Single Cell Project 2

Guanxun Li & Honggang Wang

April 27, 2021

1 Introduction

The gene expression matrix of scRNA-seq data is a sparse matrix with a large portion of zeros in it. There are two reasons why this is the case. One is the presence of technical drop out due to the low concentration of mRNAs in a given individual cell. The other reason is a biological reason—that is, the cell actively switches off the transcription of a certain set of genes at any time. Statistics are needed here to distinguish between the causes of zeros in gene expression values in scRNA-seq data.

The problem proposed here concerns the computing correlation between two genes' expression levels. The correlation analysis using Spearman or Pearson correlation tests works with scRNA-seq data but zeros in the data cause the problem. There is a zero-inflated version of correlation analysis that may solve the problem, but we have not yet tested it with real data. Moreover, if two genes have a positive correlation, it means that two genes may be positively regulated by each other and if two genes have a negative correlation, it means that two genes may be negatively regulated by each other.

Furthermore, we can also treat scRNA-seq data as the binary data and calculate the binary version correlation between each pair of genes. In this case, if two genes have a positive correlation, it means that two genes will be more likely to express at the same cells and if two genes have a negative correlation, it means that two genes won't express at the same cells. It is natural to think these two correlations should be consistent, showing the same directionality—for example, a negative correlation between two genes should produce fewer overlapped cells co-expressing the two genes.

In this project, we wanted to see if there are cases two kinds of correlations give the opposite direction and how pronounced these opposite patterns can be overserved in real data sets.

2 Correlation Calculation

2.1 Correlation for Zero-inflated Data

We calculated Spearman's correlation for Zero-inflated data based on [2]. Given random variable X and Y, suppose (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) are independent copy of (X, Y), then the population Spearman's ρ is defined as

$$\rho_S = 3(P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0]). \tag{1}$$

Given data set $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, the rank-based estimator of Spearman Correlation is

$$r_S = 1 - \frac{6\sum_{i=1}^n D_i^2}{n(n^2-1)}, \tag{2}$$

where D_i is the difference between the ranks of X_i and Y_i in their separate rankings. If there are tied ranks in X, or Y, or both, the estimator (2) has to be adjusted. For the zero-inflated data, [1] proposes a new estimator, ρ_S^*, which being an estimator of (1) for the case of pairs being tied at 0 on at least one variable.

Definition 1. *The population Spearman's ρ for zero-inflated data has a form*

$$\rho_S^* = p_{11}p_{1+}p_{+1}\rho_{S11} + 3(p_{00}p_{11} - p_{10}p_{01}), \tag{3}$$

where ρ_{S11} is the population Spearman's correlation defiend in (1) for the non-zero pairs of observations, $p_{00} = P(X = 0, Y = 0)$, $p_{11} = P(X \neq 0, Y \neq 0)$, $p_{10} = P(X \neq 0, Y = 0)$, $p_{01} = P(X = 0, Y \neq 0)$, $p_{1+} = P(X \neq 0)$ and $p_{+1} = P(Y \neq 0)$.

2.2 Correlation for Binary Data

We can also treat scRNA-seq data as 0-1 binary data and we use the ϕ coefficient, which is a measure of association for two binary variables. The phi coefficient is related to the chi-squared statistic for a 2×2 contingency table

$$\phi = \sqrt{\frac{\chi^2}{n}},$$

where n is the total number of observations and its sign equals the sign of the product of the main diagonal elements of the table minus the product of the off-diagonal elements. Since we are only interested in the sign of the association, we only need to calculate the sign of $n_{00}n_{11} - n_{01}n_{10}$, where n_{00} and n_{11} are diagonal elements of the contingency table representing the number of counts for two genes non-expressed and expressed in the same cell, and n_{10} and n_{11} are off-diagonal elements representing respectively the number of counts that first gene expressed and the second non-expressed in a cell or the opposite.

2.3 Fisher's Exact Test

For each pair of genes, we also want to test whether there is a significant association between them. Here we treat the scRNA-seq data as the binary data and use Fisher's Exact Test to test whether there is an association between each pair of genes. The null hypothesis for the test is that there is no association between the rows and columns of the 2×2 table. Since it is no meaning to talk about the anti-correlation if there is no association between two genes, we only need to focus on the pair of genes whose p-value from Fisher's Exact Test is less than 0.05.

