Supplementary Materials for Gene Co-expression Network Estimation for Spatial Transcriptomics

Satwik Acharyya, Xiang Zhou, Veera Baladandayuthapani

2022-05-19

Contents

Introduction

The spatial transcriptomics method depicts the positioning of a single cell on a spatially structured tissue. Knowledge about gene expressions and the spatial distribution of mRNA allows us to uncover cellular and subcellular heterogeneity in tissues, tumors, and immune cells. Spatial transcriptomics provides a unique opportunity to decipher both the cellular and subcellular architecture in both tissues and individual cells along with detection of gene co-expression patterns at both levels. These approaches are very insightful to study disease propagation in the field of embryology, oncology, and histology. The SpaceX method is a statistical tool to quantify spatially varying gene co-expression patterns in a tissue consists of different cell type based or sptailly contiguous clusters.

This is a supplementary file of the paper named SpaceX: Gene Co-expression Network Estimation in Spatial Transcriptomics. The sectional contents of the supplementary file is mentioned below.

1. We start with a detailed description of the methodology in section A.
2. In section B, further details of simulation study have been discussed.
3. Eploratory analysis and more findings of real data analysis have been laid out in section C.
4. Finally, we discuss the detailed steps for implementation of the SpaceX package in section D.

Appendix A

Methodology

In this section, we start with a brief discussion of the model and interpretations of each term of the model. Next, A detailed discussion of the estimation procedure for the SpaceX model in equation 1 in section 2 of the paper (also mentioned in equation (A.2)) is provided. Finally we discuss the methodological novelty of our SpaceX method.

A.1 SpaceX model

The gene expression is modeled with a Poisson distribution as

$$y_g^c(\mathbf{s}_i) \sim \text{Poi}\{M^c(\mathbf{s}_i)\lambda_g^c(\mathbf{s}_i)\}. \tag{A.1}$$

The interpretations of the notations in equation (A.1) remains same as mentioned in the section 2.2 of the paper. We denote Λ^c as a $G \times N_c$ (N_c denotes size of the c-th cluster) matrix containing the rate parameters for all genes and c-th cluster. Subsequently, we model the cluster specific and spatially dependent rate parameter $^{\mathbf{c}}$ with an additive log-linear equation, i.e.

$$\log(\ ^c) = \mathbf{B}^c\mathbf{X}^c + \mathbf{S}^c + \ \mathbf{F} + \ ^{\mathbf{c}}\mathbf{D}^{\mathbf{c}} + \mathcal{E}^c. \tag{A.2}$$

We clearly mention the shared and cluster-specific parameters and their interpretations table A.1.

Parameters

Shared

Cluster-specific

Interpretations

$\mathbf{B}^c\mathbf{X}^c$

×

✓

Covariate effect

$\mathbf{S}^c$

×

✓

Spatial information

$\mathbf{F}$

✓

×

Shared loadings and factors

$^{\mathbf{c}}\mathbf{D}^{\mathbf{c}}$

×

✓

Cluster loadings and factors

$\mathcal{E}^c$

×

✓

Error matrix

Shared and cluster-specific parameters along with their corresponding interpretations.

A full-scale MCMC will be computationally expensive on a complex hierarchical model. For computational advantage, we decompose the model into two parts (I) sPMM: spatial Poisson mixed model (Sun et al., 2018) and (II) MSFA: Multi-study factor analysis model (De Vito et al., 2021). We enable this model decomposition through a standard Gaussian random variable.

A.2 Poisson Mixed Model

We can break the SpaceX model (A.2) and write the spatial Poisson mixed model as

$$
\begin{aligned}
&\log(\lambda_g^c) = \mathbf{X}^{c^T}\beta_g^c + \mathbf{s}^c + \mathbf{z}_g^c, \\
&\lambda_g^c = (\lambda_{1g}^c, \dots, \lambda_{N_c g}^c)^T, \\
&\mathbf{s}^c = (s_1^c, \dots, s_{N_c}^c)^T \sim \text{MVN}(0, \sigma_1^2 \Omega^c(s)), \\
&\mathbf{z}_g^c = (z_{1g}^1, \dots, z_{N_c g}^c)^T \sim \text{MVN}(0, \sigma_2^2 I_{N_c \times N_c}).
\end{aligned} \tag{A.3}
$$

Here $\Omega^c(s_1, s_2) = \exp(- \mid\mid s_1 - s_2 \mid\mid^2 / 2\rho_c^2)$, $c = 1, \dots, C$. We estimate the length scale parameter of spatial kernel ρ_c based on the steps discussed in section 1 of supplementary information in Sun et al. (2020a). Here Z_g^c captures the cluster specific latent gene expressions and a multi-variate hierarchical modeling of $Z_g^c(s_i)$ will help us to identify the gene co-expression network.

A.3 Multi-Study Factor Model (MSFA)

The 2nd stage of the modeling framework is multi-study factor analysis (De Vito et al., 2021) which is provided as follows

$$
\begin{aligned}
&\hat{\mathbf{z}}_i^c = \Phi \mathbf{f}_i + \Psi^c \mathbf{d}_i^c + \mathbf{e}_i^c, \\
&\mathbf{f}_i \sim N_K(0, \mathbf{I}_K), \qquad \mathbf{d}_i^c \sim N_{K_c}(0, \mathbf{I}_{K_c}), \\
&\mathbf{e}_i^c \sim N_G(0, \Xi_c), \qquad \Xi_c = \text{diag}(\xi_1^c, \dots, \xi_G^c).
\end{aligned} \tag{A.4}
$$

The marginal distribution of $\hat{\mathbf{z}}_i^c$ is a multivariate normal distribution with mean 0 and covariance matrix Σ_c s.t.

$$
\Sigma_c = \Phi\Phi^T + \Psi^c \Psi^{c^T} + \Xi_c = \Sigma_\Phi + \Sigma_{\Psi^c} + \Xi_c \tag{A.5}
$$

$\Sigma_\Phi = \Phi\Phi^T$ and $\Sigma_{\Psi^c} = \Psi^c \Psi^{c^T}$ are covariance of shared and cluster specific factors respectively. The decomposition of Σ_c in (A.5) is not a unique since we can set $\Phi^* = \Phi Q$ and $\Psi^{*c} = \Psi^c Q_c$ where Q and Q_c are square orthonormal matrices. This will also lead to the same decomposition $\Sigma^c = \Phi^* \Phi^{*^T} + \Psi^{*c} \Psi^{*^{c^T}} = \Phi\Phi^T + \Psi_c \Psi_c^T$. To overcome the indeterminacy through orthonormal matrices, the factor loading matrices are restricted to be lower triangular matrices (Geweke and Zhou, 1996; Lopes and West, 2004).

