Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Comment on "Data Fission: Splitting a Single Data Point" Data Fission for Unsupervised Learning: A Discussion on Post-Clustering Inference and the Challenges of Debiasing

Changhu Wang, Xinzhou Ge, Dongyuan Song & Jingyi Jessica Li

To cite this article: Changhu Wang, Xinzhou Ge, Dongyuan Song & Jingyi Jessica Li (2025) Comment on "Data Fission: Splitting a Single Data Point" Data Fission for Unsupervised Learning: A Discussion on Post-Clustering Inference and the Challenges of Debiasing, Journal of the American Statistical Association, 120:549, 174-175, DOI: 10.1080/01621459.2024.2412191

To link to this article: https://doi.org/10.1080/01621459.2024.2412191

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 120, NO. 549, 174–175: Theory and Methods Discussion
https://doi.org/10.1080/01621459.2024.2412191

Comment on "Data Fission: Splitting a Single Data Point" Data Fission for Unsupervised Learning: A Discussion on Post-Clustering Inference and the Challenges of Debiasing

Changhu Wang[a], Xinzhou Ge[b], Dongyuan Song[c], and Jingyi Jessica Li[a,c]

[a]Department of Statistics and Data Science, University of California, Los Angeles, Los Angeles, CA; [b]Department of Statistics, Oregon State University, Corvallis, OR; [c]Interdepartmental Program of Bioinformatics University of California, Los Angeles, Los Angeles, CA

This article introduced data fission (Leiner et al. 2023), a valuable methodology for post-selection inference. While the article focused on the supervised setting, we explore its potential for unsupervised applications. In single-cell RNA-seq data analysis, identifying cell-type marker genes typically involves clustering cells followed by testing each gene for overexpression in each cluster, leading to a "double-dipping" issue that inflates false the discovery rate (FDR) (Zhang, Kamath, and David 2019; Neufeld et al. 2022; Song et al. 2023). One method to address this issue is countsplit (Neufeld et al. 2022), a form of data thinning (Neufeld et al. 2024) conceptually similar to data fission, which splits data points to allow clustering and marker identification on separate datasets.

Specifically, data fission requires the decomposition of data $\mathbf{X}$ into $f(\mathbf{X})$ and $g(\mathbf{X})$ to satisfy one of the following two properties:

(P1) $f(\mathbf{X})$ and $g(\mathbf{X})$ are independent with known distributions[1]; or

(P2) $f(\mathbf{X})$ has a known marginal distribution, and $g(\mathbf{X})$ has a known conditional distribution given $f(\mathbf{X})$.[2]

Then $f(\mathbf{X})$ is used for selection, and $g(\mathbf{X})$ for inference, using either $g(\mathbf{X})$ under (P1) or $g(\mathbf{X})|f(\mathbf{X})$ under (P2). In single-cell RNA-seq data analysis, a negative binomial distribution is typically assumed for each data point,[3] and only (P2) holds if data fission splits the data points using binomial sampling (see Appendix A of Leiner et al. 2023). However, deriving $g(\mathbf{X})|f(\mathbf{X})$ for single-cell RNA-seq data can be challenging.

