SC-MEB: spatial clustering with hidden Markov random field using empirical Bayes

Yi Yang[1,†], Xingjie Shi[2,†], Qiuzhong Zhou[3], Lei Sun[3], Joe Yeong[4,5,*], and Jin Liu[1,*]

[1]Program in Health Services & Systems Research, Duke-NUS Medical School, 8 College Road, 169857, Singapore

[2]Academy of Statistics and Interdisciplinary Sciences, East China Normal University, 3663 Zhongshan North Road, 200062, Shanghai, China

[3]Program in Cardiovascular and Metabolic Disorders, Duke-NUS Medical School, 8 College Road, 169857, Singapore

[4]Institute of Molecular and Cell Biology (IMCB), Agency of Science, Technology and Research (A*STAR), Street, 138673, Singaore

[5]Department of Anatomical Pathology, Singapore General Hospital, 20 College Road, 169856, Singaore

Abstract

Spatial transcriptomics (ST) has been emerging as a powerful technique for resolving gene expression profiles while retaining the tissue spatial information. These spatially resolved transcriptomics make it feasible to examine the complex mulitcellular systems under different microenvironments. To answer scientific questions with spatial transcriptomics, the first step is to identify cell clusters by integrating available spatial information which would expand the understanding of how cell types and states are regulated by tissues. Here, we introduce SC-MEB, an empirical Bayes approach for spatial clustering analysis using hidden Markov random field. We also derive an efficient expectation-maximization (EM) algorithm based on iterative conditional mode (ICM) for SC-MEB. Compared with BayesSpace, a recently developed method, SC-MEB is not only computationally efficient and scalable to the sample size increment, but also is capable of choosing the smoothness parameter and the number of clusters as

well. We then performed comprehensive simulation studies to demonstrate the effectiveness of SC-MEB over some existing methods. We first applied SC-MEB to perform clustering analysis in the spatial transcriptomics dataset from human dorsolateral prefrontal cortex tissues that were manually annotated. Our analysis results show that SC-MEB can achieve similar clustering performance with BayesSpace that uses the true number of clusters and fixed smoothness parameter. We then applied SC-MEB to analyze the dataset from a patient with colorectal cancer (CRC) and COVID-19, and further performed differential expression analysis to identify signature genes related to the clustering results. The heatmap for identified signature genes shows that the identified clusters from SC-MEB is more separable than those from BayesSpace. By using pathway analysis, we identified three immune-related clusters and further compared mean expressions of COVID-19 signature genes in immune regions with those in non-immune ones. As such, SC-MEB provides a valuable computational tool for investigating structural organizations of tissues from spatial transcriptomic data.

Keywords: spatial transcriptomics, emprical Bayes, hidden Markov random field, expectation-maximization algorithm

1 Introduction

Recent advances in spatial transcriptomics (ST) have allowed for simultaneously measuring transcriptome-wide gene expression at near-single-cell resolution while retaining spatial information of each measurement [Burgess, 2019]. These spatially resolved transcriptomics have deepened our understanding of how cell types and states are regulated by tissues, e.g., human brain [Maynard et al., 2021], mouse brain [Shah et al., 2016, Alon et al., 2021], mouse embryo [Lohoff et al., 2020], and others. Among all technologies for resolving spatial gene expression, they largely can be classified as either imaging-based or next-generation sequencing-based methods [Waylen et al., 2020]. Imaging-based methods were developed to study spatial complexity based on fluorescent in situ hybridization (FISH), e.g., sm-FISH [Lyubimova et al., 2013], seqFISH [Eng et al., 2017], and MERFISH [Xia et al., 2019]. Though FISH-based methods are capable of capturing both RNA quantity and position, these approaches are still limited by their throughput and accuracy to measure the level of gene expression. On the other hand, multiple next-generation sequencing-based methods have been developed to facilitate the high-throughput including Geo-seq [Chen et al., 2017], Slide-seq [Rodriques et al., 2019], and recently commercialized 10x Genomics Visium [Ståhl et al., 2016]. Emerging ST technologies offer great opportunities to investigate the spatial patterns of gene expression in many applications, such as cell type identification, tissue exploration, and differential expression analysis. Among all these applications, clustering is the first problem we need to address.

Similar to scRNA-seq data, excessive zeros or "drop-outs" were observed in ST data [Qiu, 2020]. Recently, many studies have argued that most drop-outs are due to biological variation such as cell-type heterogeneity rather than technical shortcomings [Svensson, 2020]. Kim et al. [2020] suggested that clustering analysis should be performed before imputing or normalizing data. To overcome the curse of dimensionality due to the high-throughput spatial gene expression, clustering is often preceded by standard dimension reduction procedures, e.g., principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) [Van der Maaten and Hinton, 2008], and uniform manifold approximation and projection (UMAP) [McInnes et al., 2018].

Most existing methods for clustering does not consider the available spatial information in ST datasets, e.g., k-means [Kriegel et al., 2017] and Gaussian mixture models (GMM) [Bishop, 2006]. To allow for incorporating the additional spatial information in ST datasets, a few methods have been developed recently including a hidden Markov random field (HMRF) model implemented in the *Giotto* package [Zhu et al., 2018, Dries et al., 2019] and a fully Bayesian model with a Markov random field (MRF), BayesSpace [Zhao et al., 2020]. In the presence of spatial coordinates for each transcriptome profiled spot, spatial clustering methods achieve better classification accuracy. For example, Zhao et al. [2020] showed that BayesSpace improves resolution and achieves better classification accuracy in manually-annotated human brain samples. However, these existing spatial clustering methods have certain limitations. First, as BayesSpace is a fully Bayesian method based on Markov chain Monte Carlo (MCMC), it is not computationally scalable for ST data with higher resolution. Second, the smoothness parameter is an essential parameter of MRF-based methods, largely determining the proximity of the neighboring spots [Tolpekin and Stein, 2009]. BayesSpace takes this smoothness parameter as fixed and thus cannot choose the optimal one that best fits a given data. Third, without optimizing the smoothness parameter, one cannot apply any model selection methods to obtain the optimal number of clusters. In practice, the number of clusters in ST datasets is usually unknown before the follow-up analysis and a preferred method should be capable of choosing the number of clusters automatically.

To address all these limitations, we propose a Spatial Clustering using the hidden Markov random field based on Empirical Bayes (SC-MEB) to model a low-dimensional representation of the gene expression matrix incorporating the spatial coordinates for each measurement. Compared with existing methods [Dries et al., 2019, Zhao et al., 2020], SC-MEB is not only computationally efficient and scalable to the sample size increment, but also is capable of choosing the smoothness parameter and the number of clusters as well. We derive an efficient expectation-maximization (EM) algorithm based on iterative conditional mode (ICM) and further choose the number of clusters for SC-MEB with the Bayesian information criterion

(BIC) [Claeskens et al., 2008]. We demonstrated the effectiveness of SC-MEB over existing methods through comprehensive simulation studies. We then applied SC-MEB to perform clustering analysis on two ST datasets. By using ST data in human dorsolateral prefrontal cortex tissues that were manually annotated, we show the better clustering performance using spatial clustering methods while the performance of SC-MEB is comparable with that of BayesSpace, even though the latter uses the "true" number of clusters and prespecified fine-tuned smoothness parameter. We further applied SC-MEB and alternative methods to analyze ST data from a patient with colorectal cancer and COVID-19. We performed the follow-up differential expressions analysis by using clustering results from SC-MEB and BayesSpace. The heatmap for identified signature genes shows that the clustering result from SC-MEB is more reasonable and interpretable. By using pathway analysis, we identified three immune-related clusters and further compared mean expressions of COVID-19 signature genes in immune regions with those in non-immune ones.

2 Materials and Methods

2.1 Problem Formulation

The SC-MEB consists of three major stages (Fig. 1A). First, we conduct principle component (PC) analysis on the log-normalized expression of the highly variable genes to obtain top PCs (Fig. 1B). Next, we perform the spatial clustering using PCs for each spot. Finally, downstream analysis such as differential expression analysis can be performed to obtain signature genes for each cluster (Fig. 1D).

Our spatial clustering method builds on a two-level hierarchical probabilistic model (Fig. 1C). Briefly, for spot i, the first level specifies the conditional probability of the low-dimensional representation (e.g., top PCs) of its gene expression y_i given an unknown label $x_i \in \{1, \ldots, K\}$, where K is the number of clusters. In SC-MEB, we assume that given labels for each spot, a d-dimensional representation y_i is mutually independent among all spots and its distribution within a given cluster k can be written as

$$p(y \mid x, \theta) = \prod_{i \in \mathcal{S}} \mathcal{N}(y_i | x_i = k, \mu_k, \Sigma_k), \tag{1}$$

where $\theta = \{\mu_k, \Sigma_k : k = 1, \ldots, K\}$, and μ_k and Σ_k denote the mean and covariance matrix for cluster k, respectively.

The second level of SC-MEB depicts the prior probability of the hidden labels and we implement Markov random field (MRF) prior to encourage the smoothness among spots. In other words, spots of the same cluster can be in close proximity. As spots in Visium are pri-

marily arranged on hexagonal lattices, we define the neighborhood of each spot by applying a proximity threshold. To promote smoothness within spot neighborhoods, we use the Potts energy $U(x)$ [Zhang et al., 2001], which ranks all possible configurations proportionally to a positive parameter β. Thus, the Potts model promotes spatial smoothness by penalizing the case that the neighboring spots are assigned to different cluster labels. The hidden random field x are assumed to be

$$P(x) = \frac{1}{Z_\beta} \exp\{-U(x)\}, \tag{2}$$

where $U(x) = \sum_{i,i' \in \mathcal{N}_i} \beta[1 - \delta(x_i, x_i')]$, δ is the delta function, and Z_β is a normalization constant that does not have a closed form. When all labels on a neighborhood take the same value, meaning that the hidden x is locally smooth, it incurs no neighborhood cost; otherwise, if they are not all the same, it incurs a positive cost, and the amount of cost is controlled by parameter β. Thus, parameter β controls the smoothness in latent labels, with larger β, the more smooth spatially in latent labels. When β is zero, SC-MEB reduces to the method without consideration of spatial information, i.e., GMM. Combining two levels of SC-MEB (1) and (2), we denote $\phi = (\theta, \beta)$ the parameter space.

