Clustering of spatial scRNA-seq data

Edward Zhao

January 8, 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

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

and known hyperparameters

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

Then

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

The conditional posterior distributions are given by

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

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

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

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

and the probability of the system is given by

$$p(\boldsymbol{z}|\boldsymbol{y}) \propto \exp\left(H(\boldsymbol{z})\right) p(\boldsymbol{y}|\boldsymbol{z})$$

where $\langle i\ j\rangle$ denotes nearest neighbors on the lattice and γ 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 $\boldsymbol{z}'$, the acceptance probability α is given by

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

The hyperparameters can be estimated using the following procedure:

1. Initialize $\mu = \bar{y}$, $\lambda = \frac{\alpha}{\beta} = 100$, $z_i = -1 \; \forall \; i \in \{1, ..., n\}$.
2. Given $\boldsymbol{y}, \boldsymbol{z}, \lambda$, sample μ from the conditional distribution given in (1).
3. Given $\boldsymbol{y}, \boldsymbol{z}, \mu$, sample λ from the conditional distribution given in (2).
4. Given $\boldsymbol{y}, \{z_2, ..., z_n\}, \mu, \lambda$, 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 $\boldsymbol{z}$.
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 $\lambda = 0.5$.

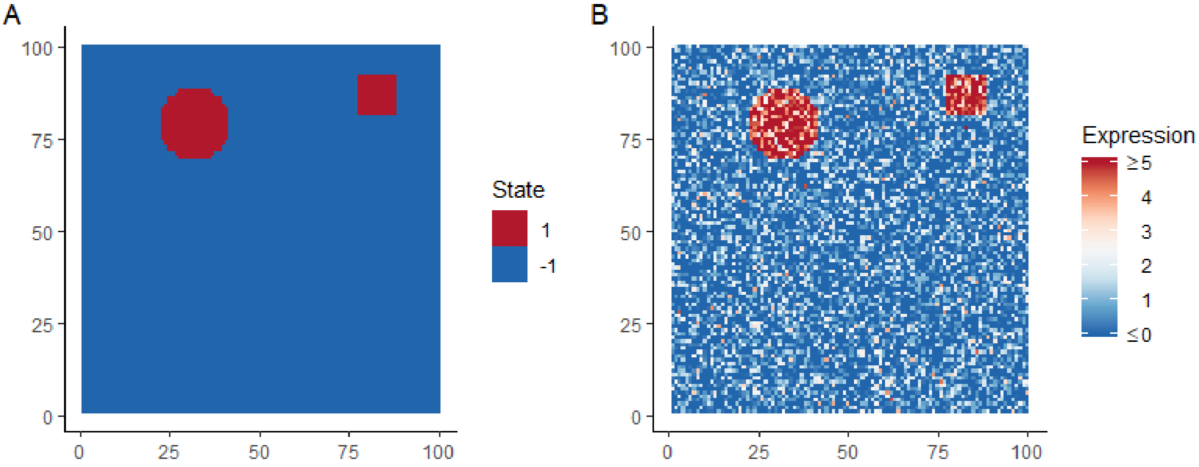

Figure 1: The ground truth (A) and simulated data (B).

One chain is generated for each of three values of γ: 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.

