Article

Not peer-reviewed version

Means and Issues for Adjusting Principal Component Analysis Results

Tomokazu Konishi *

Posted Date: 27 May 2024

doi: 10.20944/preprints202405.1445.v2

Keywords: Principal Component Analysis; rotation matrix, adjusting, unitary matrix

 doi:10.20944/preprints202405.1445.v2

Article

Means and Issues for Adjusting Principal Component Analysis Results

Tomokazu Konishi

Graduate School of Bioresource Sciences, Akita Prefectural University, Akita, 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 doing this and an issue with this are presented. However, care should be taken not to detract from the superior objectivity of PCAs.

Keywords: rotation of PCA; adjustment; unitary matrix

Introduction

Much of the data we handle consists of numerous measurement items, necessitating multivariate analysis. While several methods are known, Principal Component Analysis (PCA) is particularly suitable for scientific analysis due to its limited options and high reproducibility [1–9]. 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 is the data to be analysed, often centred or even scaled. The diagonal elements of D are sorted. The principal components (PCs) can then be obtained from these as follows: $Y = MV = UD$, $Z = M^*U = VD$, where the column vector of Y are PC for samples and those of Z are for items (PC1, PC2, etc.) (8). Since U and V are unitary matrices, with properties such as ($U^*U = I$) and ($U\,U^* = I$), these row and column vectors are independent, and their Euclidean distances are all 1: those column or row vectors have length 1 because taking the inner product with itself, sum of squares of the elements, gives 1, and taking the inner product with others gives zero, showing that each of them is orthogonal to the other. The resulting Y or Z column vectors are also independent. They can be interpreted as rotations of the original matrix M or as axes given by the unitary matrices with lengths determined by D. Since D is sorted, PC1, PC2, in that order, cover a larger percentage of the data scatter. Thus, PCA finds common directions in M and summarizes them in the order of their length of the direction. The advantage of PCA is that it can thus compress the many dimensions of multivariate data as much as possible.

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's 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 contains noise, it is expected that noise-sensitive PCA will exhibit such errors. The Robust PCA, methods that eliminate the effects of outliers have been devised [10,11], but they do not necessarily correct these rotations. In such situations, there may be a temptation to adjust PCA results or train PCA to correct rotations. In this context, I discuss potential correction methods and raise associated issues.

Materials and Methods

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2024 doi:10.20944/preprints202405.1445.v2

The test data used in this study was obtained from practical training of students on soil study. All calculations were performed using R [10], and both the data and R code are provided as supplementary material. Due to significant variability in calcium data, the data was z-normalized for each item before applying SVD. Consequently, the centring point corresponds to the mean value of the data; these centring and scaling steps are among the few options available in PCA. The calculation method for PCA is as described in the introduction. Additionally, since the number of measured items and the sample size differ 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. Robust PCA was calculated using the R package *rrcov*. All R codes are in Supplement.

Results

An example of a slight deviation of Z from the axes can be seen in Figure 1 (blue), which represents PCA of data obtained from measuring substances in soil. Samples P1 to P4 were collected from points downwards at 50 cm intervals from the surface of the padding field. Overall, soluble Na, K, and Mg appear to have infiltrated the soil, while less soluble Ca remains near the surface. In unfertilized forest areas, these materials are notably less. The measured items effectively differentiate these samples; this data demonstrates 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 counter clockwise by 14 degrees.

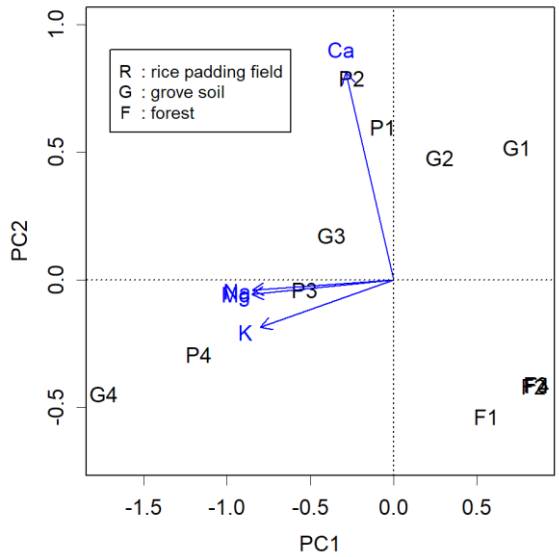

Figure 1. Results of material quantity measurements from soil material. Soil samples were obtained from rice padding field, grove soil of dry farmland and forest. The black and blue biplots show the PC of the samples, Y, and the PC of the items, Z, respectively. Soluble cations and water-insoluble Ca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