As the smoothing parameter β does not have an explicit update form, SC-MEB adaptively selects β via a grid search strategy. That is, SC-MEB model is trained with a prefixed β using an efficient ICM-EM scheme [Cuadra et al., 2005] that incorporates a pseudo-likelihood maximization step as in the iterative conditional mode (ICM) method of Besag [1974]. The optimal β is chosen by the one that maximizes the marginal log-likelihood. In a similar way, we can evaluate the marginal log-likelihood for a sequence of K. Then, we apply BIC [Forbes and Peyrard, 2002] to choose the optimal number of clusters in a data-driven manner (Fig. 1D). Please refer to Supplementary for more details about the BIC used here.

2.2 ICM-EM Algorithm

Estimation of the parameter is done through an iterative-conditional-mode based expectation-maximization (ICM-EM) algorithm [Cuadra et al., 2005]. Here, we assume (K, β) is known.

In the ICM step, the estimate of x is obtained by maximizing its posterior with respect to x_i coordinately:

$$P(x \mid y) = P(x_i, x_{\mathcal{S}-\{i\}} \mid y) = P(x_i \mid y, x_{\mathcal{S}-\{i\}}) P(x_{\mathcal{S}-\{i\}} \mid y),$$

where $i = 1, \ldots, n$, until converge [Besag, 1986]. Given initial values of x, θ, and observed y, we have the update equation:

$$\hat{x}_i = \min_{x_i} U(\hat{x}_1, \ldots, x_i, \ldots, \hat{x}_n), \tag{3}$$

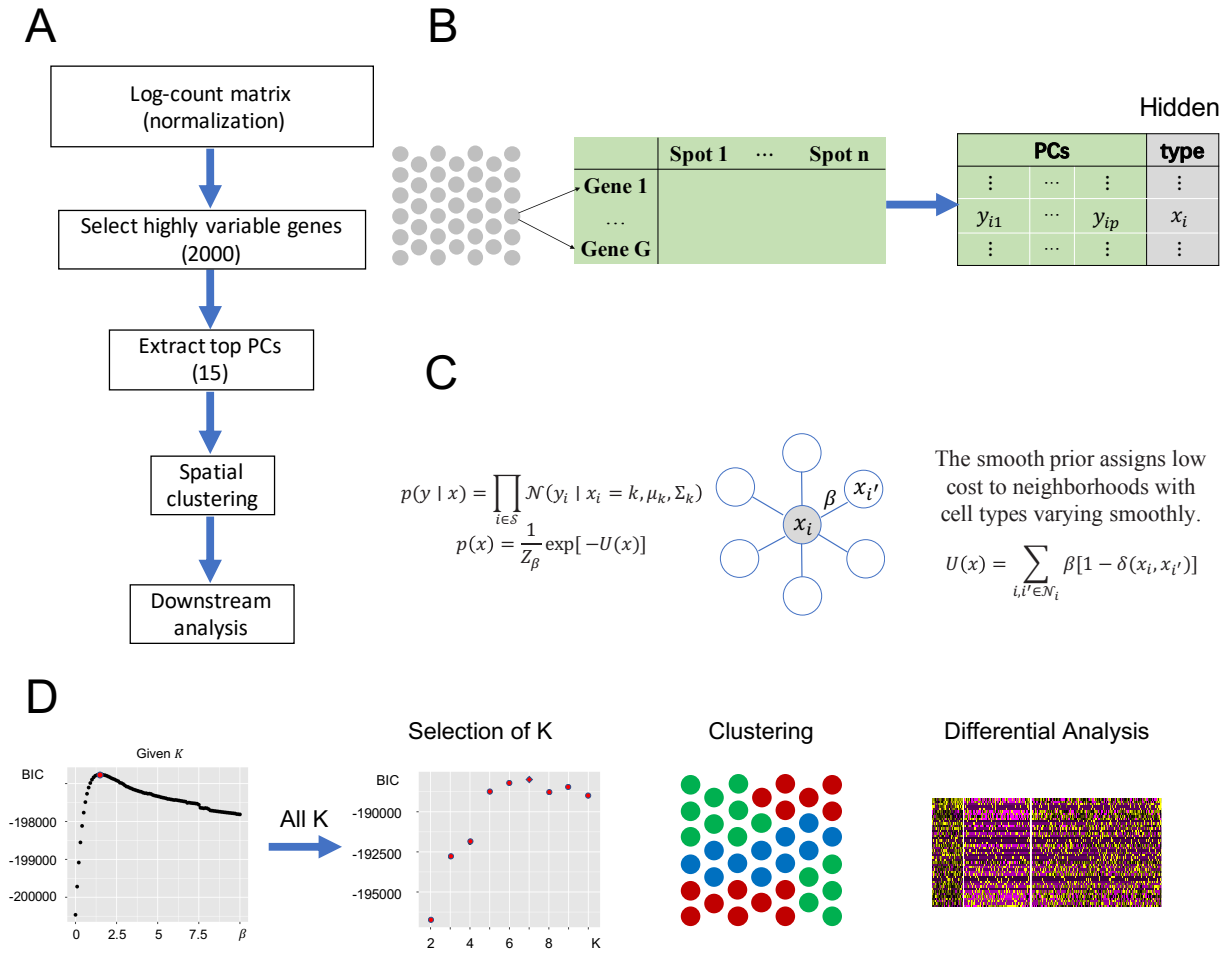

Figure 1: SC-MEB workflow. **A.** SC-MEB workflow. **B.** Data preprocessing. **C.** The hidden MRF model. For Visium dataset, we use the 6-neighborhoood for each spot. **D.** The outputs of SC-MEB.

where

$$U(x) = \{\frac{1}{2}(y_i - \mu_{x_i})^\top \Sigma_{x_i}^{-1}(y_i - \mu_{x_i}) + \frac{1}{2}\log|\Sigma_{x_i}| + \beta \sum_{i' \in \mathcal{N}_i} [1 - \delta(x_i, x_{i'})]\}.$$

In the expectation (E) step, instead of using the original complete likelihood that is difficult to evaluate, we use the following pseudo-likelihood:

$$\begin{aligned}\tilde{p}(y, x; \theta) &= p(y|x;\theta)\tilde{p}(x;\theta) \\ &= \prod_i p(y_i \mid x_i;\theta) \prod p(x_i|x_{\mathcal{N}_i};\theta)\end{aligned}$$

$$= \prod_i [p(y_i \mid x_i; \theta) p(x_i | x_{\mathcal{N}_i}; \theta)]$$
$$= \prod_i p(y_i, x_i \mid x_{\mathcal{N}_i}; \theta).$$

With the optimal conditional distribution of x (details in Supplementary), we have

$$Q(\theta) = \sum_i \sum_k \gamma_{ik} \left[\log p(y_i \mid x_i = k; \theta) + \log p(x_i = k \mid x_{\mathcal{N}_i}; \theta) \right],$$

where γ_{ik} is called responsibility that component k takes for explaining the observation y_i which is defined as follows:

$$\gamma_{ik} = \frac{P(y_i \mid x_i = k) P(x_i = k \mid X_{\mathcal{N}_i} = \hat{x}_{\mathcal{N}_i})}{\sum_{k'} P(y_i \mid x_i = k') P(X_i = k' \mid X_{\mathcal{N}_i} = \hat{x}_{\mathcal{N}_i})}. \tag{4}$$

By taking partial derivatives of $Q(\theta)$ with respect to the parameters and setting them to zero, we obtain the update equations for the maximization (M) step:

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} y_i, \tag{5}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} (y_i - \mu_k)(y_i - \mu_k)^\top. \tag{6}$$

where $N_k = \sum_{i=1}^{n} \gamma_{ik}$.

The ICM-EM algorithm iterates the ICM step and the M step until convergence. Further details on the ICM-EM algorithm are provided in Supplementary Material.

2.3 Methods for comparison

We conducted comprehensive simulations and real data analysis to gauge the performance of different methods for clustering a low-dimensional representation of gene expression matrix including both non-spatial and spatial clustering methods.

In details, we considered the following non-spatial clustering methods: (i) k-means implemented in the R package *stats* available at CRAN; (ii) GMM implemented in the R package *mclust* available at CRAN; (iii) Louvain implemented in the R package *igraph* available at https://igraph.org/r/. On the other hand, we compared the clustering performance with spatial methods: (i) SC-MEB implemented in the R package *SC.MEB* available at https://github.com/Shufeyangyi2015310117/SC.MEB; (ii) BayesSpace implemented in the R package *BayesSpace* available at Bioconductor; (iii) HMRF implemented in the *Giotto* package available at http://spatialgiotto.rc.fas.harvard.edu/.

2.4 Preprocessing for ST datasets

In all datasets, we applied log-transformation and normalization on the raw count matrix using library size [Lun et al., 2016, McCarthy et al., 2017]. Then we performed principal component analysis (PCA) on the top 2,000 highly variable genes. In the clustering analysis, we chose the top 15 principles components (PCs) from the study datasets as the input for SC-MEB and other alternative methods.

2.5 ST datasets

2.5.1 Human dorsolateral prefrontal cortex (DLPFC)

Maynard et al. [2021] used recently-released ST technology, 10x Genomics Visium platform, to generate spatial maps of gene expression matrix in the six-layered DLPFC of the adult human brain that are provided in the *spatialLIBD* package. They also provided manual annotations of layers based on cytoarchitecture. In this study, they profiled spatial transcriptomics in human postmortem DLPFC tissue sections for 12 samples, where a median depth of 291.M reads for each sample, corresponding to a mean 3,462 unique molecular indices (UMIs) and a mean 1,734 genes per spot.

2.5.2 Human colorectal cancer (CRC)

CRC tissue is from the patient who was a 45-year-old South Asian male who was diagnosed with COVID-19 on 16th April 2020. As previously described [Cheung et al., 2020], the patient had experienced mild upper respiratory tract symptoms throughout the course of the disease. He was confirmed COVID-19 negative by two consecutive nasopharyngeal swabs on 9th and 10th May 2020 and was discharged from the isolation facility from 10th May 2020. During hospital admission, further investigation involving computed tomography (CT) scan and colonoscopy revealed the presence of a large circumferential malignant mass in the caecum. Histology of the biopsies confirmed that the patient had invasive caecal adenocarcinoma stage II T3N0. He underwent laparoscopic right hemicolectomy on 18th May 2020, 9 days after being tested negative for COVID-19. He recovered uneventfully and was discharged on 21th May 2020. In this sample, we profiled spatial transcriptomics using 10x Genomics Visium platform In summary, it has a depth of 143 million reads for a total of 2988 spots within the tissue and a median 492 genes per spot.