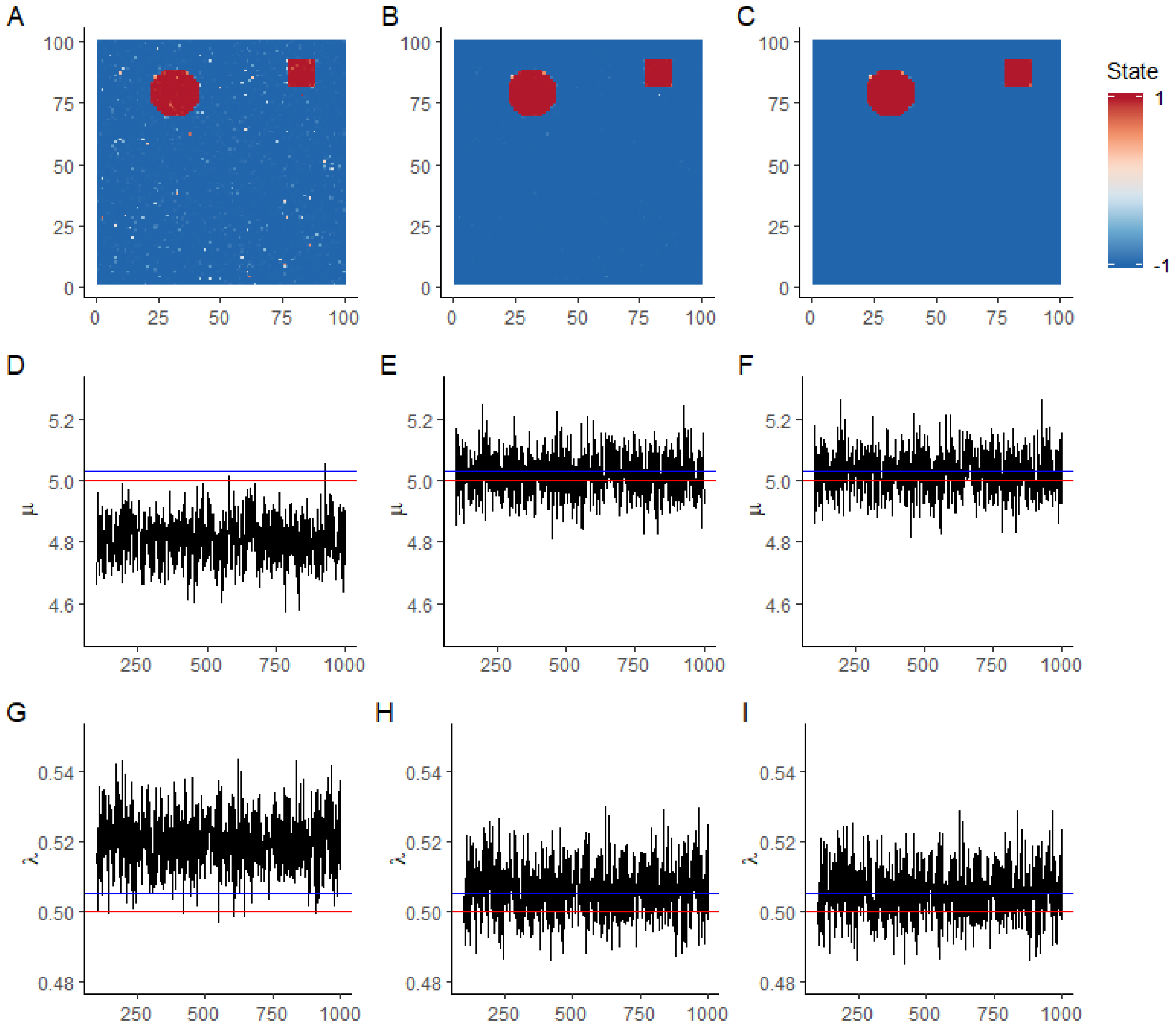

Figure 2: The left, middle, and right columns show the results for $\gamma = 2$, $\gamma = 4$, and $\gamma = 6$ respectively. A-C show the average state of each cell over 900 iterations. D-F are the trace plots for μ, with the horizontal red line denoting $\mu = 5$. G-I are the trace plots for λ, with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

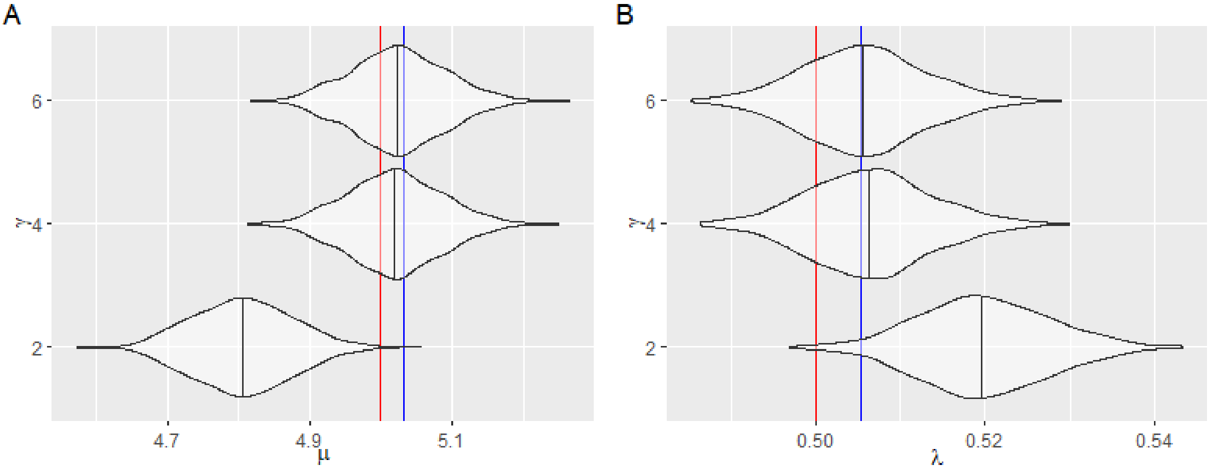

Figure 3: The posterior distributions of μ (A) and λ (B) are shown for each value of γ. The vertical red and blue lines indicate the true values and maximum likelihood estimates respectively of μ and λ.