Bayesian Regression to fit single-cell RNAseq datasets.

STAT 927 Final Project

Da Kuang

April 29 2022

Introduction

Single-cell RNA sequencing (scRNA-seq) is a powerful technique to uncover novel cell-to-cell heterogeneity in gene expression levels. However, the readout is prone to be sparse and confounded by both technical variations and biological variations. At first, the zeros in the readout are called dropout, and it is introduced to explain that several cells showed no expression for any given gene, even if it was relatively high in other cells. Naturally, many methods are built with Zero-inflated Negative Binomial to normalize and detect expression deferentially. This distribution adds a probability of observing a zero value in any given draw from the distribution of each gene. (Svensson 2020)

Since 2020, there has been a discussion about the assumption that zero inflation is an inherent property of scRNA-seq data (Svensson 2020, Cao 2021). To show the redundancy of zero-inflation, Valentine Svensson, in his paper, estimated the percentage of zeros that can be explained by gene-wise Negative binomial distribution. The results are reproduced in homogeneous single-cell data, heterogeneous single-cell data, and ERCC control data. The result supports the idea that the data is not zero-inflated for Droplet and UMI-based scRNA-seq experiments.

Valentine's paper focused on the zeros in the data, but it sparked my curiosity to see how good negative binomial distribution can fit a single-cell RNA-seq dataset. However, we must deal with technical and biological variations before applying the model. I would use Bayesian regression to estimate the mean based on the total number of mRNA molecules in the same cell for technical variation.

Hafemeister's paper (Hafemeister 2019) proposed to use GLM with either Negative Binomial or Poisson distribution to normalize single-cell RNA-seq data. His work has been integrated into Seurat as one of the popular normalization methods for clustering.

Inspired by those two papers, I would like to try to use Bayesian Negative Binomial and Poisson Regression to fit relatively simple datasets and then evaluate and compare the prediction of the models.

I prepared two datasets. The first one is a control experiment that can be seen as having no biological variation. The second one is a single-cell RNA sequencing run on a homogeneous population.

To this end, we will fit a regression model for each gene across all cells. To evaluate the Bayesian statistical model, I calculated scaled median absolute error, median absolute error, the probability of observation falling within 50% posterior prediction interval, and the probability of observation falling within 95% posterior prediction interval. But I will focus on the scaled median absolute error since it is the most robust and informative metric.

Genes at different expression levels have different biological variations, so I group the genes into four bins and describe the models' fitness among each group's genes.

This final project serves as an interesting exercise to get my hands dirty on modeling a single-cell RNA-seq dataset. I could only explore a constrained set of conditions given the limited time. But in the future, I would also love to see if the fitness would change given different methods to normalize the data. Moreover, it would also be interesting to see the model's fitness on some datasets with multiple cell types.

Method

Single-cell RNA-seq Technology

Generating single-cell data from a biological sample requires multiple steps. The following is a short notes based on Malte D Luecken's tutorials (Luecken 2019).

- Input material for a single-cell experiment is typically obtained in the form of biological tissue samples.
- As a first step, a single-cell suspension is generated in a process called single-cell dissociation in which the tissue is digested.
- To profile the mRNA in each cell separately, cells must be isolated. In this report, we focus on Single-cell isolation by droplet-based methods. Each droplet contains the necessary chemicals to break down the cell membranes and perform library construction.
- Then cellular cDNA libraries are labeled with cellular barcodes and UMI. These libraries are pooled together (multiplexed) for sequencing and then demultiplexed to produce counts of captured mRNA molecules.

Single-cell RNA-seq Data

Suppose the data is a gene by cell matrix, where N is the number of genes (G_i) and M is the number of cells (C_j). So x_{ij} is the number of mRNA molecules we observed for gene G_i in Cell C_j.

Negative Binomial Model

The negative binomial distribution is a discrete probability distribution that models the number of success in a sequence of independent and identical distributed Bernoulli trials

before a specified (non-random) number of failures (denoted r) occurs.

