

# Clustering of spatial scRNA-seq data

Edward Zhao

January 15, 2020

## 1 Ising model

### 1.1 Description

Let the expression  $y$  of cell  $i$  be determined by  $y_i = \mu I(z_i = 1) + \epsilon_i$  with priors

$$\begin{aligned}\mu|\mu_0, \lambda_0 &\sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \\ \epsilon_i|\lambda &\sim \mathcal{N}(0, \lambda^{-1}) \\ \lambda|\alpha, \beta &\sim \text{Gamma}(\alpha, \beta),\end{aligned}$$

and known hyperparameters

$$\begin{aligned}\mu_0 &= \bar{y} \\ \lambda_0 &= \frac{1}{100} \\ \alpha &= 1 \\ \beta &= 0.01.\end{aligned}$$

Then

$$y_i|z_i, \mu, \lambda \sim \mathcal{N}(\mu I(z_i = 1), \lambda^{-1}).$$

The conditional posterior distributions are given by

$$\mu|\mathbf{y}, \mathbf{z}, \lambda \sim \mathcal{N}\left(\frac{\lambda_0\mu_0 + \lambda \sum_{i=1}^n y_i I(z_i = 1)}{\lambda_0 + \lambda \sum_{i=1}^n I(z_i = 1)}, \left(\lambda_0 + \lambda \sum_{i=1}^n I(z_i = 1)\right)^{-1}\right) \quad (1)$$

$$\lambda|\mathbf{y}, \mathbf{z}, \mu \sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^n [y_i - \mu I(z_i = 1)]^2}{2}\right). \quad (2)$$

Let  $n$  be the total number of cells, arranged on a lattice. For every cell  $i \in \{1, \dots, n\}$ , the cell has state  $z_i \in \{-1, 1\}$ . Using the Ising model, the energy of the lattice system is given by

$$H(\mathbf{z}) = \frac{\gamma}{|\langle i j \rangle|} \sum_{\langle i j \rangle} z_i z_j,$$

and the probability of the system is given by

$$p(\mathbf{z}|\mathbf{y}) \propto \exp(H(\mathbf{z})) p(\mathbf{y}|\mathbf{z})$$

where  $\langle i j \rangle$  denotes nearest neighbors on the lattice and  $\gamma$  controls the magnitude of the dependence between neighbors.

The Metropolis-Hastings algorithm can be used to explore the lattice space, updating the state for one cell at a time. For each new proposal  $\mathbf{z}'$ , the acceptance probability  $\alpha$  is given by

$$\alpha(\mathbf{z}', \mathbf{z}) = \min \left\{ \frac{\exp(H(\mathbf{z}')) p(\mathbf{y}|\mathbf{z}')}{\exp(H(\mathbf{z})) p(\mathbf{y}|\mathbf{z})}, 1 \right\}. \quad (3)$$

The hyperparameters can be estimated using the following procedure:

1. Initialize  $\mu = \bar{y}$ ,  $\lambda = \frac{\alpha}{\beta} = 100$ ,  $z_i = -1 \forall i \in \{1, \dots, n\}$ .
2. Given  $\mathbf{y}, \mathbf{z}, \lambda$ , sample  $\mu$  from the conditional distribution given in (1).
3. Given  $\mathbf{y}, \mathbf{z}, \mu$ , sample  $\lambda$  from the conditional distribution given in (2).
4. Given  $\mathbf{y}, \{z_2, \dots, z_n\}, \mu, \lambda$ , sample  $z_1$  (transition from -1 to 1 or 1 to -1) with acceptance probability given by (3).
5. Repeat step 4 for all other indices of  $\mathbf{z}$ .
6. Repeat steps 2-5 for  $N$  iterations.

## 1.2 Simulation

All cells on a 100 by 100 lattice are set to have state -1, except for a radius 10 circle and side length 10 square that have state 1. For this simulation  $\mu = 5$  and  $\lambda = 0.5$ .

