Supplementary Note

Hanbin Lee

November 10, 2021

Contents

1 The conditional Poisson distribution

Our method allows arbitrary number of batch and has an analytic solution. In a theoretical perspective, our method allows batch-wise dispersion parameter but, as we will show, the final result does not depend on the specific value of the dispersion parameter. We name this method the *conditional Poisson residual*.

Let $b = 1, \ldots, m$, $i = 1, \ldots, n$ and $j = 1, \ldots, g$ be the index for batch, cell and gene (feature) respectively. Using this convention, we assume the following generative model.

$$Y_{bij} \sim \text{Poisson}(\mu_{bij}),\ \log \mu_{bij} = B_{bj} + \log m_{bi}$$

B_{bj} is the batch-wise intercept of gene j and m_{bi} is the total transcript count (or the UMI count) of a cell across $j = 1, \ldots, g$. This gives

$$Y_{bij} \left| \sum_{i \in b} Y_{bij} \sim \text{Mult}(\{\pi_{bij}\}),\ \pi_{bij} = \frac{e^{B_{bj} + \log m_{bi}}}{\sum_{i' \in b} e^{B_{bj} + \log m_{bi'}}} = \frac{m_{bi}}{\sum_{i' \in b} m_{bi'}} \ : i = 1, \ldots, n \right.$$

where B_{bj} is absent, making the inference of the batch parameters unnecessary.

Using the expectation and the variance formula of the multinomial distribution, we have

$$\mu_{bij} = \mathbb{E}\left[Y_{bij} \left| \sum_{i \in b} Y_{bij}, \log m_{bi} \right. \right] = \left[\sum_{i \in b} Y_{bij} \right] \cdot \pi_{bij}$$
$$\sigma^2_{bij} = \text{Var}\left[Y_{bij} \left| \sum_{i \in b} Y_{bij}, \log m_{bi} \right. \right] = \left[\sum_{i \in b} Y_{bij} \right] \cdot \pi_{bij}(1 - \pi_{bij})$$

This lets us to compute the Pearson residual according to

$$r_{bij} = \frac{Y_{bij} - \mu_{bij}}{\sigma_{bij}}$$

We can further generalize the model to accommodate overdispersion by using negative-binomial distribution for Y_{bij} instead of Poisson. However, this does not alter the expectation and only changes σ_{bij}.

$$\sigma^2_{bij} = \text{Var}\left[Y_{bij} \left| \sum_{i \in b} Y_{bij}, \log m_{bi} \right. \right] = \left[\sum_{i \in b} Y_{bij} \right] \cdot \pi_{bij}(1 - \pi_{bij}) \cdot \frac{\sum_{i \in b} Y_{bij} + \sum_{i \in b} m_{bi}}{1 + \sum_{i \in b} m_{bi}}$$

All the above results are exact without any approximation, and the result in the negative-binomial case is insensitive to the dispersion parameter. Hence, we do not need to impose any assumption on the value of the dispersion parameter. Note that, in practice, the additional term on the right is nearly one since $\sum_{i \in b} m_{bi} \gg \sum_{i \in b} Y_{bij}$.

2 'Regress-out' methods are biased: Theory

We refer to regression-based methods that attempt to 'regress-out' batch effects prior to estimating cell-state effect is biased if the batch is associated to cell-state composition.

Let T_{bij} be the cell-state effect of cell i's gene j. Then we assume the following log-linear model for an arbitrary function f.

$$\log \mu_{bij} = B_{bj} + f(T_{bij}) + \log m_{bi}$$

Although T_{bij} can be defined to have f included by definition, i.e. $T_{bij} = f(T_{bij})$, we explicitly state f to emphasize the generality of our statement.

The previous equation can be written to explicitly include the conditional statement in μ_{bij} and including a batch dummy variable X_{bij}:

$$\begin{aligned}\log \mathbb{E}\left[Y_{bij}\middle| T_{bij}, \log m_{bi}, B_{bj}\right] &= B_{bj} + f(T_{bij}) + \log m_{bi} \\ &= \beta_{bj} X_{bij} + f(T_{bij}) + \log m_{bi}\end{aligned}$$

The goal of 'regress-out' methods is to properly estimate β_{bj} for all b and j. This is computationally demanding if the number of batch m is large, but most importantly, running a regression without the information of T_{bij} gives a biased estimate.

To show this, we first compute the expectation of Y_{bij} not conditioned on T_{bij}.

$$\begin{aligned}\mathbb{E}\left[Y_{bij}\middle|\log m_{bi}, B_{bj}\right] &= \mathbb{E}\left[\mathbb{E}\left[Y_{bij}|T_{bij}, \log m_{bi}, B_{bj}\right]\middle|\log m_{bi}, B_{bj}\right] \\ &= \exp\left[B_{bj} + \log m_{bi}\right] \cdot \mathbb{E}\left[\exp[f(T_{bij})]\middle|\log m_{bi}, B_{b}j\right] \\ &= \exp\left[B_{bj} + \log m_{bi}\right] \cdot \mathbb{E}\left[\exp[f(T_{bij})]\middle|B_{b}j\right] \\ &= \exp\left[\beta_{bj} X_{bij} + \log m_{bi}\right] \cdot \mathbb{E}\left[\exp[f(T_{bij})]\middle|X_{bij}\right]\end{aligned}$$

