Topic model for metagenomic data

Kushal K Dey

December 29, 2015

Introduction

Topic model or admixture type models can be used for clustering metagenomic samples in 16s RNA counts data. However, these topic models while assuming all the features (the Operational Taxonomic Units or OTUs) to be independent. However, several OTUs may belong to the same species, and again several species may form one family of the microbiome. This hierarchical structure in the features is something we may be interested in exploring going forward and modifying the topic model accordingly. I present here the model which may be used to fit the topic model taking into account this hierarchical structure.

The core idea behind this model has been derived from the Multiscale Topic Tomography model described in this paper.

The Model

Let us start with the counts data $c_{N\times G}$ where N represents the number of samples and G represents the number of OTUs. Using Matt Taddy's model, we can write

$$c_{n*}|c_{n.} \sim Mult(c_{n.}, p_{n*})$$

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

Multi-resolution model for topics We build the hierarchical tree as follows. Let there be S levels in the hierarchical tree. The OTUs form the leaves of the tree, while the other levels may represent the family, species, genus etc.

$$\theta_{kl}^{(S)} = \theta_{kl} \qquad l = 0, 1, 2, \cdots, N_S - 1 \tag{1}$$

$$\theta_{kl}^{(s)} = \sum_{h:i_s(h)=l} \theta_{kh}^{(s+1)} \qquad s = 0, 1, 2, \cdots, S-1, \qquad l = 0, 1, 2, \cdots, N_s - 1 \tag{2}$$

$$\tag{3}$$

where N_s represents the number of leaves if the tree is truncated at level s and $i_s(.)$ is a function that takes the unit in level $s+1$ and maps it to the family it belongs to in level s. Note that $N_S = G$.

Latent representation of model Now if we assume that

$$c_{n.} \sim Poi(\lambda_n)$$

Then one can write

$$c_{nl} \sim Poi(\lambda_n \sum_{k=1}^{K} \omega_{nk} \theta_{kl}^{(S)})$$

Let z_{nkg} represents the number of counts from sample n and from OTU l that comes from k th subgroup or cluster. By definition,

$$\sum_{k=1}^{K} z_{nkl} = c_{nl}$$

Since the summation of two independent Poisson random variables is also a Poisson variable with mean equal to the sum of the means of the original random variables, we can infer that

$$z_{nkl} \sim Poi(\lambda_n \omega_{nk} \theta_{kl}^{(S)})$$

Let z_{kg} represents the number of latent counts coming from the k th subgroup and feature g across all the samples.

$$z_{kl} = \sum_{n=1}^{N} z_{nkl}$$

So,

$$z_{kl} \sim Poi(\theta_{kl}^{(S)} \sum_{n=1}^{N} \lambda_n \omega_{nk})$$

Multi-resolution model for latent variables Suppose we are at a particular iterative step of our model where we have plausible values of ω and θ (we can start with the same prior for these parameters as Taddy model). Given ω, we use the following step to estimate a refined θ.

From Eqn 8 of Matt Taddy's paper), we can write

$$z_{nkl} = c_{nl} \frac{\omega_{nk} \theta_{kl}}{\sum_{h=1}^{K} \omega_{nh} \theta_{hl}}$$

So,

$$z_{kl} = \sum_{n=1}^{N} c_{ng} \frac{\omega_{nk} \theta_{kl}}{\sum_{h=1}^{K} \omega_{nh} \theta_{hl}}$$

Note that z_{kg} and $z_{k^{'}g}$ for $k \neq k^{'}$ are independent. Then the multiscale framework for θ can be translated to multiscale framework for z as well. Under this framework, we have

$$z_{kl}^{(S)} = z_{kl} \qquad l = 0, 1, 2, \cdots, N_S - 1 \tag{4}$$

$$z_{k(l)}^{(s)} = \sum_{h: i_s(h) = l} z_{kh}^{(s+1)} \qquad s = 0, 1, 2, \cdots, S-1, \qquad l = 0, 1, 2, \cdots, N_s - 1 \tag{5}$$

$$\tag{6}$$

We now define

$$\mu_{kl}^{(s)} = \sum_{n=1}^{N} \lambda_n \omega_{nk} \theta_{kl}^{(s)} \qquad l = 0, 1, 2, \cdots, N_s - 1$$