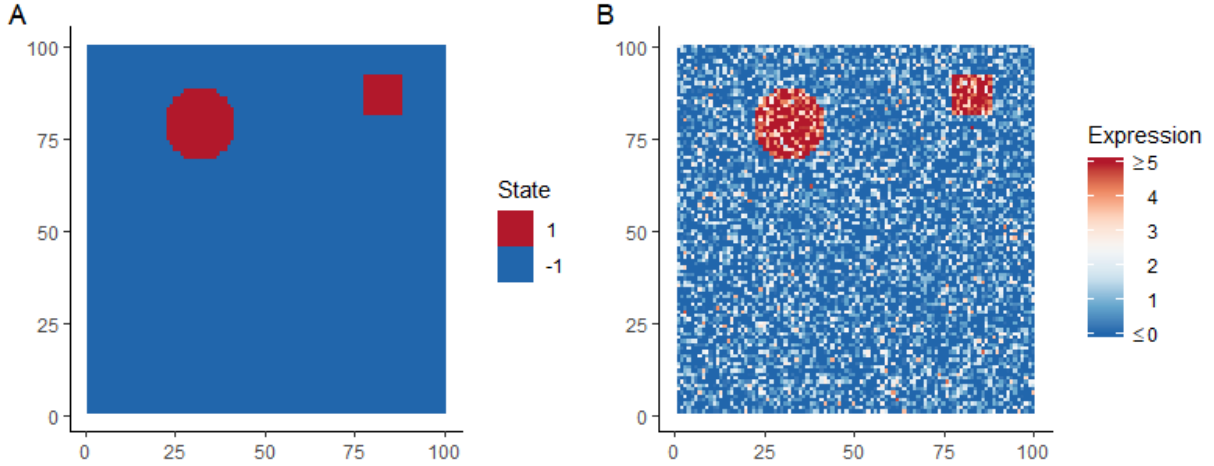


Figure 1: The ground truth (A) and simulated data (B).

One chain is generated for each of three values of  $\gamma$ : 2, 4, 6. 1000 iterations are generated, including a burn-in period of 100 iterations.

The bias is large when  $\gamma = 2$ , but negligible with a higher smoothing parameter.