2.6 Evaluation metrics

We evaluate the clustering performance by adjusted Rand index [Schütze et al., 2008]. The general form of ARI is as follows

$$\text{ARI} = \frac{(\text{RI} - \text{expected}(\text{RI}))}{(\max(\text{RI}) - \text{expected}(\text{RI}))} \tag{7}$$

where RI stands for the rand index [Rand, 1971], max(RI) and expected(RI) are the maximum value and the expected value of RI, respectively. Assume that n is the number of spots in a ST dataset. $U = \{u_1, \ldots, u_i, \ldots, u_R\} \in \mathbb{R}^n$ and $V = \{v_1, \ldots, v_j, \ldots, v_C\} \in \mathbb{R}^n$ represent two clustering labels for n spots, where R and C are the corresponding numbers of clusters in U and V, respectively. Denote n_{ij} the number of spots belonging to both classes u_i and v_j, and $n_{i\cdot}$ and $n_{\cdot j}$ the number of spots in classes u_i and v_j, respectively. Then ARI (7) is defined as

$$\text{ARI}(U, V) = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} \,/\, \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}\right] \,/\, \binom{n}{2}}. \tag{8}$$

As the expected value of RI for two random partitions does not take a constant value and is concentrated in a small interval, ARI is a corrected version of RI to avoid these drawbacks [Hubert and Arabie, 1985]. Note that ARI lies between -1 and 1 and takes a value 1 when the two partitions are equal up to a permutation. Obviously, a larger ARI value indicates a higher similarity between two partitions. In the simulation, ARI was used to measure the similarity between the estimated partition and the true one. In the analysis for DLPFC dataset [Maynard et al., 2021], the manual annotations based on additional experiments and computational results are both available. ARI was used to measure the similarity between labels from the estimated partition and the manually-annotated clusters.

3 Results

3.1 Simulation settings

We evaluate the clustering performance of SC-MEB using simulations by comparing five other methods including k-means, GMM, Louvain, BayesSpace and Giotto. For both k-means and BayesSpace, we consider the true number of clusters K, and its two nearest numbers, $K-1$ and $K+1$, as one needs to specify the number of clusters manually for these two methods. For all other methods, the number of clusters are selected automatically. The smoothness parameter β of BayesSpace is fixed at 3 by default while β of SC-MEB is optimized by a

grid search. We compared the clustering performance using the ARI for all methods, where in each setting, we run 50 replicates.

In Example I, labels of spots were randomly generated. In detail, for a 70×70 squared lattices with 4,900 spatial spots, we generated cluster labels for each spot from the K-states Potts model (as shown in Eqn. (2)) with $\beta \in [1, 1.3]$ using the R package *GiRaF*. The number of neighbors was set to be 4 and the number of true clusters K was set to be 3, 5 or 7. We then considered two distributions for low-dimensional PCs: mixture of Gaussian and mixture of Student-t distributions. The number of PCs was set to be either 10 or 15. The mean μ_k and the covariance matrix Σ_k for each component k are given in Supplementary Table S1-S4.

In Example II, labels of spots were obtained from real data analysis. In detail, we used the inferred cluster labels from SC-MEB ($K = 8$) in CRC data as the true ones for all 2,988 spots. PCs were randomly generated in the same way as Example I. The mean μ_k and the covariance matrix Σ_k for each component k are given in Supplementary Table S5-S6.

In above examples, all K components have different covariance matrices. Since BayesSpace adopts a strategy that all components have a shared covariance, we further conducted additional simulations with equal covariance matrices.

3.2 SC-MEB outperforms other methods in simulation studies

In Example I, when PCs are from mixture of Gaussian distributions, SC-MEB is more powerful than all other methods (Fig. 2A). BayesSpace has smaller ARI than SC-MEB even it takes the true number of clusters as input. The inferior performance of BayesSpace is due to its lack of adaptation to the smoothness parameter β. For other methods, Giotto, GMM, k-means and Louvain achieve lower ARI. When PCs are from mixture of Student-t distributions, assumptions of BayesSpace are satisfied. As shown in Fig. 2B, BayesSpace with the correct number of clusters performs the best. Even though SC-MEB is misspecified in this setting, it still achieves a high ARI, which is larger than that of BayesSpace with misspecificed K and other methods. This demonstrates the robustness of SC-MEB to the misspecification of distributions. Results from other settings (Supplementary Fig. S1-S2) provide similar conclusions.

In Example II, the comparison results (Fig. 2C and 2D) are largely consistent with the results obtained from Example I. Specifically, SC-MEB is more powerful than all other methods. The performance of BayesSpace is next to SC-MEB, and outperforms the rest. And k-means performs the worst in all methods. Results from other settings (Supplementary Fig. S5A-S5B, S6A-S6B) provide similar conclusions.

Finally, we considered the above two Examples under BayesSpace's assumption that all K components share a common covariance. Results are shown in (Supplementary Fig. S3

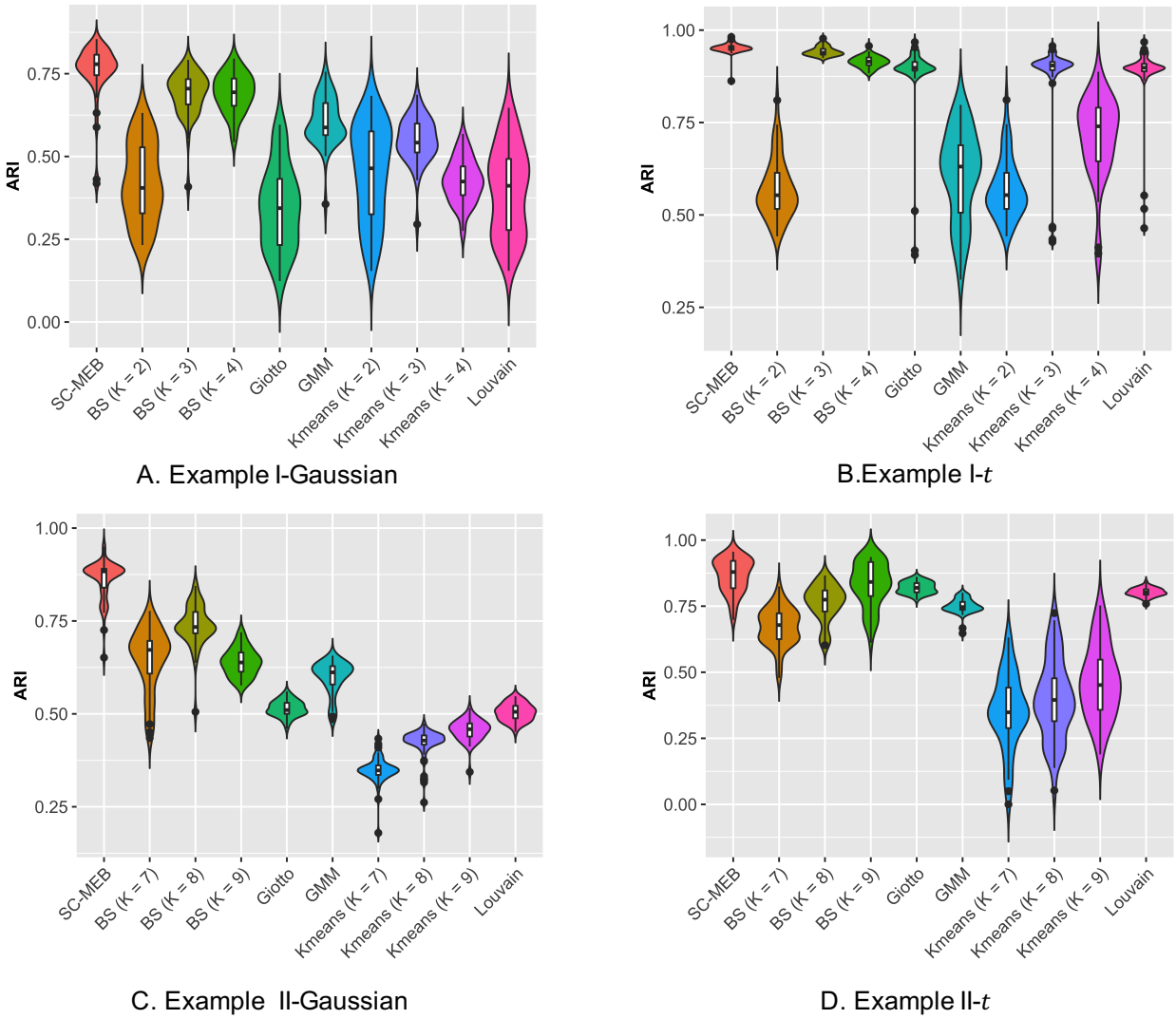

Figure 2: Summary of clustering accuracy for the six methods in simulated data. **A.** Example 1, and PCs are sampled from a Gaussian mixture model. **B.** Example 1, and PCs are sampled from a Student-t mixture model. **C.** Example 2, and PCs are sampled from a Gaussian mixture model. **D.** Example 2, and PCs are sampled from a Student-t mixture model.

- S4, S5C-D, and S6C-D). As we can see, the ARI of SC-MEB is comparable with that of BayesSpace, where both of them have better performance than that of other methods.

All simulations were conducted on a machine with 2.1 GHz Intel Xeon Gold 6230 CPU and 16G memory. SC-MEB is computationally more efficient than BayesSapce. In all simulations, SC-MEB takes about 8 minutes to complete the analysis for 200 combinatorial values in K and β, while BayesSpace takes about 25 minutes for a prefixed K and β. BayesSpace costed up to 600x more computation time than SC-MEB for a fixed combinatorial value of K and β. To better demonstrate the computational efficiency and scalability of SC-MEB, we conducted additional simulations with increasing sample size n. In Fig. 3, we can

see that the computation time of SC-MEB for a fixed number of iterations increases almost linearly with the increment of sample size, where it takes about 12.5h to run 50 iterations for a dataset with 100K spots. Thus, SC-MEB can be used to perform clustering analysis for ST datasets with higher resolution.

