Spatially-aware dimension reduction of transcriptomics data

Lauren Okamoto
Princeton University

Andrew Jones
Princeton University

Archit Verma
Gladstone Institutes

Barbara E. Engelhardt
Gladstone Institutes
Stanford University

Abstract

Spatial sequencing technologies have allowed for studying the relationship between the physical organization of cells and their functional behavior. However, interpreting these data and deriving insights from them remains difficult. Here, we present a Bayesian statistical model that performs dimension reduction for these data in a spatially-aware manner. In particular, our proposed model captures the low-dimensional structure of gene expression while accounting for the spatial variability of expression. Our model also allows us to project dissociated scRNA-seq data onto a spatial grid, as well as use scRNA-seq impute and smooth the expression of spatial sequencing data. Through simulations and applications to spatial sequencing data, we show that our model captures joint structure of spatially-resolved and dissociated sequencing data.

1 Introduction

Spatial genomics technologies provide an opportunity to study the relationship between the molecular makeup of cells and their physical organization. A variety of technologies have been developed in recent years [Ståhl et al., 2016, Rodriques et al., 2019, Stickels et al., 2021, Lee et al., 2021, Zhao et al., 2021], which have contributed to the understanding of cell types, disease, and responses to therapeutics.

While there are extensive applications of spatial genomics technologies, there remains a need for statistical methods that can distill these high-dimensional data into interpretable factors. It is of particular interest to be able to decompose the variability of gene expression into a set of patterns that are driven by the cells' spatial positions and another set of patterns that are driven solely by the cells' expression behavior. Previous methods for this task [Lopez et al., 2019, Zhu and Sabatti, 2020] have focused purely on imputation, while placing less emphasis on spatially-aware dimension reduction. Moreover, it would be beneficial to be able to leverage the vast collection of existing dissociated (non-spatially-resolved) single-cell RNA-sequencing data (scRNA-seq) to inform these factors.

Here, we build on recent work [Verma and Engelhardt, 2020a] by proposing a family of latent variable models that perform dimension reduction on spatial genomics data in a spatially-aware manner by decomposing the data variation into a set of spatial-related and non-spatial-related factors. In particular, our proposed model captures the low-dimensional structure of gene expression while accounting for the spatial variability of expression. Our model also allows us to project dissociated scRNA-seq data onto a spatial grid, as well as use scRNA-seq impute and smooth the expression of spatial sequencing data. The probabilistic framing of our approach allows dissociated scRNA-seq data to be seamlessly incorporated into the model, and even opens the door for imputation of the

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

dissociated cells' spatial coordinates. We demonstrate our model through a series of simulation studies and an application to two gene expression datasets.

2 Spatially-aware dimension reduction

Consider a set of n cells indexed by $i = 1, \dots, n$. Denote cell i's gene expression vector as $\mathbf{y}_i \in \mathbb{R}^p$, where p is the number of genes measured, and denote y_{ij} as the expression level of the jth gene in this cell. Denote the cell's spatial position as $\mathbf{x}_i \in \mathbb{R}^d$, where d is the dimension of the spatial domain (typically, $d \in \{2, 3\}$ in most applications). For data collected from spatially-resolved sequencing technologies, $\mathbf{x}_i$ is directly observed, but for data collected from dissociated cells using scRNA-seq, $\mathbf{x}_i$ is unobserved.

Model Our modeling approach is built as an extension of the Gaussian process latent variable model (GPLVM) [Lawrence, 2003, Titsias and Lawrence, 2010]. The GPLVM assumes that each feature of high-dimensional observations is generated from a GP projection of a lower-dimensional representation (latent variables) of the samples, plus noise. The latent variables can then be investigated for influential patterns in the data.

