Clustering of spatial scRNA-seq data

Edward Zhao

January 27, 2020

1 Ising model

[hidden]

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 $z_i \in \{1, ..., q\}$. The expression is determined by

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

with priors

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

and known hyperparameters

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

Then

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

The conditional posterior distributions are given by

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

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

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

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

and the probability of the system is given by

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

For each proposal $\boldsymbol{z}'$, the Metropolis-Hastings acceptance probability 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 $\mu_k = \bar{y}\ \forall\ k$, $\lambda = \left[\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}\right]^{-1}$, $z_i = 1\ \forall\ i \in \{1, ..., n\}$.
2. For each k, given $\boldsymbol{y}, \boldsymbol{z}, \lambda$, sample μ_k from the conditional distribution given in (1).
3. Given $\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\mu}$, sample λ from the conditional distribution given in (2).
4. Given $\boldsymbol{y}, \{z_2, ..., z_n\}, \boldsymbol{\mu}, \lambda$, sample z_1 (transition from state k to state $k' \in \{1, ..., q\} \setminus \{k\}$) 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.

2.2 Simulation

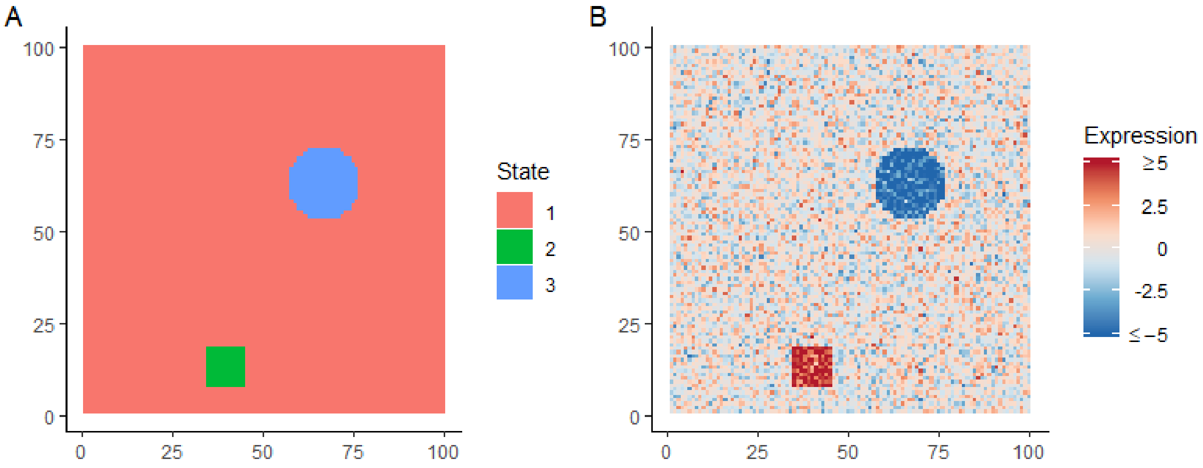

Figure 1: The ground truth (A) and simulated data (B).