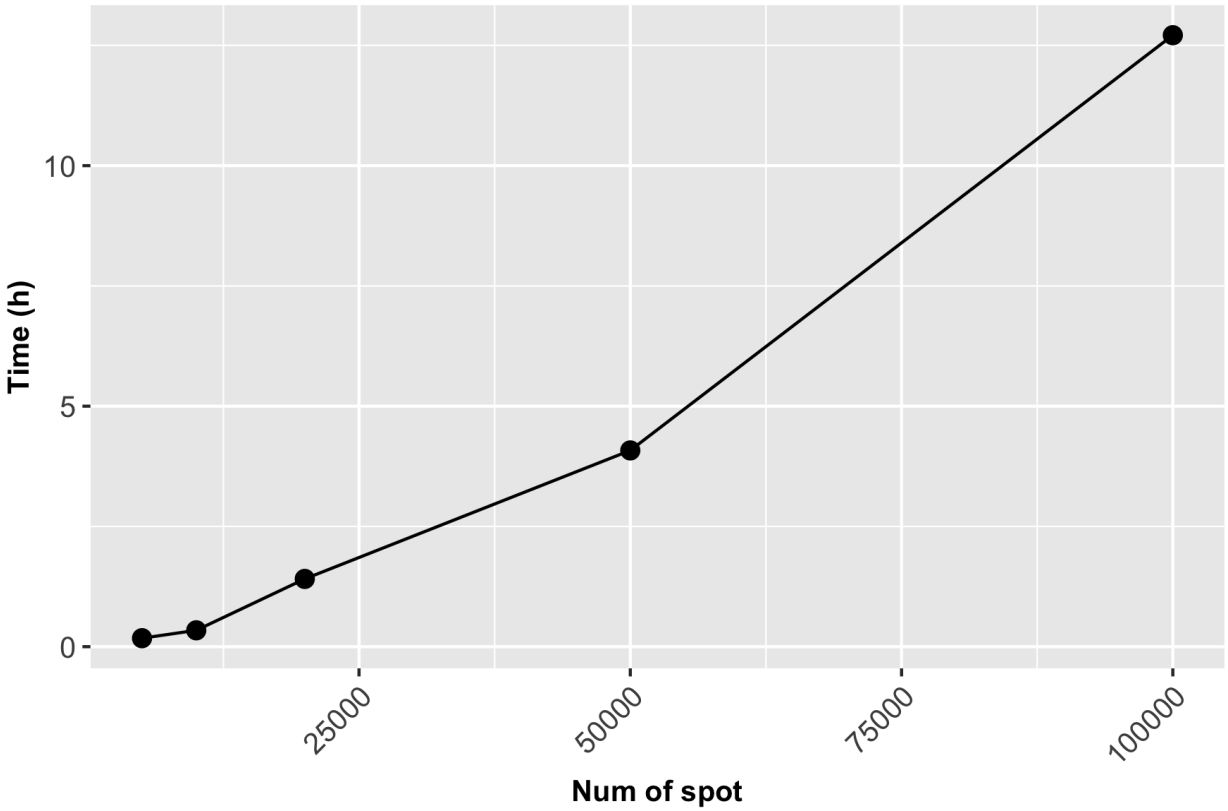

Figure 3: The computation time of SC-MEB increases linearly with the sample size. We fix the number of iterations to be 50 for different sample sizes.

3.3 Benchmark clustering performance in DLPFC dataset

To evaluate the clustering performance of SC-MEB in real datasets, we first applied it to DLPFC dataset and compared its performance with alternative methods. Specifically, we first obtained the top 15 PCs from the 2,000 most highly variable genes (HVGs) in each DLPFC sample, and then performed clustering analysis with all the five methods except k-means. As BayesSpace cannot choose the number of clusters K, we fixed it to be the number of clusters in the manual annotations. All other methods selected the number of clusters automatically.

Table 1 shows the ARI values in 12 DLPFC samples, where manually-annotated layers are taken as the "ground truth". It can be seen that SC-MEB outperforms BayesSpace in 5

Table 1: Clustering accuracy in DLPFC dataset. ARI values were evaluated by comparing manually annotations against cluster lables from SC-MEB and alternative methods for all 12 samples.

ID	SC-MEB	BayesSpace	GMM	Giotto	Louvain
151507	**0.45**	0.33	0.40	0.33	0.32
151508	**0.37**	0.36	0.33	0.33	0.25
151509	**0.45**	0.44	0.29	0.29	0.30
151510	**0.46**	0.43	0.31	0.32	0.28
151669	0.27	**0.41**	0.22	0.15	0.20
151670	0.24	**0.43**	0.19	0.13	0.26
151671	0.34	**0.38**	0.23	0.25	0.36
151672	0.44	**0.77**	0.14	0.27	0.27
151673	0.38	**0.55**	0.29	0.26	0.29
151674	0.33	0.33	0.29	**0.34**	0.33
151675	0.37	**0.41**	0.24	0.28	0.24
151676	**0.34**	0.32	0.26	0.26	0.25

samples, and vice verses in the other 6 samples. Note that in the analysis, BayesSpace took both the "true" number of clusters (from the manual annotations) and the prefixed fine-tuned β as input. In this case, the proposed SC-MEB can achieve the similar performance without the prior information. Additionally, among methods that can select (K) automatically, SC-MEB achieves the best clustering performance. Table 2 shows the comparisons of computational time for all methods. The computational time for SC-MEB is based on a sequence of values for β and K. For a fixed pair of β and K, the speed of SC-MEB is almost 2,000x faster than that of BayesSpace. Even though GMM, Giotto and Louvain are even faster than SC-MEB, the clustering performance of these methods are worse than SC-MEB and BayesSpace (Table 1).

3.4 Spatial clustering of in-house CRC sample

In addition, we used SC-MEB to analyze an in-house CRC sample. In the analysis, we first obtained top 15 PCs in the same way as in DLPFC dataset. Then we applied SC-MEB to perform spatial clustering in comparison with other methods. Since SC-MEB, Giotto and Louvain selected the optimal number of clusters (K) to be eight, we run BayesSpace with eight clusters. The computational time for SC-MEB, BayesSpace, GMM, Giotto, and Louvain are 165.80, 5324.48, 46.26, 19.75, and 0.69 seconds, respectively.

The clustering results obtained with the different methods are shown in Fig. 4. In general, the pattern of clustering assignments from SC-MEB is similar to that of GMM but the latter keeps more noisy spots. In addition, results from SC-MEB and BayesSpace have stronger

Table 2: Comparison of computational time (s) for five methods in DLPFC. Note that for SC-MEB, we evaluated on a sequence of K from 2 to 10 and a sequence of beta from 0 to 4 with equal space 0.2.

ID	SC-MEB	BayesSpace	GMM	Giotto	Louvain
151507	234.63	7015.95	46.16	21.68	2.86
151508	230.33	5471.57	36.26	41.26	6.97
151509	230.04	4660.43	53.22	33.95	6.00
151510	236.77	4288.56	39.50	42.10	7.57
151669	205.63	5840.10	20.96	16.01	3.49
151670	190.27	4917.43	37.94	26.49	5.26
151671	214.03	3889.07	31.01	38.61	8.82
151672	206.11	3793.58	22.44	33.74	5.49
151673	191.48	5984.89	40.20	18.26	1.49
151674	165.63	5344.68	36.82	38.24	3.34
151675	178.01	3863.19	27.90	31.63	4.22
151676	168.54	3609.71	37.44	28.90	3.24

spatial patterns than other methods.

By checking PanglaoDB [Franzén et al., 2019] for signature genes identified by *differential expression (DE) analysis* and with the help of H&E plot in Fig. 4A, we were able to identify regions of muscle, stroma, epithelial cells and immune cells. As shown in Fig. 4B-F, all methods except BayesSpace have good partitions for the muscle region, which can be visually verified in the H&E plot (Fig. 4A). The epithelial regions identified by BayesSpace is much smaller than those from SC-MEB, where a large proportion of epithelial region (3 o'clock in Fig. 4B-C) is classified as one of stroma regions (stroma 2) by BayesSpace. The immune regions identified by BayesSpace is larger at 9 and 12 o'clock but smaller at 6 o'clock in Fig. 4C than those identified by SC-MEB and GMM. Strikingly, a large proportion of regions identified as stroma 1 by both SC-MEB and GMM is classified as stroma 2 by BayesSpace. Even though stroma 1 and stroma 2 are both stroma regions but one can observe a clear difference in normalized expressions from signature genes for these two clusters (Fig. 5). These observations illustrate the possible oversmoothing behavior of the BayesSpace while SC-MEB is still able to recover the fine-structure of tissues.

3.4.1 Differential expression analysis for the identified clusters

As true labels for all spots are not available in the CRC dataset, we cannot quantitatively evaluate the clustering performance. For the clustering results of SC-MEB and BayesSpace, we further performed DE analysis between an identified cluster and all others using *BPSC* package [Vu et al., 2016] on log-normalized expressions. Using the partition results from SC-MEB, we identified 190, 154, 134, 102, 98, 89, 137 and 99 genes differentially expressed

for stroma 1 and 2, muscle, epithelial 1 and 2, and immune 1, 2, and 3, respectively, with false discovery rate (FDR) < 0.05. The details for all identified genes differentially expressed for SC-MEB and BayesSpace are provided in Supplementary Table S7 and S8. We further restrict the set of signature genes by choosing the ones with log fold change larger than 0.5. Finally, we got a total of 78 and 57 signature genes for SC-MEB and BayesSpace, respectively.

Fig. 5 shows the heatmap of normalized expressions for signature genes identified in the DE analysis for (A) SC-MEB and (B) BayesSpace, respectively. Clearly, for BayesSpace (Fig. 5B), normalized expressions of signature genes in stroma 2 can further be divided into two sub-clusters and the expression patterns in the second sub-cluster is much similar to that of stroma 1. This misclassification can also be observed by comparing Fig. 4B and Fig. 4C, where a large proportion of region identified as stroma 1 using SC-MEB is identified as stroma 2. The findings from SC-MEB demonstrated that stroma 1 and stroma 2 clusters as well as epithelial clusters followed by immune clusters were arranged in layers morphologically could be supported by the anatomical architecture of colonic tissue [Mills, 2019]. Namely, (from lumen to serosa) mucosal epithelium, lamina propria (immune cells are abundant and the isolated lymphoid nodules present in this tissue extend into the submucosal layer), submucosal which is the stromal layer with abundant connective tissues and lastly followed by muscularis externa which is represented as the muscle layer.