For more details about the Fisher's Exact Test, please refer to [1].

2.4 Numeric Results

We calculated the adjusted Spearman correlation based on (3), saved it as S, where S_{ij} represents the adjusted Spearman correlation between gene i and gene j. Next we generate the Spearman

sign matrix S^{01} by

$$S_{ij}^{01} = \begin{cases} +1 & \text{if } S_{ij} > 0 \\ -1 & \text{if } S_{ij} < 0 \end{cases}$$

The second step is to treat the scRNA-seq data as the $0-1$ binary data and calculate the binary sign matrix Φ, where

$$\widehat{\Phi}_{ij} = \begin{cases} +1 & \text{if } n_{00}n_{11} - n_{10}n_{01} > 0 \\ -1 & \text{if } n_{00}n_{11} - n_{10}n_{01} < 0 \end{cases}$$

After getting two sign matrices, we have the pair of genes whose signs are different between S^{01} and Φ. For these pairs of genes, the last step is to do the Fisher's exact test for them to check whether there is an association between them and we only keep the pair of genes whose p-value is less than 0.05. Finally, we get 142 pairs of genes.

3 Adjusted Correlation for Binary data based on Zero Inflated Negative Binomial Model

Since the scRNA-seq data is zero-inflated, if we treat it as the binary data, the proportion of zero is overestimated. This section is to find a way to adjust the proportion of zero in the real data.

3.1 Zero Inflated Negative Binomial Model for scRNA-seq Data

We first follow [3] to build the zero inflated negative binomial model for the scRNA-seq data. For any $\mu \geq 0$ and $\theta > 0$, let $f_{NB}(\cdot; \mu, \theta)$ denote the probability mass function (PMF) of the negative binomial (NB) distribution with mean μ and inverse dispersion parameter θ, namely:

$$f_{NB}(y; \mu, \theta) = \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\mu+\theta}\right)^{y}, \quad \forall y \in \mathbb{N}.$$

For any $\pi \in [0,1]$, let $f_{ZINB}(\cdot; \mu, \theta, \pi)$ be the PMF of the zero-inflated negative binomial (ZINB) distribution given by:

$$f_{ZINB}(y; \mu, \theta, \pi) = \pi\delta_0(y) + (1-\pi) f_{NB}(y; \mu, \theta), \quad \forall y \in \mathbb{N}, \tag{4}$$

where δ_0 is the Dirac function.

Given n samples (typically, n cells) and J features (typically, J genes), let Y_{ij} denote the count of feature j (for $j = 1, \cdots, J$) for sample i (for $i = 1, \cdots, n$). [3] models Y_{ij} as a random variable following a ZINB distribution with parameters μ_{ij}, θ_{ij} and π_{ij}, and considers the following regression models for the parameters:

$$\begin{aligned} \ln(\mu_{ij}) &= (X\beta_\mu + (V\gamma_\mu)^T + W\alpha_\mu + O_\mu)_{ij}, \\ \text{logit}(\pi_{ij}) &= (X\beta_\pi + (V\gamma_\pi)^T + W\alpha_\pi + O_\pi)_{ij}, \\ \ln(\theta_{ij}) &= \zeta_j, \end{aligned}$$

where

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$$

and elements of the regression models are as follows.

- $X = 1_n \in \mathbb{R}^n$ is corresponding to cell-level covariates and $\beta = (\beta_\mu, \beta_\pi)$ its associated J-dimension vector of regression parameters.
- $V = 1_J \in \mathbb{R}^J$ is corresponding to gene-level covariates and $\gamma = (\gamma_\mu, \gamma_\pi)$ its associated n-dimension vector of regression parameters.
- W is an unobserved $n \times K$ matrix corresponding to K unknown cell-level covariates and $\alpha = (\alpha_\mu, \alpha_\pi)$ its associated $K \times J$ matrices of regression parameters. Here we use the default parameter $K = 2$.
- O_μ and O_π are known $n \times J$ matrices of offsets.
- $\zeta \in \mathbb{R}^J$ is a vector of gene-specific dispersion parameters on the log scale.

By fitting the following ZINB model, we can get the zero-inflated probability π_{ij} for gene j at cell i.

