FIST-GT: A tool for multidimensional spatial transcriptomics data imputation via graph-regularized tensor completion

Thomas Atkins Tianci Song Rui Kuang
Department of Computer Science and Engineering, University of Minnesota

January 13, 2022

Background

Spatial transcriptomics (ST) is a highly promising new technology for measuring gene expression across a tissue section that captures spatial heterogeneity of the whole transcriptome. The technology captures two dimensional gene expression profile from tissue sections. Furthermore, parallel tissue slices can be combined to create a three dimensional gene expression map. However, data from current experimental technologies is significantly zero-inflated due to low capture efficiency, necessitating a means of data reconstruction.

Methods

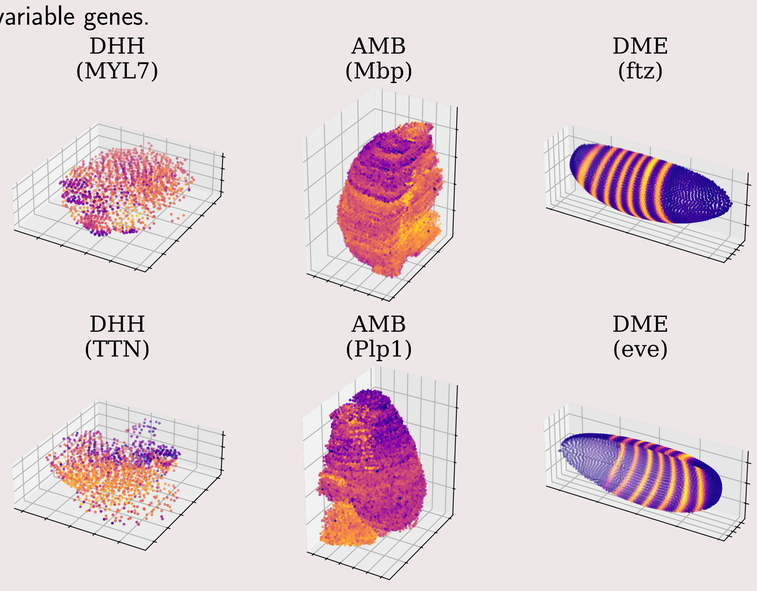
Let $\mathcal{T} \in \mathbb{R}_{+}^{n_p \times n_1 \times \ldots \times n_N}$ be the observed gene expression tensor, where n_p denotes the number of genes, and $n_1, \ldots, n_N$ represent the size of the spatial dimensions. We approximate $\mathcal{T}$ with $\hat{\mathcal{T}}$ where $\hat{\mathcal{T}}$ is represented in the Canonical Polyadic Decomposition (CPD) form, so $\hat{\mathcal{T}} = \sum_{i=1}^{r} \otimes_{i=1}^{N} [\hat{A}_p]_{:,i}$.
Now, we let the following be our objective function, similar to [1]:

$$\min_{\hat{A}_i, i \in \{x_1, \ldots, x_n, p\}} \frac{1}{2} ||\mathcal{M} \circledast (\mathcal{T} - \hat{\mathcal{T}})||_{\mathcal{F}}^2 + \frac{\lambda}{2} \mathbf{vec}(\hat{\mathcal{T}})^T \mathfrak{L}(x_1, \ldots, x_n, p) \mathbf{vec}(\hat{\mathcal{T}})$$

where $\mathcal{M}$ is a binary mask tensor of the observed values and $\mathfrak{L}(x_1, \ldots, x_n, p)$ is the graph Laplacian of the Cartesian product of G_{x_1}, .., G_{x_N}, and G_p, where G_{x_i} is a spatial chain graph of x_i, and G_p is a protein-protein interaction network. To minimize our objective function we utilize a multiplicative update rule (omitted here for brevity).

Data

We test the method on three 3D gene expression datasets. The first measures gene expression in the developing human heart at 6.5 PCW (DHH). [2] The dataset was created by mapping 9 tissue slices sequenced using ST into one tissue atlas. The second dataset is an expression atlas of the adult mouse brain (AMB). [3] This dataset was similarly prepared through ST sequencing of parallel 2-dimensional sequences, assembled to form a three-dimensional dataset. The third is an atlas of developmental genes in a stage 5/6 *Drosophila* embryo (DME). [4] Unlike this other two datasets, this dataset was obtained through fluorescent antisense RNA imaging. The three datasets are visualized below, colored based on expression of a sample spatially variable genes.