3.4.2 Pathway analysis for the signature genes identified in the DE analysis

We further conducted pathway analysis using gene ontology [Consortium, 2021] on signature genes for each identified cluster from SC-MEB. As listed in Supplementary Table S9, we show the top four pathways in each cluster. For the identified muscle region, muscle system development and muscle contraction were the top two significant pathways. For the three identified immune clusters, top significant pathways include humoral immune response and antimicrobial humoral response. For stroma 1 and 2 clusters, extracellular structure organization, extracellular matrix organization, and external encapsulating structure organization were among the top significant pathways. We also found a similar pattern in heatmap for normalized expressions of signature genes (Fig. 5) between stroma 1 and 2 clusters, among three immune clusters, and between epithelial 1 and 2 clusters. The cosine similarity between the two stroma clusters is high (0.97), as well as those among immune clusters, see Supplementary Table S10. We finally compare the mean expressions of COVID-19 signature genes [Lee et al., 2020] in the immune and non-immune regions identified by SC-MEB (Fig. 6), it is clear that COVID-19 signature genes are more expressed in the immune regions than the non-immune ones.

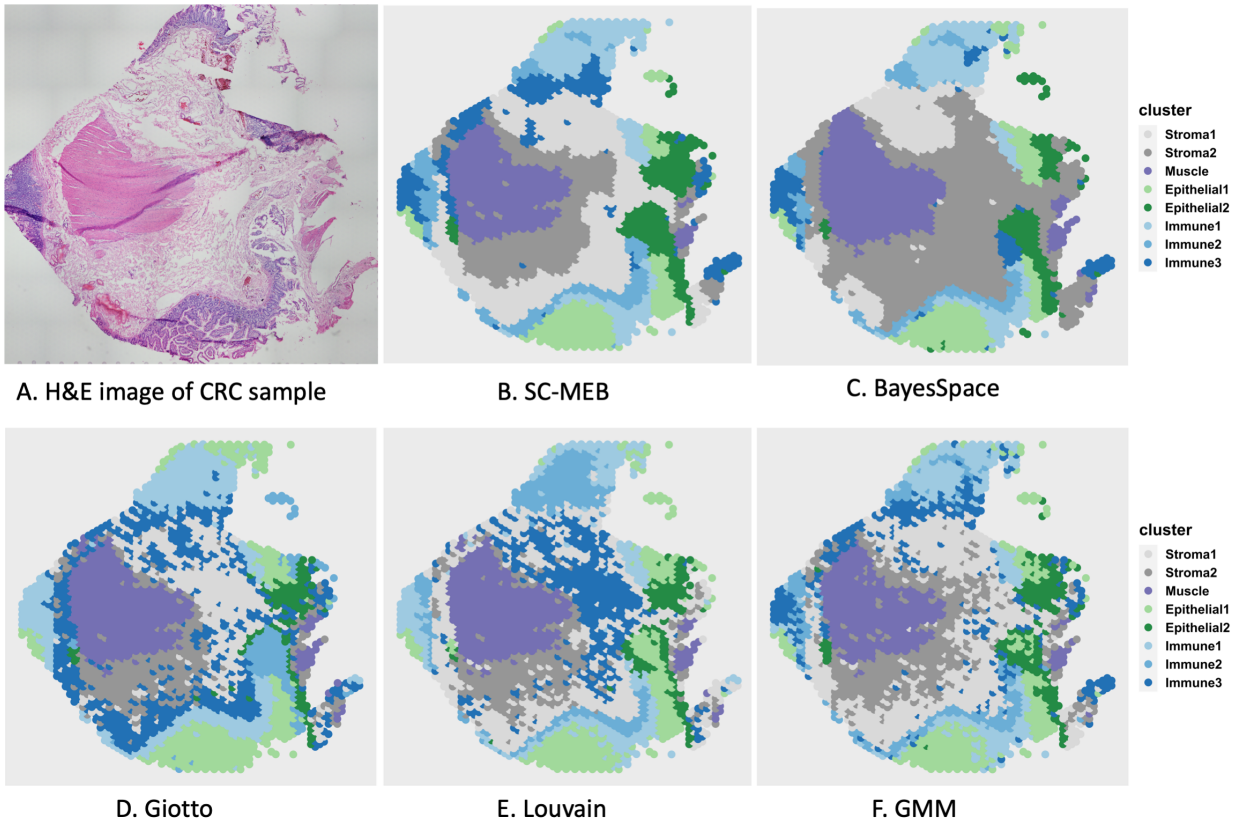

Figure 4: Clustering results for the CRC sample. (A) the original H&E stained tissue image for the CRC sample. (B) - (F) heatmaps for clustering assignments in the CRC sample using the proposed SC-MEB, BayesSpace, Giotto, Louvain and GMM, respectively. The eight clusters identified include two stroma regions, muscle region, two epithelial-cell regions, three immune-cell regions.

4 Competing interests

There is NO Competing Interest.

5 Author contributions statement

J.L. and J.Y. initiated and designed the study, X.S. and Y.Y. implemented the model and performed simulation studies and benchmarking evaluation, X.S. and J.L. wrote the manuscript, and all authors edited and revised the manuscript.

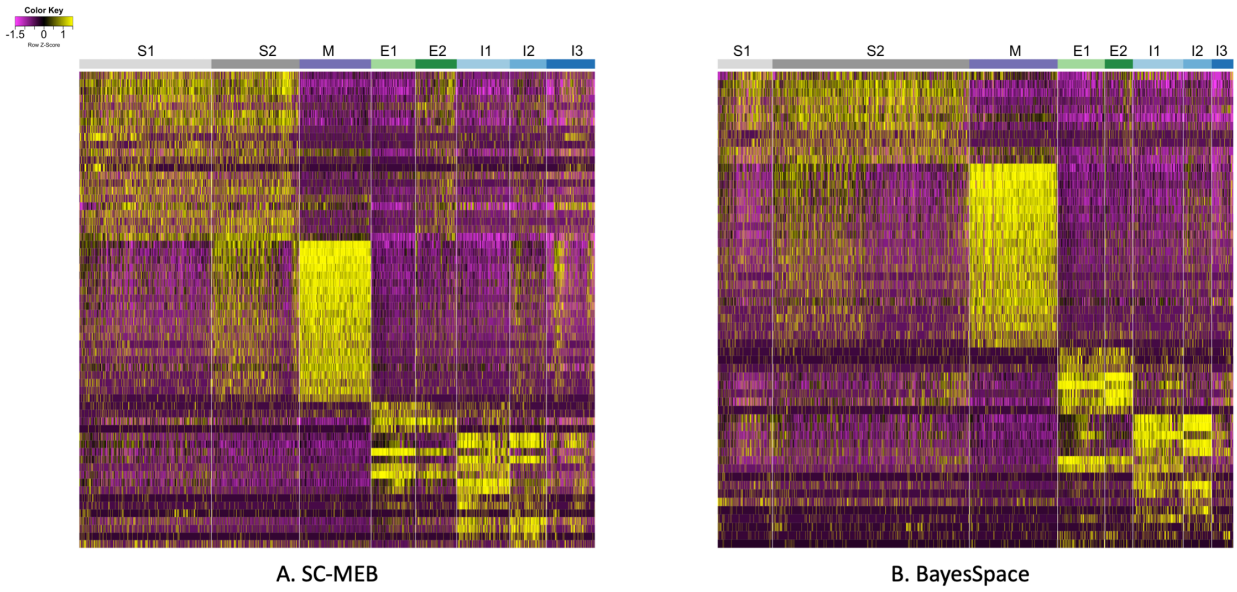

Figure 5: Heatmaps of normalized expressions for signature genes identified in the DE analysis based on two clustering analysis: (A) SC-MEB; (B) BayesSpace. In both subfigures, S1 and S2 are for Stroma 1 and 2, respectively, M is for Muscle, E1 and E2 are for Epithelial 1 and 2, respectively, and I1, I2 and I3 are for Immune 1, 2 and 3, respectively.

6 Acknowledgments

This work was supported by grant R-913-200-098-263 from the Duke-NUS Medical School, AcRF Tier 2 (MOE2018-T2-1-046 and MOE2018-T2-2-006) from the Ministry of Education, Singapore, grants (NSF: # 1636933 and # 1920920) from the National Science Foundation of China.

References

S. Alon, D. R. Goodwin, A. Sinha, A. T. Wassie, F. Chen, E. R. Daugharthy, Y. Bando, A. Kajita, A. G. Xue, K. Marrett, et al. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science*, 371(6528), 2021.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):259–279, 1986.

C. M. Bishop. *Pattern recognition and machine learning.* springer, 2006.

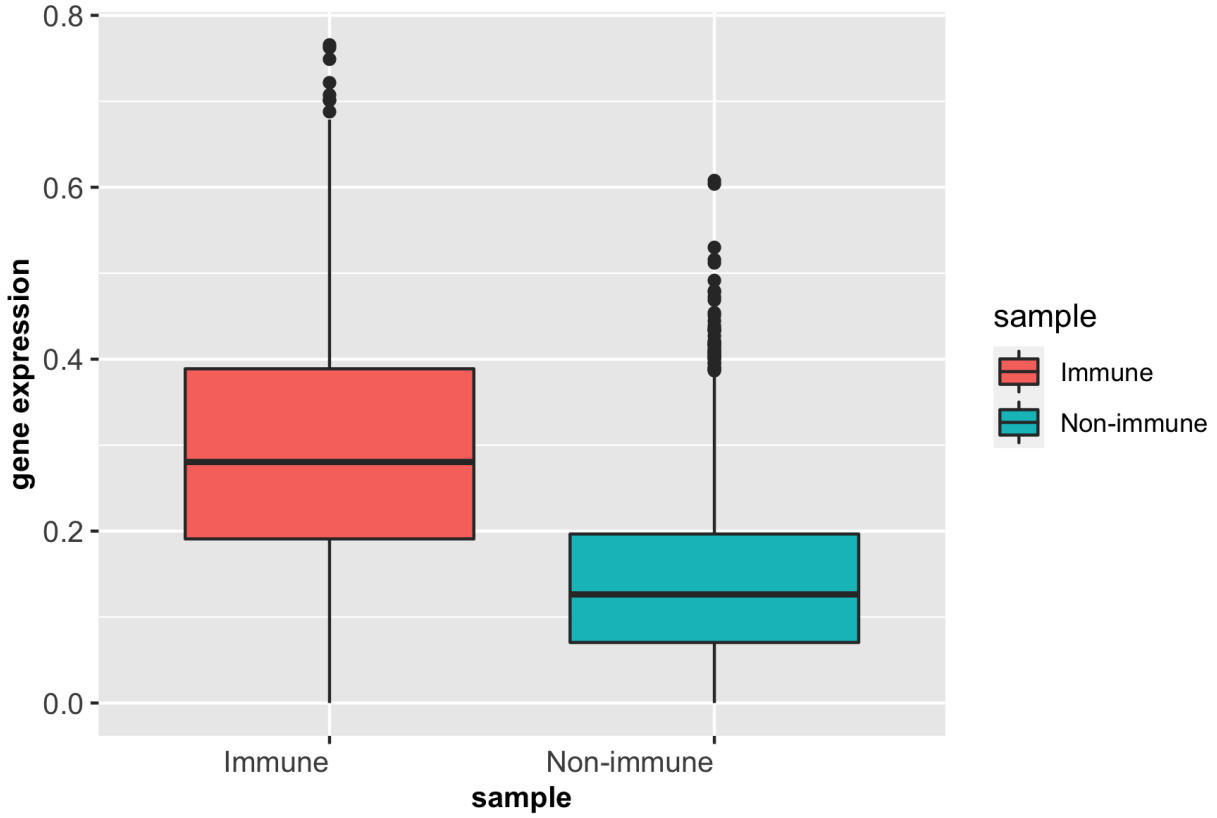

Figure 6: Boxplots of mean expressions of COVID-19 signature genes in the immune and non-immune regions.