A.4 Multiplicative gamma shrinkage prior

We follow the same steps from De Vito et al. (2021) and place multiplicative gamma shrinkage prior (Bhattacharya and Dunson, 2011) prior on the shared and cluster specific loading matrices i.e. Φ and Ψ_c $c = 1, \dots, C$. The shared and cluster specific latent factors (K and K_c respectively) are selected following methodology described in section 3.3 of De Vito et al. (2021). The multiplicative

gamma prior on elements of shared covariance matrices are provided as follows

$$\begin{aligned}
&\phi_{gk} \mid \delta_{gk}, \eta_k \sim N(0, \delta_{gk}^{-1}\eta_k^{-1}), \quad g = 1, \dots, G, \quad k = 1, \dots, \infty, \\
&\delta_{gk} \sim \Gamma(\frac{\nu}{2}, \frac{\nu}{2}) \quad \eta_k = \prod_{j=1}^{k} \zeta_j \quad \zeta_1 \sim \Gamma(a_1, 1) \quad \zeta_j \sim \Gamma(a_2, 1), \quad j \geq 2.
\end{aligned} \tag{A.6}$$

Here δ_{gk} is the local shrinkage parameter for G column elements of kth column and η_k is the global shrinkage parameter where ζ_j $(j = 1, 2. \dots)$ are independent. We repeat the same process to posit prior on the elements of cluster-specific loading matrices

$$\begin{aligned}
&\psi^c_{gk_c} \mid \delta^c_{gk_c}, \eta^c_{k_c} \sim N(0, \delta^{c^{-1}}_{gk_c}\eta^{c^{-1}}_{k_c}), \quad g = 1, \dots, G, \quad k_c = 1, \dots, \infty \text{ and } c = 1, \dots, C, \\
&\delta^c_{gk_c} \sim \Gamma(\frac{\nu^c}{2}, \frac{\nu^c}{2}) \quad \eta^c_{k_c} = \prod_{j=1}^{k_c} \zeta^c_j \quad \zeta^c_1 \sim \Gamma(a^c_1, 1) \quad \zeta^c_j \sim \Gamma(a^c_2, 1), \quad j \geq 2.
\end{aligned} \tag{A.7}$$

Here $\delta^c_{gk_c}$ $\eta^c_{k_c}$ are local and global parameters respectively and ζ^c_j $(c = 1, 2, \dots C)$ are independent of each other. We determine K and K_c following methodology described in section 3.3 of De Vito et al. (2021).

A.5 Novelty in Methodology and Estimation Procedure

The SpaceX model (A.2) incorporates the spatial information in Poisson likelihood and builds on a factor model based components to estimate the gene co-expression networks. The PQLseq algorithm (Sun et al., 2018) is based on a Poisson likelihood but it does not incorporate the spatial information. The hierarchical factor analysis model (MSFA, De Vito et al. (2021)) model considers a Gaussian likelihood and does not take into account the spatial information. The SpaceX model incorporates the discrete nature of the single cell sequencing data and builds on a Poisson likelihood. The model accounts for the spatial effect whereas other two models do not. One can infer the gene co-expression networks while considering the spatial information but MSFA fails to incorporate the spatial locations. The detailed comparison between 3 models is summarized in Table A.5.

An MCMC based algorithm for the SpaceX model will be computationally inefficient. We develop a tractable and computationally scalable Bayesian algorithm for the estimation procedure of the novel joint modeling framework. We decompose the whole model into two essential components (I) spatial Poisson model and (II) hierarchical factor analysis model.

SpaceX model

PQLseq

MSFA

Spatial information

✓

X

X

Poisson likelihood

✓

✓

X

Network inference

✓

X

✓

Gaussian likelihood

X

X

✓

Methodological comparison between SpaceX, PQLseq, MSFA models.

A.5.1 Differences from individual methods (PQLseq, SPARK and MSFA)

For spatial Poisson model, we use the scalable penalized quasi-likelihood algorithm named PQLseq. The algorithm is based on the generalized linear mixed model framework which accounts for the count nature of the single cell sequencing data. The algorithm can be used for detection of deferentially expressed genes. The PQLseq algorithm does not take into account the spatial information and the SPARK (Sun et al., 2020a) method was developed to address this limitation. Although, the SPARK method only estimates parameters while setting the spatial variance parameter to zero and develops a score test based procedure to identify spatially expressed genes. We build on the PQLseq algorithm for our estimation procedure of spatial Poisson mixed model such that the algorithm can accommodate the spatial information and estimates the parameters without setting the spatial variance parameter to zero. We use the modified version of the PQLseq algorithm to obtain the latent gene expressions and carry it forward in our next framework.

We use the multi-study factor analysis (MSFA) model (De Vito et al., 2021) on the latent gene expression matrix to estimate shared and cluster specific gene co-expressions. The MSFA model can not adapt for the count nature of the sequencing data and corresponding spatial locations. Our joint modeling framework in the SpaceX model addresses both concerns.

A.6 Modifiend version of the model for cell-type based clusters

We aimed to modify our model to explicitly accommodate the spatial correlation between cell-type based clusters. To do so, we follow the standard notations as discussed throughout the paer and Supplementary Materials. Here, $y_g^c(s_i)$ represents the gene expression for g-th gene for c-th cluster w.r.t. s_i spatial locations. The gene expression is being modeled with a Poisson distribution as where M is the normalizing constant and λ is the rate parameter. The spatially dependent rate parameter λ is modeled with a log-linear equation

$$\log(\ ^c) = \mathbf{B}^c\mathbf{X}^c + \mathbf{S}^c + \ \mathbf{F} + \ ^{\mathbf{c}}\ ^{\mathbf{c}}\mathbf{D}^{\mathbf{c}} + \mathcal{E}^c. \tag{A.8}$$

Explanation of all the parameters except for the new parameter ($^{\mathbf{c}}$) remains the same. In this model, $^{\mathbf{c}}$ now models the cell-type specific spatial process based on spatial distance between cell types. We adjust for spatial correlations for nearby cell-types through the term $\mathbf{S}^{\mathbf{C}}$ and cross cell-type specific adjustment is factored in through $^{\mathbf{C}}$. While the extended model looks appealing, unfortunately, we realized that there are several levels of difficulties that need careful thinking before being fully incorporated in the model. For example, we are working with a Poisson likelihood which is non-Gaussian and spatial information needs to be included in the model. Identifiability issue is the main challenge for the model in (A.8). Based on factor models literature, the loading matrices are usually restricted to be a lower triangular matrix for identifiability reasons in case of a Gaussian likelihood (Geweke and Zhou, 1996; Lopes and West, 2004). We need non-standard identifiable restrictions to estimate $^{\mathbf{c}}$ and $^{\mathbf{c}}$ since they are incorporated in the model in multiplicative manner. Given these methodological underpinnings that require more rigorous theoretical work to overcome such difficulties, we leave this non-trivial methodological estimation algorithm as future work.

Appendix B

Simulation Study

B.1 Induced Correlation Study

In this section, we provide more details about the simulation study. First we consider 3 different values of ρ $(0.1, 0.15, 0.2)$ and make a induced correlation plot by using the squared exponential spatial kernel. The plots are generented for all cell types and cell type specific cases. The vertical line denotes the value of induced correlation at the distance 0.01. For example the induced spatial correlations for all cell types (first Figure of B.1) w.r.t. 0.01 distance are 0.88, 0.80, 0.61 for $\mathrm{S}_{\mathrm{H}igh}$ $(\rho = 0.2)$, $\mathrm{S}_{\mathrm{M}ed}$ $(\rho = 0.15)$, $\mathrm{S}_{\mathrm{L}ow}$ $(\rho = 0.1)$ methods respectively.

Estimation of parameter ρ : We adapt the steps from the section 1.1 of supplementary information of Sun et al. (2020a) to estimate the length scale parameter of the spatial kernel ρ_c based on several grid points. The estimation procedure of ρ_c uses the pair-wise distances among spatial locations in the data to ensure scale-invariance. First the maximum (m_1) and minimum (m_2) value of those pairwise spatial distances are obtained. Next, equidistant L points are identified in the range from $\log(m_1/2)$ to $\log(m_2 * 2)$. In our case, we consider $L = 10$ as a default value for all our simulations and spatial transcriptomics data applications. We consider the middle value of those 10 grid points as the estimate of ρ_c. Table B.1 shows the true and estimated values of ρ_c which shows the grid points based estimation procedure is effective.

