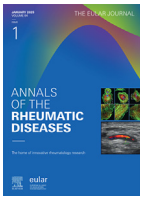
Ann Rheum Dis 000 (2025) e1–e2

Contents lists available at ScienceDirect

Annals of the Rheumatic Diseases

journal homepage: https://www.sciencedirect.com/journal/annals-of-the-rheumatic-diseases



Correspondence

Reevaluating analytical approaches in systemic sclerosis research: challenges of PCA and logistic regression

Rouvière et al [1] conducted a comprehensive study on the stratification of systemic sclerosis patients based on their autoantibody status, revealing distinct molecular signatures. Utilising principal component analysis (PCA) of transcriptomic data, the researchers identified subtle differences between the 2 patient groups, suggesting a continuous spectrum of molecular signatures that encompasses both anticentromere (ACA)- and anti-SCL70 (SCL70)-positive cohorts. To evaluate the clinical outcomes of skin fibrosis, pulmonary fibrosis, and arthritis, they employed 2 logistic regression models. One model incorporated autoantibody status along with the identified features (green curve), providing a multifactorial perspective, while the other relied solely on the identified features (blue curve) to isolate the predictive power of those variables alone. The area under the curve values for these models were calculated to assess their predictive accuracy, and these were subsequently compared with the predictive capacity of autoantibody status alone (orange curve). Additionally, to understand the relative importance of specific predictors in the model, the contributions of individual features were extracted by analysing the absolute values of the feature coefficients [1].

However, this paper raises important methodological questions about the suitability of employing PCA and logistic regression in this context, particularly due to PCA's inherent linearity and the parametric assumptions underlying logistic regression. These choices can lead to misleading interpretations, especially when analysing complex biological data that frequently exhibit nonlinear and nonparametric traits. Numerous peer-reviewed studies highlight the limitations associated with applying linear methods to data that are fundamentally nonlinear and using parametric models on nonparametric data; such misapplications can introduce significant distortions and result in skewed conclusions. This underscores the necessity for caution and methodological rigour when analysing biological data to ensure that the interpretations drawn are both valid and meaningful.

PCA, for instance, operates by emphasising linear relationships among features, potentially overlooking meaningful associations between variables that could provide insights into the underlying biological processes. Furthermore, PCA's reliance on certain assumptions—such as the necessity for linear, meaningful intervariable correlations, the use of continuous and standardised data, a sufficiently large sample size, homoscedasticity, and the presence of minimal outliers—can lead to the exclusion of critical nonlinear features that may be vital for accurately modelling biological systems. Consequently, the violation of these assumptions can significantly distort the results obtained from PCA, rendering it ill-suited for analysing the inherently nonlinear and nonparametric characteristics of biological data [2–6]. This limitation emphasises the need for alternative analytical methods that can more effectively capture the complexities and nuanced relationships present in biological systems.

Similarly, logistic regression is based on parametric assumptions regarding the data, including the independence of predictors and the absence of multicollinearity. When these assumptions are violated, particularly in biological data that often exhibit complex interactions and nonparametric characteristics, logistic regression may yield unreliable results. Just as with PCA, the reliance on these assumptions can lead to significant distortions in the analysis, potentially obscuring critical relationships and insights inherent in the data [7–10]. Therefore, the application of linear and parametric models to such multifaceted biological data warrants careful scrutiny to avoid erroneous interpretations and ensure that the underlying biological truths are accurately represented. This highlights the importance of exploring alternative analytical approaches that are better suited to accommodate the complexities of biological systems.

In this regard, the paper recommends employing nonlinear nonparametric methods such as mutual information analysis and effective transfer entropy to analyse complex interactions among multiple variables, especially when those interactions are characterised by nonmonotonic patterns. These methods can effectively capture the interdependencies and relationships that linear models may overlook, providing a richer and more accurate representation of the underlying biological phenomena.

Contributors

YT completed this research and wrote this article.

Funding

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

Handling editor Josef S. Smolen.

https://doi.org/10.1016/j.ard.2025.04.014
Received 9 April 2025; Accepted 13 April 2025

Please cite this article as: Y. Takefuji, Reevaluating analytical approaches in systemic sclerosis research: challenges of PCA and logistic regression, Ann Rheum Dis (2025), https://doi.org/10.1016/j.ard.2025.04.014

Competing interests

The author has no conflict of interest.

Patient consent for publication

Not applicable.

Ethics approval

Not applicable.

Provenance and peer review

Not commissioned; externallly peer-reviewed.

Data availability statement

Not applicable.

Orcid

Yoshiyasu Takefuji: http://orcid.org/0000-0002-1826-742X

REFERENCES

[1] Rouvière B, Le Dantec C, Bettacchioli E, Beretta L, Foulquier N, Cao C, et al. Stratification according to autoantibody status in systemic sclerosis reveals distinct molecular signatures. Ann Rheum Dis 2025;84(3):480–90. doi: 10.1136/ard-2024-225925.

[2] Dyer EL, Kording K. Why the simplest explanation isn't always the best. Proc Natl Acad Sci U S A 2023;120(52):e2319169120. doi: 10.1073/pnas.2319169120.

[3] Cristian PM, Aarón VJ, Armando ED, Estrella MY, Daniel NR, David GV, et al. Diffusion on PCA-UMAP Manifold: the impact of data structure preservation to denoise high-dimensional single-cell RNA sequencing data. Biology (Basel) 2024;13(7):512. doi: 10.3390/biology13070512.

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

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

[6] Mohseni N, Elhaik E. Biases of principal component analysis (PCA) in physical anthropology studies require a reevaluation of evolutionary insights. eLife 2024;13:RP94685. doi: 10.7554/eLife.94685.2.

[7] Dey D, Haque MS, Islam MM, Aishi UI, Shammy SS, Mayen MSA, et al. The proper application of logistic regression model in complex survey data: a systematic review. BMC Med Res Methodol 2025;25(1):15. doi: 10.1186/s12874-024-02454-5.

[8] Pinheiro-Guedes L, Martinho C, O Martins MR. Logistic regression: limitations in the estimation of measures of association with binary health outcomes. Acta Med Port 2024;37(10):697–705. doi: 10.20344/amp.21435.

[9] Wang T, Tang W, Lin Y, Su W. Semi-supervised inference for nonparametric logistic regression. Stat Med 2023;42(15):2573–89. doi: 10.1002/sim.9737.

[10] Rifada M, Chamidah N, Ningrum RA. Estimation of nonparametric ordinal logistic regression model using generalized additive models (GAM) method based on local scoring algorithm. AIP Conf Proc 2022;2668(1):070013. doi: 10.1063/5.0111771.

Yoshiyasu Takefuji *
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan
***Correspondence to** Dr Yoshiyasu Takefuji, Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan.
E-mail address: takefuji@keio.jp