D. J. Burgess. Spatial transcriptomics coming of age. *Nature Reviews Genetics*, 20(6): 317–317, 2019.

J. Chen, S. Suo, P. P. Tam, J.-D. J. Han, G. Peng, and N. Jing. Spatial transcriptomic analysis of cryosectioned tissue samples with geo-seq. *Nature protocols*, 12(3):566–580, 2017.

C. C. L. Cheung, D. Goh, X. Lim, T. Z. Tien, J. C. T. Lim, S. N. Nerurkar, S. Loong, P. C. Cheow, C. Y. Chan, Y. X. Koh, et al. Residual sars-cov-2 viral antigens detected in gastrointestinal and hepatic tissues from two recovered covid-19 patients. *medRxiv*, 2020.

G. Claeskens, N. L. Hjort, et al. Model selection and model averaging. *Cambridge Books*, 2008.

T. G. O. Consortium. The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021.

M. B. Cuadra, L. Cammoun, T. Butz, O. Cuisenaire, and J.-P. Thiran. Comparison and

validation of tissue modelization and statistical classification methods in t1-weighted mr brain images. *IEEE transactions on medical imaging*, 24(12):1548–1565, 2005.

R. Dries, Q. Zhu, C.-H. L. Eng, A. Sarkar, F. Bao, R. E. George, N. Pierson, L. Cai, and G.-C. Yuan. Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data. *BioRxiv*, page 701680, 2019.

C.-H. L. Eng, S. Shah, J. Thomassie, and L. Cai. Profiling the transcriptome with rna spots. *Nature methods*, 14(12):1153–1155, 2017.

F. Forbes and N. Peyrard. *Hidden Markov Models Selection Criteria based on Mean Field-like approximations*. PhD thesis, INRIA, 2002.

O. Franzén, L.-M. Gan, and J. L. Björkegren. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019, 2019.

L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

T. H. Kim, X. Zhou, and M. Chen. Demystifying "drop-outs" in single-cell umi data. *Genome biology*, 21(1):1–19, 2020.

H.-P. Kriegel, E. Schubert, and A. Zimek. The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems*, 52(2): 341–378, 2017.

J. S. Lee, S. Park, H. W. Jeong, J. Y. Ahn, S. J. Choi, H. Lee, B. Choi, S. K. Nam, M. Sa, J.-S. Kwon, et al. Immunophenotyping of covid-19 and influenza highlights the role of type i interferons in development of severe covid-19. *Science immunology*, 5(49), 2020.

T. Lohoff, S. Ghazanfar, A. Missarova, N. Koulena, N. Pierson, J. A. Griffiths, E. S. Bardot, C.-H. L. Eng, R. C. Tyser, R. Argelaguet, et al. Highly multiplexed spatially resolved gene expression profiling of mouse organogenesis. *bioRxiv*, 2020.

A. T. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, 17(1):1–14, 2016.

A. Lyubimova, S. Itzkovitz, J. P. Junker, Z. P. Fan, X. Wu, and A. Van Oudenaarden. Single-molecule mrna detection and counting in mammalian tissue. *Nature protocols*, 8 (9):1743, 2013.

K. R. Maynard, L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, J. L. Catallini, M. N. Tran, Z. Besich, M. Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, pages 1–12, 2021.

D. J. McCarthy, K. R. Campbell, A. T. Lun, and Q. F. Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33(8):1179–1186, 2017.

L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

S. Mills. *Histology for pathologists*. Lippincott Williams & Wilkins, 2019.

P. Qiu. Embracing the dropouts in single-cell rna-seq analysis. *Nature communications*, 11 (1):1–9, 2020.

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

S. Shah, E. Lubeck, W. Zhou, and L. Cai. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, 92(2):342–357, 2016.

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

V. Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.

V. A. Tolpekin and A. Stein. Quantification of the effects of land-cover-class spectral separability on the accuracy of markov-random-field-based superresolution mapping. *IEEE transactions on geoscience and remote sensing*, 47(9):3283–3297, 2009.

L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

T. N. Vu, Q. F. Wills, K. R. Kalari, N. Niu, L. Wang, M. Rantalainen, and Y. Pawitan. Beta-poisson model for single-cell rna-seq data analyses. *Bioinformatics*, 32(14):2128–2135, 2016.

L. N. Waylen, H. T. Nim, L. G. Martelotto, and M. Ramialison. From whole-mount to single-cell spatial assessment of gene expression in 3d. *Communications biology*, 3(1):1–11, 2020.

C. Xia, H. P. Babcock, J. R. Moffitt, and X. Zhuang. Multiplexed detection of rna using merfish and branched dna amplification. *Scientific reports*, 9(1):1–13, 2019.

Y. Zhang, M. Brady, and S. Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.

E. Zhao, M. R. Stone, X. Ren, T. Pulliam, P. Nghiem, J. H. Bielas, and R. Gottardo. Bayesspace enables the robust characterization of spatial gene expression architecture in tissue sections at increased resolution. *bioRxiv*, 2020.

Q. Zhu, S. Shah, R. Dries, L. Cai, and G.-C. Yuan. Identification of spatially associated subpopulations by combining scrnaseq and sequential fluorescence in situ hybridization data. *Nature biotechnology*, 36(12):1183–1190, 2018.

Supplementary: SC-MEB: spatial clustering with hidden Markov random field using empirical Bayes

1 SC-MEB Method

1.1 the Hidden Markov random field model

Let $\mathcal{S}$ be the set of all spots with a neighborhood system defined on it. The cell types in each spot can be modeled by a hidden random field

$$\{x = (x_1, \ldots, x_N) \mid x_i \in \mathcal{L}, i \in \mathcal{S}\},$$

where N is the number of spots, and $\mathcal{L} = \{1, \ldots, K\}$ is the set of all possible cell types. The p-dimensional representation of the gene expression is modeled by an observed random field

$$\{y = (y_1, \ldots, y_N) \mid y_i \in \mathbb{R}^p, i \in \mathcal{S}\}.$$

For spot i, its neighbors we considered are the spots surrounding it, and is denoted by $\mathcal{N}_i$.

The representation of the gene expression y is assumed to be a Gaussian MRF model, with two prefixed parameters, the number of cell types K, and the smoothing parameter $\beta(> 0)$. Specifically, the observed random field y are assumed to be mutually independent given the hidden random field x, and normally distributed with:

$$p(y \mid x, \theta) = \prod_{i \in \mathcal{S}} \mathcal{N}(y_i | x_i = k, \mu_k, \Sigma_k),$$

where $\theta = \{\mu_k, \Sigma_k : k = 1, \ldots, K\}$, μ_k and Σ_k denote the mean and Covariance matrix for cell type k. The hidden random field x are assumed to be the Potts model:

$$\begin{aligned} P(x) &= \frac{1}{Z_\beta} \exp\{-U(x)\}, \\ U(x) &= \sum_{i, i' \in \mathcal{N}_i} \beta[1 - \delta(x_i, x_i')], \end{aligned} \tag{1}$$

where δ is the delta function, and Z_β is a normalization constant that lacks a closed form.

With prefixed (K, β), the question of interest is to recover the unknown cell types x, interpreted as a clustering into a finite number K of types. This clustering usually is conducted based on the posterior distribution of x, and requires the estimation of parameter θ.

1.2 ICM-EM algorithm

Estimation of the parameter is done using an iterative conditional mode based expectation-maximization (ICM-EM) algorithm [Cuadra et al., 2005]. We assume (K, β) is known at present, and will discuss their selection in next section.

In ICM step, the estimate of x is obtained by maximizing its posterior with respect to x_i coordinately:

$$P(x \mid y) = P(x_i, x_{\mathcal{S}-\{i\}} \mid y) = P(x_i \mid y, x_{\mathcal{S}-\{i\}}) P(x_{\mathcal{S}-\{i\}} \mid y),$$

where $i = 1, \ldots, n$, until converge [Besag, 1986]. Given initial values of x, θ, and observed y, we have the update equation:

$$\hat{x}_i = \min_{x_i} U(\hat{x}_1, \ldots, x_i, \ldots, \hat{x}_n), \tag{2}$$

where

$$\begin{aligned} U(x) = & \Big\{ \frac{1}{2}(y_i - \mu_{x_i})^\top \Sigma_{x_i}^{-1} (y_i - \mu_{x_i}) + \frac{1}{2} \log |\Sigma_{x_i}| \\ & + \beta \sum_{i' \in \mathcal{N}_i} [1 - \delta(x_i, x_{i'})] \Big\}. \end{aligned}$$

In EM step, instead of the original complete likelihood which is difficult to perform the E-step, the following pseudo-likelihood is used:

$$\begin{aligned} \tilde{p}(y, x; \theta) &= p(y|x; \theta) \tilde{p}(x; \theta) \\ &= \prod_i p(y_i \mid x_i; \theta) \prod p(x_i | x_{\mathcal{N}_i}; \theta) \\ &= \prod_i [p(y_i \mid x_i; \theta) p(x_i | x_{\mathcal{N}_i}; \theta)] \\ &= \prod_i p(y_i, x_i \mid x_{\mathcal{N}_i}; \theta). \end{aligned}$$