We extend the GPLVM by augmenting the latent space to also include a set of fixed covariates, which we specify as the spatial locations of the cells. The high-dimensional outcome space is then a function of both unobserved latent variables and the spatial locations. By placing a GP prior on this function, we obtain our model, which we call a semi-supervised Gaussian process latent variable model (ssGPLVM). Specifically, let $\mathbf{z}_i \in \mathbb{R}^{d'}$ be the vector of unobserved latent variables for cell i, and let $\mathbf{t}_i = (\mathbf{x}_i^\top, \mathbf{z}_i^\top)^\top \in \mathbb{R}^{d+d'}$ be the concatenation of spatial coordinates and latent variables. Then the ssGPLVM for cells $i = 1, \dots, n$ and genes $j = 1, \dots p$ is given by

$$y_{ij} = f_j(\mathbf{t}_i) + \epsilon_i, \;\; f_j \sim GP(0, k(\cdot, \cdot)), \;\; \mathbf{z}_i \sim N(\mathbf{0}, \mathbf{I}_{d'}), \;\; \epsilon_i \sim \pi_{\text{noise}}, \tag{1}$$

where π_{noise} is the prior for the noise, and $k(\cdot, \cdot)$ is a covariance function. When cell i does not have observed spatial coordinates, $\mathbf{x}_i$ is unobserved, and we place a Gaussian prior on it, $\mathbf{x}_i \sim N(\mathbf{0}_d, \mathbf{I}_d)$. In our work, we explore two choices for π_{noise}: Gaussian noise $\pi_{\text{noise}} = N(0, \sigma^2)$ and Student's t-distributed noise $\pi_{\text{noise}} = \mathcal{T}(0, \nu, \sigma^2)$. For the t-distributed noise, we optimize the noise variance σ^2 during inference, and we set the degrees of freedom $\nu = 4$ based on prior work [Verma and Engelhardt, 2020b].

Thus, the ssGPLVM allows for modeling gene expression from both spatially-resolved and dissociated cells under a single modeling framework.

Inference Our goal in inference for the ssGPLVM is to compute the posterior for the latent variables, $p(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$. When a subset of the spatial coordinates are unobserved, we also wish to compute the posterior for those spatial coordinates. This posterior is intractable, so we use a variational inference approach using inducing points to approximate the posterior for the latent variables. See Appendix A.2 for details.

3 Results

3.1 ssGPLVM enables spatially-aware dimension reduction

We now show that the ssGPLVM can be used to perform spatially-aware dimension reduction. We first demonstrate this with synthetic data before applying it to spatially-resolved sequencing data.

Simulations We generated synthetic univariate gene expression data with a one-dimensional spatial coordinate. We generated the data with a quadratic function, $y = ax^2 + b$. We randomly sampled $n = 200$ spatial coordinates from the interval $[-3, 3]$ and sampled half of the outcomes with $a = 1, b = 0$ and the other half with $a = 1, b = 4$. We fit both the GPLVM and the ssGPLVM on these data and visualized the latent variables. Each parabola is colored by its spatial coordinates, and the latent coordinates are colored by which parabola they are from. We find that the GPLVM latent variables capture all variation, including the variation between the two parabolas and the spatial shape (Figure 1). On the other hand, the SSPGLVM latent variables remove the variation due to the spatial coordinates, and we are better able to isolate the variation between the parabolas (Figure 1).

