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 pixels/voxels in the assemblage map and view them as as features and each pixel/voxel contains a count value representing the abundances. Let c_{nv} be the counts of reads for sample n and voxel v. Let c_{n+} be the sum of reads for sample n, also called the *library size*.

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

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

Here $\omega_{n.}$ represents the topic proportions for n th samples. On the other hand $\theta_{k.}$ represents the probability distribution on the genes for the kth topic or cluster.