1 Grade of Membership Model with covariates - covGoM

1.1 Standard Grade of Membership (GoM) model

In a general Grade of Membership model, $c_{n.}$, the vector of read counts across genes (G many) for each sample n can be modeled as following

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

where c_{n+} is the sequencing depth for sample/cell n.

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

Here ω_{nk} represents the membership proportion of the sample n in cluster k and θ_{kg} represents the weight on gene g for the cluster k.

We assume priors on ω and θ as follows

$$\omega_{n.} \sim Dir_K\left(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}\right)$$

$$\theta_{k.} \sim Dir_G\left(\alpha_1, \alpha_2, \cdots, \alpha_G\right)$$

where as default $\alpha_g = 1/KG$ for each g.

1.2 Grade of Membership model with covariates (covGoM) model

In this modified model, we assume that the cluster that was previously represented by θ_{kg} has a sample specific component that takes into account the sample metadata information θ_{nkg}. The full model can be expressed as following

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

where c_{n+} is the sequencing depth for sample/cell n.

$$p_{ng} = \sum_{k=1}^{K} \omega_{nk}\theta_{nkg} \qquad \sum_{k=1}^{K} \omega_{nk} = 1 \qquad \sum_{g=1}^{G} \theta_{nkg} = 1$$

$$\omega_{n.} \sim Dir_K\left(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}\right)$$

$$\theta_{nkg} = exp\left(\mu_g + \beta_{kg} + \gamma_{b(n):g} + \nu_{b(n):k,g}\right) / \left\{ \sum_{g=1}^{G} \left(exp\left(\mu_g + \beta_{kg} + \gamma_{b(n):g} + \nu_{b(n):k,g}\right)\right) \right\}$$

where μ_g is the mean profile for gene g. This is an important feature because it takes care of the gene length biases. $beta_{kg}$ is the cluster k specific effect, whereas $\gamma_{b(n):g}$ is the batch specific effect. $\nu_{b(n):k,g}$ represents the interaction between batch and cluster for gene g.

We are flexible in choosing the prior formulations for the effect sizes μ, γ and β. As of now, we are inclined to use the gamma lasso prior for each of these parameters.

1.3 Model fit

We assume latent variables to be T_{nkg}, the number of reads mapping to gene g and cluster k from sample or cell n.

To write down the complete log-likelihood, one will have to account for the following two conditional probabilities

$$(T_{n1+}, T_{n2+}, \cdots, T_{nK+}) \sim Mult\,(c_{n+}, \omega_{n1}, \omega_{n2}, \cdots, \omega_{nK})$$

Also for any cluster k,

$$(T_{nk1}, T_{nk2}, \cdots, T_{nkG} |\, T_{nk+}) \sim Mult\,(T_{nk+}, \theta_{nk1}, \theta_{nk2}, \cdots, \theta_{nkG})$$

In the E-step, we determine the expectation of these latent variables T_{nkg} given the data c_{ng} and the parameters θ and ω.

$$\begin{aligned} E\left(T_{nkg}|c_{n+}, \theta, \omega\right) &= E\left(E\left(T_{nkg}|T_{nk+}, c_{n+}, \theta, \omega\right)\right) \\ &= E\left(T_{nk+}\theta_{nkg}\right) \\ &= c_{n+}\omega_{nk}\theta_{nkg} \end{aligned}$$

We know that

$$c_{ng} = \sum_{k=1}^{K} T_{nkg}$$

We can write

$$\left(T_{n1g}, T_{n2g}, \cdots, T_{nKg}|c_{ng}\right) \sim Mult\left(c_{ng} : v_{n1g}, v_{n2g}, \cdots, v_{nKg}\right)$$

where

$$v_{nkg} = \frac{\omega_{nk}\theta_{nkg}}{\sum_{h=1}^{K}\omega_{nh}\theta_{nhg}}$$

The iterate of v_{nkg} at the t th iteration is as follows

$$v_{nkg}^{(t)} := \frac{\omega_{nk}^{(t)}\theta_{nkg}^{(t)}}{\sum_{h=1}^{K}\omega_{nh}^{(t)}\theta_{nhg}^{(t)}}$$

Under the Standard GoM model,

$$(\theta_{nk1}, \theta_{nk2}, \cdots, \theta_{nkG}) \sim Dir_G(\alpha_1, \alpha_2, \cdots, \alpha_G)$$

Then the MAP for θ was

$$\theta_{nkg}^{(t+1)} = \frac{E\left(T_{nkg}|c_{ng}, \omega^{(t)}, \theta^{(t)}\right) + \alpha}{E\left(T_{nk+}|c_{ng}, \omega^{(t)}, \theta^{(t)}\right) + G\alpha}$$

So, the EM update for θ after filling in the expectation is

$$\theta_{nkg}^{(t+1)} = \frac{c_{ng}v_{nkg}^{(t)} + \alpha}{\sum_{g=1}^{G} c_{ng}v_{nkg}^{(t)} + G\alpha}$$

However here the parameters are μ, β, γ and ν. To estimate these, we use the afollowing relation from the EM complete likelihood set up

$$(T_{nk1}, T_{nk2}, \cdots, T_{nkG}|\, T_{nk+}) \sim Mult\,(T_{nk+}, \theta_{nk1}, \theta_{nk2}, \cdots, \theta_{nkG})$$

We consider the estimate $E\left(T_{nkg}|c_{ng}, \omega^{(t)}, \theta^{(t)}\right)$ and then we perform Multinomial Logistic regression with the covariates as present in the model. In order to perform this, we want a fast Multinomial model because there are $B \times G + K \times G + G$ many parameters for the batch effects model. This can be computationally extensive. I am planning on trying the **distrom** package due to Matt Taddy as it performs parallel implementations of this model.

For ω, we assume the Dirichlet distribution prior

$$\omega_{n.} \sim Dir\left(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}\right)$$

The EN update for ω is as follows

$$\omega_{nk}^{(t+1)} = \frac{E\left(T_{nk+}|c,\omega,\theta\right) + \frac{1}{K}}{c_{n+}+1}$$

where

$$E\left(T_{nk+}|c,\omega,\theta\right) := c_{n+}\omega_{nk}^{(t)}$$

Therefore the update equation can be written as

$$\omega_{nk}^{(t+1)} = \frac{c_{n+}\omega_{nk}^{(t)} + \frac{1}{K}}{c_{n+}+1}$$

We additionally update the $\omega^{(t+1)}$ and $\theta^{(t+1)}$ by Quasi-Newton acceleration so that the convergence is quicker. Also we use an active set method as well to update the ω as well.

However we assume a form for the θ as follows