Simulated Data Generative Model

SMP Gym

Simulated data are generated from three artificial spatial patterns and two real spatial patterns. For each dataset, there are $p = 100$ genes, among which $p_\gamma = 15$ are spatially variable (SV) genes.

To characterize the excess zeros and over-dispersion in the sequencing count data, we sample the observed gene expression counts y_{ij}'s from a zero-inflated negative binomial (ZINB) distribution,

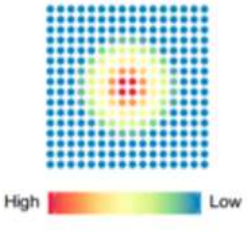
$$y_{ij} \sim \pi_i I(y_{ij} = 0) + (1 - \pi_i)\mathrm{NB}(s_i\lambda_{ij}, \phi_j).$$

The size factors s_i's are independent and identically distributed from $\mathrm{LN}(0, 0.2^2)$ and the dispersion parameters ϕ_j's are independent and identically distributed from $\mathrm{Exp}(0.1)$. For the choice of the false zero proportion π_i, we randomly select 0%, 10%, 30%, or 50% counts and force their values to zero.

For each gene j, the latent normalized expression level at spot i is generated via,

$$\log \lambda_{ij} = \begin{cases} \beta_0 + e_i + \varepsilon_{ij}, & \text{gene } j \text{ is an SV gene} \\ \beta_0 + \varepsilon_{ij}, & \text{gene } j \text{ is not an SV gene} \end{cases},$$

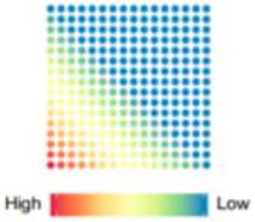
where β_0 is the baseline of log normalized expression levels and ε_{ij} is the non-spatial errors following $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$. We set $\beta_0 = 2$ and $\sigma = 0.3$. For a non-SV gene, the normalized expression levels are independent and identically distributed from a log-normal (LN) distribution with mean and variance being 2 and 0.3^2. Consequently, no spatial correlation should be observed. For SV genes with different patterns, the procedures to generate the spot-specific fold-change between SV and non-SV genes e_i are different.



Spot pattern

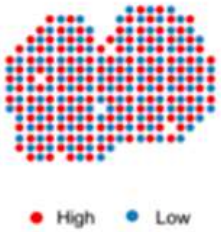
SV genes with the spot pattern, which is an artificial pattern, are on a 16×16 square lattice ($n = 256$ spots). The values of e_i 's of the four center spots, i.e., (8,8), (8,9), (9,8), and (9,9), is set to $\log 6$, while all others are linearly decreased to 0 within the radius of 5 spots.

Linear pattern



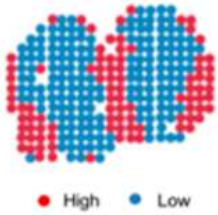
The artificial pattern, linear pattern, is on a 16×16 square lattice ($n = 256$ spots). For the linear pattern, the value of e_i's of the most bottom-left spot, i.e., (1,1), is set to $\log 6$, while all others are linearly decreased to 0 along the diagonal line.

MOB I pattern



MOB I pattern are on the $n = 260$ spots and location of spots is from the mouse olfactory bulb (MOB) study. This spatial pattern is generated based on the Ising model with positive interaction parameter between two states, which exhibits a complete attraction spatial pattern (the clustering of points with different expression levels). Each spot is dichotomized into high and low expression level groups with $e_i = \log 3$ and 0, respectively.

MOB II pattern



SV genes with the MOB II pattern are on the $n = 260$ spots. This pattern is constructed based on the spatial pattern identification results of various methods in the mouse olfactory bulb (MOB) study. Each spot is dichotomized into two group, high and low expression levels with $e_i = \log 3$ and 0, respectively.

BC pattern

SV genes with the BC pattern are on the $n = 250$ spots. This pattern is constructed based on the spatial pattern identification results of various methods in the breast cancer (BC) study. Each spot is dichotomized into two group, high and low expression levels with $e_i = \log 3$ and 0, respectively.

Combined with the five patterns (i.e., spot, linear, MOB I, MOB II, and BC) and four zero-inflation settings, there were $5 \times 4 = 20$ different scenarios in total. For each of the scenarios, we independently repeat the above steps to generate 30 datasets.