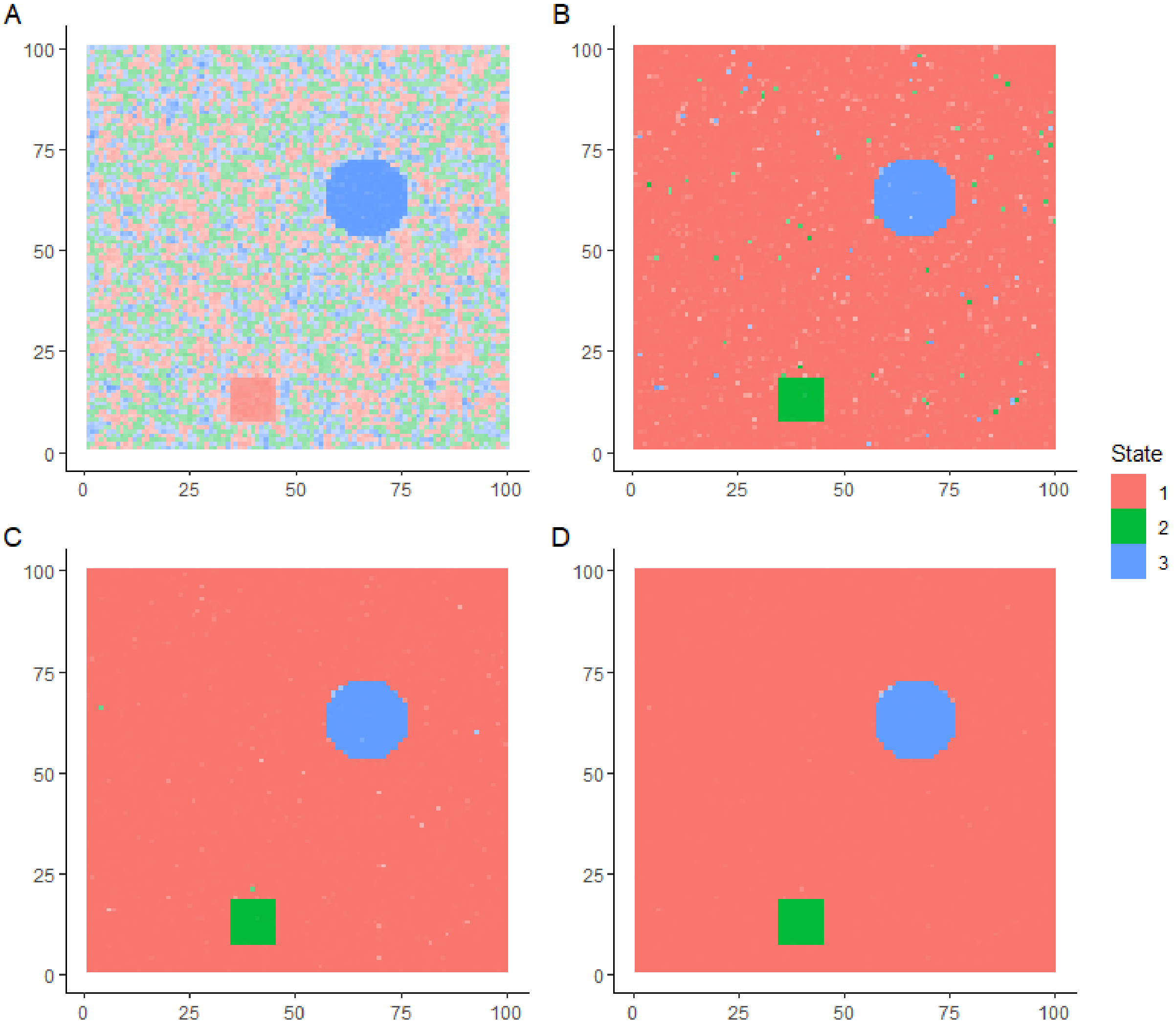

Figure 2: The average state of each cell over 900 iterations are shown for $\gamma = 1$ (A), $\gamma = 2$ (B), $\gamma = 3$ (C), and $\gamma = 4$ (D). The cell's color is determined by the most frequent state for the cell and the saturation is determined by the percentage of iterations the cell was in its most frequent state.