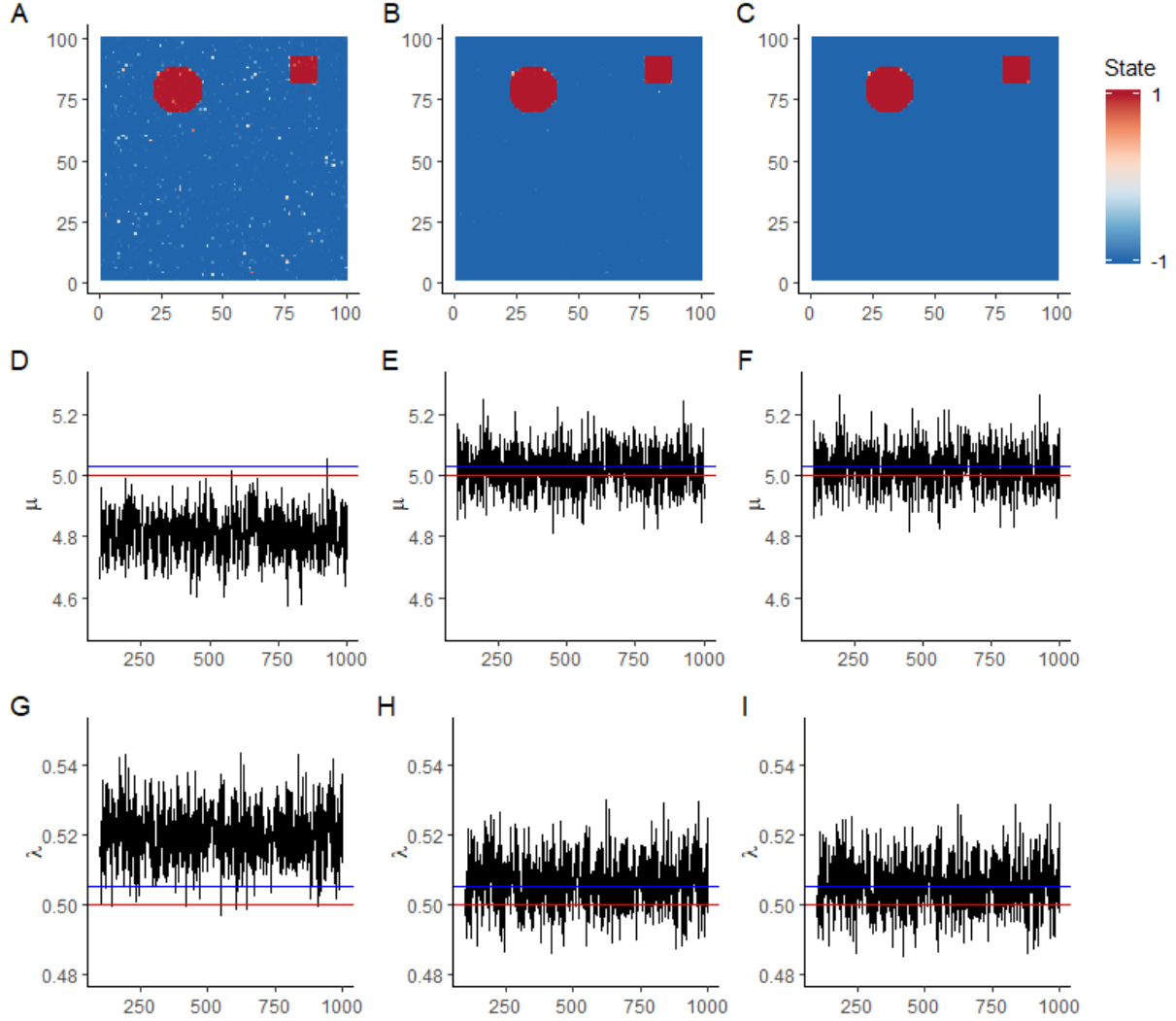


Figure 2: The left, middle, and right columns show the results for  $\gamma = 2$ ,  $\gamma = 4$ , and  $\gamma = 6$  respectively. A-C show the average state of each cell over 900 iterations. D-F are the trace plots for  $\mu$ , with the horizontal red line denoting  $\mu = 5$ . G-I are the trace plots for  $\lambda$ , with the horizontal red line denoting  $\lambda = 0.5$ . The horizontal blue lines denote the maximum likelihood estimates.

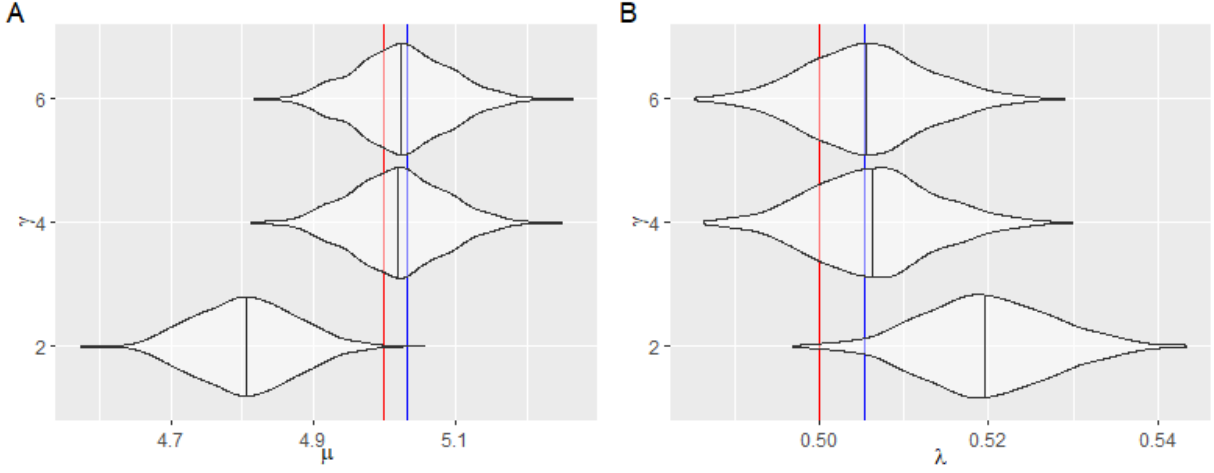


Figure 3: The posterior distributions of  $\mu$  (A) and  $\lambda$  (B) are shown for each value of  $\gamma$ . The vertical red and blue lines indicate the true values and maximum likelihood estimates respectively of  $\mu$  and  $\lambda$ .

