Patterns

CellPress
OPEN ACCESS

Preview

A supervised take on dimensionality reduction via hybrid subset selection

Javad Rahimikollu[1,2] and Jishnu Das[2,*]
[1]CMU-Pitt Program in Computational Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[2]Center for Systems Immunology, Departments of Immunology and Computational & Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
*Correspondence: jishnu@pitt.edu
https://doi.org/10.1016/j.patter.2022.100563

Amouzgar et al. present HSS-LDA, a supervised dimensionality reduction approach for single-cell data that outperforms existing unsupervised techniques. They couple hybrid subset selection to linear discriminant analysis and identify interpretable linear combinations of predictors that best separate predefined biological groups.

With the advent of multi-omic technologies that can generate deep cellular and molecular profiles, there has been an explosion in the generation of high-dimensional biological datasets both at and across scales of organization. However, while these datasets can help provide multi-faceted insights into complex biological processes, they are often noisy, and the high dimensionality necessitates the choice of appropriate analytical approaches. These include dimensionality reduction as these techniques aim to facilitate visualization of high-dimensional datasets by mapping large numbers of features to lower-dimensional subspaces. These approaches help better visualize and interpret primary components of variability in the data. Popular dimensionality reduction techniques include PCA (principal-component analysis),[1] t-SNE (t-distributed stochastic neighbor embedding),[2] UMAP (uniform manifold approximation and projection),[3] and PHATE (potential of heat-diffusion for affinity-based trajectory embedding).[4] However, these methods are unsupervised, and each comes with its own set of limitations. PCA captures components of highest variance in the data and successfully recapitulates only overall global structure. t-SNE attempts to preserve local structure but is poor at capturing global structure. Further, as a stochastic embedding, the results are highly sensitive to initialization conditions, and pairwise distances in an embedding do not have an intuitive Cartesian interpretation. Alternate hybrid approaches such as a t-SNE with a PCA-based initialization[5] as well as UMAP and PHATE have also been recently proposed as alternate approaches. But these too could be quite sensitive to initialization values; i.e., the initial solutions can make significant difference in the obtained optimal solutions. Further, attempting to preserve pairwise relationships or distances from very high-dimensional data in two or three dimensions is prone to major distortions.[6] Based on the Johnson-Lindenstrauss lemma, it would take hundreds of dimensions to preserve the structure of typical single-cell datasets within reasonable error bounds.[6] Thus, semi-supervised or supervised dimensionality reduction approaches that use group labels may better capture biological structure corresponding to these labels of interest, rather than unsupervised approaches that attempt to preserve all components of variance in the data but end up suffering from large distortions. Motivated by some of these considerations, in this issue of *Patterns*, Bendall and colleagues propose a new method[7]—hybrid subset selection coupled to linear discriminant analyses (HSS-LDA)—to address some of these limitations (Figure 1).

Linear discriminant analysis (LDA) is used as a supervised classifier to stratify data based on *a priori* labels corresponding to groups/classes of interest. The LDA classifier identifies a linear combination of features that maximizes the ratio of the between and within classification group variance. This offers several advantages over current unsupervised dimensionality reduction techniques, which solely exploit variance in the data and are often less sensitive to specific components of covariance between data and labels. However, applying just LDA on high-dimensional multi-collinear data can be sub-optimal, and removing redundant features is often necessary. To this end, the authors used hybrid subset selection (HSS), an iterative feature selection tool that utilizes forward and backward selection to determine the most relevant set of features for LDA and removes unnecessary features. And unlike sparse LDA with L2 regularization, which drives feature coefficients close to zero but does not explicitly remove features, HSS-LDA actually removes features. Distinct from other reduction techniques, HSS-LDA does not require computationally expensive hyperparameter optimization. Further, unlike t-SNE, UMAP, or PHATE, HSS-LDA does not require initialization; hence, the selection of optimal hybrid subsets is not affected by initialization. Further although chosen based on an objective function that seeks to maximize predictive accuracy, the authors also demonstrate that the use of this approach leads to at least partially interpretable hybrid subsets selected for the LDA.

The authors compared HSS-LDA, PCA, UMAP, and PHATE on an array of single-cell datasets with different structural properties and corresponding to a range of biological contexts. The different datasets included discrete and continuous data types, cyclical systems, and imbalanced groups/classes. For example, a mass cytometry dataset that used scatterbodies to discriminate between bone marrow-derived immune and hematopoietic cells

