scRNA and Spots Data Analysis

Jinge Yu

Institute of Statistics
Renmin University of China

Introduction

- Explore heterogeneity at celluar level.
- Cluster cells using scRNA-seq.
- Spatial organization.
- Aims at finding a cluster method detecting the heterogeneity between different spots, and cluster different cells at the same time.

Data Preprocessing

- Dimensions. scRNA: 19831×2000, ST: 16528×996.
- Data normalization(remove scale effects).

  $$x = x * \frac{median\{ss\}}{ss}$$

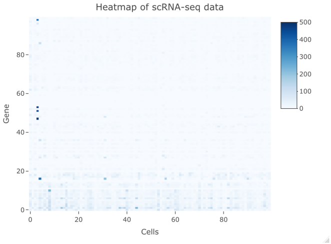
  where ss is sum of gene expression level of oen cell, x vector is the gene expresseion levels of one cell.
- Continuous transformation:

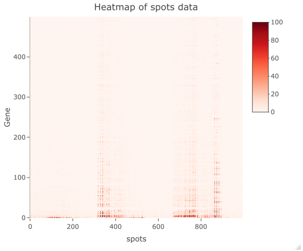
  $$x = log(2(x + 1))$$

  where x represents every single value in the data set.

Heatmap

We choose the most 1000 variable genes to plot maps and do data analysis, where 'the most variable' means the genes has larger standard deviation. Notice that we chose them based on normalization data instead of original ones. Plot heatmap of two data sets:

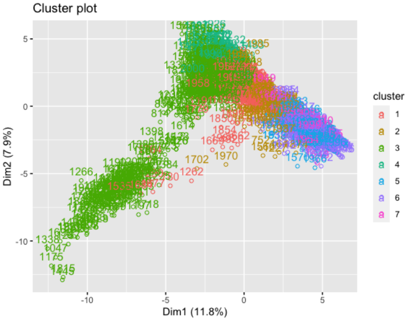


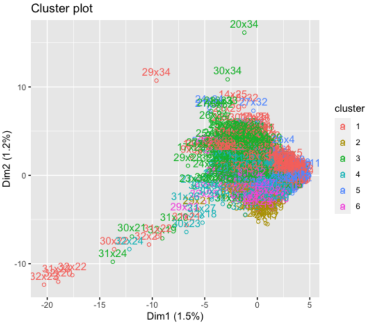


- Heatmaps revels that scRNA data is easier to cluster than spots data.
- Ordinary method may fail in identifying spots heterogeneity.

K-means Cluster

Try K-means cluster on both data sets.





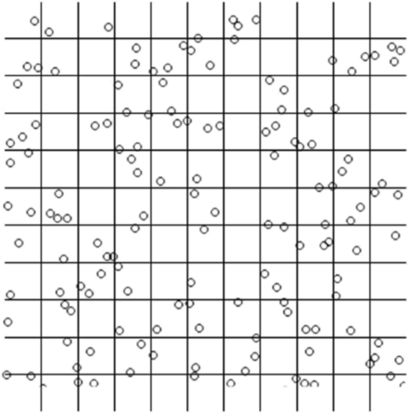
- K-means cluster did poor in both data sets, especailly in the spots data.

Spot

- scRNA-seq can not capture spatial features of cells, since the cells are dissolved one by one.
- We use spots to explore the spatial structure of cells.
- Spot can be viewed as a tiny region of several cells close to each other.
- The gene expression level in one spot is the sum of gene expression of all cells in the spot.

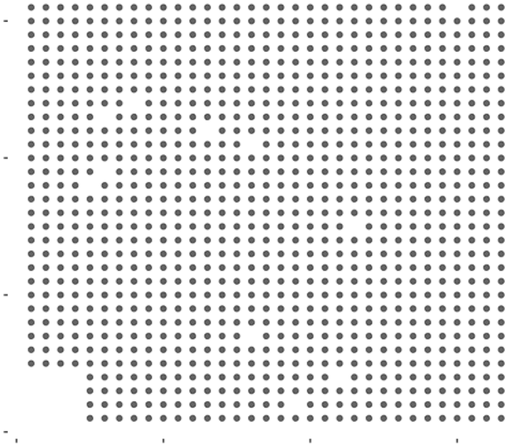
Spot

Sketch map of spots:

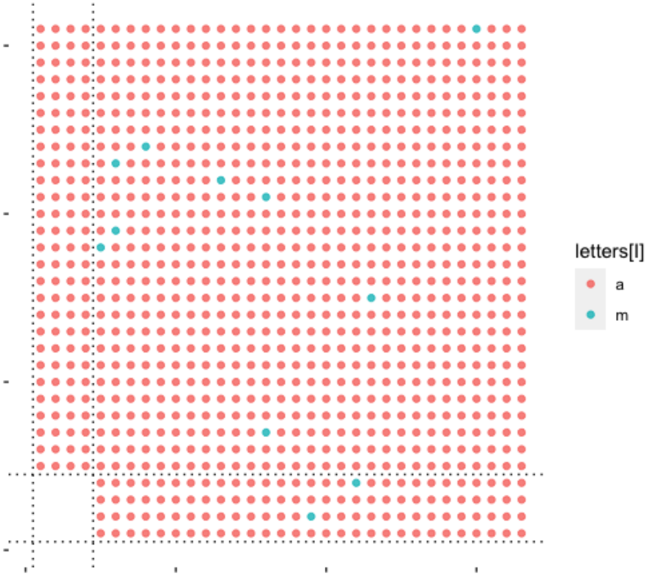


Spot

Visualization of spots in the ST data:



Spot



Null Value Spots

There are two types of null valu spots:

- Left bottom area, continuously ordered.
- Discretely spread.

For the sake of calculation simplication, we fill null value spots respectively in two ways.

- Set all gene expressions levels to 0, and then transform them to continuous values.
- Find the neighbors of null values spots and take the average value of the neighbors.

Notations

Notations	Explanation
M	Number of cells
N	Number of spots
L	Row number of spots
W	Column number of spots
G	Number of genes
K	Cell types
S	Region numbers

Model

As for scRNA data, we developed the following cluster method:

- Model:

$$Z_{gi} | C_i = k \quad \sim \quad N(h_{gk}, \sigma_g^2).$$

  Z_{gi} is the g-th gene expression level in cell i of scRNA data

  C_i represents cell type of the i-th cell.

- Unknown parameters:
  $C_i,\ h_{gk},\ \sigma_g^2, \quad i = 1, \ldots M,\ g = 1, \ldots, G, k = 1, \ldots, K.$

- Update h_{gk}, C_i and σ_g^2 in turn using Bayesian method.

Update parameters

Priors:

- $h_{gk} \sim N(\eta_h, \tau_h^2)$, $g = 1, \ldots, G, k = 1, \ldots, K$.
- $P(C_i = k) = \pi_k, \quad \pi_{1:K} \sim Dirichlet(\gamma_1, \ldots, \gamma_1)$.
- $\sigma_g^2 \sim Inv - gamm(\alpha_1, \beta_1)$.

Posteriors:

- $h_{gk}|- \sim N(\hat{h}_{gk}, (N_k/\sigma_g^2 + 1/\eta_h^2)^{-1})$, $g = 1, \ldots, G, k = 1, \ldots, K$.
  where $\hat{h}_{gk} = \frac{N_k \overline{Z_{gk}}/\sigma_g^2 + \eta_h/\tau_h^2}{N_k/\sigma_g^2 + 1/\eta_h^2}$,
  $N_k = \sum_{i=1}^M \mathbb{I}(C_i = k)$, $\overline{Z_{gk}} = \frac{1}{N_k} \sum_{i=1}^M Z_{gi} \mathbb{I}(C_i = k)$.
- $\frac{1}{\sigma_g^2}|- \sim gamma(\frac{M}{2} + \alpha_1, \beta_1 + \frac{1}{2} \sum_{i=1}^M (Z_{gi} - h_{c_i g})^2)$
- $P(C_i = k|-) = \frac{\prod_{g=1}^G N(Z_{gi} : h_{gk}, \sigma_g^2)}{\sum_{k=1}^K \prod_{g=1}^G N(Z_{gi} : h_{gk}, \sigma_g^2)}$

Remark

The notation '-' stands for given other paramters.

Potts Model

We consider potts model to obtaion spaital relations between spots.
Notatioins:

- (l, w), coordinate of a spot. We divide the spatial into a matrix of $L \times W$.
- R_{lw}, the region which spot (l, w) belongs to. $R = R_{lw}$, $1 \leq l \leq L$, $1 \leq w \leq W$. $R_{lw} \in \{1, 2, ..., S\}$
- $\Theta = (\theta ss')$, interaction energy matrix, whose diagnoal element are zero.
- $X = \{x_{lwg]} : 1 \leq l \leq L, \quad 1 \leq w \leq W, \quad 1 \leq g \leq G\}$, observed spots data.