and it can be shown easily that

$$z_{k(l)}^{(s)} \sim Poi(\mu_{kl}^{(s)}) \qquad l = 0, 1, 2, \cdots, N_s - 1$$

Transformation of variables on hierarchy Instead of using $\mu_{kg}^{(s)}$ along the multi-resolution tree, we transform the parameters as follows

$$\beta_{k(l,h_l)}^{(s)} = \frac{\mu_{kh_l}^{(s+1)}}{\mu_{kl}^{(s)}} \qquad s = 0, 1, 2, \cdots, S-1, \qquad l = 0, 1, 2, \cdots, N_s - 1, \qquad i_s(h_l) = l, \qquad l = 1, 2, \cdots, c(l)$$

where $c(l)$ is the number of children of the node l.

We only need the highest level wavelet parameter $\mu_{k0}^{(0)}$ and $\beta_{k(l,h_l)}^{(s)}$ instead of $\mu_{kl}^{(s)}$. We work on these transformed parameter space. The transformed parameters are easy to work with as they are independent. We assume the priors to be

$$\mu_{k0}^{(0)} \sim Gamma(.|\nu_\mu, \delta_\mu)$$

$$\beta_{k(l,.)}^{(s)} \sim Dir_{c(l)} \left(.|\frac{1}{c(l)}, \frac{1}{c(l)}, \cdots, \frac{1}{c(l)} \right)$$

where $c(l)$ is the number of children for the node l.

Prior on wavelet parameters The prior distribution is therefore given by

$$P(\mu|\delta) = \prod_{k=1}^{K} Gamma(\mu_{k0}^{(0)}|\nu_\mu, \delta_\mu) \times \prod_{k=1}^{K} \prod_{s=0}^{S-1} \prod_{l=0}^{N_s-1} Dir_{c(l)} \left(\beta_{k(l,.)}^{(s)}|\frac{1}{c(l)}, \frac{1}{c(l)}, \cdots, \frac{1}{c(l)} \right)$$

Loglikelihood given wavelet parameters The loglikelihood of μ is given as follows

$$L(\mu) = \sum_{l=0}^{2^S-1} \sum_{k=1}^{K} logPoi(z_{k(l)}^{(S)}|\mu_{kl}^{(S)}) \tag{7}$$

$$= \sum_{s=0}^{S-1} \sum_{l=0}^{N_s-1} \sum_{k=1}^{K} logMult \left(z_{k,.}^{(s+1)}|\beta_{k(l,1)}^{(s)}, \beta_{k(l,2)}^{(s)}, \cdots, \beta_{k(l,c(l))}^{(s)} \right) + \sum_{k=1}^{K} logPoi(z_{k(0)}^{(0)}|\mu_{k0}^{(0)}) \tag{8}$$

$$\tag{9}$$

The z values estimated may not always be integers but we assume that they are approximated to the nearest integer. This is the same policy also adopted by the authors in the multiscale Topic Tomography paper.

MAP estimates of wavelet parameters Given the prior and the log likelihood functions reported above, one can compute th log posterior of the μ and then one can update the parameters using their MAP estimates.

$$\beta^{(s)}_{k(l,h_l)} = \frac{z^{(s+1)}_{kh_l} + \delta_\beta - 1}{z^{(s)}_{kl} + 2(\delta_\beta - 1)} \qquad h_l = 1, 2, \cdots, c(l) \qquad \forall k$$

$$\mu^{(0)}_{k0} = \frac{z^{(0)}_{k(0)} + \nu_\mu - 1}{\delta_\mu + 1} \qquad \forall k$$

This helps us generate the $\mu^{(s)}_{kl}$ for all s, k, l and most importantly $\mu^{(S)}_{kl}$. Given that we know $\mu^{(S)}_{kl}$, we can compute the variables of interest θ as

$$\theta^{(S)}_{kl} = \frac{\mu^{(S)}_{kl}}{\sum_{r=1}^{G} \mu^{(S)}_{kr}}$$

Updating topic proportions These are the θ update of the step. The $\theta^{(S)}$ values updated this way can then be used to update the ω parameters, which incidentally depend only on the leaf node parameters $\theta^{(S)}$. The approach to estimating ω is similar to the one used by Matt Taddy, using active set strategy.