algorithms

MDPI

Communication

Means and Issues for Adjusting Principal Component Analysis Results

Tomokazu Konishi

Graduate School of Bioresource Sciences, Akita Prefectural University, Akita 010-0195, Japan; konishi@akita-pu.ac.jp

Abstract: Background: Principal component analysis (PCA) is a method that identifies common directions within multivariate data and presents the data in as few dimensions as possible. One of the advantages of PCA is its objectivity, as the same results can be obtained regardless of who performs the analysis. However, PCA is not a robust method and is sensitive to noise. Consequently, the directions identified by PCA may deviate slightly. If we can teach PCA to account for this deviation and correct it, the results should become more comprehensible. Methods: The top two PCA results were rotated using a rotation unitary matrix. Results: These contributions were determined and compared with the original. At smaller rotations, the change in contribution was also small and the effect on independence was not severe. The rotation made the data considerably more comprehensible. Conclusions: The methods for achieving this and an issue with this are presented. However, care should be taken not to detract from the superior objectivity of PCA.

Keywords: rotation of PCA; adjustment; unitary matrix

check for updates

Academic Editors: Dunhui Xiao and Shuai Li

Received: 7 January 2025
Revised: 14 February 2025
Accepted: 17 February 2025
Published: 25 February 2025

Citation: Konishi, T. Means and Issues for Adjusting Principal Component Analysis Results. *Algorithms* **2025**, *18*, 129. https://doi.org/10.3390/a18030129

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Much of the data we handle consist of numerous measurement items, requiring multivariate analysis. Among the various methods available, principal component analysis (PCA) is particularly well-suited for scientific studies due to its limited methodological flexibility and high reproducibility [1–9]. The simplicity of PCA ensures that its results are consistent, regardless of the individual performing the analysis. This property is highly valuable in scientific research as it guarantees objectivity.

In PCA, we use singular value decomposition (SVD) to separate the matrix into unitary matrices (U and V) and a diagonal matrix D,

$$M = UDV^*,$$

where V^* denotes the conjugate transpose of V, and M represents the data to be analyzed, often centred or even scaled. The diagonal elements of D are sorted in descending order. Using this decomposition, the principal components (PCs) can be derived as follows:

$$Y = MV = UD \text{ and } Z = M^*U = VD$$

where the column vector of Y is PC for the samples and that of Z represents the PCs for the measurement items (e.g., PC1, PC2, etc.) [8]. Since U and V are unitary matrices, they satisfy the properties $U^*U = I$ and $UU^* = I$. This ensures that the row and column vectors are orthogonal and have a Euclidean norm of 1. Specifically, the inner product of a

vector with itself equals 1, while the inner product with other vectors equals 0, confirming orthogonality. As a result, the column vectors of Y and Z are also independent. These vectors can be interpreted as rotations of the original matrix M, with the unitary matrices defining the axes and the diagonal elements of D determining their lengths.

Since the diagonal elements of D are sorted, PC1, PC2, and subsequent components account for the decreasing proportions of the total data variance. PCA thus identifies common directions within M and summarizes them in order of their significance. This ability to effectively reduce the dimensionality of multivariate data is a key advantage of PCA.

The advantage of PCA as a scientific method lies in its objectivity. Due to the limited degree of choice involved in its application, the results remain consistent regardless of who conducts the analysis. Given that science is inherently a collaborative endeavour, it is essential that participants are interchangeable at all times and that the implications of the data are clear and accessible to all. While there are other methods of multivariate analysis, such as cluster analysis and multiple regression analysis, none appear to be as well suited to scientific research as PCA in terms of objectivity. For instance, cluster analysis is inherently less suitable for scientific purposes due to the abundance of options available, which can lead to varying results depending on the choices made. This is because cluster analysis typically presents only one possible interpretation of the data. Additionally, PCA can be viewed as an upwardly compatible method, as it is capable of revealing multiple regression analysis outcomes. Furthermore, PCA has the ability to condense a vast amount of information into a well-defined and interpretable form. To illustrate this, consider a study examining changes in the mammary glands of rodents during pregnancy and childbirth. This study involved dozens of sample measurements and a microarray analysis of tens of thousands of genes [8]. Despite the sheer volume of data, PCA facilitated a clear and concise summary, elucidating the directional trends associated with pregnancy and childbirth, as well as the relationships between individual genes.

