JOURNAL
OF HEPATOLOGY

Limitations of sparse partial least squares in multiomics: A critical analysis of linear methods applied to non-linear biological data

To the Editor:

Gadd *et al.* investigated how host hepatocyte senescence influences the success of hepatocyte transplantation in a mouse model of liver injury using a sparse partial least squares (sPLS) model for dimensionality reduction of paired multiomic data.[1]

Understanding fundamental theoretical principles of machine learning tools is crucial for biological analysis of multiomic data. While supervised machine learning provides ground truth values for target prediction accuracy validation, feature importance and reduction methods lack such validation metrics. The application of linear methods such as sPLS to non-linear data, or parametric approaches to non-parametric data, can lead to distorted outcomes and erroneous conclusions.[2–4]

The application of sPLS to multiomic data, as utilized by Gadd *et al.*, introduces significant analytical challenges. sPLS is based on the assumption of linear relationships ($y = X\beta + \varepsilon$), which is fundamentally at odds with the inherently non-linear dynamics observed in gene regulatory networks and metabolic pathways. This basic discord forces the method to oversimplify complex, non-linear patterns into linear models, potentially obscuring critical biological insights and leading to erroneous conclusions.[5–8]

Moreover, the parametric constraints of sPLS impose additional limitations on biological data analysis. While sPLS relies on fixed parameters and predefined statistical distributions, biological data often follows unknown or non-normal distributions. Additionally, by primarily focusing on linear feature combinations, sPLS overlooks complex synergistic or antagonistic interactions that are common in biological systems. The assumption of normally distributed errors further exacerbates bias, especially when the actual data exhibit skewed or multimodal patterns, resulting in biased feature selection.

These methodological limitations can manifest in several detrimental ways, including incorrect feature importance rankings, missed biological interactions, and ultimately misleading interpretations of molecular mechanisms. In practice, crucial non-linear relationships between genes and proteins might be overlooked, regulatory network interactions oversimplified, and essential pathway dependencies misinterpreted, thus substantially compromising the validity and reliability of multiomic analyses using sPLS.

To address these limitations, this paper advocates for the use of non-linear and non-parametric robust statistical methods. For instance, rank-based correlation measures such as Spearman's correlation and Kendall's tau[9] are noted for their ability to effectively capture monotonic relationships without assuming linearity. Unlike traditional methods that rely on absolute data values, these techniques assess the rank order of the data, making them less sensitive to outliers and skewed distributions – a common characteristic in biological datasets. Additionally, ordinal association measures like Goodman-Kruskal gamma and Somers' D[10] provide robust alternatives for analyzing ranked data by quantifying the strength and directionality of associations between ordinal variables. These measures are particularly advantageous when the underlying data do not meet the criteria of normality or exhibit multimodal patterns. Accompanied by appropriate *p* values for statistical significance, these methods offer a more reliable approach for uncovering and validating complex biological relationships, ensuring both statistical rigor and enhanced biological relevance.

Yoshiyasu Takefuji*
Musashino University, Data Science Department, Japan
*Corresponding author. Address: 3-3-3 Ariake, Koto-Ku, Tokyo 135-8181, Japan.
E-mail address: takefuji@keio.jp

Received 13 March 2025; Accepted 20 March 2025; available online 25 March 2025
https://doi.org/10.1016/j.jhep.2025.03.021

Financial support

The authors did not receive any financial support to produce this manuscript.

Conflict of interest

The author has no conflict of interest.

Please refer to the accompanying ICMJE disclosure forms for further details.

Authors' contributions

Yoshiyasu Takefuji completed this research and wrote this article.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhep.2025.03.021.

References

[1] Gadd VL, Ferreira-Gonzalez S, Man TY, et al. Host hepatocyte senescence determines the success of hepatocyte transplantation in a mouse model of liver injury. J Hepatol 2025;83(1):161–176. https://doi.org/10.1016/j.jhep.2024.12.039.

[2] Chen M, Papadikis K, Jun C, et al. Linear, nonlinear, parametric, and nonparametric regression models for nonstationary flood frequency analysis. J Hydrol 2023;616:128772. https://doi.org/10.1016/j.jhydrol.2022.128772.

[3] Janse RJ, Hoekstra T, Jager KJ, et al. Conducting correlation analysis: important limitations and pitfalls. Clin Kidney J 2021;14(11):2332–2337. https://doi.org/10.1093/ckj/sfab085.

[4] Jarantow SW, Pisors ED, Chiu ML. Introduction to the use of linear and nonlinear regression analysis in quantitative biological assays. Curr Protoc 2023;3(6):e801. https://doi.org/10.1002/cpz1.801.

[5] Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc Ser B Stat Methodol 2010;72(1):3–25. https://doi.org/10.1111/j.1467-9868.2009.00723.x.

[6] Ahmad NA. Numerically stable locality-preserving partial least squares discriminant analysis for efficient dimensionality reduction and classification of high-dimensional data. Heliyon 2024;10(4):e26157. https://doi.org/10.1016/j.heliyon.2024.e26157.

[7] Wüthrich K, Zhu Y. Omitted variable bias of lasso-based inference methods: a finite sample analysis. Rev Econ Stat 2023;105(4):982–997. https://doi.org/10.1162/rest_a_01128.

[8] Jain R, Xu W. HDSI: high dimensional selection with interactions algorithm on feature selection and testing. PLoS One 2021;16(2):e0246159. https://doi.org/10.1371/journal.pone.0246159. Published 2021 Feb 16.

[9] Okoye K, Hosseini S. Correlation tests in R: pearson cor, Kendall's tau, and Spearman's Rho. In: R Programming. Singapore: Springer; 2024. https://doi.org/10.1007/978-981-97-3385-9_12.

[10] Metsämuuronen J. Directional nature of Goodman–Krus gamma and some consequences: Identity of Goodman–Kruskal gamma and Somers delta, and their connection to Jonckheere–Terpstra test statistic. Behaviormetrika 2021;48:283–307. https://doi.org/10.1007/s41237-021-00138-8.