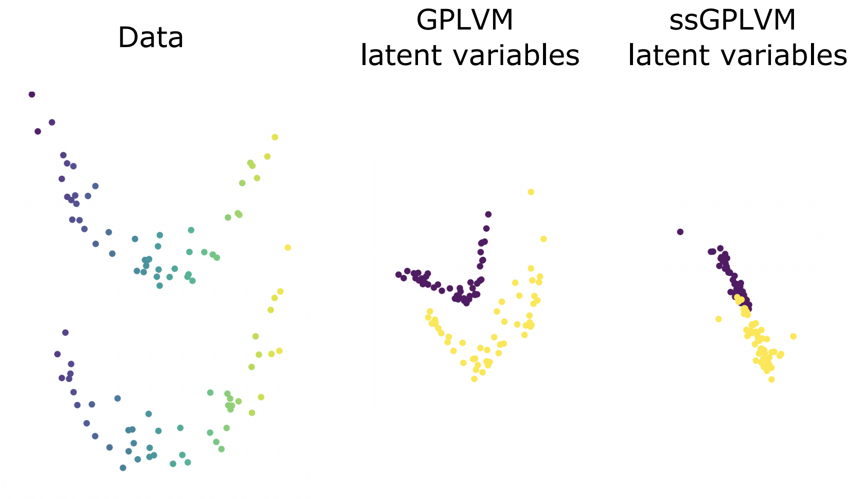

Figure 1: **Demonstration of the GPLVM and ssGPLVM on synthetic data.** *Left:* Synthetic data with two-dimensional outcome (represented by each point's location) and one-dimensional spatial coordinate (represented by each point's color). *Middle:* Latent variables recovered by the GPLVM. Each point's color corresponds to which parabola it belongs to. *Right:* Latent variables recovered by the ssGPLVM. Each point's color corresponds to which parabola it belongs to.

Application to Visium data We next applied the ssGPLVM to a spatial transcriptomics dataset collected from the mouse cortex using the Visium platform [10x Genomics, 2020]. We fit the model with $d' = 2$ latent variables and extracted the latent variables. Visualizing the latent variables, we find that there is some natural clustering of the cells in the latent space (Figure 2). To more precisely define these clusters, we use the K-means algorithm to assign each point to one of $K = 5$ clusters in the latent space. It is then possible to use this data to see what the latent variables capture in terms of biology (see Figure 6 in Appendix). We next checked whether the ssGPLVM latent space captured any information about the spatial coordinates. To check this, we again visualized the latent variables and colored each point by the first spatial coordinate. We find that the spatial coordinates show only minor clustering on the latent variables, suggesting that the latent variables capture information apart from the spatial location (Figure 2).

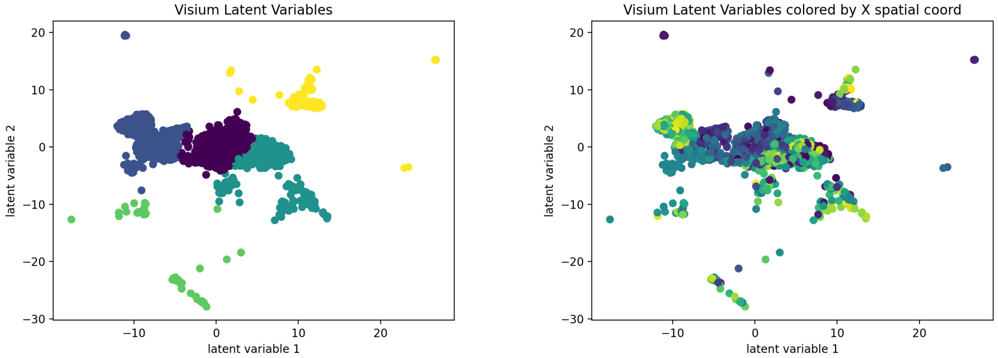

Figure 2: **ssGPLVM applied to Visium mouse cortex data.** *Left:* ssGPLVM latent variables colored by a K-means clustering. *Right:* ssGPLVM latent variables colored by the first spatial coordinate.

3.2 ssGPLVM enables imputation of gene expression and spatial coordinates

We next sought to validate that our model can accurately impute partially missing data.

Simulations We first generated non-spatial synthetic data with a two-dimensional outcome (Figure 3). We then randomly dropped a single outcome variable for two samples and used the ssGPLVM to impute the dropped values. Specifically, we fit the ssGPLVM and computed the approximate posterior distribution over the missing values. We find that the model provides a reasonable estimate of both the mean and variance for the imputed values (Figure 3). As a similar experiment, we next tested whether we could impute a missing spatial coordinate. We find that we are also able to impute the spatial coordinate within reasonable accuracy (Figure 3). This suggests that our model is a feasible approach to jointly modeling partially-observed data in a semi-supervised manner.