However, one of the drawbacks of PCA is its lack of robustness. Even a single outlier can cause errors in SVD. Naturally, noise affects the results. It is common for the identified directions to be slightly off. For instance, group positions or individual Y values may deviate slightly from the axes due to rotation. Since experimental data inherently contain noise, it is expected that noise-sensitive PCA will exhibit such errors. In such situations, there may be a temptation to adjust PCA results or train PCA to correct rotations.

Accordingly, PCA now offers various options for data manipulation. One approach, for instance, involves reducing the dataset to be subjected to SVD by selecting it in a specific way [10]. Another involves altering the direction of rotation derived from SVD [11,12]. Both orthogonal and oblique rotations are available, each offering several methods. Naturally, these variations result in different solutions, thereby compromising the objectivity of PCA. This subjectivity arises because the outcomes depend on the analyst's choices, which poses a significant challenge when analyzing scientific data. Oblique rotations, in particular, introduce a high degree of freedom, potentially leading to arbitrary results.

This study introduces a straightforward method to rotate data in relation to the two principal components while maintaining their orthogonality. The impact of such rotations on the original PCA results is evaluated in terms of the contributions of the principal components. The findings reveal that mild rotations do not significantly alter the results.

2. Materials and Methods

The test data used in this study were obtained from the practical training of students on a soil study. All calculations were performed using R (4.4.0) [13], and both the data and R code are provided as Supplementary Materials. Due to significant variability in the

calcium data, the data were z-normalized for each item before applying SVD. Consequently, the centring point corresponded to the mean value of the data; these centring and scaling steps are among the few options available in PCA. The calculation method used for PCA is described in the Introduction. Additionally, since the number of measured items and the sample size differed considerably, I scaled the PCs by the square root of the sample size to facilitate biplots on a common axis [8], specifically $Y_s = Y/\sqrt{n_i}$, where n_i is number of items, and so on.

The angle of rotation, however, needed to be determined on a basis that was intuitive and easily understandable for all. In this case, an angle of 14 degrees was selected because it aligned the potassium (K) component of PC1 horizontally while simultaneously positioning the calcium (Ca) component of PC2 closer to the vertical axis. Alternatively, an averaging method could be employed to achieve an overall alignment of the components.

3. Results

An example of a slight deviation of Z from the axes can be seen in Figure 1 (blue), which represents the PCA of data obtained from measuring substances in soil. Samples P1 to P4 were collected from points found downward, at 50 cm intervals from the surface of the padding field. Overall, soluble Na, K, and Mg appeared to have infiltrated the soil (P3, P4, G3, and G4), while less-soluble Ca remained near the surface (P1, P2, G1, and G2). In unfertilized forest areas, these materials were notably less (F1–F4). The measured items effectively differentiated these samples, demonstrating the impact of human activity on the land—increased calcium levels may alter surface soil, and cations that have penetrated underground could contribute to salinity. However, upon closer examination, the figure appears to be rotated counterclockwise by 14 degrees.

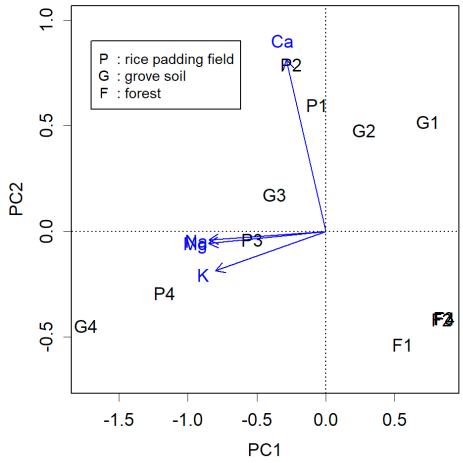


Figure 1. Results of material quantity measurements from soil material. Soil samples were obtained from a rice padding field, the grove soil of dry farmland, and a forest. The black and blue biplots show the PC of the samples, Y, and the PC of the items, Z, respectively. Soluble cations and water-insoluble Ca guided PC1 and PC2, respectively.

For example, when creating a two-dimensional plot using PC1 and PC2 for a matrix of four columns, these two column vectors can be rotated by taking the inner product of them with a rotation matrix, as follows:

$$R(\theta) = \begin{pmatrix} \cos(\theta) & \sin(\theta) & 0 & 0 \\ -\sin(\theta) & \cos(\theta) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (1)$$

