HW1

Jiaqian Yu-jy2880
October 10, 2018

1 Problem 1

I use Senate data as dataset. Given instruction about this dataset, I decided to generate a Bernoulli mixture model with Beta priors to analyze this dataset.

I consider senators as data point and we assume each senator has a proportional membership in K blocks, given by a probability vector $(\theta_1, \theta_2, .., \theta_k)$ that sums to one. Each block k has its own voting pattern represented by a parameter $beta$ which is generated by a Beta distribution with a suitable prior. Then $beta$ will be treated as the parameter for a Bernoulli distribution as the 'vote for probability'. So senators from different membership will have a diffferent probability to 'vote for' bills.

The mixture model I apply on this dataset is:

1. Draw proportions $\theta \sim Dir_K(\alpha)$.
2. For each compenent $k \in [1..., K]$.
(a) Draw component $\beta_k \sim g(\beta; \mu)$.
3. For each data point $i \in [1, 2, ..., n]$
(a) Draw assignment $z_i|\theta \sim Cat(\theta)$.
(b) Draw data point $x_i|\beta, z_i \sim f(x; \beta_{z_i})$.

Here $n = 103$ which is the number of senators. $g(\beta; \mu)$ is Beta distribution with $\mu = (1, 1)$ as a prior and $f(x; \beta_{z_i})$ is Bernoulli distribution with β as a parameter. And we choose α as 1 for the Dirichlet distribution.

I have known there are three political parties for these senators. I choose $K = 3$.

Then I implement Gibbs sampling for this model, the basic algorithm could be found on page 51 in the text. In every iteration step, the parameter β is updated by a new Beta distribution (which is a posterior distribution after gaining information from data points) where I calculate the total 'vote for' numbers and 'vote against' numbers within different groups and add these to the former parameters for Beta distribution.

I plot the parameters θ and β against iteration step.

I also plot the log of joint distribution $\log p(\theta^t, \beta^t, z^t, x)$ with the formula in page 46 against iteration step.

The plots are shown below:

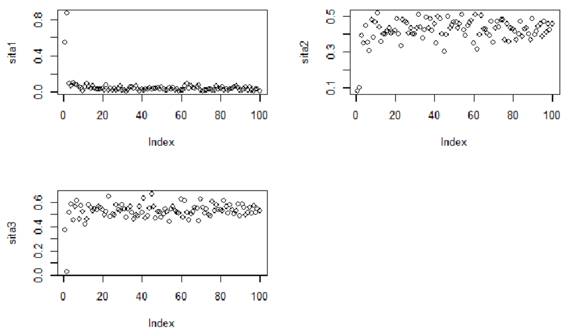

Figure 1: θ vs iteration step

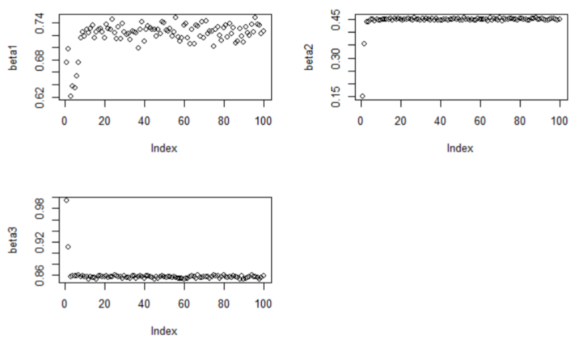

Figure 2: β vs iteration step

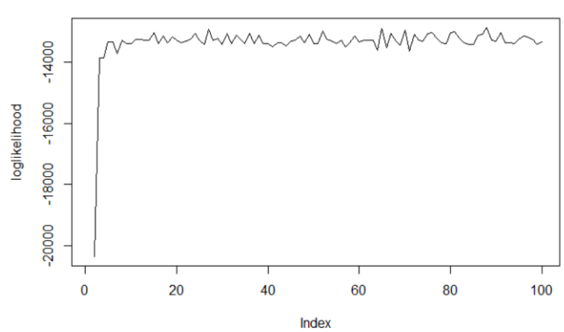

Figure 3: log joint distritbuion vs iteration step

From the plot we can see that the logarithm of joint distribution do iterate after several steps. So it means we can find a pattern However, it seems that the sampling algorithm converges too fast. And I think the reason is we choose a prior that contains least information so that our the parameter is greatly influenced by the data which has some patterns in different groups. Also, the number of groups K is really important, here I choose K based on the real number of groups in the data. So this may also occur the fast convergence of this algorithm.

2 Problem 2

I have some interest in social network data. I have large social network dataset from SNAP. And I want to see whether there are some groups in the social network.

a) The original dataset is all the directed edges in the social network. So I think I can treat every nodes as a single data point (which means every individual) and treat the characters of each nodes as variables, such as in-degree and out-degree. I think the in-degree and out-degree are independent.
b) The latent variables are the group assignments for every nodes(every individual).
c)
1. Is there any groups in the social network?
2. How can we choose a suitable distribution for those characters?
3. Can we use other characters, especially can we treat subgraphs as variables?