The probability mass function of negative binomial distribution has different versions based on the choice of parameters. In this report we use NB(μ, ϕ) with the following pmf.

$$\Pr(x|\mu,\phi) = \frac{\Gamma(x+\phi_i)}{\Gamma(\phi_i)x!}(\frac{\phi_i}{\mu_i+\phi_i})^{\phi_i}(\frac{\mu_i}{\mu_i+\phi_i})^x.$$

Bayesian Negative Binomial Linear Regression

Inspired by SCTransform paper[], suppose m is the vector of molecules assigned to the cell j, i.e. $m_j = \sum_i x_{ij}$. We have

$$\log(\mu_i) = \beta_0 + \beta_1 \log_{10} m$$

For each gene i, the posterior distribution is

$$f(\beta_1,\beta_0,\phi|x,m) \propto f(\beta_1)f(\beta_0) \times \sum_{j=1}^{M}[\phi \ln(\frac{\phi}{\phi+\mu}) + \ln\Gamma(\phi+x_j) + x_j \ln(\frac{\mu}{\mu+\phi}) - \ln\Gamma(\phi) - \ln x_j!]$$

We use flat uniform prior and use MCMC sampler with 4 chains to construct the posterior distribution. For each chain, we iterate 2000 times and the first 1000 rounds is used for warmup.

The conditional posterior distribution for each parameter is as follows.

- β_1

$$\begin{aligned} P(\beta_1|\beta_0,\phi,x,m) &= \sum_{j=1}^{M} \phi \ln\frac{\phi}{\phi+\mu} + x_j \ln\frac{\mu}{\mu+\phi} \\ &= \sum_{j=1}^{M} \phi \ln\frac{\phi}{\phi+\beta_0+\beta_1\log m} + x_j \ln\frac{\beta_0+\beta_1\log m}{\phi+\beta_0+\beta_1\log m} \\ &= \sum_{j=1}^{M} \frac{-\log m}{\phi+\beta_0+\beta_1\log m} + x_j \frac{\phi\log m}{(\beta_0+\beta_1\log m)(\phi+\beta_0+\beta_1\log m)} \end{aligned}$$

- β_0

$$P(\beta_0|\beta_1,\phi,x,m) = \sum_{j=1}^{M} \frac{-1}{\phi+\beta_0+\beta_1\log m} + x_j \frac{\phi}{(\beta_0+\beta_1\log m)(\phi+\beta_0+\beta_1\log m)}$$

- ϕ

$$P(\phi|\beta_0,\beta_1,,x,m) = \sum_{j=1}^{M} \frac{\beta_0+\beta_1\log m}{\phi+\beta_0+\beta_1\log m} + \ln\frac{\phi}{\phi+\beta_0+\beta_1\log m} - \frac{x_j}{\phi+\beta_0+\beta_1\log m}$$

Bayesian Poission Linear Regression

Same as Negative Binomial linear regression, suppose we have $\log(\mu_i) = \beta_0 + \beta_1 \log_{10} m$. For each gene i, the posterior distribution is

$$f(\beta_1, \beta_0 | x, m) \propto f(\beta_1) f(\beta_0) \times \sum_{j=1}^{M} x_j \ln \lambda - \lambda - \ln x_j!$$

The conditional posterior distribution for each parameter is as follows.

$$P(\beta_0 | \beta_1, x, m) = \sum_{j=1}^{M} \frac{x_j}{\beta_0 + \beta_1 \log_{10} m}$$
$$P(\beta_1 | \beta_0, x, m) = \sum_{j=1}^{M} \frac{x_j \log_{10} m}{\beta_0 + \beta_1 \log_{10} m} - \log_{10} m$$

Experiment design

Inspired by Valentine's paper, I selected two datasets.

One is from Svensson et al. with Endogenous RNA from K562 cell line and ERCC spike-ins. This is a negative control experiment to characterizes the single-cell RNA sequencing techniques. To reduce the computation, I randomly pick 100 cells/droplets. Then I group the genes into 4 bins as follows.

- Bin 1 (150 < x): 95 genes
- Bin 2 (100 < x < 150): 53 genes
- Bin 3 (50 < x < 100): 160 genes
- Bin 4 (x < 50): 9252 genes

I observed that bin 4 is too sparse to fit so in this report I will focus on Bin 1 - 3.

The second dataset is from 10X Chromium v3 with both NIH3T3 cells and HEK293T cells. I first filtered cell to have at least 2000 UMIs and then selected the cell with 20 times more UMI's from human than mouse. To this end, I create a dataset only containing HEK293T cells. Similarly, I also grouped the genes into 4 bins as follows.

- Bin 1 (150 < x): 3575 genes
- Bin 2 (100 < x < 150): 1263 genes
- Bin 3 (50 < x < 100): 2353 genes
- Bin 4 (x < 50): 16119 genes

I fit Bayesian regression for the first three bins since the last one is too sparse.