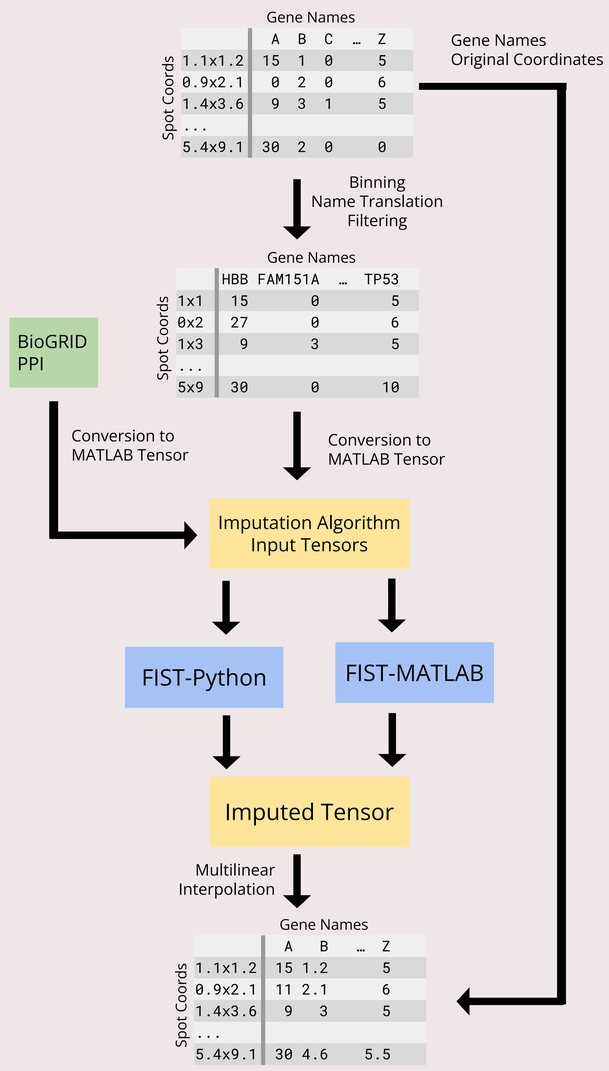


Additionally, the number of genes obtained in every dataset can be found in the table below:

Dataset	Method	n_p
DHH	Stacked ST	13850
AMB	Stacked ST	14035
DME	mRNA Imaging	84

Pipeline

Our method takes 3D data as a table of genes by spots (continuous in 3D space), converts it to a discrete tensor representation, imputes this tensor, and then interpolates it back into the continuous spot data.



Results

We compare our model (FIST-GT) to a spatial nearest-neighbor model (SNN) using 5-fold cross validation. Here, we plot the cumulative distribution function (CDF) of absolute errors for the two methods, and see that the error CDF of FIST-GT is generally less than the error CDF of SNN.

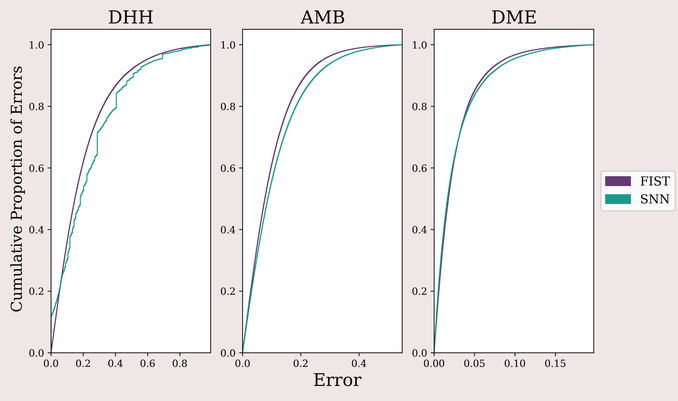


Figure: Cumulative absolute error distributions for both models on all datasets.

A one-sided paired Wilcoxon signed-rank test on the error distributions of the two methods produced the following:

Dataset	Statistic	p
DHH	$6.4 \cdot 10^{11}$	< 0.001
AMB	$4.8 \cdot 10^{13}$	< 0.001
DME	$1.6 \cdot 10^{9}$	0.79

Results (cont.)