This is almost the identity matrix, but an alteration has been made in the upper left. This rotation matrix is unitary, so $R(\theta)R(\theta)^* = I$, where I is the identity matrix. Consequently, if we rotate two columns of U, the adjusted U is $U_a = UR(\theta)$. The resulting rotated matrix U_a is also unitary, as $UR(\theta)\ \{UR(\theta)\}^* = UR(\theta)\ R(\theta)^*U^* = UIU^* = I$. Since U and V are related as mirror images, if we rotate one column vector, we need to rotate the corresponding column vector by the same amount. From $M = UDV^*$, we have

$$M = UR(\theta)R(\theta)^*D\left(V\,R(\theta)R(\theta)^*\right)^* = U_aR(\theta)^*DR(\theta)V_a^*. \quad (2)$$

The matrix that is sandwiched by the two unitary matrixes,

$$R(\theta)^*DR(\theta) = D_a, \quad (3)$$

is not necessarily a diagonal matrix, according to the twice rotations of D. Using D_a, $Y_a = YR(\theta) = MV_a = U_aD_a$ and $Z_a = ZR(\theta) = M^*U_a = V_aD_a$. Hence, the resulting column vectors of Y_a and Z_a are not necessarily independent. The degree of dependence depends on the rotation angle.

The actual result after rotating by 14 degrees is shown in Figure 2. Each item is now displayed closer to the axes (blue). Notably, the high calcium content in P1 and P2, as well as the abundance of soluble cations in G4 and P4, is more clearly visible. The question now is how much independence has been compromised. This becomes evident when examining the contribution of PC1.

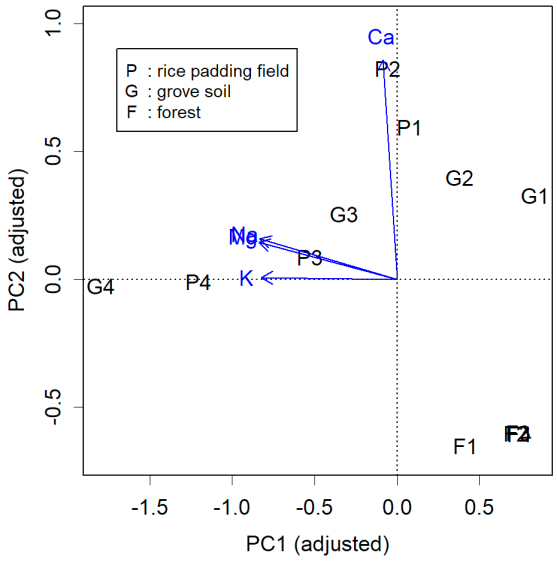


Figure 2. Adjusted results from Figure 1 by rotating the plot 14° clockwise. Each Z (blue) is more along the axis, with the characteristic G4 and P4 in PC1 of the Y (black), and the characteristic P1 and P2 in PC2 also appearing more along their respective axes.

Contributions are often derived from the diagonal elements of D. As D is a diagonal matrix, the contributions can be obtained from the component $\text{diag}(D)/\Sigma(D)$. However, since D_a is not diagonal, we need to calculate the Euclidean distance using the squared sum of each column vector. The contribution is then given by the distance/total distance.

From the perspective of PCA, the goal is to collect contributions primarily from the top PCs. In the case of rotating by 14 degrees, the loss appeared to be relatively small (Figure 3, dashed line). In this context, the most significant change occurred when rotating by 45 degrees, $\pi/4$ (dotted line). At this point, PC1 completely lost its advantage over PC2, and both had equal contributions. In addition, of course, if rotated by 90 degrees,

PC1 and PC2 would have effectively swapped places, and their contributions would have naturally switched as well. As the data's dispersion remained preserved, the sum of all total distances was maintained regardless of the degree of rotations.

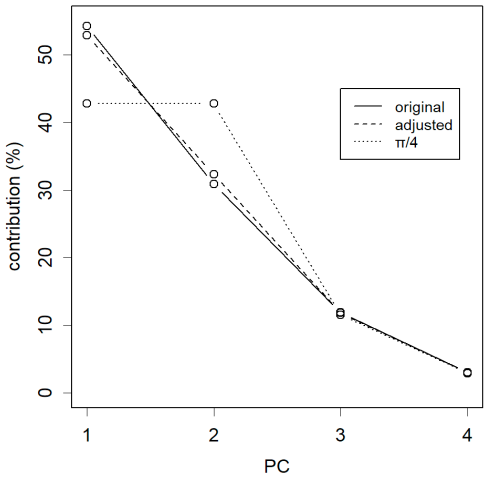


Figure 3. Contribution of each method. The original PCA is the most ideal diagram, with the greatest concentration on PC1; adjusting by 14° rotation weakens this concentration somewhat; finally, rotating by 45°, $4/\pi$, the advantage of PC1 disappears and becomes the same contribution as that of PC2. Rotation by 90 degrees, not shown here, replaces PC1 and PC2.