True value of ρ_c

0.2

0.15

0.1

Estimated value of ρ_c

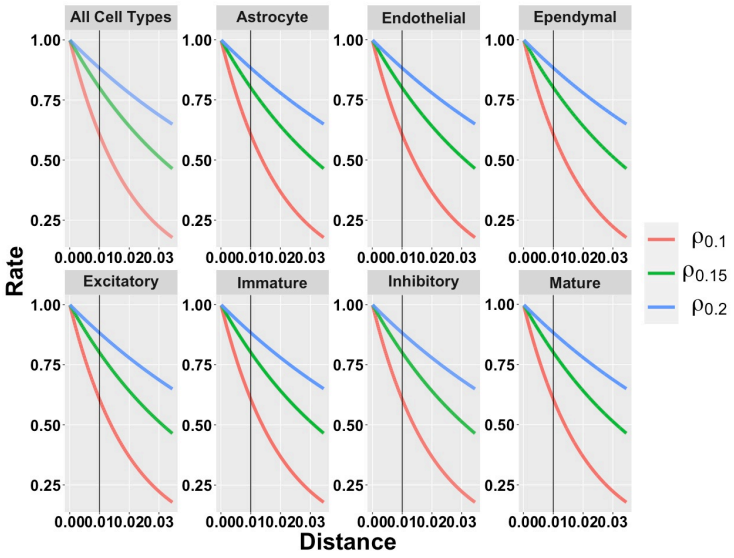

Figure B.1: Induced correlation plot for the Merfish data

0.18

0.16

0.11

Standard Errors

0.11

0.13

0.14

True, estimated values and standard errors of ρ_c.

B.2 Data generation with spatial correlation

Simulation Design: The details of the simulation design is provided in section 3 of the paper.

Comparative Models: We discussed the 5 comparative models in section 3 of the paper. A summary of the comparative models is provided in Table B.2

SpaceX model

Non-spatial Poisson model

Gaussian model

Spatial information

✓

X

X

Poisson likelihood

✓

✓

X

Gaussian likelihood

X

X

✓

Overview of comparative models.

Metrics for comparison: To measure the co-expression estimation accuracy, we use the following metrics in Table B.2 such as RV coefficient (Robert and Escoufier, 1976) and 4 Euclidean distance based norms (**Frobenius, Log-Euclidean, Root Euclidean, Riemanian**, Dryden et al. (2009)) as defined in Table B.2. These metrics are used to quantify the similarity between true and estimated covariance (co-expression) matrices. RV values close to 1 (0) implying higher (lower) level of similarity. For rest of the norms on Table B.2, values closer to 0 indicates higher level of similarity between true and estimated matrices.

Name

Notation

Form

RV

$$\mathrm{RV}(S_1, S_2)$$

$$\frac{tr(S_1^T S_2)}{\sqrt{tr(S_1^T S_1) tr(S_2^T S_2)}}$$

Euclidean (Frobenius)

$$d_E(S_1, S_2)$$

$$|| S_1 - S_2 ||$$

Log-Euclidean

$$d_L(S_1, S_2)$$

$$|| \log(S_1) - \log(S_2) ||$$

Root-Euclidean

$$d_H(S_1, S_2)$$

$$|| S_1^{1/2} - S_2^{1/2} ||$$

Riemanian

$$d_R(S_1, S_2)$$

$$|| S_1^{-1/2} S_2 S_1^{-1/2} ||$$

Definition of different norms. Here $|| X ||= \sqrt{\text{trace}(X^T X)}$.

B.2.1 Comparative analysis with different norm measures

Figure B.2 represents the boxplot of distances between true (Σ_{True}) and estimated (Σ_{Est}) covariance matrices where the distances are measured in **Euclidean, root Euclidean, log Euclidean and Riemanian** norms respectively. In all the norms we observe that spatial settings are performing better in terms of estimation than the no-spatial settings. Among the spatial settings the estimation accuracy increase with an increment in induced spatial correlation.

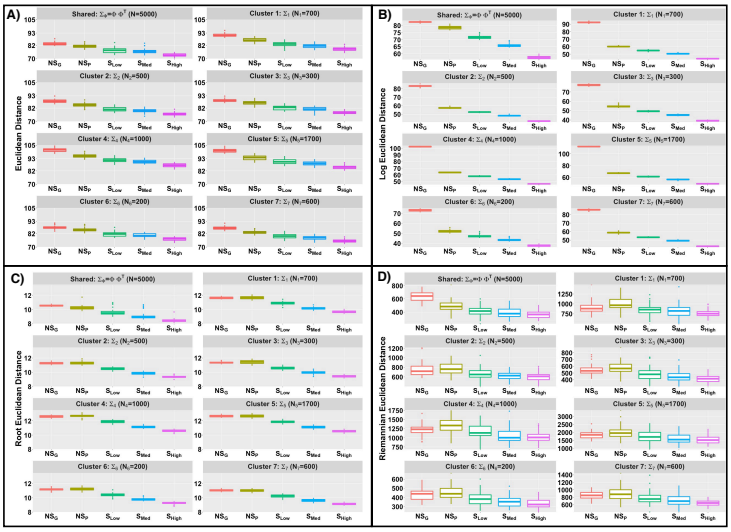

Figure B.2: **Comparison of different methods based on several norms for estimation of gene co-expression in simulation study while the data is generated from a spatially correlated setting**. Boxplot of Euclidean, log-Euclidean, root-Euclidean, Riemannian distance (Figure A, B, C and D respectively) across 50 replicates for $\Sigma_\Phi = \Phi\Phi^T$ and Σ_l $(l = 1, \dots, L)$. We compare the norm distances for different settings for data generation with spatial correlation.

B.2.2 Estimation of latent factors

We follow same procedure from section 3.3 of De Vito et al. (2021) to estimate shared and cluster specific number of factors i.e. K and K_c $(c = 1, 2, \dots C)$. Figure B.3 shows shared and cluster specfic estimated factor loadings accross 50 replicates for 5 different methods. Figure B.4 shows the median estimate of shared and cluster specfic factor loadings for 5 different methods. From both figures one can observe that spatial settings are estimating the loadings more precisely than the non-spatial settings.

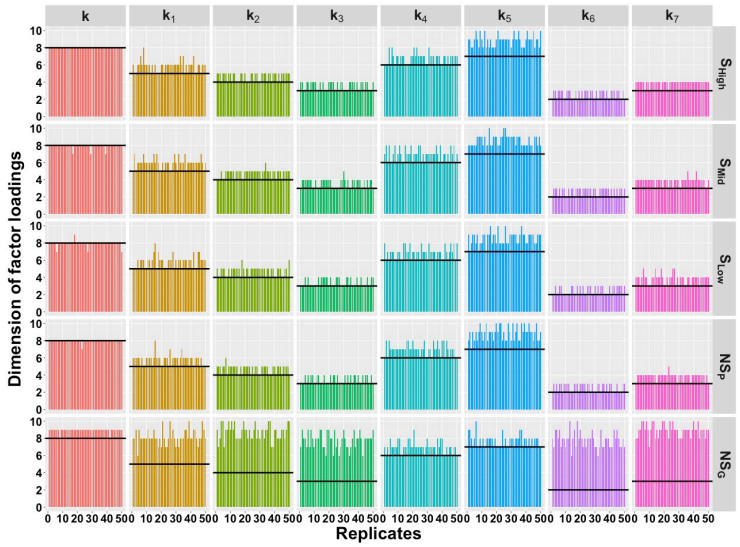

Figure B.3: Estimated dimension of factor loadings for shared and cluster specific cases accross 50 replicates. Black solid line denotes the true dimensions.