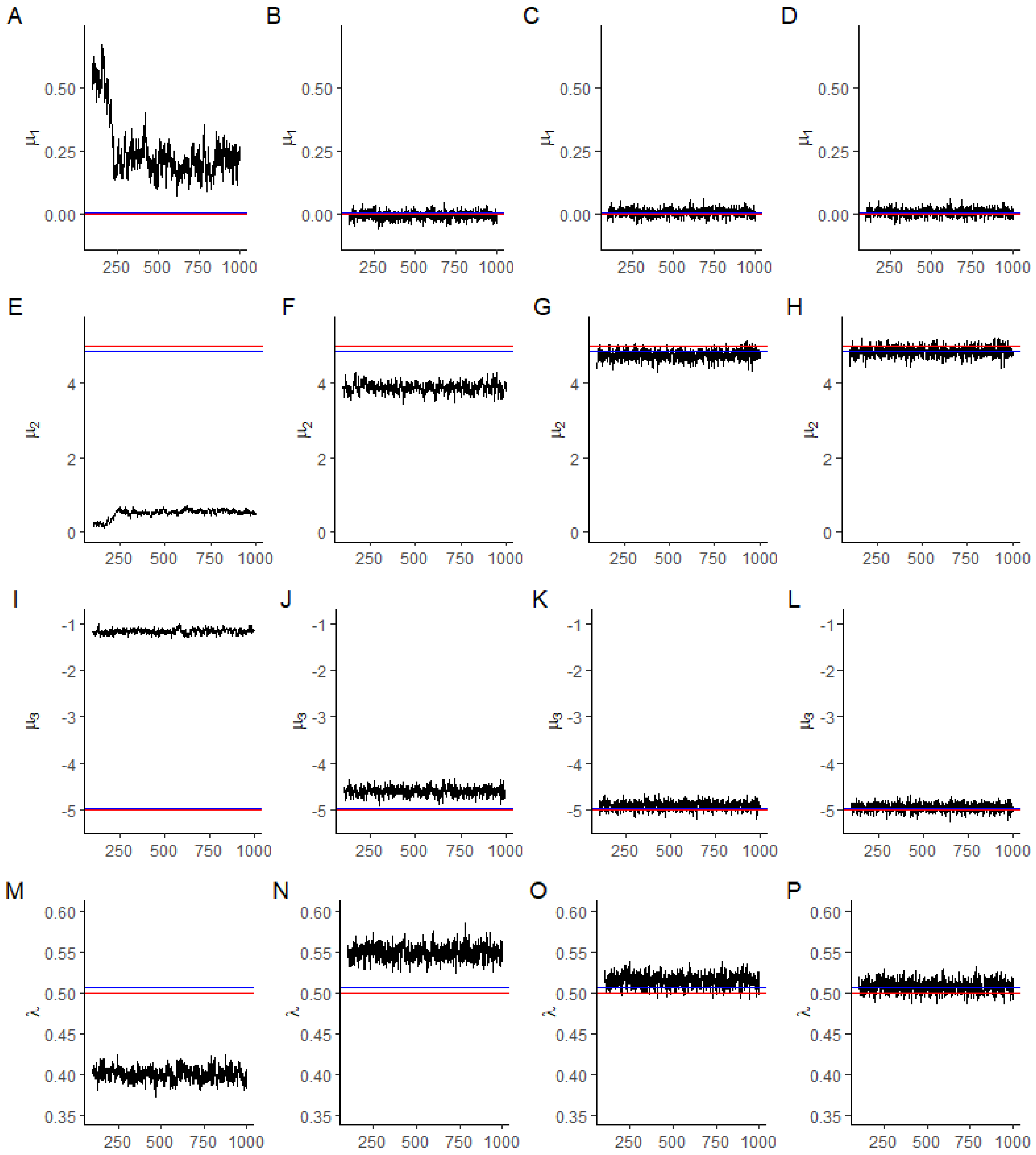

Figure 3: From left to right, the columns show the results for $\gamma = 1$, $\gamma = 2$, $\gamma = 3$, and $\gamma = 4$ respectively. A-D are the trace plots for μ_1, with the horizontal red line denoting $\mu_1 = 0$. E-H are the trace plots for μ_2, with the horizontal red line denoting $\mu_2 = 5$. I-L are the trace plots for μ_3, with the horizontal red line denoting $\mu_3 = -5$. M-P are the trace plots for λ, with the horizontal red line denoting $\lambda = 0.5$. The horizontal blue lines denote the maximum likelihood estimates.