3.2 Adjusted Correlation for Binary data

The estimator we adopted is the Φ matrix for the binary data, where

$$\widehat{\Phi}_{ij} = \begin{cases} +1 & \text{if } n_{00}n_{11} - n_{10}n_{01} > 0 \\ -1 & \text{if } n_{00}n_{11} - n_{10}n_{01} < 0. \end{cases}$$

The i,j denote different genes here and n_{11} counts how many cells the two genes are expressed together. The left variables n_{10}, n_{01} and n_{00} are defined similarly. This method is able to accurately give the direction estimation for the correlation of any two genes. However, when we consider the data are generated from a more complex zero-inflated generative model, the situation becomes more complicated. For example, the n_{00} may be overestimated. Avoiding this phenomenon or, in another word, adjusting this case, we need to exert an appropriate modification over the estimators based on the generating assumption. In above model, we assumed variant π for different genes and cells, and the formulation can follow as

$$X_{ic} = \pi_{ic}\delta_0 + (1 - \pi_{ic})f_{NB}(X_{ic}; \mu, \theta), \tag{5}$$

where i and c denote the gene and cell respectively. Based on this assumption, we can think of the 0 coming from the Negative binomial model with the probability $\tau_{ic} = P(X_{ic} \in f_{NB} \mid X_{ic} = 0)$. Then for each gene pair (i, j), we have the modified counting estimator as

$$\begin{aligned} n'_{00} &= \sum_c (1 - \tau_{ic})(1 - \tau_{jc})1_{(0,0)} \\ n'_{01} &= \sum_c (1 - \tau_{ic})1_{(0,1)} \\ n'_{10} &= \sum_c (1 - \tau_{jc})1_{(1,0)} \\ n'_{11} &= \sum_c 1_{(1,1)}. \end{aligned}$$

This adaption comes from our understanding for the occur of the 0's in this model. We think of that part of 0 comes from the meaningless δ_0 and the left part is generated from the picked trained model M which in our case is a negative binomial distribution learned by MLE. For two mutually independent genes, assuming the generative distribution as above, we can easily find the adjusted estimator is unbiased for the true $n_{00}, n_{01}, n_{10}, n_{11}$. Then following this idea, we can plug them back in and get the direction estimator Φ_{ij},

$$\widehat{\Phi}'_{ij} = \begin{cases} +1 & \text{if } n'_{00}n'_{11} - n'_{10}n'_{01} > 0 \\ -1 & \text{if } n'_{00}n'_{11} - n'_{10}n'_{01} < 0. \end{cases}$$

Until now, we only need to calculate τ_{ic} for each gene and cell. Since we have already got our estimation for all the parameters in last step using MLE. We can write down the formula for τ_{ic} here directly by the most basic Bayes rule,

$$\tau_{ic} = P(X_{ic} \in f_{NB} \mid X_{ic} = 0) = \frac{\pi_{ic}}{\pi_{ic} + P(X_{ic} = 0 \mid f_{NB})(1 - \pi_{ic})}.$$

Now the algorithm is complete. Furthermore, we notice that if we assume π_{ic}'s are the same for all genes and cells, we get no change for deciding on the direction after the modification. Through this, we can see that the most simplest zero-inflated model is included as a submodel in our case.

3.3 Numeric Results

Follow the same steps in section 2.4, we can find 764 pairs of genes. These pairs of genes include 141 of 142 pairs we found in the original method.

4 Summary

We use two methods in this project to calculate the association between pairs of genes and find anti-correlation pairs. The standard correlation analysis for scRNA-seq gene expression data is Spearman or Pearson correlation, but zeros in the data cause problems. As a result, for the zero-inflated data, we try an adjusted Spearman correlation. Then, by treating scRNA-seq data as binary data, we calculate the binary version similarity between each pair of genes. In this case, the positive and negative relationship is described by the sign of the *Phi* coefficients. We choose pairs of genes whose signs differ between two correlations after obtaining two association relationships. Finally, we perform the Fisher's exact test on selected pairs of genes, keeping only those with p-values less than 0.05.

In the future, we'll need to see if there are any biological implications for this pair of genes.

References

[1] JV Freeman and MJ Campbell. The analysis of categorical data: Fisher's exact test-tutorial. 2007.

[2] Ronald Silva Pimentel. Kendall's tau and spearman's rho for zero-inflated data. 2009.

[3] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. Zinb-wave: A general and flexible method for signal extraction from single-cell rna-seq data. *BioRxiv*, page 125112, 2017.