B.3 Data generation with no spatial correlation

Simulation design: We consider the same simulation design, comparative models and metrics for comparison. Now, we generate the from NS_{P} model where we do not consider any spatial correlation. We model the spatial parameter with a multi-variate normal distribution with mean 0 and identity as covariance matrix. We generate the data with no spatial correlation model NS_{P} and fit all the comparative models as mentioned in Table B.2. We summarize our results based on 50 replicated simulation study w.r.t co-expression estimation and network recovery.

Co-expression estimation: We display the boxplot of RV coefficients for shared ($\mathbf{G_s}$) and cluster-specific ($\mathbf{G_c}$, $c = 1, \dots, C$) covariance matrices in Figure B.5A across 3 comparative models in Table B.2. Based on RV coefficients, the highest level of precision in estimation is obtained for the non-spatial setting

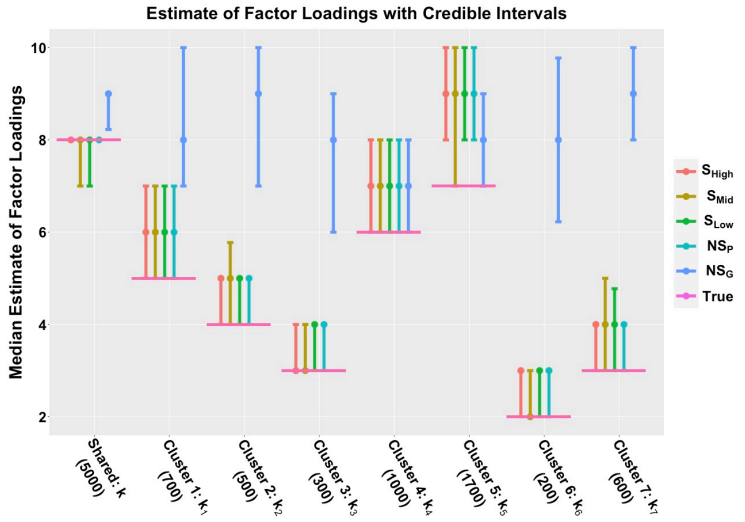

Figure B.4: Estimated Factor loadings with credible intervals.

in (IV) i.e. NS_{P}. For spatial settings, we do not observe a significant loss in estimation accuracy. For example, we only loose 1.3% accuracy in estimation for spatial settings with the SpaceX method in case of the shared network. A similar inference can be drawn based on different norm measures (Euclidean, log-Euclidean, root-Euclidean, Riemannian as defined in Table B.2 as provided in Figure B.6 (in clock-wise manner).

Network recovery: AUC based comparisons for shared and cluster specific networks are shown in Figure B.5B. The Figure leads us to infer that no significant reduction in network recovery while comparing the spatial and non-spatial settings while data is generated without spatial correlation.

In summary, we do not observe a significant loss in precision while applying the SpaceX model to the simulated data without any spatial correlation.

B.4 Hub gene detection based simulation

Given the lack of ground truth for real data, we used a simulation study mimicking the real data structure, to evaluate the accuracy of the SpaceX method to recover true hub genes. To this end, we generate the data consisting of 160 genes and 5000 locations with 7 clusters; additional design parameters and details about the simulation settings are provided in Section 3 of the paper. Next, we apply our SpaceX model to the simulated data to estimate the hub genes for each of the 7 spatial clusters. We classify the hub gene into two categories: one with more than 40% connectivity (named “High”) and another with less

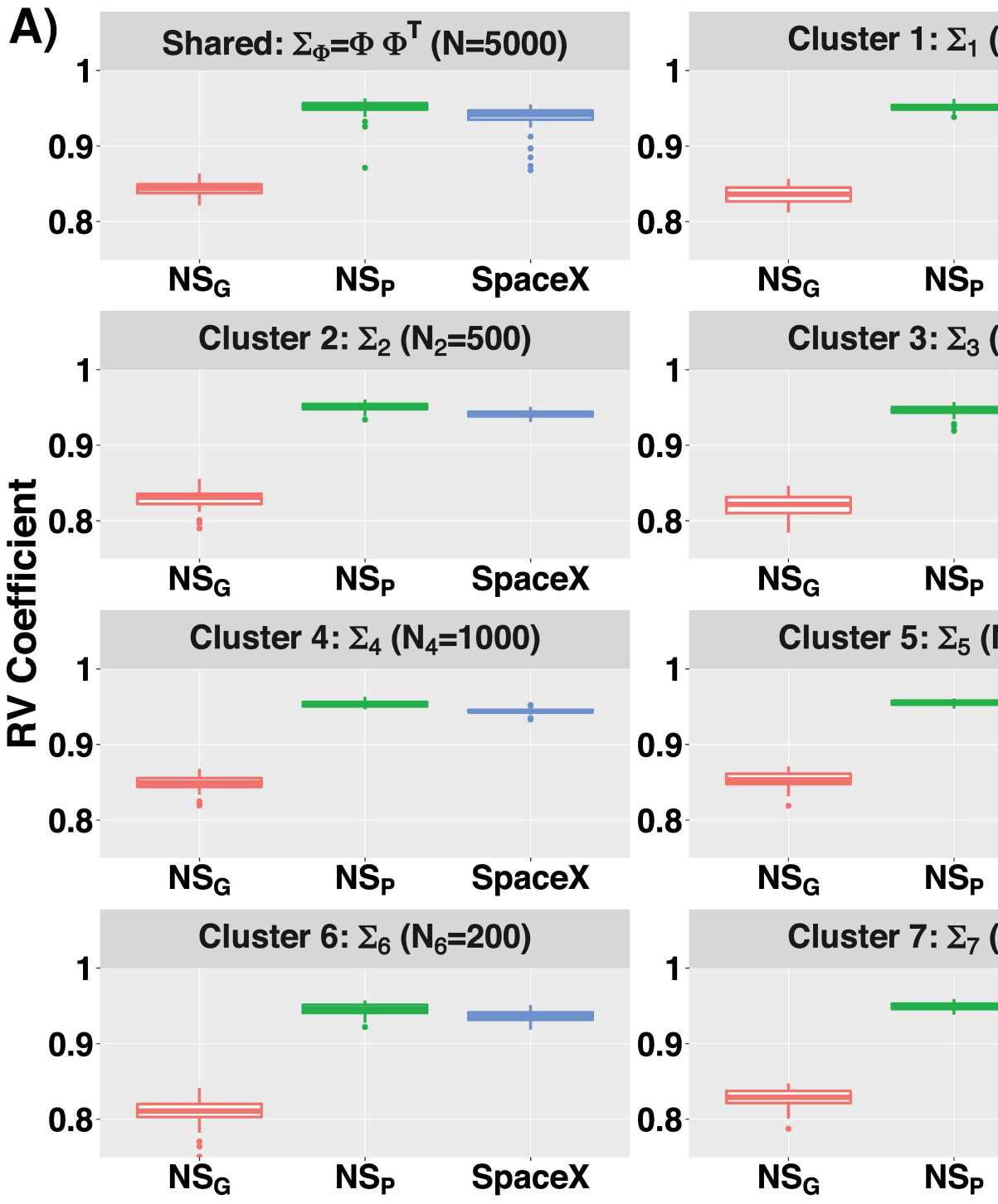

Figure B.5: **Accuracy comparison of different methodological settings in estimation and recovery of gene co-expression networks in simulation study while the data is generated without any spatial correlation. A)** The RV coefficient measures the distance between the true and estimated networks. In the left panel, we have boxplot of RV coefficients across 50 replicates for shared and cluster-specific networks. We compare the RV coefficients for 3 different methods (I) SpaceX, (IV) NS_{P} ($\rho = 0$) and (V) NS_{G} (the PMM and spatial informations are not taken under consideration). **B) In the right panel**, we use AUC metric as a measure of network recovery. The Figure represents ROC curves

