Parts-based decomposition of spatial genomics data finds distinct tissue regions

Dimension reduction is a cornerstone of exploratory data analysis; however, traditional methods fail to preserve the spatial context of spatial genomics data. In this work, we develop a nonnegative spatial factorization (NSF) model that allows interpretable, parts-based decomposition of spatial single-cell count data. NSF allows label-free annotation of regions of interest in spatial genomics data and identifies genes and cells that can be used to define those regions.

This is a summary of:
Townes, F. W. & Engelhardt, B. E. Nonnegative spatial factorization applied to spatial genomics. *Nat. Methods* https://doi.org/10.1038/s41592-022-01687-w (2022)

Published online:
Published online: 7 January 2023

The mission

Dimension reduction aims to find a low-dimensional set of components that capture variation in high-dimensional data such as genomics data. Constraining components to be nonnegative[1], as opposed to real-valued[2], can often improve interpretability as nonnegative components are forced to represent constituent parts and include only a small number of features, as opposed to real-valued methods that generally capture global patterns using all features. Our mission in this work was to develop a model with this parts-based behavior when applied to count-based, single-cell spatial genomics data derived from many different available technologies. Our aims for the model were that nearby cells should be included in the same components when they had similar cell states and that each component should represent unique functional patterns and regions of interest in the tissue. We also wanted to be able to directly identify the genes that make up each component. If possible, we hoped to use results from our model to distinguish spatially associated variability from variability that is intrinsic to the cell and does not depend on its location.

The solution

We built this behavior in nonnegative spatial factorization (NSF) using a Gaussian process-regularized, nonnegative latent factor model with a non-smooth kernel to allow sharp region boundaries. The Gaussian-process prior induces a correlation structure in the component values of cells based on their distance in space, whereby nearby cells have highly correlated component values and distant cells have little to no correlation in their component values. Inference in this model uses variational inducing point methods that scale to large numbers of cells. In an extension of our model that we named NSF hybrid, we included both spatial components and nonspatial components that absorb spatially uncorrelated patterns of gene expression in tissue samples. This allowed us to compare spatial and nonspatial variation in gene expression and cell state through spatial importance scores, which compare the normalized weight of the spatial components that capture the expression of a gene or a cell state against the weight of the nonspatial components that capture the expression of a gene or a cell state.

Using simulated spatial genomic data, we found that the nonnegative model more accurately recovered the unique components of the simulations than real-valued alternatives such as factor analysis or MEFISTO. Spatially aware models had higher prediction accuracy for masked cells. When applied to spatial transcriptomics datasets of mouse brain samples gathered using the Visium and Slide-seqV2 platforms, the NSF hybrid model identified distinct brain regions (Fig. 1) and also gene markers identifying those regions. For example, the meninges, a thin boundary separating brain regions only a few cells thick, were identified in a Slide-seqV2 mouse hippocampus sample (Fig. 1, panel 10). We found that the genes with the highest weight for each component are generally a poor representation of that brain region, suggesting that many genes should be used as markers for brain regions for better accuracy. We used reference single-cell datasets to identify the most abundant cell types in each component. The nonspatial components identified cell subtypes that appeared diffusely throughout the sample. Moreover, the cell types with the largest weights in distinct brain regions and model components have distinct marker genes, suggesting that there are many distinct and identifiable subtypes of cells that define specific brain regions.

Future directions

As spatial genomics extends into three and four dimensions, our modular implementation will be able to analyze these 3D or spatiotemporal datasets. NSF may be used to annotate novel organ atlases and developmental atlases by identifying functional regions and regional boundaries and capturing marker genes and cell types.

NSF is limited in tractability to tens of thousands of cells; however, we suggest that nearest-neighbor Gaussian process methods[3] would improve scaling to hundreds of thousands of cells. Future work will also enable analysis of spatial variation within specific cell types by leveraging new and existing large single cell atlases[4].

For spatial genomic data with cell type labels, we plan to extend NSF using multi-group Gaussian processes[5] to encourage nearby cells with the same or similar cell types to share components.