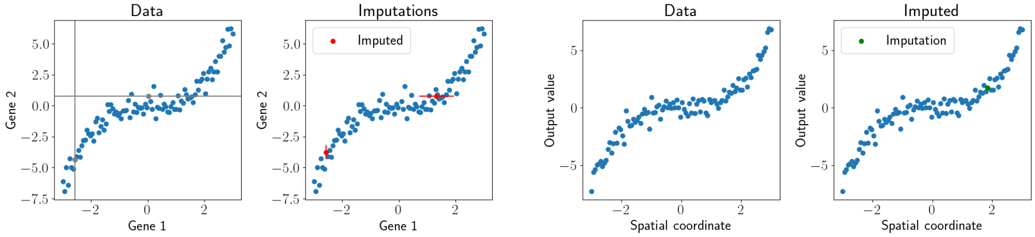

Figure 3: **Imputation of synthetic gene expression and spatial coordinates with the ssGPLVM.** *Left:* Original two-dimensional dataset with two partially observed samples. Gray lines indicate the observed outcome of each sample, while the other outcome is unobserved. *Middle left:* Imputation of synthetic gene expression using a GPLVM. Red points and bars show the approximate posterior mean plus or minus twice the standard deviation. *Middle right:* Original dataset with one-dimensional outcome (y-axis) and one-dimensional spatial coordinate (x-axis). *Right:* Green dot shows imputed spatial coordinate for one point.

Application to seqFISH+ data We then applied our model to a dataset from a mouse cortex collected using the seqFISH+ technology [Eng et al., 2019]. We first impute missing gene expression of the seqFISH+ data set. We find that the imputed gene expression values lie reasonably within the range of what we expect to see (Figure 4).

Next, we impute spatial coordinates. We randomly drop 10% of both of the spatial coordinates (x and y) in seqFISH+, and we use ssGPLVM to try to accurately impute them back. In (Figure 4), we plot the imputed x and y spatial coordinates against the true x and y. If our imputation was perfect, the two plots would show a straight line since the imputed x and y would be equal to the true x and y. However, we do see a positive linear trend in both graphs that closely follow the identity line. This result suggests that ssGPLVM is working properly, and is imputing within a reasonable accuracy.

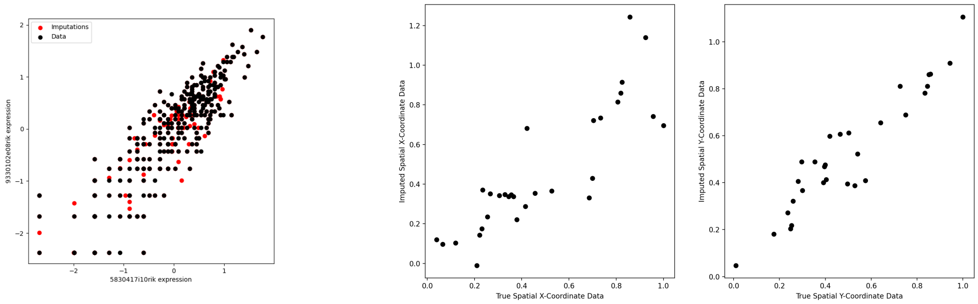

Figure 4: **Imputation of gene expression and spatial coordinates in seqFISH+ data.** *Left:* Black points show observed gene expression values for two genes, and red points show ssGPLVM-imputed gene expression for points whose values were randomly dropped. *Middle:* True spatial coordinate (x-axis) and imputed spatial coordinate (y-axis) for the x spatial coordinate. *Right:* True spatial coordinate (x-axis) and imputed spatial coordinate (y-axis) for the y spatial coordinate.

4 Discussion

We propose a Bayesian statistical model, which we call the semi-supervised Gaussian process latent variable model (ssGPLVM) that captures low-dimensional structure of gene expression data while accounting for the spatial variability. We demonstrate that the ssGPLVM can be used across modalities by jointly fitting spatial spatially-resolved and dissociated gene expression data to learn the low-dimensional structure of expression, as well as impute partially missing gene expression and spatial coordinates from sequencing data. With this knowledge, we can uncover genetic patterns and patterns across space that may not be accessible in one of the modalities separately. As single cell technologies evolve, the ability to integrate data sets will become more vital part of any analysis pipeline. The ssGPLVM offers a principled method for integrating single-cell data and potentially other types of multi-modal molecular data. Future work includes fitting the ssGPLVM on other existing data sets, such as from the Visium platform, and comparing imputation results with related methods (such as PCA).

References

Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82, 2016.

Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature biotechnology*, 39(3):313–319, 2021.