Models:

- $x_{lwg}|R_{lw} = s \sim N(\mu_{sg}, \sigma_g'^2)$.
- $Pr(R|\Theta) = \frac{1}{C(\Theta)} exp\{-H(R|\Theta)\}$.
- $H(R|\Theta) = -\sum_{(l,w)\sim(l',w')} \theta_{R_{lw}R_{l'w'}} \mathbb{I}(R_{lw} \neq R_{l'w'})$
- $H(R_{lw}|R_{-l,-w}, \Theta) = -\sum_{(l',w')\in Nei(l,w)} \theta_{R_{lw}R_{l'w'}} \mathbb{I}(R_{lw} \neq R_{l'w'})$

Priors:

- $\theta ss' \sim N(\eta_\theta, \tau_\theta^2)$
- $\mu_{sg} \sim N(\eta_\mu, \tau_\mu^2)$
- $\sigma_g'^2 \sim$Inv-gamma(α_2, β_2)

Update parameters by Gibbs sampling:

- $\mu_{sg}(1 \leq s \leq S, \quad 1 \leq g \leq G)$
- $\sigma_g^2(1 \leq g \leq G)$
- $R_{lw}(1 \leq s \leq S, \quad 1 \leq w \leq W)$,
- $\Theta = (\theta ss'), \quad (1 \leq s \leq S, \quad 1 \leq s' \leq S, \quad s \neq s')$

Double MH

Intractable normalizing constant $C(\Theta)$ in the posterior distributioin of $\theta_{ss'}$.
We consider Double Metropolis–Hastings algorithm to sample $\theta_{ss'}$.
The DMH algorithm iterates between the following steps:

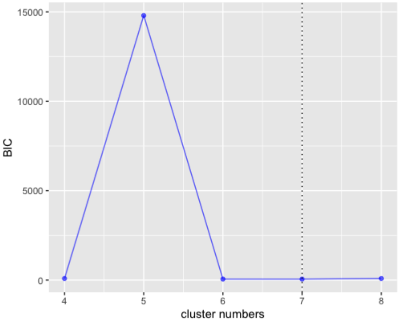
- Sample $\theta^\star_{ss'}$ from prposal distribution $N(\theta_{ss'}, \tau_0^2)$.
- Simulate auxiliary variable $y \sim \mathbb{P}_{\Theta^\star}(y|R)$.
- Calculate the MH ratio:

$$r = \frac{N(\theta^\star_{ss'}; \eta_\theta, \tau_\theta^2) N(\theta_{ss'}; \theta^\star_{ss'}, \tau_0^2) \mathbb{P}(R|\Theta^\star) \mathbb{P}(y|\Theta)}{N(\theta_{ss'}; \eta_\theta, \tau_\theta^2) N(\theta^\star_{ss'}; \theta_{ss'}, \tau_0^2) \mathbb{P}(R|\Theta) \mathbb{P}(y|\Theta^\star)}$$

- Accept $\theta^\star_{ss'}$ with probability $\min\{1, r\}$.

Results

We use BIC to define the best cluster number of both models. As for scRNA model:

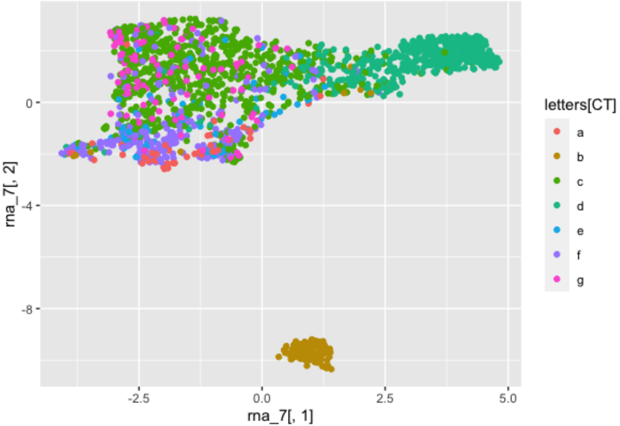


Clusters	4	5	6	7	8
BIC value	87.88	14794.44	56.19	55.55	88.09

So that 7 is the best cluster number.

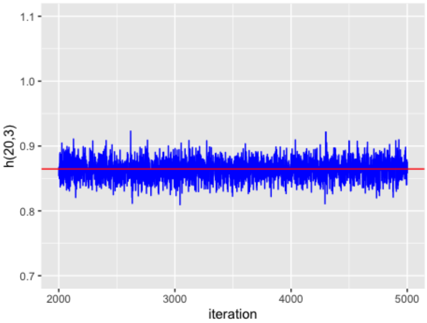
Cluster result

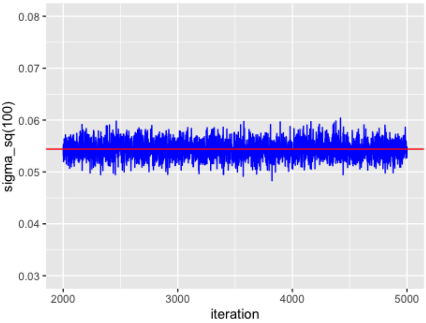
I used umap method to perform dimension reduction on genes, and the result is as follows:



Cluster index	1	2	3	4	5	6	7
Number of cells	48	152	796	548	90	268	98

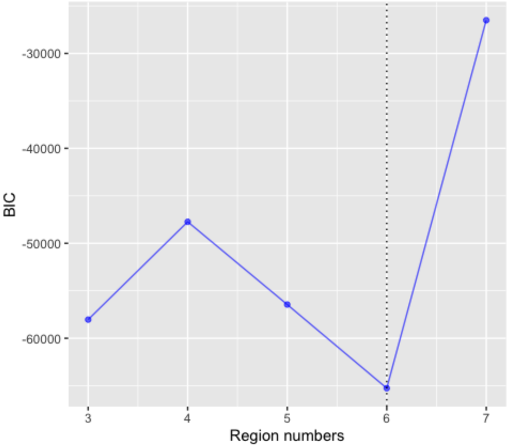
Iteration of parameters





Results

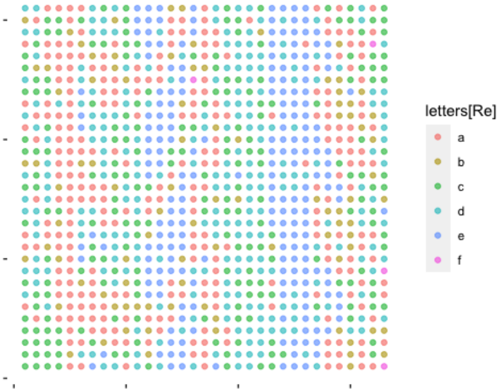
We use BIC to define the best cluster number of both models. As for spots model:



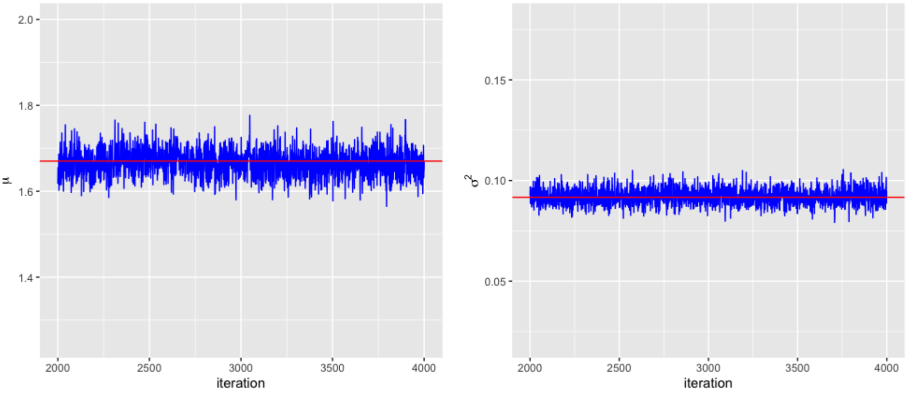
So that 5 is the best cluster number.

Cluster result

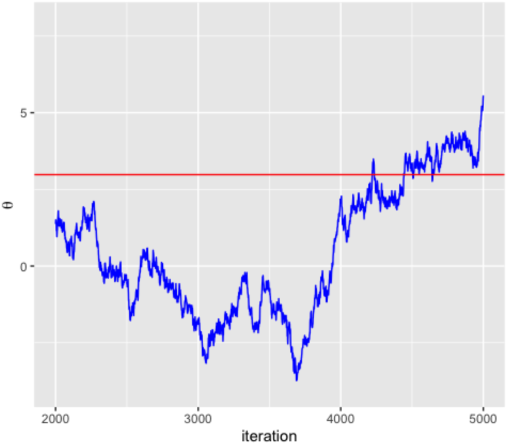
The result is as follows:



Iteration of parameters



Iteration of parameters



- It seems that Θ did not reach stable state after 5000 iterations. But we cares about *R* most, Θ is only an a auxiliary variable.

Summarize

- The two models have satisfied the demand of clustering in geeral, though the spatial one is less satisfactory.
- Take spatial orgination into consideration.
- Used Potts model and Double MH method when it comes to spatial models.
- However, every iteration took about 30-50 minutes.

Future Plan

- Find a link between scRNA-seq data and spots data, integrate single-cell RNA-seq methods and microarray-based spatial models.
- Improve the speed of iteration.
- Improve the performance of spatial model.
- Figure out the deep meaning of different regions.