Applications of Zero-inflated Model for Single-cell RNA-seq Data

Fanny Perraudeau

2016-06-18

Contents

1. Supervised Differential Expression Analysis

We want to find differentially expressed genes between two groups (for example, two cell types) using single-cell RNA-sequencing measurements. We observe Y_{ij} the read count for gene $j = 1, \dots, J$ in cell $i = 1, \dots, n$, and we do not observe the detection indicator $Z_{ij} \in \{0, 1\}$, such that $Z_{ij} = 1$ when gene j is not detected in sample i (i.e., technical zero inflation) and 0 otherwise. We want to use the zero-inflated negative binomial (ZINB) model described in vignette model2 to account for zero inflation and over-dispersion of the observed counts. Then, the mean, ZI probability and dispersion parameter for the count of the gene j in cell i are modeled as follows:

$$\log(\mu_{ij}) = \left(X\beta_\mu + (V\gamma_\mu)^\top + W\alpha_\mu\right)_{ij} , \tag{1}$$

$$\text{logit}(\pi_{ij}) = \left(X\beta_\pi + (V\gamma_\pi)^\top + W\alpha_\pi\right)_{ij} , \tag{2}$$

$$\phi_{ij} = \phi_j , \tag{3}$$

See vignette model2 for details.

Parameter of interest

To perform supervised differential expression analysis, we want to estimate matrices β_μ and β_π, and more specifically the m^{th} row of β_μ and β_π corresponding to covariate m of interest (for example cell type).

Steps

1. Filter. Low quality samples should be removed. For example, we want number of reads $> 15,000$, percentage of aligned reads $> 50\%$.
2. Estimate β_μ and β_π using model ZINB (package pscl, then when ready team JP package).
3. Estimate β_μ using NB (package prefered?)
4. Compare NB and ZINB. Metric?
5. Plot. Goodness of fit of ZINB compared to NB: MD-plots, heatmaps of DE genes (see Sandrine's slides from APBC conference, from slide 75).

2. Unsupervised Differential Expression Analysis

We want to find differentially expressed genes between groups when groups are not known a priori using single-cell RNA-sequencing measurements. The data struture is the same as in supervised differential expression analysis except that the covariate for cell type is unknown. It means that the column in the design matrix corresponding to cell type is unknown. To determine the groups and identify genes that show the greatest differences amongst the groups, we use the model in clusterExperiment (is it written somewhere?) and the zinb model described in vignettes model and model2.

Parameter of interest

- Covariate corresponding to the groups (i.e. cell types), so a column of W?
- Matrices β_μ and β_π.

Steps

Unsupervised differential expression analysis is implemented in package clusterExperiment where function getBestFeatures can be used to perform the DE analysis. Under the hood, package limma is used with an option to use the "voom" correction to account for overdispersion. We want to implement a function to perform the DE analysis using the zinb model. We could then have an additional argument in function getBestFeatures where the user could choose to use limma or zinb.

1. Understand what the inputs and outputs of our function should be.
2. Implement it. Note that (if I understood well), we do have the covariates corresponding to the groups/clusters at this points of the analysis.

3. Imputation

From the observed read counts y_{ij}, we want to estimate $y^\star_{ij}$ the corrected read counts that take into account the excess of zeros (i.e. technical zero-inflation). Data structure and model are the same as in vignettes model and model2. We have

$$y^*_{ij} = y_{ij}\, Pr(Z_{ij} = 0|Y_{ij} = y_{ij}) + \mu_{ij}\, Pr(Z_{ij} = 1|Y_{ij} = y_{ij}).$$

Hence,

$$y^*_{ij} = \begin{cases} y_{ij} & y_{ij} > 0, \\ \mu_{ij}\, Pr(Z_{ij} = 1|Y_{ij} = 0) & y_{ij} = 0. \end{cases}$$

where

$$Pr(Z_{ij} = 1|Y_{ij} = 0) = \frac{\pi_{ij}}{\pi_{ij} + (1 - \pi_{ij})(1 + \phi_j \mu_{ij})^{-1/\phi_j}}.$$

Parameter of interest

Parameters of interest are the $y^\star_{ij}$.

Steps

One of the arguments of function Scone is *imputation*. For the moment, this argument can only be *identity*. The goal here would be to write a function that could be passed to argument *imputation*. The function would take as arguments the matrice of observed counts Y and all the matrices needed to run JP team optimization function. It would return $Y^{\star}$. Then, Scone function would run their usual pipeline to perform normalization and compute metrics to compare normalization methods.