At the same time, optimizing the computational cost and efficiency of NR-CPCA for handling high-dimensional data is crucial. This includes considering algorithm improvements and the introduction of parallel processing techniques to reduce computation time for high-dimensional datasets.

In conclusion, the proposed NR-CPCA offers researchers dealing with HDLSS multi-view data a more precise and reliable tool for dimensionality reduction. We anticipate that this method will propel research and development in fields that handle high-dimensional data, such as data science, biology, and neuroscience, thereby making a significant contribution to the advancement of these domains.

## Acknowledgements

This work was supported by JST, PRESTO Grant Number JPMJPR21S4, Japan.

## References.

[1] J. Rahnenführer, R. D. Bin, A. Benner, F. Ambrogi, L. Lusa, A. Boulesteix, E. Migliavacca, H. Binder, S. Michiels, W. Sauerbrei, L. McShane, “Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges,” *BMC Medicine*, vol. 21, no. 182, 2023.

[2] J. Liu and M. Vinck, “Improved visualization of high-dimensional data using the distance-of-distance transformation,” *PLOS Computational Biology*, vol. 18, no. 12, pp. 1-19, 2022.

[3] K. Yata and M. Aoshima, “Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations,” *J. Multivariate Analysis*, vol. 105, no. 1, pp. 193-215, 2012.

[4] K. Yata and M. Aoshima, “Principal component analysis based clustering for high-dimension, low-sample-size data,” arXiv preprint arXiv:1503.04525, 2015.

[5] B. N. Flury, “Common principal components in k groups,” *J. the American Statistical Association*, vol. 79, no. 388, pp. 892-898, 1984.

[6] Y. Takane and M. A. Hunter, "A new family of constrained principal component analysis (CPCA)," *Linear Algebra and Its Applications*, vol. 434, Issue 12, pp. 2539-2555, 2011.

[7] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *J. the American Statistical Association*, vol. 104, Issue 486, pp. 682-693, 2009.