Additionally, we measure performance using the following three metrics:

- Mean absolute error (MAE)$= \frac{1}{n} \sum_i |\mathcal{T}_i - \hat{\mathcal{T}}_i|$.
- Symmetric mean absolute percentage error (SMAPE)$= \frac{1}{n} \sum_i \frac{|\mathcal{T}_i - \hat{\mathcal{T}}_i|}{|\mathcal{T}_i| + |\hat{\mathcal{T}}_i|}$.
- Coefficient of determination (R^2): $1 - \left(\sum_i (\mathcal{T}_i - \hat{\mathcal{T}}_i)^2\right) \left(\sum_i (\mathcal{T}_i - \overline{\mathcal{T}})^2\right)^{-1}$

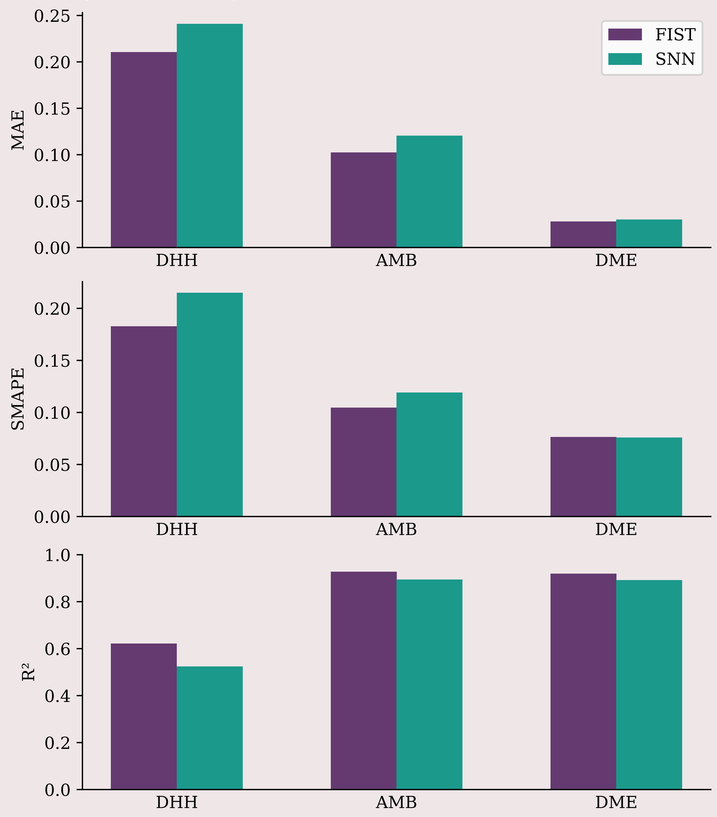


Figure: Comparison of MAE, SMAPE, and R^2 for FIST-GT and SpatialNN on 5-fold cross validation of three datasets.

From the chart above, we see that FIST-GT clearly outperforms SNN for almost every metric on nearly every dataset (the exception being SMAPE on the DME dataset, which is approximately equal for both methods).

Conclusions and Future Directions

Here we have shown that FIST imputes three-dimensional spatial expression data from a variety of datasets more accurately than a spatial nearest-neighbor model. The heterogeneity of the datasets tested demonstrates that the method is widely applicable. Future work will measure errors between the original unbinned data and output interpolated data to ensure the method's real-world utility.

References

1. Li, Z., Song, T., Yong, J. & Kuang, R. Imputation of spatially-resolved transcriptomes by graph-regularized tensor completion. *PLOS Computational Biology* **17** (Apr. 2021).
2. Asp, M. *et al.* A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. *Cell* **179,** 1647–1660.e19 (Dec. 2019).
3. Ortiz, C. *et al.* Molecular atlas of the adult mouse brain. *Science Advances* **6** (2020).
4. Fowlkes, C. C. *et al.* A Quantitative Spatiotemporal Atlas of Gene Expression in the Drosophila Blastoderm. *Cell* **133,** 364–374 (Apr. 2008).

Acknowledgements

This research work is supported by a grant from the National Science Foundations, USA (NSF BIO DBI-IIBR 2042159).