4. Discussion

In this analysis, PCA revealed two principal directions: PC1, which represented water-soluble ions that permeated the ground in a specific direction, and PC2, which represented a water-insoluble ion that remained on the surface. The samples were differentiated based on their respective depths within the soil, while forested areas that had not been treated were distinctly separated from the treated regions. The application of artificial rotation enhanced the clarity of this representation.

The results of the rotated PCA, albeit by a small angle, were more understandable. This somewhat compromises the priority of higher levels of PCs, but if the angle is not too large, the damage will not be too great. Understanding PCA plots can often be challenging, as PCA merely provides a summary, without revealing the specifics. Interpretation of the plot is left to the analyst. In this regard, enhanced readability could be highly appreciated. Notably, PCA is sensitive to noise. If this effect could be easily mitigated, this technique would be worth considering. To address this, it would be advisable to employ outlier detection methods provided by exploratory data analysis (EDA) [14,15]. This approach could potentially eliminate the need for artificial rotation, which would be a preferable outcome. In any case, introducing rotation complicates the process and reduces its objectivity. Unless there is a clear and compelling justification for its use, it would be more prudent to avoid it altogether.

It should be noted that this constitutes an active intervention by the analyst, potentially compromising PCA's objectivity. The beauty of PCA lies in its impartiality. Diluting this aspect would be regrettable and might even lead to data manipulation. If such adjustments were made, it would be essential to keep the original, unadjusted plot available. This original information is also necessary for others to check whether a rotation was necessary. Alternatively, the rotated plot could serve as supplementary material for explaining the results. Although this analysis focuses primarily on the two main principal components (PCs), there may be instances where it is necessary to rotate subordinate PCs as well. However, this would further diminish the objectivity of the results. In such cases, it is

recommended to present visual representations, such as those illustrated in Figure 3, to provide a clear and transparent depiction of the data.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/a18030129/s1: Rscript.txt: R code; and test.txt: the sample data.

Funding: This research received no external funding.

Data Availability Statement: The sample data and R codes are supplied in the Supplementary Data.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PC	Principal component
PCA	Principal component analysis
SVD	Singular value decomposition

References

1. Abdi, H.; Williams, L.J. Principal component analysis. *WIREs Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
2. Aluja, T.; Morineau, A.; Sanchez, G. Principal Component Analysis for Data Science. Available online: https://github.com/pca4ds/pca4ds.github.io (accessed on 22 February 2025).
3. Bartholomew, D.J. Principal Components Analysis. In *International Encyclopedia of Education*, 3rd ed.; Peterson, P., Baker, E., McGaw, B., Eds.; Elsevier: Oxford, UK, 2010; pp. 374–377.
4. David, C.C.; Jacobs, D.J. Principal component analysis: A method for determining the essential dynamics of proteins. *Methods Mol. Biol.* **2014**, *1084*, 193–226. [CrossRef] [PubMed]
5. Jackson, E. *A Use's Guide to Principal Components*; Wiley: Hoboken, NJ, USA, 1991.
6. Jolliffe, I.T. *Principal Component Analysis*; Springer: Berlin/Heidelberg, Germany, 2002.
7. Jolliffe, I.T.; Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150202. [CrossRef] [PubMed]
8. Konishi, T. Principal component analysis for designed experiments. *BMC Bioinform.* **2015**, *16*, S7. [CrossRef] [PubMed]
9. Ringnér, M. What is principal component analysis? *Nat. Biotechnol.* **2008**, *26*, 303–304. [CrossRef] [PubMed]
10. Jolliffe, I.T.; Trendafilov, N.T.; Uddin, M. A modified principal component technique based on the LASSO. *J. Comput. Graph. Stat.* **2003**, *12*, 531–547. [CrossRef]
11. Brown, J.D. Choosing the Right Type of Rotation in PCA and EFA. *JALT Test. Eval. SIG Newsl.* **2009**, *13*, 20–25.
12. Jolliffe, I.T. Rotation of principal components: Choice of normalization constraints. *J. Appl. Stat.* **1995**, *22*, 29–35. [CrossRef]
13. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024.
14. National Institute of Standards and Technology. NIST/SEMATECH e-Handbook of Statistical Methods. Available online: https://www.itl.nist.gov/div898/handbook/index.htm (accessed on 14 February 2025).
15. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley Pub. Co.: Reading, MA, USA; London, UK, 1977.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.