Topographical Topic Models

Kushal K Dey

January 24, 2016

Introduction

In phylogenetic studies, one often encounters assemblage maps that plot counts/abundances of different bird species on geographical maps. Recently, bird species abundance data was collected from 35 forest spots in Eastern and Western Himalayas by Trevor Price's lab. Assemblage map was then computed for the bird species in each forest spot depending on the presence/absence and/or the relative abundance patterns. The idea then is to cluster these assemblage maps using a graded membership model and represent each map by a weighted combination of a set of base maps.

Model

Let us assume that each map is a grid with R rows and C columns. Therefore there are $V = R \times C$ cells/ voxels. We can pool the V cells/voxels in the assemblage map and view them as as features. Each cell/voxel contains a count value representing the total number of species present in that cell or the total abundance of all species in that cell. Let c_{nv} be the counts of reads number of bird species/ total number of birds observed across different species in forest spot n and cell/voxel v. Let c_{n+} be the sum of the counts across the cells for forest spot n (this would be equal to the total counts in the nth map corresponding to the assemblage map for n th forest spot).

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

$$p_{nv} = \sum_{k=1}^{K} \omega_{nk} f_{kv} \qquad \sum_{k} \omega_{nk} = 1 \quad \forall n \qquad \sum_{v} f_{kv} = 1 \quad \forall k$$

Here $\omega_{n.}$ represents the topic proportions for n th forest spot. On the other hand $f_{k.}$ represents the probability distribution or relative intensity across the different cells for the k th cluster or topic. Note that $f_{k.}$ may also be visualized as the base image for the k th cluster/topic. We assume the priors

$$(\omega_{n1}, \omega_{n2}, \cdots, \omega_{nK}) \sim Dir_K \left(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K} \right) \qquad \forall n$$

$$(f_{k1}, f_{k2}, \cdots, f_{kV}) \sim Dir_V \left(\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{kV} \right)$$

We define

$$\alpha_{kv} := \frac{exp(-\frac{||r_v - \mu_k||^2}{\lambda_k})}{\sum_v exp(-\frac{||r_v - \mu_k||^2}{\lambda_k})} \tag{1}$$

This function is called *Gaussian radial basis function.* Note that by definition $\sum_v \alpha_v = 1$ and this can be viewed as the weight we are assigning to each cell or voxel for the k th base map/cluster.

We assume μ_k and λ_k to be hyperparameters in the model although in TFA, prior distributions are assumed for these two parameters.

Model Specifications

We can assume that

$$c_{n+} \sim Poi(\lambda_n) \tag{2}$$

Then given the model specifications, we get

$$c_{nv} \sim Poi\left(\lambda_n \sum_k \omega_{nk} f_{kv}\right) \tag{3}$$

Let z_{nkv} represents the number of counts from forest spot n and cell v that comes from k th subgroup or base map. By definition,

$$\sum_{k=1}^{K} z_{nkv} = c_{nv}$$

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

$$z_{nkv} \sim Poi\left(\lambda_n \omega_{nk} f_{kv}\right)$$

We define

$$z_{kv} := \sum_n z_{nkv}$$

z_{kv} is the number of counts of bird species corresponding to k th base map and v th cell. We can write

$$z_{kv} \sim Poi\left(f_{kv} \sum_n \lambda_n \omega_{nk}\right)$$

Model Estimation

We start at iteration n and suppose we have the iterates $\mu_k^{(m)}$, $\lambda_k^{(m)}$, $\omega_{nk}^{(m)}$ and $f_{kv}^{(m)}$. We want to update the parameters for the $m+1$ th step. We update f_{kv} using the EM algorithm.

$$\mathcal{Q}\left(f_{k.}|C_{N\times G}, f^{(m)}, \mu_k^{(m)}, \lambda_k^{(m)}, \omega^{(m)}\right) = \mathbb{E}_{Z|C_{N\times G}, f^{(m)}, \mu_k^{(m)}, \lambda_k^{(m)}, \omega^{(m)}}\left(log\ \mathcal{L}(z_{k1}, z_{k2}, \cdots, z_{kV}|f_{k.}) + log\ \pi(f_{k.}|\mu_{k.}^{(m)}, \lambda_{k.}^{(m)})\right) \tag{4}$$

where

$$\pi(f_{k.}|\mu_k, \lambda_k) \propto \prod_{v=1}^{V} f_{kv}^{\alpha_{kv}} \tag{5}$$

where α_{kv} is defined as per Equation 1.

and

$$\mathcal{L}(z_{k1}, z_{k2}, \cdots, z_{kV} | f_{k.}) := Mult\,(z_{k+}, f_{k1}, f_{k2}, \cdots, f_{kV}) \tag{6}$$

Using EM algorithm, the parameter updates we get are

$$f_{kv}^{(m+1)} := \frac{\mathbb{E}\left(z_{kv} | C_{N \times V}, f^{(m)}, \mu_k^{(m)}, \lambda_k^{(m)}, \omega^{(m)}\right) + \alpha_{kv}^{(m)}}{\sum_v \mathbb{E}\left(z_{kv} | C_{N \times V}, f^{(m)}, \mu_k^{(m)}, \lambda_k^{(m)}, \omega^{(m)}\right) + \sum_v \alpha_{kv}^{(m)}} \tag{7}$$

where

$$\mathbb{E}\left(z_{kv} | C_{N \times V}, f^{(m)}, \mu_k^{(m)}, \lambda_k^{(m)}, \omega^{(m)}\right) := c_{nv} \frac{\omega_{nk}^{(m)} f_{kv}^{(m)}}{\sum_{h=1}^{K} \omega_{nh}^{(m)} f_{hv}^{(m)}}$$

Given that we have the new updates $f_{kv}^{(m+1)}$, we can obtain new estimates for $\mu_k^{(m+1)}$ and $\lambda_k^{(m+1)}$. We define

$$\mu_k^{(m+1)} = \frac{\sum_v v f_{kv}^{(m)}}{\sum_v f_{kv}^{(m)}}$$

$$\lambda_k^{(m+1)} = \frac{\sum_v v^2 f_{kv}^{(m)}}{\sum_v f_{kv}^{(m)}} - (\mu_k^{(m)})^2$$

Given the new updates for f, we can update the ω parameters in the same way as done by Matt Taddy in his paper using active set strategy and convex optimization.

At the end of these steps, we will have $\omega_{nk}^{(m+1)}$, $f_{kv}^{(m+1)}$, $\mu_k^{(m+1)}$ and $\lambda_k^{(m+1)}$. We can use these to update the parameters further and we continue till the log-likelihood converges.