1 Introduction

Single cell RNA-sequencing (scRNA-seq) is a powerful and relatively young technique which allows for characterizing molecular states of individual cells through their transcriptional profiles. It represents a major advance with respect to the standard RNA-sequencing which is only capable of detecting gene expressions averaged over millions of cells. Such averaged gene expression profiles may describe and characterize the global state of the tissue but cannot afford to study its heterogeneity and completely mask signals coming from individual cells. Accessing cell-to-cell variability is crucial for understanding many important biological processes such as tissue development and cancer. To be continued...

2 ZINB model

Zero-inflated count models provide a flexible statistical tool for modeling the over-dispersed count data with excessive number of zeros. The overdispersion is taken into account by the Negative Binomial distribution with two parameters: the mean and the dispersion parameter, allowing the variance to be greater than the mean. According to zero inflated negative binomial model (ZINB), zero counts may be generated by two different processes, one of which being the main signal modeled by Negative Binomial distribution and another being a source of extra zeros which is well adapted to the structure of single cell data where zero counts may correspond either to zero expressions in some cells or to the technical errors of detection.

We propose the following statistical model for the number of counts E_{ij}, observed in cell i for gene j:

$$P(E_{ij}=k)=\begin{cases}\Pi_{ij}+(1-\Pi_{ij})P(NB(M_{ij},\sigma_j)=0), & \text{if } k=0\\ (1-\Pi_{ij})P(NB(M_{ij},\sigma_j)=k), & \text{for any k}\neq 0\end{cases}$$

Here Π_{ij} is the probability of non detection of gene j in cell i, M_{ij} is its true expression value and σ_j is the gene specific dispersion parameter considered to be constant across cells. The elements $\{M_{ij}\}$ form a matrix M of size $n\times J$ which may be considered as the de-noised version of the observed matrix of counts $E=\{E_{ij}\}$. Linear dimension reduction is based on the hypothesis that data lies along a low dimensional linear space, or equivalently that the effective dimension (rank) of the data matrix is low. This means that it may be factorized in a product of low dimensional matrices. We suppose that

$$\begin{aligned}\log(M) &= X_M\alpha+UV^T\\ \text{logit}(\Pi) &= X_\Pi\beta+UW^T\end{aligned}$$

Here M is $n \times J$ matrix with cells in rows and genes in columns, X_M and X_Π are known design matrices of sizes $n \times k_M$ and $n \times k_\Pi$ respectively, and α, β are unknown matrices of respective sizes $k_M \times J$ and $k_\Pi \times J$. U is $n \times p$ matrix of latent factors (pseudo cells), where p is the number of factors. The matrix V is $J \times P$ matrix with the columns representing gene expression profiles in pseudo cells. W plays a similar role for the matrix of probabilities of zeros. Matrix U is shared between two parts of the model according to the idea that factors influencing the expression of gene also influences its dropout probability.

3 Experiences with simulated data

4 Study of biological data