1 Zero Inflated Grade of Membership Models - ZIGoM

Typically in a grade of membership model, we model the count vector $c_{n\star}$ as follows

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

where n is the sample index, c_{n+} represents the sequencing depth for sample n and G is the total number of genes. This formulation of the model conditions on the library size c_{n+}. Without that conditioning, the same model can be rephrased as follows

$$c_{ng} \sim Poi\left(c_{n+} \sum_{k=1}^{K} \omega_{nk}\theta_{kg}\right)$$

While this model fits the counts nature of the data well, it does not effectively deal with the zero inflation caused by the dropouts, technical noise, bursting genes etc. The other issue that is of concern for a model like this is that it does not account for the over-dispersion typically observed in the single cell data. A negative binomial model would have been a better candidate choice, but also complicates the model inference.

A generalization of the above model to take into account the sparsity of the data is as follows

$$c_{ng} \sim \pi_{ng}\delta_0 + (1-\pi_{ng})Poi\left(c_{n+} \sum_{k=1}^{K} \omega_{nk}\theta_{kg}\right)$$

where

$$logit(\pi_{ng}) := (X\beta + \gamma Y)$$

where X is a sample level covariates matrix with its columns including covariates like sequencing depth, while Y is a gene level covariates matrix with the gene level covariates representing gene length, percentage GC content etc.

For the non zero observations, the sequencing depth is already taken care of, but not gene level features.