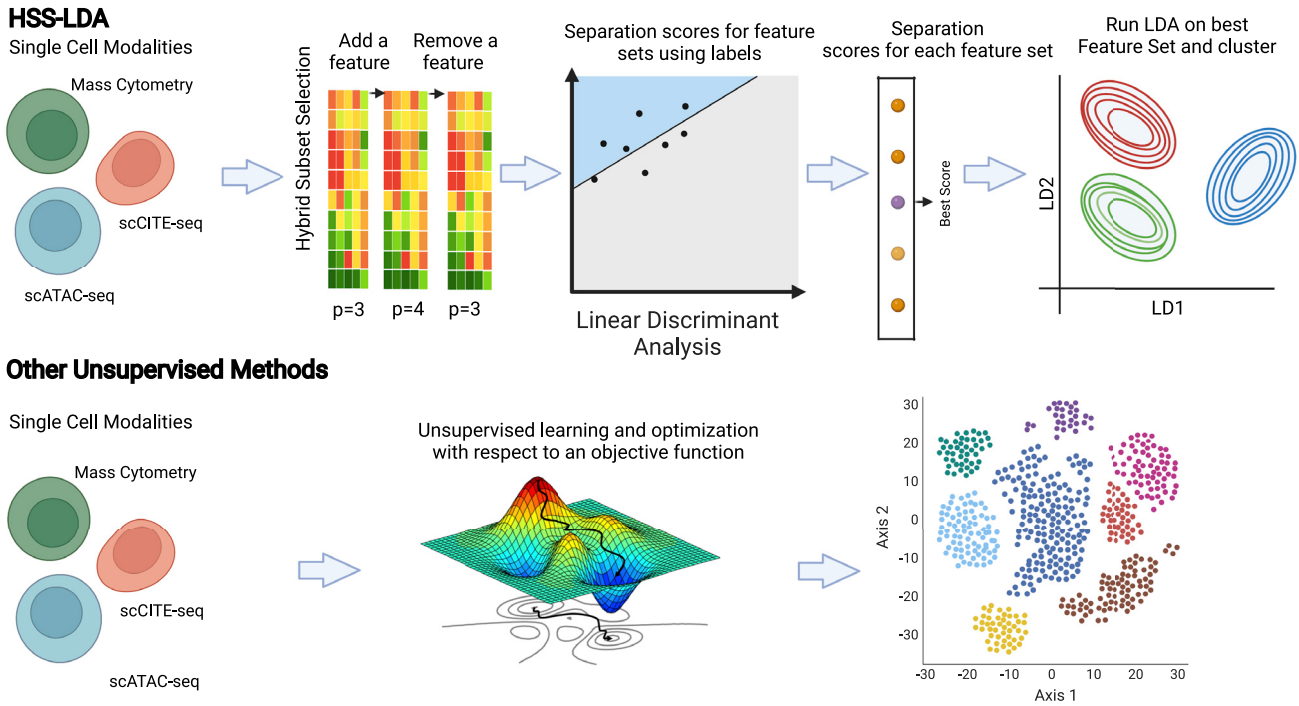

Figure 1. Conceptual comparison of the supervised HSS-LDA approach to existing unsupervised dimensionality reduction techniques

across both healthy and disease samples was efficiently analyzed by HSS-LDA. While existing approaches produced results that were biased by the cellular frequency imbalances (neutrophils are over-represented compared to other cell types), HSS-LDA was not impacted by these frequency differences and accurately predicted identities of cells from patients with hematopoietic malignancies, with high accuracy. In another analysis, the authors analyzed the metabolic state of CD8 T cells obtained from peripheral blood mononuclear cells (PBMCs) across multiple timepoints after *ex vivo* T cell receptor (TCR) stimulation. While both HSS-LDA and UMAP separated cells across timepoints, HSS-LDA captured biologically interpretable linear discriminant axes corresponding to specific senescence and terminal differentiation profiles. In another intriguing example, HSS-LDA captured cyclic biological processes with multi-label data corresponding to both cell type and cell-cycle phase labels. While UMAP was confounded by imbalances in the distribution of the dataset, HSS-LDA selected features that separated both cell-cycle labels and generated an interpretable feature combination that stratified cellular labels. HSS-LDA also outperformed existing approaches in a range of other contexts including those that involved multi-omic integration and capturing cellular differentiation as well as pseudotime trajectories.

Overall, HSS-LDA opens new avenues of exploration for the development of supervised and semi-supervised dimensionality reduction approaches for multi-omic biological datasets, especially those corresponding to single-cell modalities. Importantly, methods like HSS-LDA are constrained by the availability of suitable group labels, and when labels are not available or noisy, existing unsupervised dimensionality approaches may be more appropriate than HSS-LDA. However, when labels are available, approaches like HSS-LDA may outperform existing unsupervised dimensionality reduction approaches to capture complex cellular processes and trajectories along interpretable axes.

ACKNOWLEDGMENTS

J.D. was partially supported by NIAID DP2AI164325.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Jolliffe, I.T., and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philos. Trans. A Math Phys. Eng. Sci. *374*, 20150202. https://doi.org/10.1098/rsta.2015.0202.

2. Maaten, L.v.d., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

3. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.

4. Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., Elzen, A.V.D., Hirn, M.J., Coifman, R.R., et al. (2019). Visualizing structure and transitions in high-dimensional biological data. Nat. Biotechnol. *37*, 1482–1492. https://doi.org/10.1038/s41587-019-0336-3.

5. Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. Nat. Commun. *10*, 5416. https://doi.org/10.1038/s41467-019-13056-x.

6. Chari, T., Banerjee, J., and Pachter, L. (2021). The specious art of single-cell genomics. Preprint at bioRxiv. https://doi.org/10.1101/2021.08.25.457696.

7. Amouzgar, M., Glass, D.R., Baskar, R., Averbukh, I., Kimmey, S.C., Tsai, A.G., Hartmann, F.J., and Bendall, S.C. (2022). Supervised dimensionality reduction for exploration of single-cell data by HSS-LDA. Patterns *3*, 100536. https://doi.org/10.1016/j.patter.2022.100536.