Robust PCA [11, 12] was applied to the data (Figure 2). This method aims to recover a low-rank matrix while minimizing the impact of outliers. From the figure, it appears that the rotation has actually increased; in reality, it has expanded by approximately 25 degrees.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2024 doi:10.20944/preprints202405.1445.v2

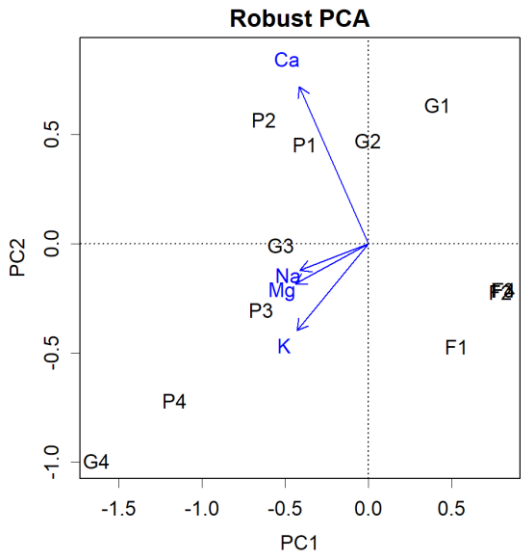

Figure 2. Robust PCA is applied in Hubert's procedure [11]. The outlier disturbance should have been reduced, but it is clear that the rotation became greater.

It was decided to try to solve this problem in a more proactive way. That is, artificially reverse the rotation to compensate for the results. 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:

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

This is almost the identity matrix, but alteration was made at 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 $UR(\theta)\{UR(\theta)\}^* = UR(\theta)\,R(\theta)^*U^*$. 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

$$M = UR(\theta)R(\theta)^*D(VR(\theta)R(\theta)^*)^* = U_aR(\theta)^*DR(\theta)V_a^*.$$

The matrix that is sandwiched by the two unitary matrixes,

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

is not necessarily a diagonal matrix, according to the twice rotations of D. By 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 and D.

The actual result after rotating by 14 degrees is shown in Figure 3. Each item is now displayed closer to the axes. 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.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2024 doi:10.20944/preprints202405.1445.v2

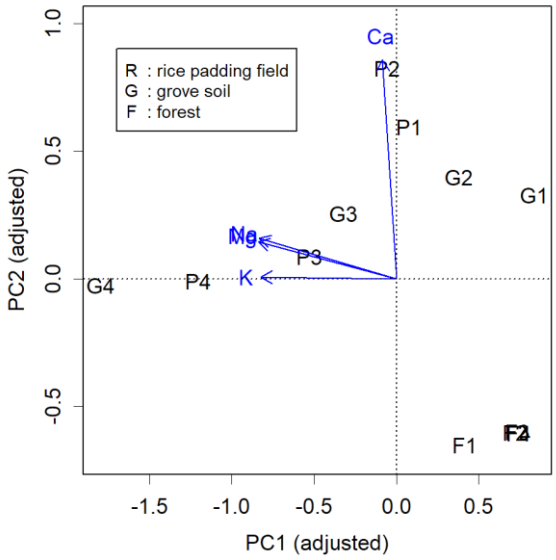

Figure 3. Adjusted results from Figure 1 by rotating the plot by 14° clockwise. Each Z (blue) is more along the axis, with the characteristic G4 and P4 in PC1 of the Y (black), and the characteristic P1 and P2 in PC2 also appearing more along their respective axes.

