Journal of Geochemical Exploration 282 (2026) 107962

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/gexplo

Beyond principal component analysis: Nonparametric and nonlinear approaches for robust analysis of Gafsa basin groundwater

Ko Kumada [a,*], Yoshiyasu Takefuji [a]

[a] Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan

ARTICLE INFO

Keywords
Contamination
Radioactivity
Nitrates
PCA
Nonlinearity

ABSTRACT

This paper critically examines methodological limitations in hydrochemical contamination studies, focusing on Principal Component Analysis (PCA) applications to the Gafsa basin in Tunisia. While PCA, as employed by Boschetti et al. (2025), effectively identified primary contamination sources from phosphate mining and agriculture, its inherent linearity assumptions fundamentally constrain its ability to represent complex environmental processes. We demonstrate how complementary methodologies—Feature Agglomeration, Independent Component Analysis, and High Variance Gene Selection—create a more comprehensive analytical framework capable of capturing nonlinear relationships, hierarchical structures, and statistically independent variation sources that PCA might overlook. This integrated approach enhances result reliability through methodological triangulation, providing environmental managers with more accurate contamination profiles that reflect the true complexity of groundwater systems.

This paper critiques common AI misapplications of principal component analysis (PCA), emphasizing that its inherently linear structure can yield distorted insights when imposed on nonlinear data manifolds. It argues that PCA's conclusions are reliable only when its underlying assumptions—such as linearity, global Euclidean geometry, and variance-as-signal—are reasonably satisfied. To promote robust inference, this paper discourages exclusive reliance on linear PCA and instead advocates a multifaceted, stability-focused workflow that combines PCA with complementary unsupervised methods, including feature agglomeration (FA) to capture hierarchical structure and highly variable gene selection (HVGS) to enrich signal-to-noise, thereby improving interpretability and safeguarding against spurious patterns.

Boschetti et al. (2025) studied groundwater contamination by radioactivity and nitrates in southern Tunisia's Gafsa basin, impacted by phosphate mining and agriculture. They analyzed 33 samples using unsupervised machine learning models such as PCA to identify contamination sources. Elevated radioactivity levels were linked primarily to phosphate mining and secondarily to the North Western Sahara Aquifer System. Nitrate pollution mainly derived from agricultural runoff, with additional mining contributions. PCA classified samples into groups reflecting dominant influences: phosphate mining, combined agriculture and mining, fossil geothermal waters, and low agricultural zones. Anthropogenically affected samples showed higher radium and nitrate levels. These results highlight the multi-source nature of Gafsa basin contamination and underscore the need for targeted management.

Nevertheless, the paper highlights significant methodological concerns regarding PCA applications to hydrochemical datasets. Due to its inherently linear nature, PCA can lead to erroneous interpretations and flawed conclusions when analyzing environmental systems. All PCA variants fundamentally operate on assumptions of linearity and multivariate normality—premises that often fail to capture the nonlinear phenomena characterizing environmental processes (Dyer and Kording, 2023; Cristian et al., 2024; Yao and Ochoa, 2023; Elhaik, 2022; Mohseni and Elhaik, 2024; Lenz et al., 2016; Dey and Lee, 2019; Mehta et al., 2019; Prasad and Bruce, 2008; Alanis-Lobato et al., 2015; Black et al., 2024; Caprihan et al., 2008; Nyamundanda et al., 2010).

PCA fundamentally struggles with nonlinear, nonparametric data because it projects high-dimensional information onto linear subspaces that maximize variance. When applied to real-world environmental data residing on curved manifolds, this linear transformation inevitably distorts underlying relationships. Consequently, critical dynamic processes may be overlooked or misrepresented, particularly when analyzing variables exhibiting skewness, multimodality, or heavy-tailed distributions.

The technique's sensitivity to outliers presents another significant

* Corresponding author.
E-mail addresses: g2550003@stu.musashino-u.ac.jp (K. Kumada), takefuji@keio.jp (Y. Takefuji).

https://doi.org/10.1016/j.gexplo.2025.107962
Received 17 July 2025; Received in revised form 20 November 2025; Accepted 15 December 2025
Available online 18 December 2025

limitation for environmental datasets, which frequently contain extreme values due to episodic events or sampling anomalies. Furthermore, PCA's variance-maximization objective can overemphasize high-variance features while potentially neglecting subtle yet ecologically significant patterns with lower variance. The orthogonality constraint on principal components further restricts PCA's ability to represent interconnected nonlinear environmental processes that don't align with perpendicular axes.

These limitations are particularly problematic in hydrochemical studies, where complex biogeochemical interactions, threshold effects, and seasonal dynamics create inherently nonlinear relationships. For robust environmental data analysis, researchers should consider complementary or alternative approaches more adept at capturing nonlinear relationships.