For any $q_i(x_i)$, we have

$$\begin{aligned}
\log \tilde{p}(y;\theta) &= \sum_i \log \sum_k p(y_i, x_i = k \mid x_{\mathcal{N}_i};\theta) \\
&= \sum_i \log \sum_k q_i(x_i = k) \frac{p(y_i, x_i = k \mid x_{\mathcal{N}_i};\theta)}{q_i(x_i = k)} \\
&= \sum_i \log \mathbf{E}_{x_i} \left[\frac{p(y_i, x_i \mid x_{\mathcal{N}_i};\theta)}{q_i(x_i)} \right] \\
&\geq \sum_i \mathbf{E}_{x_i} \left[\log \frac{p(y_i, x_i \mid x_{\mathcal{N}_i};\theta)}{q_i(x_i)} \right] := \mathrm{ELBO}(\theta).
\end{aligned}$$

The above inequality is an equality if and only if

$$q_i(x_i) = \frac{p(y_i, x_i \mid x_{\mathcal{N}_i};\theta)}{\sum_{x_i} p(y_i, x_i \mid x_{\mathcal{N}_i};\theta)} = p(x_i \mid y_i, x_{\mathcal{N}_i}).$$

Then we have

$$\begin{aligned}
\mathrm{ELBO}(\theta) =& \sum_i \sum_k \gamma_{ik} \log p(y_i, x_i = k \mid x_{\mathcal{N}_i};\theta) - \sum_i \sum_k \gamma_{ik} \log \gamma_{ik} \\
=& \sum_i \sum_k \gamma_{ik} \left[\log p(y_i \mid x_i = k;\theta) + \log p(x_i = k \mid x_{\mathcal{N}_i};\theta) \right] \\
& - \sum_i \sum_k \gamma_{ik} \log \gamma_{ik}, \\
:=& Q(\theta) + \mathrm{Const},
\end{aligned}$$

where γ_{ik} is called responsibility that component k takes for explaining the observation y_i, and is defined as follows:

$$\gamma_{ik} = \frac{P(y_i \mid x_i = k) P(x_i = k \mid X_{\mathcal{N}_i} = \hat{x}_{\mathcal{N}_i})}{\sum_{k'} P(y_i \mid x_i = k') P(X_i = k' \mid X_{\mathcal{N}_i} = \hat{x}_{\mathcal{N}_i})}. \tag{3}$$

By taking partial derivatives of $\mathrm{ELBO}(\theta)$ with respect to the parameters and setting them to zero, we obtain the update equations for the maximization step:

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} y_i, \tag{4}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} (y_i - \mu_k)(y_i - \mu_k)^\top. \tag{5}$$

where $N_k = \sum_{i=1}^{n} \gamma_{ik}$.

The ICM-EM algorithm iterates the ICM step and maximization steps until convergence. Implementations are summarized in Algorithm 1 for clarity.

Algorithm 1: ICM-EM with prefixed (K, β)

Initialize $x, \theta = \{\mu_k, \Sigma_k : k = 1, \ldots, K\}$.
repeat
  ICM-step:
  while *the change of* $U(x)$ *is larger than a threshold* **do**
    for $i = 1, \ldots, n$ **do**
      $\hat{x}_i = \min_{x_i} U(\hat{x}_1, \ldots, x_i, \ldots, \hat{x}_n)$, ;
    end
  end
  E-step:
  for $i = 1, \ldots, n, k = 1, \ldots, K$ **do**
    $\gamma_{ik} = \frac{P(y_i|x_i=k)P(x_i=k|X_{\mathcal{N}_i}=\hat{x}_{\mathcal{N}_i})}{\sum_{k'} P(y_i|x_i=k')P(X_i=k'|X_{\mathcal{N}_i}=\hat{x}_{\mathcal{N}_i})}$;
  end
  M-step:
  for $k = 1, \ldots, K$ **do**
    $N_k = \sum_{i=1}^{n} \gamma_{ik}$;
    $\mu_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} y_i$;
    $\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik}(y_i - \mu_k)(y_i - \mu_k)^\top$;
  end
until *the change of* $\log \tilde{p}(y; \theta)$ *is smaller than a threshold*;
return $\hat{x}$ *and* θ.

1.3 Spatial clustering model selection with BIC

Since the number of cell types and the smoothing parameter β are not known in advance, we use the Bayesian Information Criterion (BIC) to choose them.

Let M_r denote a MRF model with prefixed (K_r, β_r), where $K_r = K_{\min}$ to $K_r = K_{\max}$, and $\beta_r \in \{2^{-2}, 2^{-1}, \ldots, 2^4\}$. Denote θ_r the parameters in M_r. BIC is defined as follows:

$$\mathrm{BIC}(\hat{\theta}_r) = 2 \log p(y \mid \hat{\theta}_r) - d_r \log n, \tag{6}$$

where d_r the number of free parameters in M_r, and equals to $K \times p + K \times \frac{p(p+1)}{2}$ in a K-labels MRF model. Here the logarithm of the observed likelihood $p(y \mid \hat{\theta}_r)$ is approximated with the $\log \tilde{p}(y; \hat{\theta}_r)$.

2 Simulation Results

2.1 Additional results for Example I

In Example I, lables of spots are radomly generated. We generated 4900 spatial spots defined on a 70×70 square lattices, and simulated the cluster label for each spot from the K-states Potts model with $\beta \in [1, 1.3]$ using the R package *GiRaF*. The number of neighbors is set to 4 and the number of true clusters K is set to be 3, 5 or 7. At last, two distributions are considered for PC representations of gene expression: a Gaussian mixture model, and a Student-t mixture model. The number of PCs was set to be either 10, or 15. The component mean μ_k and the component covariance matrix Σ_k are described in Supplementary Table S1-S4.

Figure S1. All component densities have different covariances, and the number of PCs was set to be 10.

Figure S2. All component densities have different covariances, and the number of PCs was set to be 15.

Figure S3. All components have a shared covariance, and the number of PCs was set to be 10.

Figure S4. All components have a shared covariance, and the number of PCs was set to be 15.

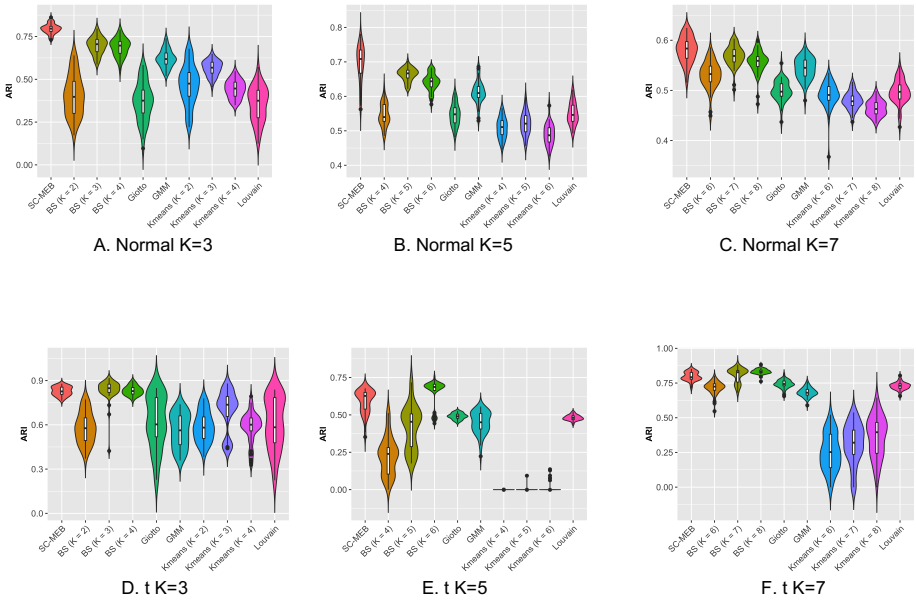

Figure S1: Additional results for Example I. All component densities have different covariances, and the number of PCs was set to be 10. **A-C.** PCs are sampled from a Gaussian mixture model, and the number of true clusters K is set to be be 3, 5 or 7. **D-F.** PCs are sampled from a Student-t mixture model, and the number of true clusters K is set to be be 3, 5 or 7.

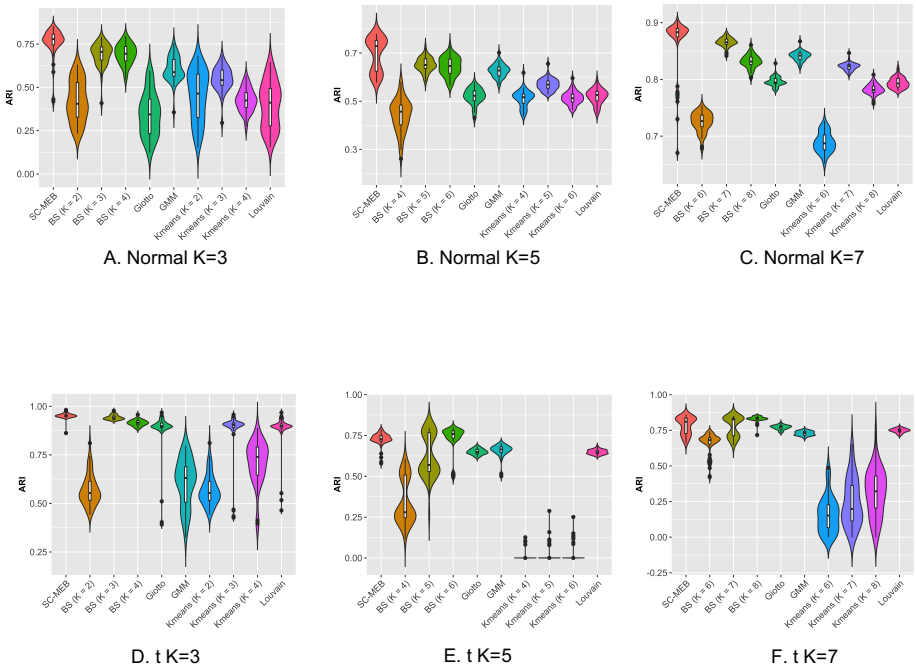

Figure S2: Additional results for Example I. All component densities have different covariances, and the number of PCs was set to be 15. **A-C.** PCs are sampled from a Gaussian mixture model, and the number of true clusters K is set to be be 3, 5 or 7. **D-F.** PCs are sampled from a Student-t mixture model, and the number of true clusters K is set to be be 3, 5 or 7.