## 2 Potts model

### 2.1 Description

The Potts model is a generalization of the Ising model that allows for more than two states. A cell's state can take values  $z_i \in \{1, \dots, q\}$ . The expression is determined by

$$y_i = \sum_{k=1}^q \mu_k I(z_i = k) + \epsilon_i,$$

with priors

$$\begin{aligned} \mu_k | \mu_0, \lambda_0 &\sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \quad \forall k \in \{1, \dots, q\} \\ \epsilon_i | \lambda &\sim \mathcal{N}(0, \lambda^{-1}) \\ \lambda | \alpha, \beta &\sim \text{Gamma}(\alpha, \beta), \end{aligned}$$

and known hyperparameters

$$\begin{aligned} \mu_0 &= \bar{y} \\ \lambda_0 &= \frac{1}{100} \\ \alpha &= 1 \\ \beta &= 0.01. \end{aligned}$$

Then

$$y_i | z_i, \boldsymbol{\mu}, \lambda \sim \mathcal{N}\left(\sum_{k=1}^q \mu_k I(z_i = k), \lambda^{-1}\right).$$

The conditional posterior distributions are given by

$$\mu_k | \mathbf{y}, \mathbf{z}, \lambda \sim \mathcal{N}\left(\frac{\lambda_0 \mu_0 + \lambda \sum_{i=1}^n y_i I(z_i = k)}{\lambda_0 + \lambda \sum_{i=1}^n I(z_i = k)}, \left(\lambda_0 + \lambda \sum_{i=1}^n I(z_i = k)\right)^{-1}\right) \quad (4)$$

$$\lambda | \mathbf{y}, \mathbf{z}, \boldsymbol{\mu} \sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{\sum_{k=1}^q \sum_{i=1}^n I(z_i = k) [y_i - \mu_k]^2}{2}\right). \quad (5)$$

Using the Potts model, the energy of the lattice system is given by

$$H(\mathbf{z}) = \frac{\gamma}{|\langle i j \rangle|} \times 2 \sum_{\langle i j \rangle} I(z_i = z_j) - 0.5,$$

and the probability of the system is given by

$$p(\mathbf{z}|\mathbf{y}) \propto \exp(H(\mathbf{z})) p(\mathbf{y}|\mathbf{z}).$$

For each proposal  $\mathbf{z}'$ , the Metropolis-Hastings acceptance probability is given by

$$\alpha(\mathbf{z}', \mathbf{z}) = \min \left\{ \frac{\exp(H(\mathbf{z}')) p(\mathbf{y}|\mathbf{z}')}{\exp(H(\mathbf{z})) p(\mathbf{y}|\mathbf{z})}, 1 \right\}. \quad (6)$$

The hyperparameters can be estimated using the following procedure:

1. Initialize  $\mu_k = \bar{y} \forall k$ ,  $\lambda = \left[ \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n} \right]^{-1}$ ,  $z_i = 1 \forall i \in \{1, \dots, n\}$ .
2. For each  $k$ , given  $\mathbf{y}, \mathbf{z}, \lambda$ , sample  $\mu_k$  from the conditional distribution given in (4).
3. Given  $\mathbf{y}, \mathbf{z}, \boldsymbol{\mu}$ , sample  $\lambda$  from the conditional distribution given in (5).
4. Given  $\mathbf{y}, \{z_2, \dots, z_n\}, \boldsymbol{\mu}, \lambda$ , sample  $z_1$  (transition from state  $k$  to state  $k' \in \{1, \dots, q\} \setminus \{k\}$ ) with acceptance probability given by (6).
5. Repeat step 4 for all other indices of  $\mathbf{z}$ .
6. Repeat steps 2-5 for  $N$  iterations.