Youjin Lee, Derek Bogdanoff, Yutong Wang, George C Hartoularos, Jonathan M Woo, Cody T Mowery, Hunter M Nisonoff, David S Lee, Yang Sun, James Lee, et al. XYZeq: Spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. *Science advances*, 7(17), 2021.

Tongtong Zhao, Zachary D Chiang, Julia W Morriss, Lindsay M LaFave, Evan M Murray, Isabella Del Priore, Kevin Meli, Caleb A Lareau, Naeem M Nadaf, Jilong Li, et al. Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature*, pages 1–7, 2021.

Romain Lopez, Achille Nazaret, Maxime Langevin, Jules Samaran, Jeffrey Regier, Michael I Jordan, and Nir Yosef. A joint model of unpaired data from scrna-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv preprint arXiv:1905.02269*, 2019.

Junjie Zhu and Chiara Sabatti. Integrative spatial single-cell analysis with graph-based feature learning. *Biorxiv*, 2020.

Archit Verma and Barbara E Engelhardt. A Bayesian nonparametric semi-supervised model for integration of multiple single-cell experiments. *bioRxiv*, 2020a.

Neil Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16, 2003.

Michalis Titsias and Neil D Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 844–851. JMLR Workshop and Conference Proceedings, 2010.

Archit Verma and Barbara E Engelhardt. A robust nonlinear low-dimensional manifold for single cell rna-seq data. *BMC bioinformatics*, 21(1):1–15, 2020b.

10x Genomics. *Mouse Brain Serial Sections (Sagittal-Posterior)*, spatial gene expression dataset by space ranger 1.1.0, 10x genomics, (2020, june 23)., 2020.

Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751):235–239, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Data

A.1.1 seqFISH+

The seqFISH+ dataset we used was collected from the mouse visual cortex. It measures the gene expression levels for a set of *spatially-resolved* cells (the "spatial" data set). The full spatial data matrix is dimensions (913 cells, 9541 genes) and also includes spatial data for each cell. Before imputing, we manipulate the data to be easier to work with, namely log transforming the gene expression data and normalizing the spatial data. We use the manipulated data, and refer to it as the seqFISH+ data set.

A.1.2 Visium

This data set includes data on mouse brain slices generated using Visium v1 chemistry. Tissue sections of 10 µm thickness from a sagittal slice of the posterior were placed on Visium Gene Expression Slides. Under the tissues, 3,355 spots were detected with approximately 4,772 genes per spot. Visium spatial gene expression maintains spatial information, but the resolution of each spot can cover is limited as each spot can cover multiple cells. We refer to this set as the Visium data set in this paper as we run our model on it.

A.2 Methods

Variational inference for the tGPLVM

The joint probability model for the tGPLVM is given by

$$\begin{aligned} p(\mathbf{Y}, \mathbf{Z}, \mathbf{F}, \theta, \sigma^2) &= p(\mathbf{Y}|\mathbf{F}, \sigma^2) p(\mathbf{F}|\mathbf{Z}, \theta) p(\mathbf{Z}) \\ &= \underbrace{\left[\prod_{j=1}^{D} \prod_{i=1}^{n} \mathcal{T}(y_{ij} | f_j(\mathbf{x}_i), \sigma^2, \nu) \right]}_{\text{Noise model}} \underbrace{\left[\prod_{j=1}^{D} N(\mathbf{f}_j | \mathbf{0}_n, \mathbf{K}_{ff}) \right]}_{\text{GP prior}} \underbrace{\left[\prod_{i=1}^{n} N(\mathbf{x}_i | \mathbf{0}_d, \mathbf{I}_d) \right]}_{\text{Latent prior}}. \end{aligned}$$

The posterior of interest is

$$p(\mathbf{F}, \mathbf{Z}|\mathbf{Y}) = p(\mathbf{F}|\mathbf{Z}, \mathbf{Y}) p(\mathbf{Z}|\mathbf{Y}). \tag{2}$$