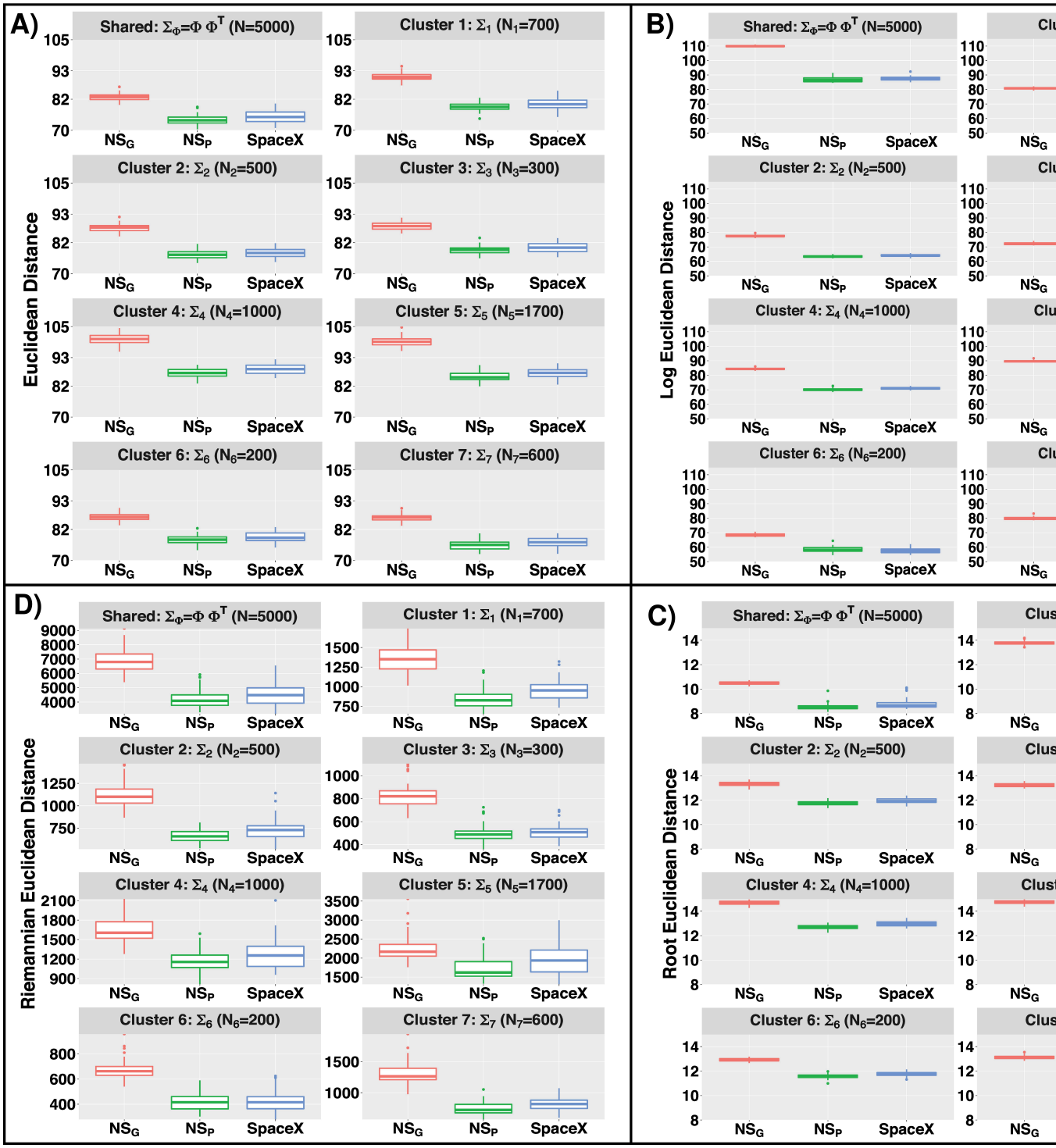

Figure B.6: **Comparison of different methods based on several norms for estimation of gene co-expression in simulation study while the data is generated without any spatial correlation**. Boxplot of Euclidean, log-Euclidean, root-Euclidean, Riemannian distances (Figure A, B, C and D respectively) across 50 replicates for $\Sigma_\Phi = \Phi\Phi^T$ and Σ_l $(l = 1, \dots, L)$. We provide the definitions of all the norms in Table
reftab:Normtable and compare the distances for different methodological settings.

than 40% connectivity (named "Low"). We choose this cut-off based on degree connectivity we found in the real data example (Section 4). Table B.4 shows the percentage of hub gene recovered for each cluster by setting 3 different values of the spatial correlation parameter: $\rho = 0.2, 0.15$ and 0.1 corresponding to low, medium and high levels of spatial correlation. The denominator of recovery rate is the total number of hub genes calculated from the true simulation settings whereas the numerator denotes the estimated number of hub genes obtained from the SpaceX method. From Table B.4, we observe that the percentage hub gene recovery increases with an increase as the level of spatial correlations increase ρ and size of the cluster. For example, cluster 5 (size = 1700) has the following recovery rates 97.5% , 95.2%, 92.3% for highly connected hub genes corresponding to different level of spatial correlations i.e. $\rho = 0.2, 0.15$ and 0.1 respectively. The recovery rates are 97.5%, 93.2%, 90.3% for cluster 5, 4 and 1 respectively with different cluster sizes ($N_5 = 1700$, $N_4 = 1000$, $N_1 = 700$). As expected, the recovery rate of the hub genes with higher connectivity is more than in the ones with low connectivity.

	S_{High} ($\rho = 0.2$)		S_{Med} ($\rho = 0.15$)		S_{Low} ($\rho = 0.1$)	
	High	Low	High	Low	High	Low
Cluster 1 ($N_1 = 700$)	90.3	86.3	87.1	84.7	83.5	81.6
Cluster 2 ($N_2 = 500$)	88.6	83.9	85.4	80.2	81.8	78.7
Cluster 3 ($N_3 = 300$)	84.7	77.4	81.3	74.5	79.4	71.9
Cluster 4 ($N_4 = 1000$)	93.2	90.7	91.6	88.9	88.6	86.7
Cluster 5 ($N_5 = 1700$)	97.5	94.1	95.2	93.5	92.3	91.3
Cluster 6 ($N_6 = 200$)	82.0	75.9	80.7	72.6	78.2	70.2
Cluster 7 ($N_7 = 600$)	89.2	85.8	86.7	83.1	84.4	80.5

Table B.4: Hub gene recover percentages for different simulation settings. High and Low denotes hub genes with high and low levels of connectivity.