The last term in the right-hand side is dependent on X_{bij} as long as T_{bij} and X_{bij} is associated. For example, the cell-state distribution might depend on the batch (e.g., there are more B cells in one batch than another). Thus, regressing X_{bij} on Y_{bij} will be affected not only by β_{bj} but also $\left[\exp[f(T_{bij})]\middle|X_{bij}\right]$. To explicitly show this, the quantity that is being estimated unconditional on T_{bij} is

$$\log \frac{\mathbb{E}\left[Y_{bij}\middle|\log m_{bi}, X_{bij}=1\right]}{\mathbb{E}\left[Y_{bij}\middle|\log m_{bi}, X_{bij}=0\right]} = \beta_{bj} + \log \frac{\mathbb{E}\left[\exp[f(T_{bij})]\middle|X_{bij}=1\right]}{\mathbb{E}\left[\exp[f(T_{bij})]\middle|X_{bij}=0\right]} \neq \beta_{bj}$$

Our method is robust to this problem because it estimates the residuals on a different conditional set (sum of Y_{bij}). Another way to understand our approach is to see that, as mentioned earlier, it does not estimate β_{bj} at all. Hence, the bias of estimating β_{bj} cannot exist by definition.

3 'Regress-out' methods are biased: Experiment

We demonstrate our theoretical finding using real data. We sub-sampled *fluorescence activated cell sorting* (FACS) B cells and monocytes obtained from 10X Genomics. We divided cells into two batches: batch 1 with 2000 monocytes and 2000 B cells and batch 2 with 2500 monocytes and 1500 B cells. Then for cells in batch 1, randomly selected genes with probability 0.5 were downsampled with probability 0.3 to apply gene-wise batch effect. The Pearson residuals were calculated using analytic Pearson residual, Poisson regression residuals with fixed effects dummies (for batch) and our method.

The features were selected in a decreasing order of Pearson residual variance. Total 2000 genes that were present in more than 5 cells in each batch were selected. Then PCA was applied followed by t-SNE (OpenTSNE was used) to top 20 dimensions of the PCA coordinate.

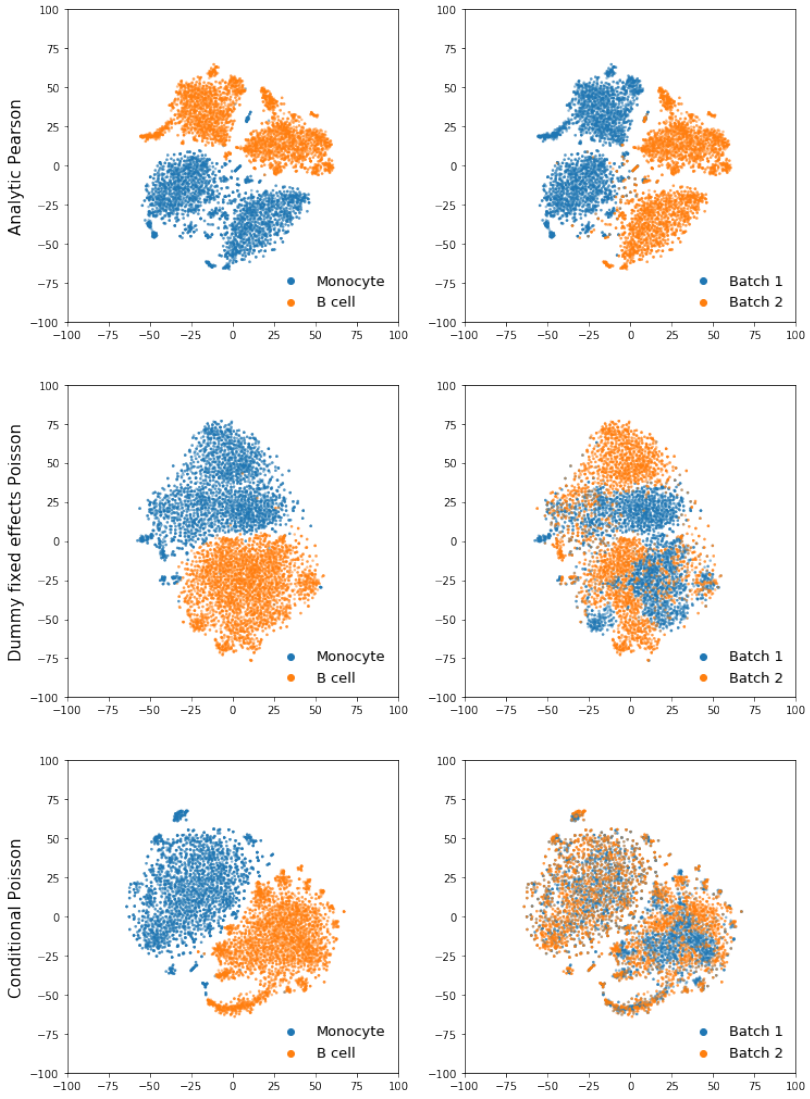

Figure 1: t-SNE visualizaiton using Analytic Pearson residual, Poisson regression residuals with fixed effects and our conditional Poisson residuals

References