Topic model with ordered features

Kushal K Dey

December 28, 2015

Introduction

In genetics, we often come across scenarios where the features have an inherent order relationship among themselves. For example, in population genetics, due to recombination, the SNPs close to each other are more likely to come from same ancestor. In fact the SNPs can be arranged on a linear scale based on their positions on the chromosomes. In phylogenetic applications, we may have data on abundances of different bird species for different forest spots studied (as in case of Alex White and Trevor Price's data). The usual topic model approach (as in the **maptpx** software due to Matt Taddy) does not take into account this ordering among the different features.

The core idea behind this model has been derived from the Multiscale Topic Tomography model described in this paper. I recommend going through this article before proceeding to the next section.

The Model

Let us start with the counts data $c_{N \times G}$ where N represents the number of samples and $G \approx 2^S$ represents the number of features. 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 Now we follow a multi-resolution analysis approach. We assumed there are 2^S features. We now define the multiscale wavelet parameters (assuming Haar wavelet) as

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

$$\theta^{(s)}_{k(l)} = \theta^{(s+1)}_{k(2l)} + \theta^{(s+1)}_{k(2l+1)} \qquad s = 0, 1, 2, \cdots, S-1, \qquad l = 0, 1, 2, \cdots, 2^s - 1 \tag{2}$$

$$\tag{3}$$

The indices s is called the scale and represents the depth of the tree. The highest scale is S which corresponds to the leaves of the trees and these leaves represent the actual features individually in this case.

Latent representation of model Now if we assume that

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

Then one can write

$$c_{ng} \sim Poi(\lambda_n \sum_{k=1}^{K} \omega_{nk} \theta^{(S)}_{kg})$$

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

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

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

$$z_{nkg} \sim Poi(\lambda_n \omega_{nk} \theta^{(S)}_{kg})$$

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

$$z_{kg} = \sum_{n=1}^{N} z_{nkg}$$

So,

$$z_{kg} \sim Poi(\theta^{(S)}_{kg} \sum_{n=1}^{N} \lambda_n \omega_{nk})$$

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

$$\mathbb{E}\left(z_{nkg}\right) = \mathbb{E}\left(c_{ng} \frac{\omega_{nk}\theta_{kg}}{\sum_{h=1}^{K} \omega_{nh}\theta_{hg}} \right)$$

So,

$$\mathbb{E}\left(z_{kg}\right) = \mathbb{E}\left(\sum_{n=1}^{N} c_{ng} \frac{\omega_{nk}\theta_{kg}}{\sum_{h=1}^{K} \omega_{nh}\theta_{hg}} \right)$$

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

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

$$z^{(s)}_{k(l)} = z^{(s+1)}_{k(2l)} + z^{(s+1)}_{k(2l+1)} \qquad s = 0, 1, 2, \cdots, S-1, \qquad l = 0, 1, 2, \cdots, 2^s - 1 \tag{5}$$

$$\tag{6}$$

We now define

$$\mu^{(s)}_{kg} = \sum_{n=1}^{N} \lambda_n \omega_{nk} \theta^{(s)}_{kg}$$

and it can be shown easily that

$$z_{k(l)}^{(s)} \sim Poi(\mu_{kl}^{(s)})$$

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

$$\beta_{k(l)}^{(s)} = \frac{\mu_{k(2l)}^{(s+1)}}{\mu_{k(l)}^{(s)}} \qquad s = 0, 1, 2, \cdots, S-1, \qquad l = 0, 1, 2, \cdots, 2^s - 1$$

We only need the highest level wavelet parameter $\mu_{k0}^{(0)}$ and $\beta_{kl}^{(s)}$ 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

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

$$\beta_{k(l)}^{(s)} \sim Beta(.|\delta_\beta, \delta_\beta)$$

Prior on wavelet parameters The prior distribution is therefore given by

$$\pi(\mu|\delta) = \prod_{k=1}^{K} Gamma(\mu_{k0}^{(0)}|\nu_\mu, \delta_\mu) \times \prod_{k=1}^{K} \prod_{s=0}^{S-1} \prod_{l=0}^{2^s-1} Beta(\beta_{k(l)}^{(s)}|\delta_\beta, \delta_\beta)$$

$$\pi(\omega|\alpha) \sim Dir_K(\alpha_1, \alpha_2, \cdots, \alpha_K)$$

Loglikelihood given wavelet parameters The full model loglikelihood of μ assuming we know the hidden variables z is given as follows