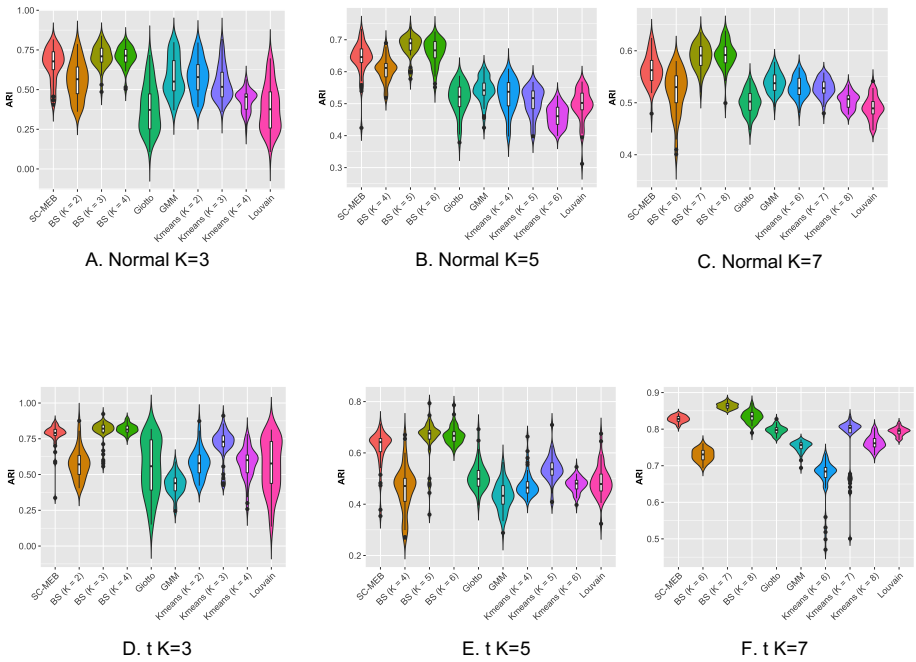

Figure S3: Additional results for Example I. All components have a shared covariance, and the number of PCs was set to be 10. **A-C.** PCs are sampled from a Gaussian mixture model, and the number of true clusters K is set to be be 3, 5 or 7. **D-F.** PCs are sampled from a Student-t mixture model, and the number of true clusters K is set to be be 3, 5 or 7.

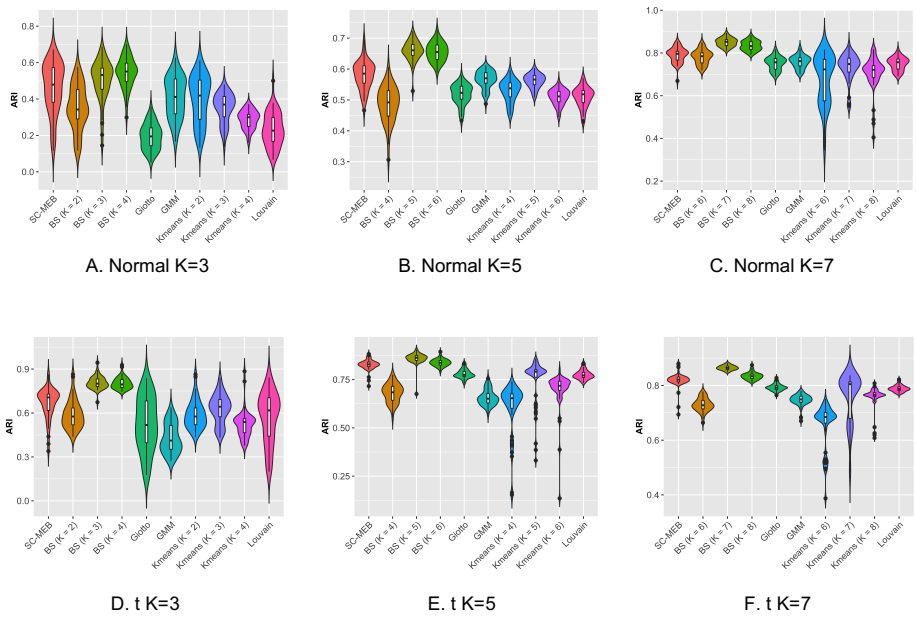

Figure S4: Additional results for Example I. All components have a shared covariance, and the number of PCs was set to be 15. **A-C.** PCs are sampled from a Gaussian mixture model, and the number of true clusters K is set to be be 3, 5 or 7. **D-F.** PCs are sampled from a Student-t mixture model, and the number of true clusters K is set to be be 3, 5 or 7.

2.2 Additional results for Example II

In Example II, lables of spots are obtained from our real data. Specifically, we obtained labels of 2,988 spots in the CRC data inferred from the corresponding real data with SC-MEB ($K = 8$). PCs were sampled in the same way as Example I. The component mean μ_k and the component covariance matrix Σ_k are described in Supplementary Table S5-S6.

Figure S5. The number of PCs was set to be 10.

Figure S5. The number of PCs was set to be 15.

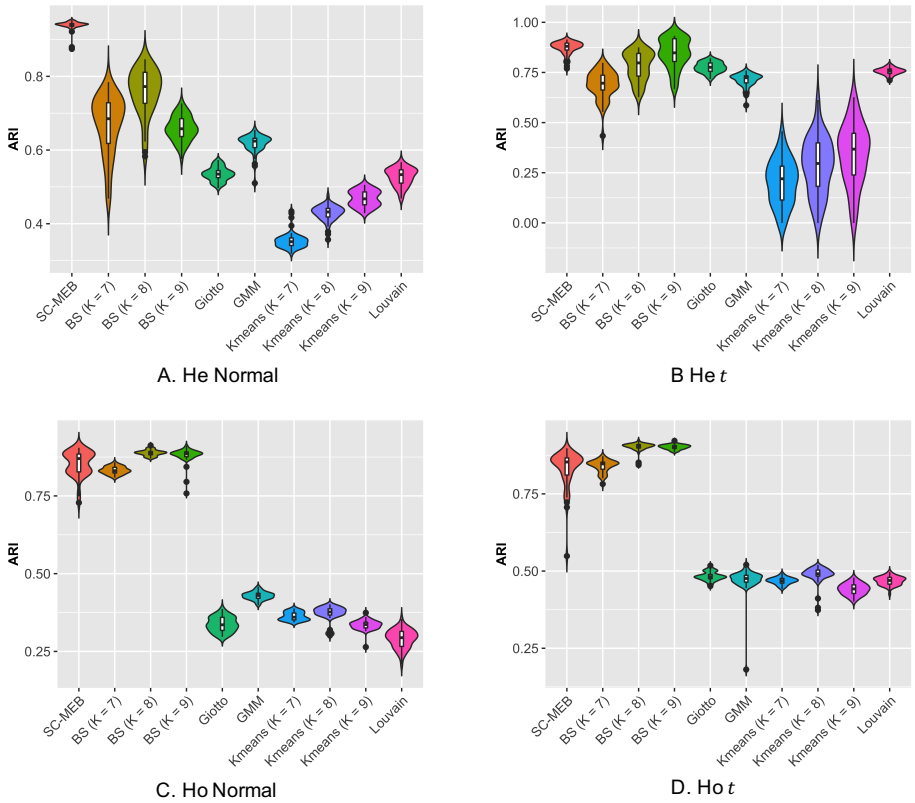

Figure S5: Additional results for Example II. The number of PCs was set to be 10. **A-B.** PCs are sampled from a Gaussian mixture model, and all component densities have different covariances. **C-D.** PCs are sampled from a Gaussian mixture model, and all components have a shared covariance.

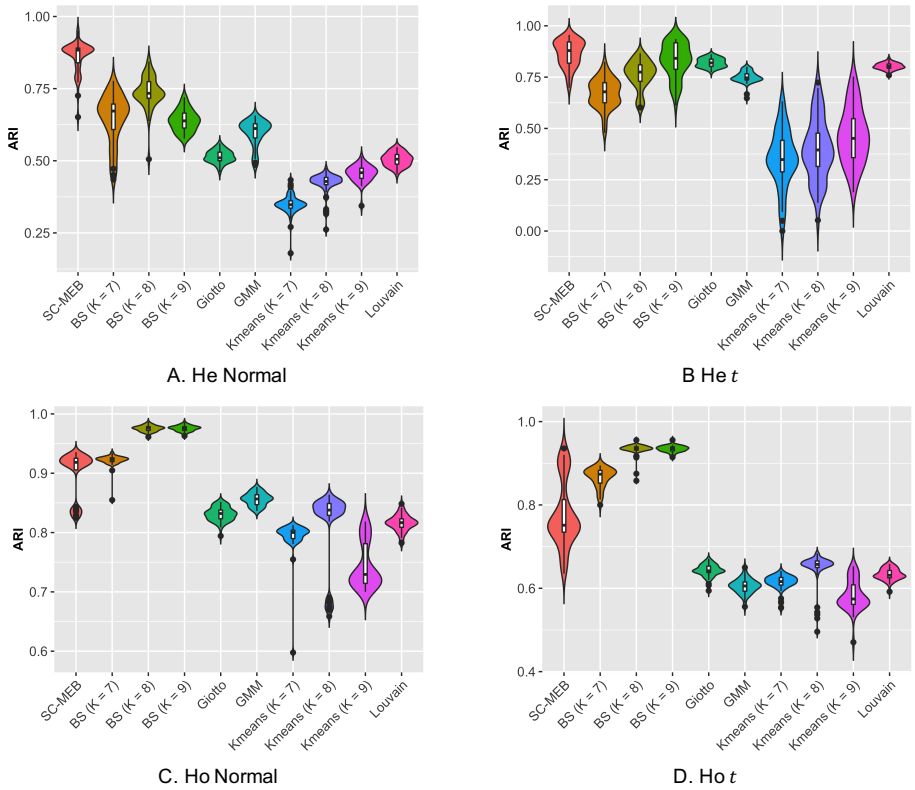

Figure S6: Additional results for Example II. The number of PCs was set to be 15. **A-B.** PCs are sampled from a Gaussian mixture model, and all component densities have different covariances. **C-D.** PCs are sampled from a Gaussian mixture model, and all components have a shared covariance.

References

Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):259–279, 1986.

Meritxell Bach Cuadra, Leila Cammoun, Torsten Butz, Olivier Cuisenaire, and J-P Thiran. Comparison and validation of tissue modelization and statistical classification methods in t1-weighted mr brain images. *IEEE transactions on medical imaging*, 24(12):1548–1565, 2005.