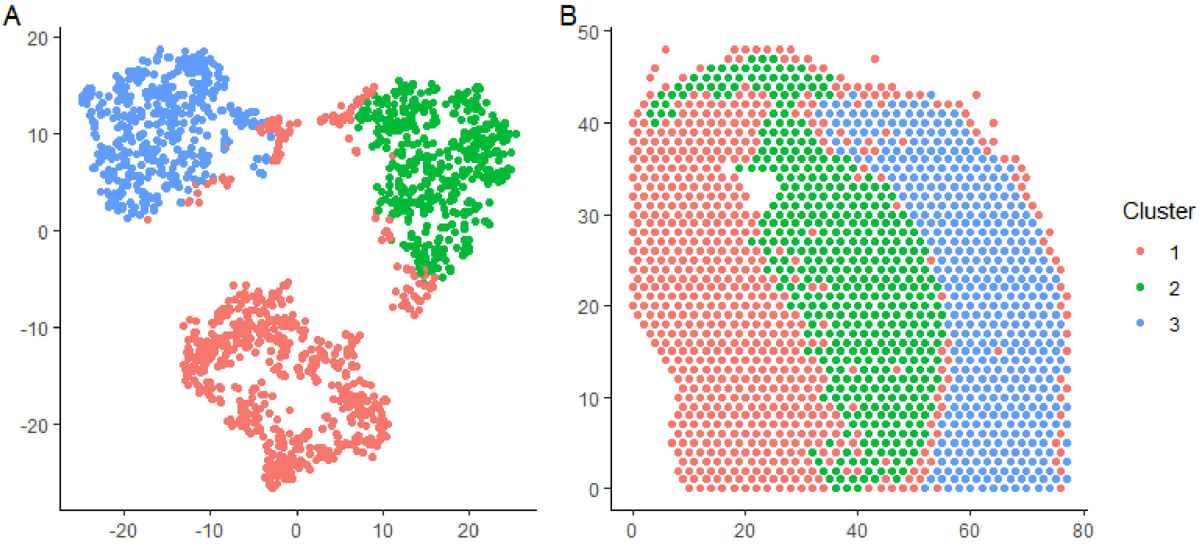

Figure 4: clustering by k=100 nearest neighbors A. tSNE B. spatial, MOUSE example

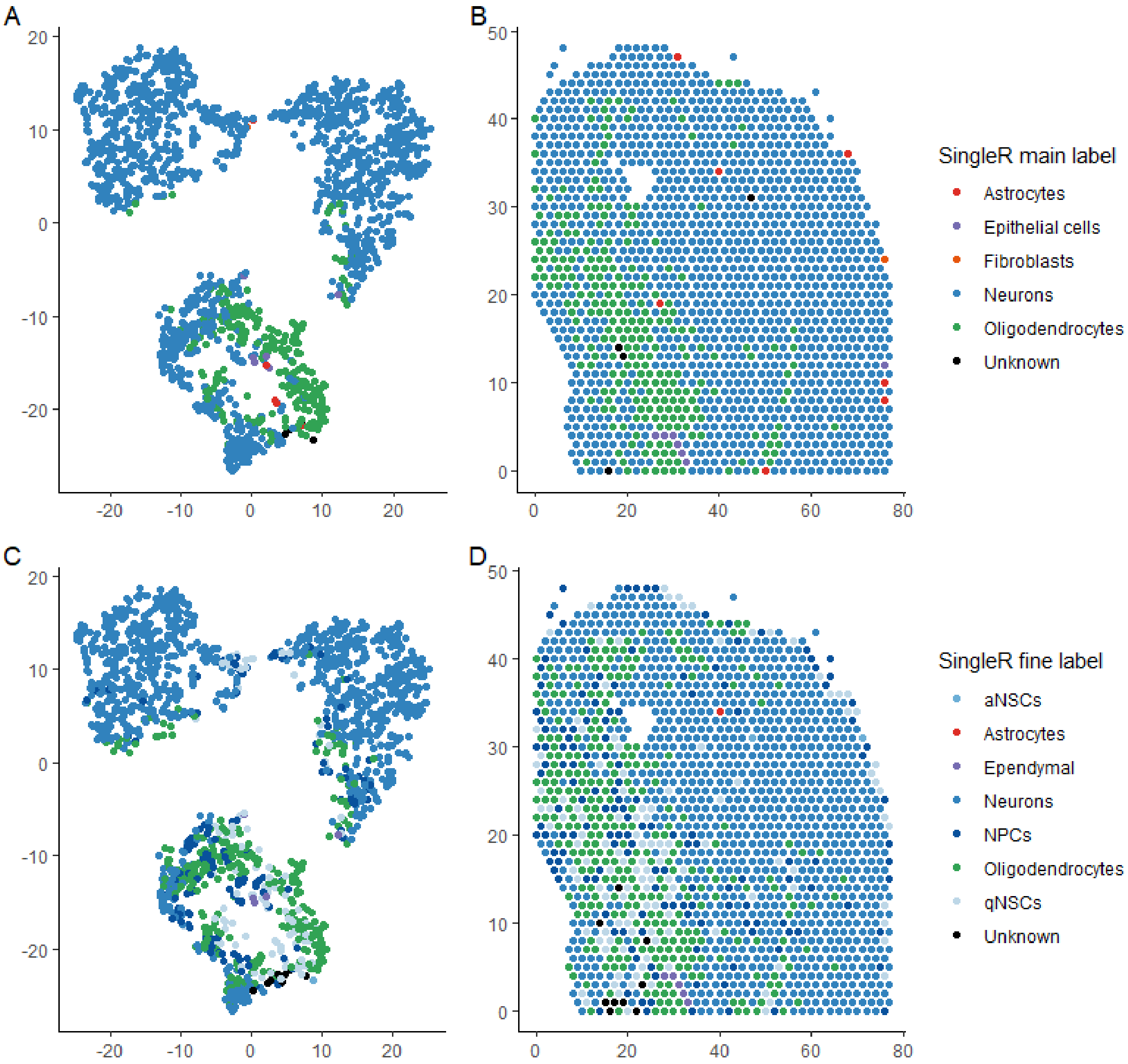

Figure 5: main labels (A,B) and fine labels(C,D) from SingleR using Benayoun 2019 mouse bulk RNA-seq reference