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

$$= \sum_{s=0}^{S-1} \sum_{l=0}^{2^s-1} \sum_{k=1}^{K} logBin(z_{k(2l)}^{(s+1)}|z_{k(l)}^{(s)}, \beta_{k(l)}^{(s)}) + \sum_{k=1}^{K} logPoi(z_{k(0)}^{(0)}|\mu_{k0}^{(0)}) \tag{8}$$

$$\tag{9}$$

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.

EM algorithm Assume after the previous step of iteration, we have obtained iterates β_{old}, $\mu_{old}^{(0)}$ and ω_{old}. We define

$$\mathcal{Q}\left(\beta, \mu^{(0)}|C_{N\times G}, \beta_{old}, \mu_{old}^{(0)}, \omega_{old}\right) = \mathbb{E}_{Z|C_{N\times G}, \beta_{old}, \mu_{old}^{(0)}, \omega_{old}}\left(logL(\beta, \mu^{(0)}, X, Z) + log(\pi(\beta, \mu^{(0)}|\delta)\right)$$

The first term in the sum expression above does not contain ω because given that we know z, the information on ω is redundant. We maximize this with respect to $\beta_{kl}^{(s)}$ for each resolution and cluster and $\mu_{k0}^{(0)}$. We obtain the following MAP updates for these parameters.

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.

$$\beta^{(s)}_{k(l)} = \frac{\mathbb{E}\left(z^{(s+1)}_{k(2l)} | C, \beta_{old}, \mu^{(0)}_{old}, \omega_{old}\right) + \delta_{\beta} - 1}{\mathbb{E}\left(z^{(s)}_{k(l)} | C, \beta_{old}, \mu^{(0)}_{old}, \omega_{old}\right) + 2(\delta_{\beta} - 1)}$$

$$\mu^{(0)}_{k0} = \frac{\mathbb{E}\left(z^{(0)}_{k(0)} | C, \beta_{old}, \mu^{(0)}_{old}, \omega_{old}\right) + \nu_{\mu} - 1}{\delta_{\mu} + 1}$$

where C reprrsents the counts matrix. Then, the expectations above can be written as

$$\mathbb{E}\left(z^{(S)}_{kl} | C, \beta_{old}, \mu^{(0)}_{old}, \omega_{old}\right) = c_{ng} \frac{\omega_{old,nk}\theta_{old,kg}}{\sum_{h=1}^{K} \omega_{old,nh}\theta_{old,hg}}$$

Then use the relation

$$\mathbb{E}\left(z^{(S)}_{kl} | C, \beta_{old}, \mu^{(0)}_{old}, \omega_{old}\right) = \mathbb{E}\left(z^{(s+1)}_{k(2l)} | C, \beta_{old}, \mu^{(0)}_{old}, \omega_{old}\right) + \mathbb{E}\left(z^{(s+1)}_{k(2l+1)} | C, \beta_{old}, \mu^{(0)}_{old}, \omega_{old}\right)$$

This multiscale approach is then used coupled with the equations

$$\mu^{(s)}_{k(2l)} = \beta^{(s-1)}_{k(l)} \mu^{(s-1)}_{k(l)}$$

$$\mu^{(s)}_{k(2l+1)} = \mu^{(s-1)}_{kl} - \mu^{(s)}_{k(2l)}$$

Updating the topic distribution parameters 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.

Questions

- Choice of hyperparameters. This question has not been addressed much in the multiscale topic tomography paper.
- Not sure if the distance between the ordered points should play a part in the model. Three features could be ordered next to each other but the first two may be pretty close (say bodymass 5 and 6 in birds data) and the third one is far apart (say bodymass 100 in birds data).
- What is the interpretation of the topics under this set up and how it differs from the original topic model.
- Can we compare this model with the original model likelihoodwise and how to test the hypothesis whether the ordering matters for the clustering or not.