We introduce a set of m auxiliary variables (inducing points) to facilitate variational inference. Specifically, let $\mathbf{B} \in \mathbb{R}^{m \times d'}$ be a matrix of pseudo-inputs (inducing locations), and let $\mathbf{U} \in \mathbb{R}^{m \times p}$ be a corresponding set of pseudo-outputs. We attempt to configure these variables such that $\mathbf{Y}$ and $\mathbf{F}$ become conditionally independent given $\mathbf{U}$ in our posterior (Equation 2):

$$p(\mathbf{F}, \mathbf{Z}, \mathbf{U}|\mathbf{Y}) = p(\mathbf{F}|\mathbf{Z}, \mathbf{U}) p(\mathbf{U}|\mathbf{Y}) p(\mathbf{Z}|\mathbf{Y}). \tag{3}$$

We approximate this posterior with a variational distribution of the form

$$q(\mathbf{F}, \mathbf{Z}, \mathbf{U}) = p(\mathbf{F}|\mathbf{Z}, \mathbf{U}) q(\mathbf{U}) q(\mathbf{X}).$$

We let $q(\mathbf{U})$ and $q(\mathbf{Z})$ be free-form multivariate Gaussians,

$$q(\mathbf{U}) = N(\mathbf{m}_U, \mathbf{S}_U), q(\mathbf{Z}) = N(\mathbf{m}_Z, \mathbf{S}_Z).$$

Our variational parameters are

$$\phi = \{\mathbf{B}, \mathbf{m}_U, \mathbf{S}_U, \mathbf{m}_Z, \mathbf{S}_Z\}.$$

To optimize this approximation, we maximize a lower bound on the log marginal likelihood (the evidence lower bound, or ELBO) with respect to the variational parameters:

$$\log p(\mathbf{Y}) \geq \mathbb{E}_q \left[\log \frac{p(\mathbf{Y}, \mathbf{Z}, \mathbf{F}, \mathbf{U})}{q(\mathbf{F}, \mathbf{Z}, \mathbf{U})} \right].$$

We can simplify this expression as follows.

$$\begin{aligned} \mathbb{E}_q \left[\log \frac{p(\mathbf{Y}, \mathbf{Z}, \mathbf{F}, \mathbf{U})}{q(\mathbf{F}, \mathbf{Z}, \mathbf{U})} \right] &= \mathbb{E}_q \left[\log \frac{p(\mathbf{Y}|\mathbf{F}) p(\mathbf{F}|\mathbf{Z}, \mathbf{U}) p(\mathbf{U}) p(\mathbf{Z}))}{p(\mathbf{F}|\mathbf{Z}, \mathbf{U}) q(\mathbf{U}) q(\mathbf{Z})} \right] \\ &= \mathbb{E}_q \left[\log \frac{p(\mathbf{Y}|\mathbf{F}) p(\mathbf{U}) p(\mathbf{Z}))}{q(\mathbf{U}) q(\mathbf{Z})} \right] \\ &= \mathbb{E}_q \left[\log p(\mathbf{Y}|\mathbf{F}) \right] - D_{KL}(q(\mathbf{U}) \| p(\mathbf{U})) - D_{KL}(q(\mathbf{Z}) \| p(\mathbf{Z})). \end{aligned}$$

The final two terms (the KL divergences) can be computed in closed-form given our Gaussian assumptions. However, the first term cannot be computed analytically. Instead, we estimate this term with a Monte Carlo approximation,

$$\mathbb{E}_q \left[\log p(\mathbf{Y}|\mathbf{F}) \right] \approx \frac{1}{L} \sum_{\ell=1}^{L} \log p(\mathbf{Y}|\widehat{\mathbf{F}}_\ell)$$

where we draw $\widehat{\mathbf{F}}_\ell$ from $p(\mathbf{F}|\mathbf{Z}, \mathbf{B})$, where $\mathbf{U}$ has been marginalized out. This distribution is given by

$$p(\mathbf{f}|\mathbf{Z}, \mathbf{B}) = N(\widetilde{\mu}, \widetilde{\Sigma})$$

where

$$\begin{aligned}\widetilde{\mu} &= \mu(\mathbf{Z}) + \mathbf{K}_{uf}^\top \mathbf{K}_{uu}^{-1}(\mathbf{m} - \mu(\mathbf{Z}))\\ \widetilde{\Sigma} &= \mathbf{K}_{ff} - \mathbf{K}_{uf}^\top \mathbf{K}_{uu}^{-1}(\mathbf{K}_{uu} - \Omega)\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}.\end{aligned}$$

This ultimately becomes a two-step sampling procedure. First, we sample $\widehat{\mathbf{Z}} \sim q(\mathbf{Z})$, and then we sample $\widehat{\mathbf{F}} \sim p(\mathbf{F}|\widehat{\mathbf{Z}}, \mathbf{B})$.

