Topic model with Batch effects

Kushal K Dey

January 22, 2016

Introduction

In RNA-seq experiments, we often encounter samples coming from different batches. The batches may be determined by the amplification procedures used, or the sequencing machine or even the sequencing lane effects. When these batch effects or technical effects are present in the samples, it becomes difficult to often separate out the biological information from the technical information (the latter is often relatively stronger). The topic model or the grade-of membership model has been used to cluster the samples based on their RNA-seq reads counts data (see paper). In the paper, we have shown that the topic model is sensitive to the presence of batch effects, however we have not been able to present a solution to that problem. We address the issue of how one can tackle batch effects in a topic model type framework.

We first present the standard topic model framework

Standard Topic Model

Let c_{ng} be the counts of reads for sample n and gene g. Let c_{n+} be the sum of reads for sample n, also called the *library size*.

$$(c_{n1}, c_{n2}, \cdots, c_{nG}) \sim Mult(c_{n+}, p_{n1}, p_{n2}, \cdots, p_{nG})$$

$$p_{ng} = \sum_{k=1}^{K} \omega_{nk}\theta_{kg} \qquad \sum_{k} \omega_{nk} = 1 \quad \forall n \qquad \sum_{g} \theta_{kg} = 1 \quad \forall k$$

Here $\omega_{n\cdot}$ represents the topic proportions for n th samples. On the other hand $\theta_{k\cdot}$ represents the probability distribution on the genes for the kth topic or cluster.

Topic model with Batch effects

One way batch effects may be incorporated in the above model would be to make the topic distribution for each cluster/ topic a function of the batch the sample is coming from, as well. Then we can write the above model as

$$(c_{n1}, c_{n2}, \cdots, c_{nG}) \sim Mult(c_{n+}, p_{n1}, p_{n2}, \cdots, p_{nG}) \tag{1}$$

$$p_{ng} = \sum_{k=1}^{K} \omega_{nk}\theta_{b(n):k,g} \qquad \sum_{k} \omega_{nk} = 1 \quad \forall n \qquad \sum_{g} \theta_{b(n):k,g} = 1 \quad \forall k, \quad b(n) \in \{1, 2, \cdots, B\} \tag{2}$$

Prior Specification

Note that the above the model is analogous to applying topic model separately for each batch. The problem with that approach is that we will not be able to track which cluster is Batch 1 corresponds to that cluster in Batch 2. Also, we expect each cluster distribution to have some common features across different batches despite getting effected by batch effects. In order to tackle this, we make the following assumption.

For each cluster k

$$(\theta_{b:k,1}, \theta_{b:k,2}, \cdots, \theta_{b:k,G}) \sim Dir_G\left(\theta_{k1}, \theta_{k2}, \cdots, \theta_{kG}\right) \qquad b \in \{1, 2, \cdots, B\} \tag{3}$$

Which is same as saying that for each batch, we are generating a sample from the cluster with mean $(\theta_{k1}, \theta_{k2}, \cdots, \theta_{kG})$, which represents the cluster k. This is analogous to the assumption in the normal linear models with batch effects,

$$y_{ng} = \mu_{t(n):b(n),g} = \mu + \tau_{t(n)} + \beta_{b(n)} + e_{ng}$$

where $t(n)$ in the treatment effect and $b(n)$ is the batch effect. We often assume that

$$\beta_b \sim N(0, \sigma_b^2)$$

Then

$$\mu_{t(n):b(n)} \sim N(\mu + \tau_n, \sigma_b^2) := N(\mu_{t(n)}, \sigma_b^2)$$

You can see that the treatment effects under the different batches are then a random sample from a distribution whose mean is the treatment effect without the batch information. Note that σ_b term is there in normal models to tune the variance for each effect. We can also put such a scaling parameter in our model Equation 3.

Then for each k,

$$(\theta_{b:k,1}, \theta_{b:k,2}, \cdots, \theta_{b:k,G}) \sim Dir_G\left(\alpha_b\theta_{k1}, \alpha_b\theta_{k2}, \cdots, \alpha_b\theta_{kG}\right) \qquad b \in \{1, 2, \cdots, B\}$$

However, as of now, I am assuming that $\alpha_b = 1$ for all batches and working with the simpler model.

We assume a prior for θ_{kg}.

$$(\theta_{k1}, \theta_{k2}, \cdots, \theta_{kG}) \sim Dir_G\left(\frac{1}{KG}, \frac{1}{KG}, \cdots, \frac{1}{KG}\right) \qquad \forall k \tag{4}$$

So, essentially we have a hierarchical structure in the θ's, on combining Equation 3 and Equation 4. The goal would be to get hold of the ω_{nk} and θ_{kg}.

We can assume the same prior for ω as in standard topic model, given by

$$(\omega_{n1}, \omega_{n2}, \cdots, \omega_{nK}) \sim Dir_K\left(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}\right) \qquad \forall n$$

Model estimation

We can assume that

$$c_{n+} \sim Poi(\lambda_n) \tag{5}$$

Then from