1 Model

Assume that we have counts data c_{ng} where n runs across the sample indices from 1 to N and g runs across the feature indices (genes in the single Cell and RNA Seq experiments) from 1 to G. Assume that each sample n has a batch label $b(n)$ which can is a factor label that result from technical effects - may be lane effect, amplifier effect, sequencing machine effect. We would ideally want to remove the batch effects and then apply the clustering model on the residuals but at the same time we also want to restore the count structure of the data so as to apply the topic model or admixture model clustering algorithm. As of now, we are first fitting the model

$$log(c_{ng} + 0.5) = \alpha_g + \beta_{b(n):g} + e_{ng} \qquad \sum_b \beta_{b:g} = 0 \qquad \forall g$$

We fit this fixed effect model under the sum constraint and obtain the effect size estimates $\hat{\alpha}_g$, $\hat{\beta}_{b:g}$ and $r_{ng} = \hat{e}_{ng}$, the residuals. Then we transform to the original space using the reverse function

$$y_{ng} = |exp\left(\hat{\alpha}_g + r_{ng}\right) - 0.5|$$

However this quantity is not a count, so in order to have counts, we generate counts using

$$z_{ng} \sim Poi(y_{ng})$$

These z's would form a new count data that would be batch effect corrected and would also retain the nice variational properties of the counts data.