In detail, we use $\mathbf{X} = [X_{ij}] \in \mathbb{R}^{n\times p}$ to represent a single-cell RNA-seq gene expression matrix, where n is the number of cells, p is the number of genes, and $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ denotes the unobserved cell types. Ideally, $\mathbf{Z}$ would be used to test whether gene j is a marker gene. However, in practice, the test is based on the cluster labels $\widehat{\mathbf{Z}} = (\widehat{Z}_1, \ldots, \widehat{Z}_n)^\top$. With data fission, clustering is performed on $f(\mathbf{X})$, resulting in $\widehat{\mathbf{Z}} = C(f(\mathbf{X}))$, where $C(\cdot)$ represents the clustering function, and tests are conducted on $g(\mathbf{X})$. Since $\widehat{\mathbf{Z}} \neq \mathbf{Z}$, bias may arise, affecting FDR control. With knowledge of $g(\mathbf{X})|f(\mathbf{X})$, or $g(\mathbf{X})|C(f(\mathbf{X}))$, debiasing can be performed. However, in single-cell RNA-seq data, the complex dependency structure of genes (e.g., gene-gene correlations (Song et al. 2023)) often makes $g(\mathbf{X})|f(\mathbf{X})$ and $g(\mathbf{X})|C(f(\mathbf{X}))$ intractable to compute in practice. The only exceptions where debiasing is unnecessary are the ideal case where clustering is perfect, meaning $\widehat{\mathbf{Z}} = \mathbf{Z}$, and a less ideal, yet hardly realistic, case where $\widehat{\mathbf{Z}}$ is independent of the non-marker genes. In this latter case, the non-marker genes are not involved in the clustering process, and such independence ensures no FDR inflation.

Our discussion aligns to some extent with the literature: debiasing can be performed using post-selection inference without data fission (Chen, and Witten 2023; Gao, Bien, and Witten 2024), restricting the clustering function $C(\cdot)$ to be K-means or hierarchical clustering; debiasing is unnecessary in the data thinning paper (Neufeld et al. 2024) because it does not consider gene dependencies and thus assumes $\widehat{\mathbf{Z}}$ to be independent of the non-marker genes.

In summary, debiasing is crucial in data fission but challenging to implement for post-clustering inference with general clustering algorithms and in realistic scenarios. Further research is needed to explore debiasing in the context of data fission for unsupervised learning.

Disclosure Statement

The authors declare that they have no known competing interests.

CONTACT Jingyi Jessica Li jli@stat.ucla.edu Department of Statistics and Data Science, University of California, Los Angeles, 520 Portola Pl., 8125 Math Sciences Bldg., Los Angeles, CA 90095-1554.

[1]up to the unknown θ

[2]up to the unknown θ

[3]The negative binomial distribution is assumed in the data thinning paper (Neufeld et al. 2024), while a Poisson distribution was assumed in the same authors' previous countsplit paper (Neufeld et al. 2022).

Funding

This work was supported by the following grants: National Science Foundation DBI-1846216 and DMS-2113754, NIH/NIGMS NIH funding, and the Chan-Zuckerberg Initiative Single-Cell Biology Data Insights Grant (to J.J.L.).

References

Chen, Y. T., and Witten, D. M. (2023), "Selective Inference for k-means Clustering," *Journal of Machine Learning Research*, 24, 1–41. [174]

Gao, L. L., Bien, J., and Witten, D. (2024), "Selective Inference for Hierarchical Clustering," *Journal of the American Statistical Association*, 119, 332–342. [174]

Leiner, J., Duan, B., Wasserman, L., and Ramdas, A. (2023), "Data Fission: Splitting a Single Data Point," *Journal of the American Statistical Association* DOI: 10.1080/01621459.2023.2270748. [174]

Neufeld, A., Dharamshi, A., Gao, L. L., and Witten, D. (2024), "Data Thinning for Convolution-Closed Distributions," *Journal of Machine Learning Research*, 25, 1–35. [174]

Neufeld, A., Gao, L. L., Popp, J., Battle, A., and Witten, D. (2022), "Inference after Latent Variable Estimation for Single-Cell RNA Sequencing Data," *Biostatistics*, 12, kxac047. [174]

Song, D., Li, K., Ge, X., and Li, J. J. (2023), "ClusterDE: A Post-Clustering Differential Expression (DE) Method Robust to False-Positive Inflation Caused by Double Dipping," bioRxiv. DOI:10.1101/2023.07.21.550107. [174]

Zhang, J. M., Kamath, G. M., and David, N. T. (2019), "Valid Post-Clustering Differential Analysis for Single-Cell RNA-seq," *Cell Systems*, 9, 383–392. [174]