## 2.2 Simulation

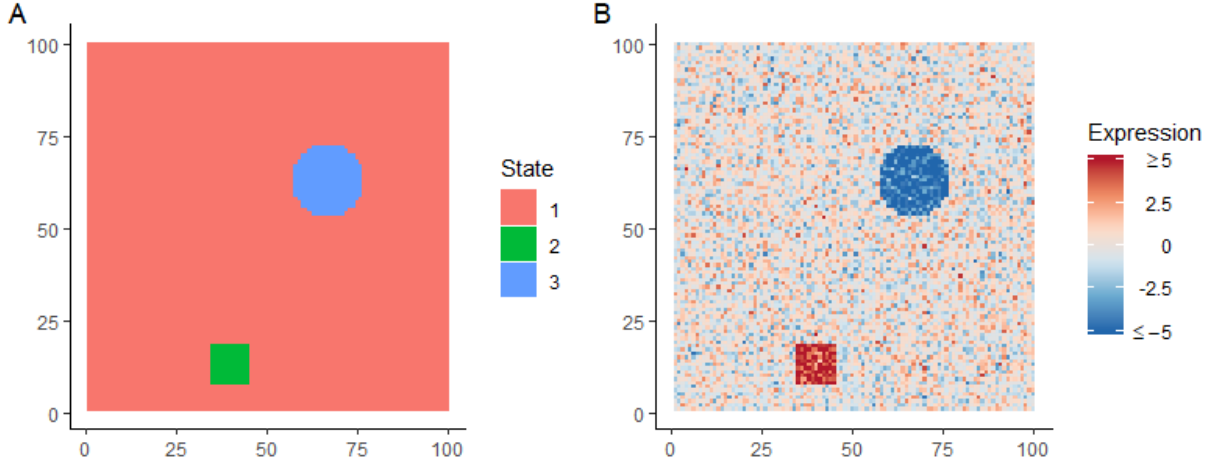


Figure 4: The ground truth (A) and simulated data (B).

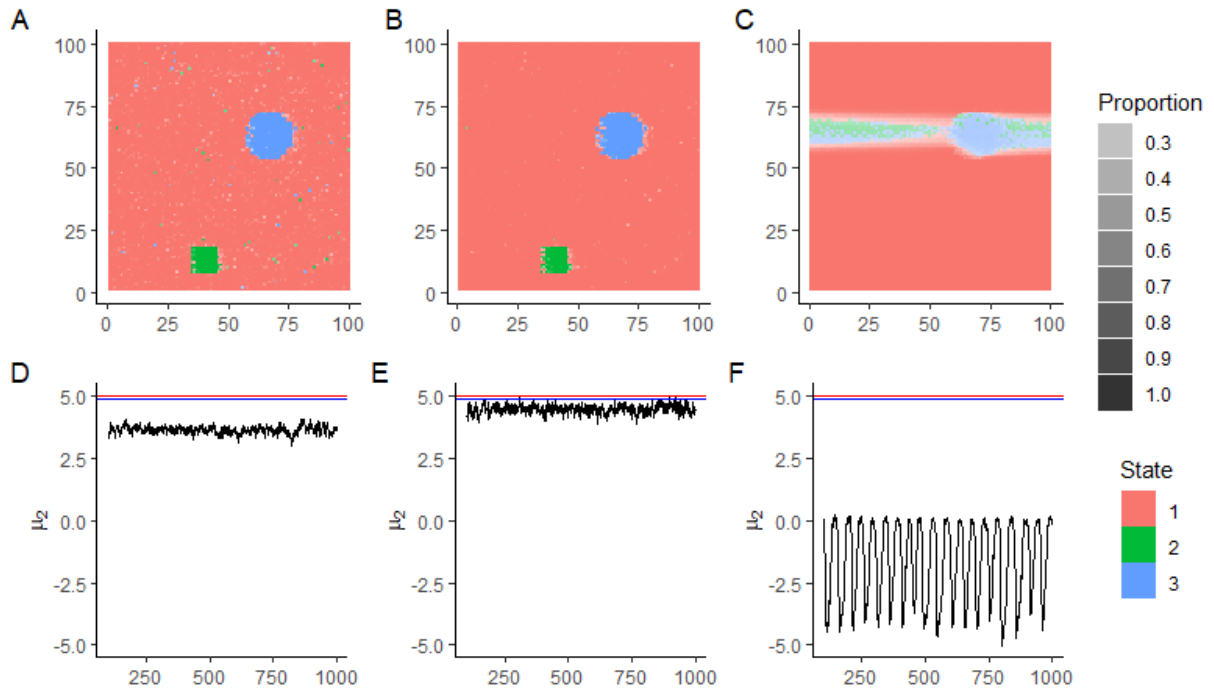


Figure 5: The left, middle, and right columns show the results for  $\gamma = 2$ ,  $\gamma = 3$ , and  $\gamma = 4$  respectively. A-C show the average state of each cell over 900 iterations. D-F are the trace plots for  $\mu_2$