Appendix C

Real Data Analysis

We applied the SpaceX method on two spatial transcriptomics datasets which are obtained from the preoptic region of the mouse hypothalamus (Moffitt et al., 2018) and the human breast cancer dataset (Ståhl et al., 2016). Here we provide details of preprocessing and exploratory analysis of both datasets in section C.1. We illustrate the detailed application of the community detection algorithm on those two datasets in section C.2.

C.1 Exploratory analysis of the datasets

C.1.1 Merfish Data

The MERFISH dataset is obtained from the preoptic area of the mouse hypothalamus (Moffitt et al., 2018). The dataset consists of 160 genes and corresponding gene expressions are measured in 4975 spatial locations. There are 7 pre-determined spatial clusters in the dataset named Astrocyte, Endothelial, Ependymal, Excitatory, Inhibitory, Immature, Mature, and the corresponding sizes are 724, 503, 314, 1024, 1694, 168, 385 respectively. The dataset consists of 2 more clusters named Microglia, Pericytes with cluster sizes 90, 73 respectively which are less than 100. Those two clusters are removed from the dataset. After removing those two clusters, we have gene expressions from 4812 locations corresponding to 160 genes. There are no genes with more than 95% zeros reads. The left panel of Figure C.1 shows the violin plot of the percentage of zero reads among the genes for each cluster in the MERFISH dataset. The Umap representation of the Merfish data has been provided on the right panel of Figure C.1.

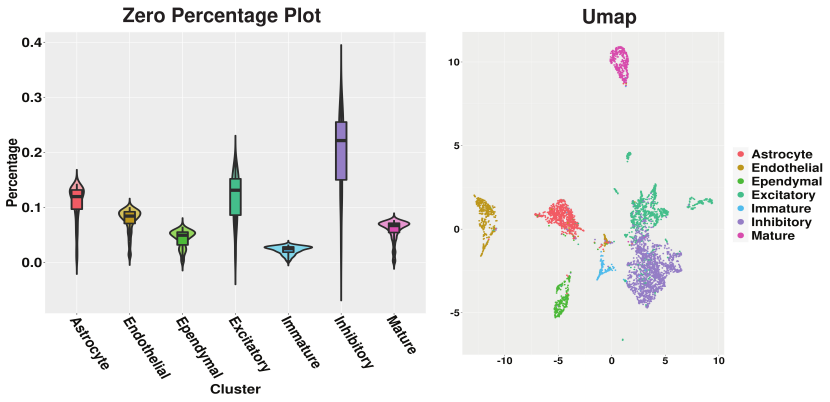

Figure C.1: Left panel shows the violin plot of percentage of zero reads among the genes for each cluster w.r.t. Merfish data and the right panel shows the Umap.

C.1.2 Breast Cancer Data

The human breast cancer dataset contains expression levels from 5262 genes measured at 250 locations (Ståhl et al., 2016). We use the SPARK method with 5% FDR cut-off on p-values to detect 290 spatially expressed genes to carry forward our analysis. The violin plot of the percentage of zero reads among the genes for each spatially contiguous cluster in the Breast cancer dataset is shown in the left panel of Figure C.2. On the right panel of Figure C.2, we have provided the Umap.

C.2 Community detection

The community detection is a downstream analysis of the shared and cluster-specific networks which are obtained from the SpaceX method. The communities are detected by optimizing modularity over partitions in a network structure (Brandes et al., 2007). Figure C.3 and C.4 show the detected community modules from shared and cluster-specific co-expression networks for MERFISH and breast cancer data respectively.

C.3 Benchmarking on real spatial transcriptomics data

In this section, we benchmark our models on two real spatial transcriptomics datasets based on model fitting criteria. To this end, we use information-based criteria – a standard and well-established technique to compare the model fits between hierarchical Bayesian models (Gelman et al., 2014). In this case, we use

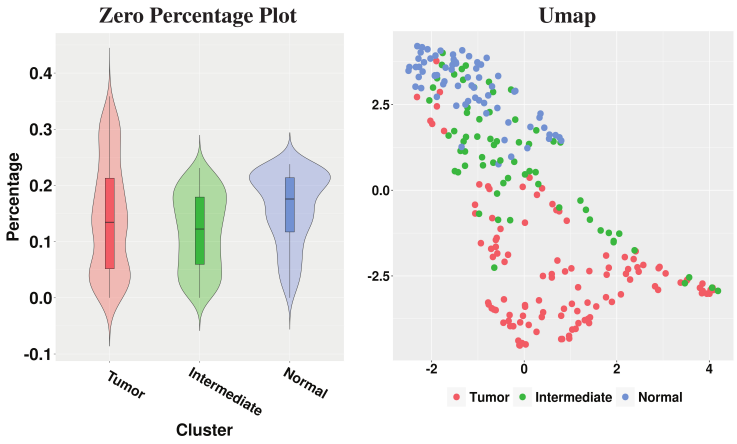

Figure C.2: On the left panel, we have violin plot of percentage of zero reads among the genes for each cluster w.r.t. Breast Cancer data and the Umap is shown on the right panel.