The complexity of environmental data demands analytical sophistication beyond singular methodologies. While PCA provides an accessible entry point for dimensionality reduction in groundwater contamination studies, it inherently assumes linear relationships that may oversimplify hydrochemical interactions. Real-world environmental systems rarely conform to such constraints, exhibiting complex, nonlinear behaviors across multiple scales. A multifaceted analytical framework using powerful unsupervised machine learning models incorporating FA and HVGS—offers crucial advantages by addressing these limitations. FA captures hierarchical relationships between contaminants and HVGS highlights particularly dynamic parameters that signal environmental change. Together, these complementary approaches create a methodological triangulation that enhances result reliability through cross-validation. This integrated strategy enables researchers to differentiate between spurious correlations and meaningful patterns, particularly in datasets plagued by outliers, missing values, or temporal inconsistencies. By embracing methodological diversity, investigators can construct a more complete contamination profile that captures both obvious pollution signatures and subtle hydrochemical interactions that might otherwise remain hidden. The resulting insights offer environmental managers and policymakers a scientifically rigorous foundation for remediation efforts, regulatory decisions, and long-term monitoring strategies that reflect the true complexity of groundwater systems.

CRediT authorship contribution statement

Ko Kumada: Writing – review & editing, Writing – original draft, Validation, Investigation, Conceptualization. **Yoshiyasu Takefuji:** Writing – review & editing, Supervision, Conceptualization.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Ethics approval

Not applicable.

Code availability

Not applicable.

This study provides guidance for integrating analytical methods such as PCA, ICA, and feature aggregation, without providing code or data.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used OpenAI/o4-mini in order to refine English sentences. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to express their sincere gratitude to Professor Takefuji for his invaluable guidance and support in the preparation of this manuscript.

Data availability

Not applicable.

References

Alanis-Lobato, G., Cannistraci, C.V., Eriksson, A., Manica, A., Ravasi, T., 2015. Highlighting nonlinear patterns in population genetics datasets. Sci. Rep. 5, 8140. https://doi.org/10.1038/srep08140.

Black, D., Byrne, D., Walke, A., et al., 2024. Towards machine learning-based quantitative hyperspectral image guidance for brain tumor resection. Commun. Med. (Lond.) 4 (1), 131. https://doi.org/10.1038/s43856-024-00562-3.

Boschetti, T., Hamed, Y., Hadji, R., et al., 2025. Using principal component analysis to distinguish sources of radioactivity and nitrates contamination in Southern Tunisian groundwater samples. J. Geochem. Explor. 163, 107670. https://doi.org/10.1016/j.gexplo.2025.107670.

Caprihan, A., Pearlson, G.D., Calhoun, V.D., 2008. Application of principal component analysis to distinguish patients with schizophrenia from healthy controls based on fractional anisotropy measurements. Neuroimage 42 (2), 675–682. https://doi.org/10.1016/j.neuroimage.2008.04.255.

Cristian, P.M., Aarón, V.J., Armando, E.D., et al., 2024. Diffusion on PCA-UMAP manifold: the impact of data structure preservation to denoise high-dimensional single-cell RNA sequencing data. Biology (Basel). 13 (7), 512. https://doi.org/10.3390/biology13070512.

Dey, R., Lee, S., 2019. Asymptotic properties of principal component analysis and shrinkage-bias adjustment under the generalized spiked population model. J. Multivar. Anal. 173, 145–164. https://doi.org/10.1016/j.jmva.2019.02.007.

Dyer, E.L., Kording, K., 2023. Why the simplest explanation isn't always the best. Proc. Natl. Acad. Sci. U. S. A. 120 (52), e2319169120. https://doi.org/10.1073/pnas.2319169120.

Elhaik, E., 2022. Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Sci. Rep. 12 (1), 14683. https://doi.org/10.1038/s41598-022-14395-4.

Lenz, M., Müller, F.J., Zenke, M., Schuppert, A., 2016. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. Sci. Rep. 6, 25696. https://doi.org/10.1038/srep25696.

Mehta, P., Wang, C.H., Day, A.G.R., et al., 2019. A high-bias, low-variance introduction to machine learning for physicists. Phys. Rep. 810, 1–124. https://doi.org/10.1016/j.physrep.2019.03.001.

Mohseni, N., Elhaik, E., 2024. Biases of Principal Component Analysis (PCA) in physical anthropology studies require a reevaluation of evolutionary insights. eLife 13, RP94685. https://doi.org/10.7554/eLife.94685.2.

Nyamundanda, G., Brennan, L., Gormley, I.C., 2010. Probabilistic principal component analysis for metabolomic data. BMC Bioinformatics 11, 571. https://doi.org/10.1186/1471-2105-11-571.

Prasad, S., Bruce, L.M., 2008. Limitations of principal components analysis for hyperspectral target recognition. IEEE Geosci. Remote Sens. Lett. 5 (4), 625–629. https://doi.org/10.1109/LGRS.2008.2001282.

Yao, Y., Ochoa, A., 2023. Limitations of principal components in quantitative genetic association models for human studies. eLife 12, e79238. https://doi.org/10.7554/eLife.79238.