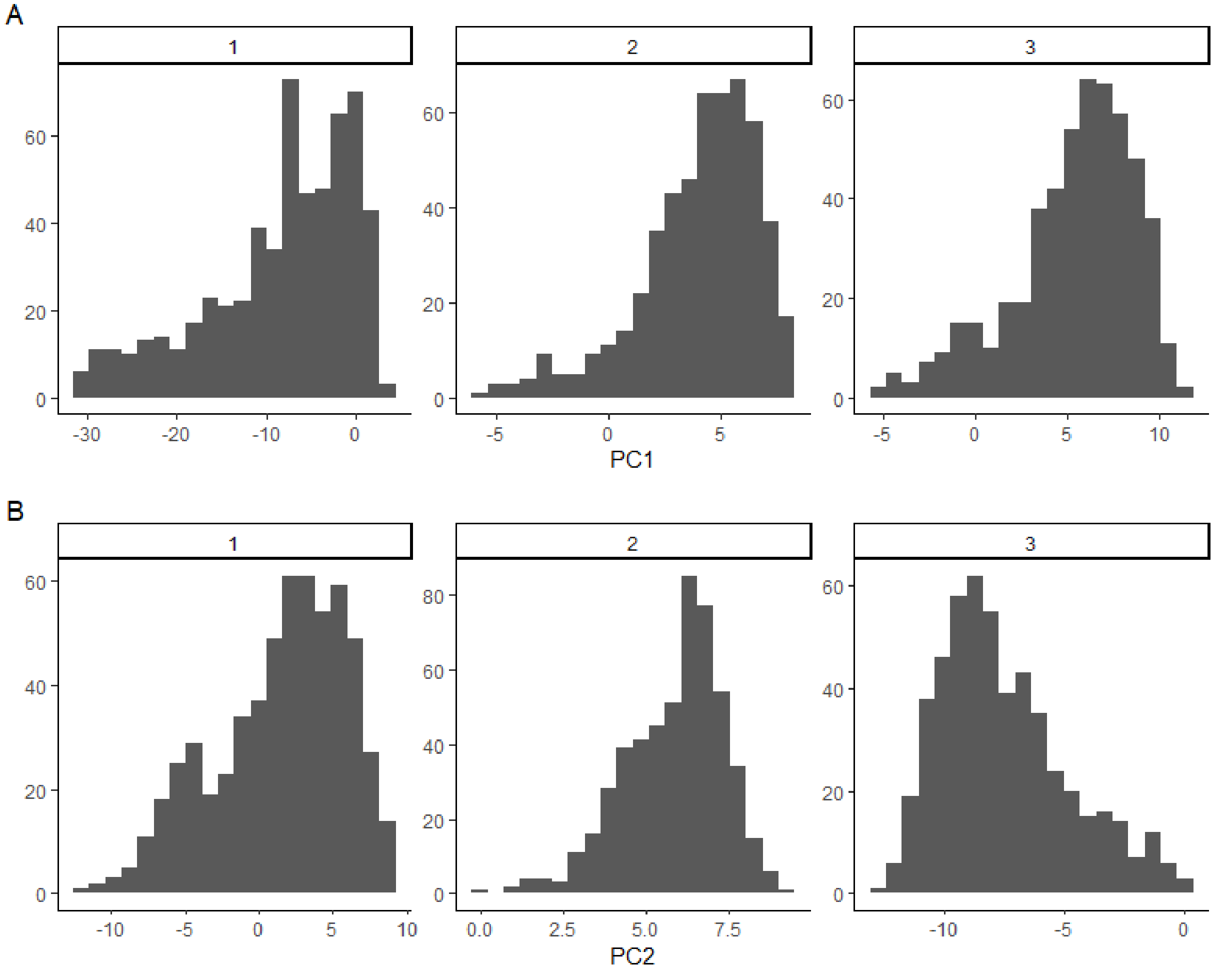

Figure 6: A. PC1, B. PC2, distributions by nn cluster