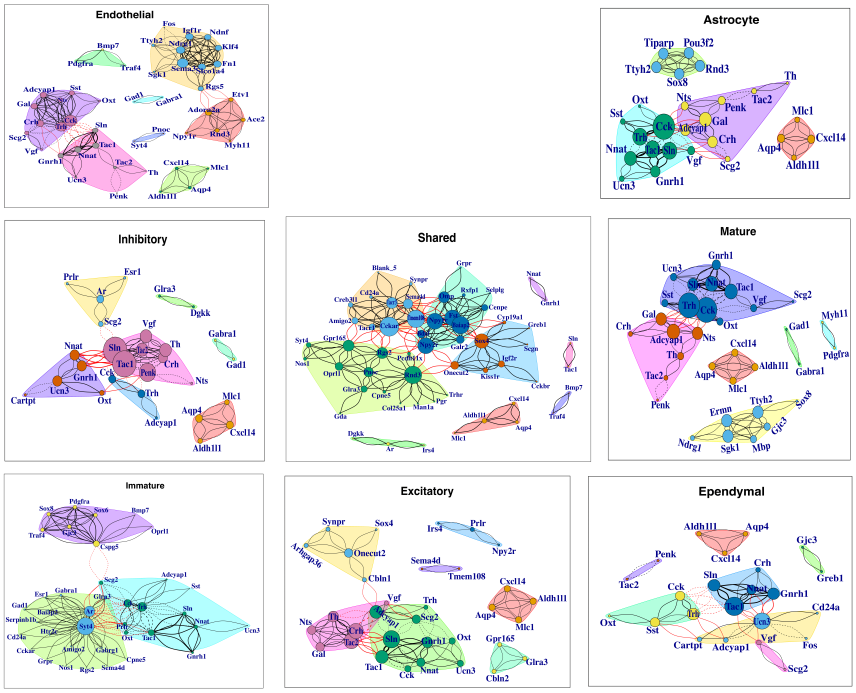

Figure C.3: Shared and cell-type specific community detection for Merfish data

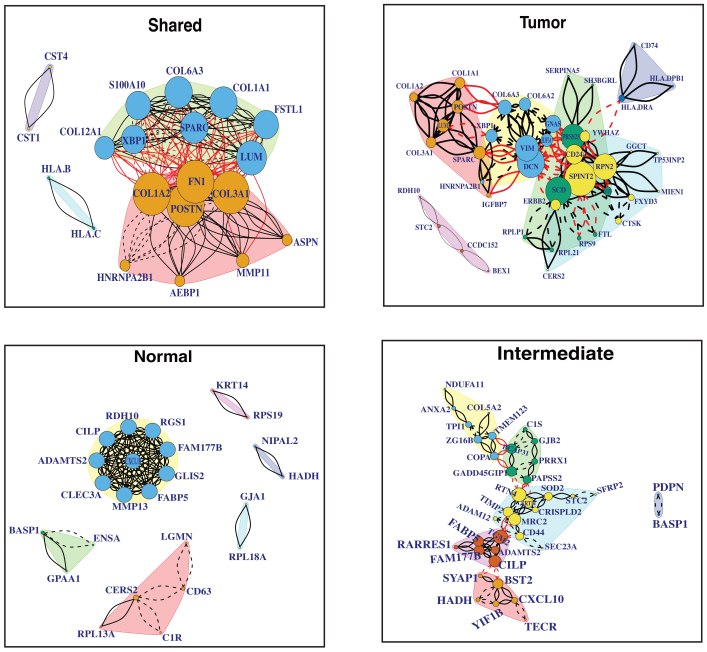

Figure C.4: Shared and cell-type specific community detection for Breast cancer data

two information criteria-based metrics to assess our model fitting: (i) Bayesian analogue of AIC (Akaike, 1998), defined as the Bayesian information criteria (BIC, Watanabe (2013)); and (ii) Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010), an improvement on the AIC and a fully Bayesian approach to measure model accuracy computed with log pointwise posterior predictive density and then adding a correction for the effective number of parameters to adjust for over-fitting. These criterion based methods are often used for model selection and specifically for spatial datasets (Banerjee et al., 2003, 2000; Lee and Ghosh, 2009). In both cases, lower (relative) values indicate better model fits.

Table C.3 shows the BIC and WAIC values for the SpaceX and non-spatial Poisson model for both the mouse hypothalamus and breast cancer data. Based on the criteria based values from the Table C.3, we can conclude that the SpaceX model is a better fit to both spatial transcriptomics datasets than the non-spatial Poisson model. For example, there is 64.7% and 46.6% of relative gain in accuracy of model fitting of the SpaceX model and non-spatial model w.r.t. BIC and WAIC respectively in case of Merfish data. A similar inference can be drawn for the breast cancer data where the relative gains are 66.4% and 45.5% for BIC and WAIC respectively in case of model fitting.

BIC (Merfish)

WAIC (Merfish)

BIC (Breast cancer)

WAIC (Breast cancer)

SpaceX Model

13520

43783

24346

54179

Non-spatial Poisson model

38274

82045

72523

99474

Criteria based values for application of the SpaceX and non-spatial Poisson model to spatial transcriptomics data i.e. mouse hypothalamus and Breast cancer data.

C.4 List of hub genes and edges

A detailed list of hub genes and top edges for both the datasets can be found at https://github.com/SatwikAch/SpaceX.

C.5 Corroboration with TCGA Breast Cancer Data

To corroborate some of our findings, we consider the TCGA-based gene expression from 67 breast cancer tissues and 20,000 genes using parallel high-throughput sequencing (Wirth et al., 2011; Weinstein et al., 2013). To make a fair "apples-to-apples" comparison, we used the same intersecting gene set from the spatial transcriptomics based breast cancer data used in our paper (Ståhl et al., 2016). We used a network-based algorithm: personalized cancer-specific integrated network estimation (PRECISE, Ha et al. (2018)) to obtain gene networks. PRECISE is Bayesian method for gene-network reconstruction for bulk-sequencing data that uses a regression-based approach. The PRECISE method detected 77 hub genes out of total 290 genes compared to the SpaceX method, which detected 59 hub genes – with 19 intersecting hub genes using both methods. The list of all the hub genes detected from each method and intersection hub genes from both method can be found at the webiste mentioned below under the name **BC_Hub_genes_TCGA.csv** (https://github.com/SatwikAch/SpaceX/tree/main/Hub%20genes).

Interestingly, multiple collagens genes (COL16A1, COL6A2, COL5A1) are detected as hub genes by both methods. Collagen biosynthesis can be regulated by cancer cells through mutated genes, transcription factors and signaling pathways (Xu et al., 2019). Understanding of the structural properties and functions of collagen in cancer will lead to anticancer therapy. The LUM gene is associated with collagen genes and effectively regulates estrogen receptors and function properties of breast cancer cells (Karamanou et al., 2017). Upregulation in FN1 gene indicates development various types of tumors (Sun et al., 2020b). XBP1 can induce cell invasion and metastasis in breast cancer cells by promoting high expression (Chen et al., 2020). VIM gene is used as a biomarker for the early detection of cancer (Mohebi et al., 2020).

C.6 Network similarity between cell-type specific networks

We further evaluated the performance of SpaceX to detect similarity between cell-cell interactions. To this end, we used Hamming distance, a well-established similarity measure between two networks, which has been used in several network topology based research studies (Tian and Shen, 2005, 2006; Ehounou et al., 2020). In our case, the Hamming distance is equivalent to the distance

between their two co-expression networks, i.e., the number of elements having a similar (or different) values in each of the two networks. A low (high) value in Hamming distance between two networks implies those two networks are more (less) similar to each other.

The mouse hypothalamus data consists of 7 cell-type based clusters among 4812 spatial locations. The SpaceX method provides gene co-expression networks specifically for each cell types. Using the Hamming distance as similarity metric, we measure the network similarity between cell-type specific networks obtained from the Mouse hypthalamaous data analyses in Section 4.1 of the paper. The heatmap of the Hamming distances between cell-type specific networks is shown in Figure C.5. We can observe that the co-expression network of immature cell-type is further apart than other cell type specific network in terms of Hamming distance. We rescale the Hamming distance with maximum value such that the distances are in [0,1] interval. Specifically, the Hamming distances of immature cell type network with other cell type (Endothelial, Astrocyte, Mature, Inhibitory, Excitatory, Ependymal) networks are 1, 0.73, 0.83, 0.79, 0.82, 0.73 respectively. Based on Figure C.5, network of endothelial cell type is distant from other cell-type based networks except for the immature cell-type. The distance of Astrocyte cell-type netwrok from Ependymal and Excitatory are 0.27 and 0.32 respectively. The neuronal cell type specific networks have lower distance than others which leads to infer a higher level of similarity between cell-type based networks than others. This similarity and disparity based finding aligns with multiple prior works which discuss about hypothalamic cell diversity (Chen et al., 2017; Mickelsen et al., 2020).

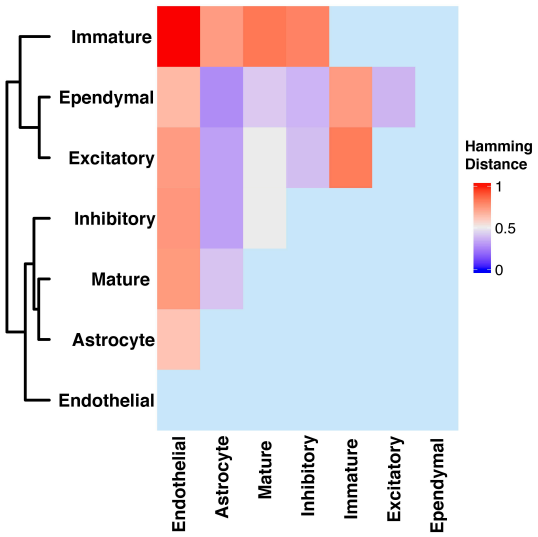

Figure C.5: The Figure shows heatmap of Hamming distances between cell-type specific networks.