F. William Townes[1] & Barbara E. Engelhardt[2,3]
[1]Carnegie Mellon University, Pittsburgh, PA, USA. [2]Gladstone Institutes, San Francisco, CA, USA. [3]Stanford University, Stanford, CA, USA.

EXPERT OPINION

"Applying both of their methods to simulated and real data, the authors demonstrate favorable performance compared to factor analysis, MEFISTO and probabilistic NMF. This paper will be useful to the field and I am excited to see it published; the supplied code was sufficient to reproduce the analyses as described in the manuscript (with acceptable deviations arising from system differences and the stochastic nature of some of the decomposition procedures)." **Genevieve Stein-O'Brien, Johns Hopkins University, Baltimore, MA, USA.**

FIGURE

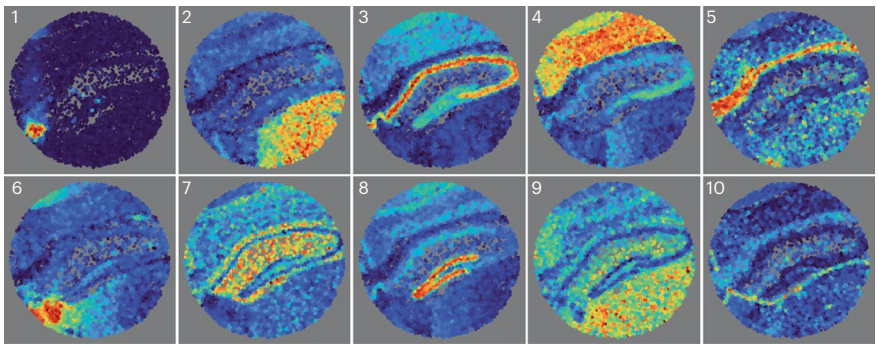

Fig. 1 | The nonnegative spatial factorization hybrid (NSFH) model identifies distinct regions in Slide-seqV2 mouse hippocampus data. Each panel shows a different latent spatial component as a surface in two-dimensional space; red and yellow capture higher values, blue captures values near zero. Regions identified include the choroid plexus (1), thalamus (2), dentate gyrus (8) and meninges (10).

BEHIND THE PAPER

Our work was inspired by MEFISTO[2], which pioneered the combination of spatial correlation with real-valued dimension reduction in genomics applications, as well as the classic nonnegative matrix factorization paper[1] that introduced the concept of parts-based representation. Additional inspiration came from the environmental science and neuroscience research communities. I spent a lot of time and energy on the implementation of all the models in TensorFlow, and in some ways the engineering details may be more consequential to performance than modeling choices. The biggest challenge of the project was the COVID-19 pandemic. This created frequent life disruptions that delayed progress, and I am grateful to my co-author Prof. Engelhardt, the editors and the reviewers for their patience. My wife Tina Townes deserves enormous credit for making it possible for me to finish this work. **F.W.T.**

REFERENCES

1. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
Original nonnegative matrix factorization paper that introduces the concept of an interpretable parts-based representation.
2. Velten, B. et al. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat. Methods* **19**, 179–186 (2022).
This paper showed how to combine real-valued dimension reduction with spatial correlation using Gaussian processes.
3. Wu, L. et al. Variational nearest neighbor Gaussian process. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2202.01694 (2022).
This study combines variational inference for nonconjugate likelihoods (also used by us) with nearest-neighbor approximations to enable greater scalability to large numbers of observations.
4. Cable, D. M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol.* **40**, 517–526 (2022).
This study uses single-cell reference data to assign cell types to spatial transcriptomics data and then examines spatial variation within each cell type.
5. Li, D. et al. Multi-group Gaussian processes. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2110.08411 (2021).
This study extends Gaussian processes to include both spatial locations and categorical labels such as cell or tissue type.

FROM THE EDITOR

"High-dimensional spatial data such as those generated by spatially resolved transcriptomics technologies are challenging to analyze owing to complex underlying forces and noise. Using nonnegative spatial factorization (NSF) and its extension, NSF hybrid, various spatial and nonspatial patterns of variation in such data can be identified and quantified, facilitating interpretation and further downstream analysis." **Lin Tang, Senior Editor, *Nature Methods*.**