Often, contributions are derived from the diagonal elements of D. As D is a diagonal matrix, the contributions can be obtained from the components $\mathrm{diag}(D)/\Sigma(D)$. 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 appears to be relatively small (Figure 4, dashed line). The Robust PCAs had the same degree of change as the rotation (blue dotted line). Incidentally, the most significant change occurs when rotating by 45 degrees, $\pi/4$ (dotted line). At this point, PC1 completely loses its advantage over PC2, and both have equal contributions. And of course, if rotated by 90 degrees, PC1 and PC2 would effectively swap places, and their contributions would naturally exchange as well. As the data's dispersion remains preserved, the sum of all total distance is maintained regardless of the degree of rotations; however, this is not always the case for the Robust PCA.

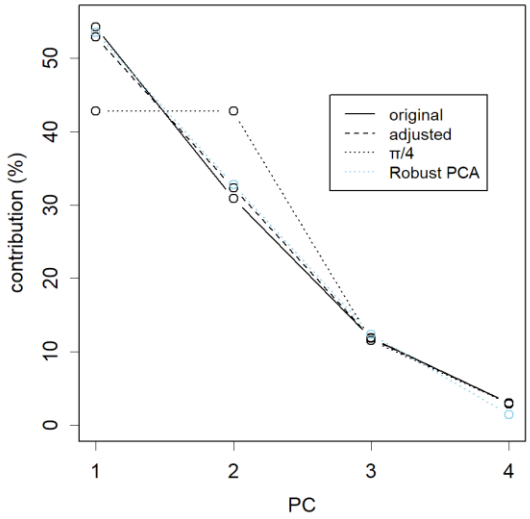

Figure 4. Contribution of each method. The original PCA is the most ideal diagram, with the greatest concentration on PC1; adjusting by 14° rotation weakens this concentration somewhat; rotating by

 doi:10.20944/preprints202405.1445.v2

45°, 4/π, 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.

Discussion

The results of the rotated PCA, albeit by a small angle, were more comprehensible (Figure 3). 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 (Figure 4). 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 can be easily mitigated, it's worth considering. 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.

However, the rotation 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 are made, it's 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.

Robust PCA may be useful when there is some clear outlier. But this sophisticated calculation also undermines the simplicity of PCA. And above all, it is not always effective, as shown here (Figure 2). The method of rotation shown here is a clear setback to objectivity, but it is obvious to everyone whether it is necessary or not, and how it worked when it was done, and in this respect objectivity remains.

References

1. Abdi H, Williams LJ. *Principal component analysis.* WIREs Computational Statistics. 2010;2(4):433-459. doi: https://doi.org/10.1002/wics.101.
2. Aluja T, Morineau A, Sanchez G. *Principal Component Analysis for Data Science*. 2018. Available from: https://pca4ds.github.io
3. Bartholomew DJ. *Principal Components Analysis.* In: Peterson P, Baker E, McGaw B, editors. International Encyclopedia of Education (Third Edition). Oxford: Elsevier; 2010. p. 374-377.
4. David CC, Jacobs DJ. Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol Biol.* 2014;1084:193-226. doi: 10.1007/978-1-62703-658-0_11. PubMed PMID: 24061923; PubMed Central PMCID: PMCPMC4676806. eng.
5. Jackson E. *A Use's Guide to Principal Components*. Wiley 1991. (Wiley Series in Probability and Statistics).
6. Jolliffe IT. *Principal Component Analysis*. Springer; 2002. (Springer Series in Statistics (SSS)).
7. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2016;374(2065):20150202. doi: doi:10.1098/rsta.2015.0202.
8. Konishi T. Principal component analysis for designed experiments. *BMC Bioinformatics.* 2015 2015/12/09;16(18):S7. doi: 10.1186/1471-2105-16-S18-S7.
9. Ringnér M. What is principal component analysis? *Nat Biotechnol.* 2008 Mar;26(3):303-4. doi: 10.1038/nbt0308-303. PubMed PMID: 18327243; eng.
10. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2024.
11. Hubert M, Rousseeuw P, Verdonck T. Robust PCA for skewed data and its outlier map. *Computational Statistics & Data Analysis.* 2009 2009/04/15/;53(6):2264-2274. doi: https://doi.org/10.1016/j.csda.2008.05.027.
12. Todorov V, Filzmoser P. An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software.* 2009 10/14;32(3):1 - 47. doi: 10.18637/jss.v032.i03.