Appendix D

Implementation of SpaceX

SpaceX function extimates shared and cluster specfic gene co-expression networks for spatial transcriptomics data. More details about the SpaceX method can be found in the main manuscript. See below for detailed discussion on installation of SpaceX package, Data inputs and outputs followed by an example.

D.1 Installation

The package requires a dependency that is not available on CRAN. Install it with:

```
remotes::install_github("rdevito/MSFA")
```

You can install the released version of SpaceX from (https://github.com/SatwikAch/SpaceX) with:

```
devtools::install_github("SatwikAch/SpaceX")
library(SpaceX)
```

D.2 Data inputs

Please make sure to provide both inputs as dataframe.

The first input is **Gene_expression_mat** which is $N \times G$ dataframe. Here N denotes the number of spatial locations and G denotes number of genes.

The second input is **Spatial_locations** is a dataframe which contains spatial coordinates.

The third input is **Cluster_annotations**.

The fourth input is **sPMM**. If TRUE, the code will return the estimates of sigma1_sq and sigma2_sq from the spatial Poisson mixed model.

The fifth input is **Post_process**. If TRUE, the code will return all the posterior samples, shared and cluster specific co-expressions. Please make sure to request for large enough memory to work with the posterior samples. Default is FALSE and the code will return the posterior samples of Φ and Ψ^c (based on definition in equation 1 of the SpaceX paper) only.

D.3 Output

You will obtain a list of objects as output.

Posterior_samples contains all the posterior samples.

Shared_network provides the shared co-expression matrix (transformed correlation matrix of $G_s = \Phi\Phi^T$).

Cluster_network provides the cluster specific co-expression matrices (transformed correlation matrices of $G_c = \Phi\Phi^T + \Psi^c\Psi^{cT}$).

D.4 Example

Here we provide an example to run the SpaceX method.

```
# Reading the Breast cancer data

# Spatial locations
head(BC_loc)

# Gene expression for data
head(BC_count)

# Data processing
G <-dim(BC_count)[2] # number of genes
N <-dim(BC_count)[1] # number of locations
```

Next, we'll apply the SpaceX method on the Breast cancer dataset.

```
# Application to SpaceX method
BC_fit <- SpaceX(BC_count,BC_loc[,1:2],BC_loc[,3],sPMM=FALSE,Post_process = TRUE)

# Shared_network :: Shared co-expression matrix
# Cluster_network :: Cluster specific co-expression matrices
```

Bibliography

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer.

Banerjee, S., Gelfand, A. E., and Polasek, W. (2000). Geostatistical modelling for spatial interaction data with application to postal service performance. *Journal of statistical planning and inference*, 90(1):87–105.

Banerjee, S., Wall, M. M., and Carlin, B. P. (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics*, 4(1):123–142.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse bayesian infinite factor models. *Biometrika*, pages 291–306.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2007). On modularity clustering. *IEEE transactions on knowledge and data engineering*, 20(2):172–188.

Chen, R., Wu, X., Jiang, L., and Zhang, Y. (2017). Single-cell rna-seq reveals hypothalamic cell diversity. *Cell reports*, 18(13):3227–3241.

Chen, S., Chen, J., Hua, X., Sun, Y., Cui, R., Sha, J., and Zhu, X. (2020). The emerging role of xbp1 in cancer. *Biomedicine & Pharmacotherapy*, 127:110069.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2021). Bayesian multistudy factor analysis for high-throughput biological data. *The Annals of Applied Statistics*, 15(4):1723–1741.

Dryden, I. L., Koloydenko, A., Zhou, D., et al. (2009). Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123.

Ehounou, W. J., Barth, D., De Moissac, A., Watel, D., and Weisser, M.-A. (2020). Minimizing the hamming distance between a graph and a line-graph to discover the topology of an electrical network. *J. Graph Algorithms Appl.*, 24(3):133–153.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6):997–1016.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2):557–587.

Ha, M. J., Banerjee, S., Akbani, R., Liang, H., Mills, G. B., Do, K.-A., and Baladandayuthapani, V. (2018). Personalized integrated network modeling of the cancer proteome atlas. *Scientific reports*, 8(1):1–14.

Karamanou, K., Franchi, M., Piperigkou, Z., Perreau, C., Maquart, F.-X., Vynios, D. H., and Brezillon, S. (2017). Lumican effectively regulates the estrogen receptors-associated functional properties of breast cancer cells, expression of matrix effectors and epithelial-to-mesenchymal transition. *Scientific reports*, 7(1):1–15.

Lee, H. and Ghosh, S. K. (2009). Performance of information criteria for spatial models. *Journal of statistical computation and simulation*, 79(1):93–106.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, pages 41–67.

Mickelsen, L. E., Flynn, W. F., Springer, K., Wilson, L., Beltrami, E. J., Bolisetty, M., Robson, P., and Jackson, A. C. (2020). Cellular taxonomy and spatial organization of the murine ventral posterior hypothalamus. *Elife*, 9:e58901.

Moffitt, J. R., Bambah-Mukku, D., Eichhorn, S. W., Vaughn, E., Shekhar, K., Perez, J. D., Rubinstein, N. D., Hao, J., Regev, A., Dulac, C., and Zhuang, X. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416).

Mohebi, M., Ghafouri-Fard, S., Modarressi, M. H., Dashti, S., Zekri, A., Kholghi-Oskooei, V., and Taheri, M. (2020). Expression analysis of vimentin and the related lncrna network in breast cancer. *Experimental and molecular pathology*, 115:104439.

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25(3):257–265.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82.

Sun, S., Zhu, J., Mozaffari, S., Ober, C., Chen, M., and Zhou, X. (2018). Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. *Bioinformatics*, 35(3):487–496.

Sun, S., Zhu, J., and Zhou, X. (2020a). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature methods*, 17(2):193–200.

Sun, Y., Zhao, C., Ye, Y., Wang, Z., He, Y., Li, Y., and Mao, H. (2020b). High expression of fibronectin 1 indicates poor prognosis in gastric cancer. *Oncology Letters*, 19(1):93–102.

Tian, H. and Shen, H. (2005). Hamming distance and hop count based classification for multicast network topology inference. In *19th International Conference on Advanced Information Networking and Applications (AINA'05) Volume 1 (AINA papers)*, volume 1, pages 267–272. IEEE.

Tian, H. and Shen, H. (2006). Multicast-based inference for topology and network-internal loss performance from end-to-end measurements. *Computer Communications*, 29(11):1936–1947.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120.

Wirth, H., Löffler, M., von Bergen, M., and Binder, H. (2011). Expression cartography of human tissues using self organizing maps. *Nature precedings*, pages 1–1.

Xu, S., Xu, H., Wang, W., Li, S., Li, H., Li, T., Zhang, W., Yu, X., and Liu, L. (2019). The role of collagen in cancer: from bench to bedside. *Journal of translational medicine*, 17(1):1–22.