Variational inference for the ssGPLVM

In order to perform inference in the semisupervised version of the model, we make one minor change: we append the fixed spatial dimensions to both the latent variables $\mathbf{X}$ and the inducing locations $\mathbf{Z}$. Denote the matrix of spatial coordinates as $\mathbf{S} \in \mathbb{R}^{n\times 2}$, and let the augmented latent variables be given by

$$\mathbf{T} = [\mathbf{X} \quad \mathbf{Z}] \in \mathbb{R}^{n\times(d+d')}.$$

To incorporate the spatial coordinates into the inducing locations $\mathbf{Z}$, we take one of two approaches. First, similar to above, we can append a subset of the spatial locations to the inducing locations such that

$$\widetilde{\mathbf{B}} = \left[\widetilde{\mathbf{X}} \quad \mathbf{B}\right] \in \mathbb{R}^{m\times(d+2)}$$

where $\widetilde{\mathbf{X}} \in \mathbb{R}^{m\times 2}$ is a subset of the spatial coordinates. This subset can be chosen by, for example, k-means clustering of the spatial coordinates (where $k = m$). As a second approach, we can allow $\widetilde{\mathbf{S}}$ to be an extra set of variational parameters, and optimize them as part of our approximate inference strategy.

A.3 Supplementary figures

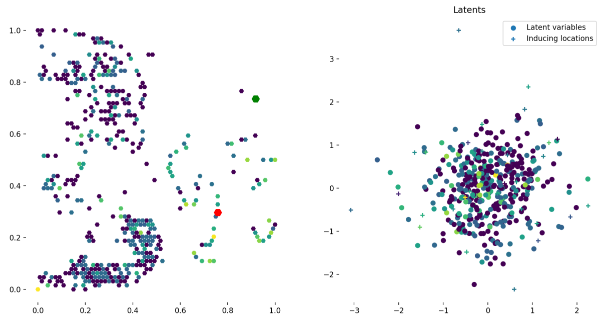

Figure 5: **Visualizing imputing spatial location in Visium data set**: If we look at the left graph, we can see the faint outline of the tissue. Each dot on the graph represents a cell and it is colored by the molecular counts. For this particular graph, we only took the 15000 cells with the most variance (highest molecular count) and dropped one of those cells to impute. The true location is in red, and the green dot is our imputed guess. We can see that as the green dot approaches the red dot, our imputation gets better. The right graph shows the latent variables.

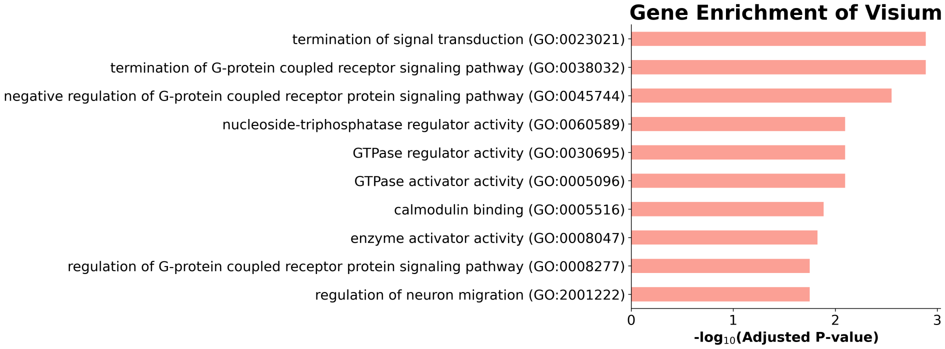

Figure 6: **Gene set enrichment analysis on the ssGPLVM latent variables obtained from the Visium data**