To evaluate the model, scaled median absolute error, median absolute error, the probability of observation falling within 50% posterior prediction interval, and the probability of observation falling within 95% posterior prediction interval are calculate. But in the result section, I will focus on scaled median absolute errors since it is the most robust and informative metric.

Software

Follow the discussion in Piazza @51, I learned about Stan and found two useful packages `rstanarm` and `bayesrules`. `rstanarm` is used for fitting Bayesian Negative Binomial Regression. `bayesrules` is used to calculate `mae`, `mae_scaled`, `within_50`, and `within_95` for evaluation.

Thanks to the efficient of the software, I was able to focus on the biological question I was interested in. Also `rstanarm` has a build-in R-shiny App to visualize the construction and prediction of the model, which is a good material to build intuition.

Result

For each gene, we applied both Negative Binomial and Poisson Regression and get the Scaled median absolute error. It measures the typical number of standard deviations that the observed Y_i fall from their posterior predictive means Y_i'.

$$\text{MAE scaled} = \text{median}\frac{|Y_i - Y_i'|}{\text{sd}_i}$$

For each bin of genes, we have a group of scaled median absolute error. Here we list the Negative Binomial and Poisson Regressions' posterior predictions on two datasets by the median and stander deviation of the scaled median absolute error within each bin.

	Control Data				HEK293T			
	Negative Binomial		Poission		Negative Binomial		Poission	
	median	sd	median	sd	median	sd	median	sd
Bin1	0.544	0.086	0.719	0.156	0.568	0.056	0.725	0.303
Bin2	0.658	0.086	0.753	0.101	0.535	0.064	0.631	0.092
Bin3	0.671	0.055	0.755	0.070	0.615	0.066	0.703	0.080

Discussion

On both control and HEK293T dataset, NB Regression fits better than Poisson Regression. Because the predictions from NB regression have smaller scaled median absolute error with less variance.

Within the same dataset, regression fits better in Bin 1 for genes with higher expression. Low expressed genes are more vulnerable to sequencing capture efficiency, which makes them harder to estimate.

Before the experiment, I expect to observe the NB regression has a better fit on Control Data than HEK239T due to the within-cell-type biological variation. For instance, life-cycles and

environmental stimulation during the experiment. But we observe the opposite. In other words, NB regression has slightly smaller errors on HEK239T than on the control dataset. It can be explained by two reasons. (1) The HEK239T cells are in a stable and homogeneous status so there is not much biological variations among the cells. (2) The control experiment is sequenced by Chromium v1 in 2017 while the HEK239T cells are sequenced by Chromium v3, which is more advanced technique. So it is possible that HEK239T dataset has smaller technical variations.

Reference

- Svensson, Valentine. "Droplet ScRNA-Seq Is Not Zero-Inflated." Nature Biotechnology 38, no. 2 (February 2020): 147–50. https://doi.org/10.1038/s41587-019-0379-5.
- Hafemeister, Christoph, and Rahul Satija. "Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression." Genome Biology 20, no. 1 (December 23, 2019): 296. https://doi.org/10.1186/s13059-019-1874-1.
- Luecken, Malte D, and Fabian J Theis. "Current Best Practices in Single-Cell RNA-Seq Analysis: A Tutorial." Molecular Systems Biology 15, no. 6 (June 1, 2019): e8746. https://doi.org/10.15252/msb.20188746.
- Tang, Wenhao, François Bertaux, Philipp Thomas, Claire Stefanelli, Malika Saint, Samuel Marguerat, and Vahid Shahrezaei. "BayNorm: Bayesian Gene Expression Recovery, Imputation and Normalization for Single-Cell RNA-Sequencing Data." Bioinformatics 36, no. 4 (February 15, 2020): 1174–81. https://doi.org/10.1093/bioinformatics/btz726.
- Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." Genome Biology 15, no. 12 (2014): 550. https://doi.org/10.1186/s13059-014-0550-8.
- Cao, Yingying, Simo Kitanovski, Ralf Küppers, and Daniel Hoffmann. "UMI or Not UMI, That Is the Question for ScRNA-Seq Zero-Inflation." Nature Biotechnology 39, no. 2 (February 2021): 158–59. https://doi.org/10.1038/s41587-020-00810-6.