Comparison of Gibbs Sampler and Collapsed Gibbs Sampler under Normal Mixture Model

Hongxuan Zhai & Ashley Lu

Abstract

Normal mixture models are useful tools for both density estimation and clustering. A Gibbs sampler is used for making inference about the models. Besides the "classical" Gibbs sampler, the collapsed gibbs sampler is also tractable under conjugate assumptions. Under clustering problem, the collapsed gibbs sampler, by reducing number of sampled variables, usually yields better" estimation of the parameters. In our final project, we implement both Gibbs sampler and collapsed gibbs sampler and compare these two MCMC algorithms empirically.

Introduction

Bayesian finite normal mixture with common σ:

$$x_i|z_i \overset{ind}{\sim} N(\mu_{z_i}, \sigma^2)$$

$$\mu_k \overset{iid}{\sim} N(\mu_0, \sigma_0^2)$$

$$z_i|\pi \sim \sum_{k=1}^{K} \pi_k \delta_i(\cdot)$$

$$\pi \sim Dir(\alpha_1, ..., \alpha_K)$$

A Gibbs sampler can be used to make inference about this problem. In principle, we sample $\boldsymbol{\mu}$, $\boldsymbol{\pi}$ and $\mathbf{Z}$ iteratively from their complete conditionals and repeat this process S times until "convergence". Gibbs sampler is widely used when the joint posterior is intractable but the conditionals are easy to draw samples from. Collapsed Gibbs sampler further considers the conditional random variables in the complete conditionals. By "integrating" out those random variables, we can make the sampler just sample the random variable of our interest. Collapsed Gibbs sampler usually leads to faster convergence of the Markov chain. The tradeoff is that collapsed Gibbs sampler is more costly per iteration. If we use bayesian mixture model for clustering, the variable of our interest is $\mathbf{Z}$, the mixture assignment variable and by collapsed Gibbs sampler, we only sample those $\mathbf{Z^{(s)}}, \quad \mathbf{s} = \mathbf{1}, ..., \mathbf{S}$ at each iteration without sampling $\boldsymbol{\mu}$, $\boldsymbol{\pi}$.

Before introducing collapsed gibbs sampler for normal mixture model, we need to pay more attention to some problems associated with bayesian mixture model. The main problem is label switching problem, which makes some inference unreliable.

Lable switching problem

The lable switching problem in bayesian finite mixture model is described as swapping components and its probability in a mixture model leads to finitely many posterior maxima. For instance, given a mixture model with K components, the likelihood function,

$$f(x_i) = \sum_{k=1}^{K} \pi_k \cdot p(x_i|\theta_k),$$

remains unchanged under any permutation $\boldsymbol{\pi}^*$ of $\boldsymbol{\pi}$, where $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$. This effect will not inflencce our inference when the inferencial target is posterior predictive distribution, but this issues will play a role when

we want to learn the clusterwise parameters and also the cluster assignments. In finite mixture of normal model, the effect of label switching leads to the marginal distribution of μ_k's from the MCMC output to be multimodel and every μ_k's marginal may look pretty similar, which make it invalid for making inference about all the μ_k's. Since our project focus on the clustering functionality of bayesian normal mixture model, it will be affected by the label switching issue. As you may expect, the lable switching issue generally get more complicated as the number of components K grows. Given K, there will be $K!$ identical models which yeild exact the same likelihood. As the first step of exploring the difference between Gibbs sampler and collapsed Gibbs sampler, in this project, we do experiment with two component normal mixture model, the simplest case.

Gibbs Sampler and Collapsed Gibbs Sampler

Given $K = 2$, our model can be rewritten as

$$x_i|z_i \overset{ind}{\sim} N(\mu_{z_i}, \sigma^2)$$
$$\mu_k \overset{iid}{\sim} N(\mu_0, \sigma_0^2)$$
$$z_i|\pi \sim \sum_{k=1}^{2} \pi_k \delta_i(\cdot)$$
$$\pi_1 \sim Beta(\alpha, \beta),$$

with $\pi_1 + \pi_2 = 1$.

Gibbs sampler

Complete conditionals for π_1:

$$p(\pi_1|\boldsymbol{Z}) \propto \pi_1^{(\alpha-1)}(1-\pi_1)^{(\beta-1)}\pi_1^{n_1}(1-\pi_1)^{n_2} \sim Beta(\alpha + n_1, \beta + n_2),$$

where $n_k = ||i : z_i = k||$.

Complete conditionals for z_i:

$$p(z_i = 1|\boldsymbol{\pi}, \boldsymbol{\mu}, x_i) \propto \pi_1 p(x_i|\mu_1, \sigma^2)$$
$$p(z_i = 2|\boldsymbol{\pi}, \boldsymbol{\mu}, x_i) \propto \pi_2 p(x_i|\mu_2, \sigma^2)$$

and we draw $z_i|\boldsymbol{\pi}, \boldsymbol{\mu}, x_i$ from bernoulli distribution.

Complete conditionals for μ_k:

$$p(\mu_k|\boldsymbol{\pi}, \boldsymbol{Z}, \boldsymbol{X}) \sim N(\frac{\sigma^2}{n_k\sigma_0^2 + \sigma^2}\mu_0 + \frac{n_k\sigma_0^2}{n_k\sigma_0^2 + \sigma^2}\bar{X_k}, (n_k/\sigma^2 + 1/\sigma_0^2)^{-1}),$$

where $\bar{X_k} = \frac{\sum_{i:z_i=k} x_i}{n_k}$.

Collapsed Gibbs sampler

Given the form of

$$p(z_i|Z_{-i}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{X}), \quad where \quad Z_{-i} = (z_1, ..., z_{i-1}, z_{i+1}, z_n),$$

we can further integrate out all the "nuisance" parameters and simplyfy the MCMC algorithm. If our problem is clustering, the collapsed Gibbs sampler is done by intergrating out $\boldsymbol{\pi}, \boldsymbol{\mu}$ in the condition part. After that we are left with only $p(z_i|Z_{-i}, \boldsymbol{X})$ and we do immediately updating the mixture components after

resampling the mixture component. In fact, this process is feasiable under bayesian normal mixture model under our setting.

To derive the collapsed Gibbs sampler for two component mixture model, we first factorize the desired conditional distribution

$$p(z_i|Z_{-i}, \boldsymbol{X})$$

as

$$p(z_i|Z_{-i}, \boldsymbol{X}) = p(z_i|x_i, Z_{-i}, X_{-i}), \quad where \quad X_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n).$$

Then by the conditional independency and the fact that conditional distribution is proposional to the jiont, we obtain

$$\begin{aligned} p(z_i|x_i, Z_{-i}, X_{-i}) &\propto p(z_i|Z_{-i}, X_{-i})p(x_i|Z_{-i}, X_{-i}, z_i) \\ &\propto p(z_i|Z_{-i})p(x_i|Z_{-i}, X_{-i}, z_i) \end{aligned} \tag{1}$$

Under this kind of factorization, we can identify that the two component in our desired conditional distribution are nothing but two posterior predictive distribution.

Since we know that z_i's are discrete random variable that takes value from $\{1, 2\}$, we can further calculate the proposional weights for sampling each z_i's in the collapsed Gibbs sampler.

Derivation for posterior predictives

Posterior predictive for z_i:

$$\begin{aligned} p(z_i = 1|Z_{-i}) &= \int p(z_i = 1|\pi_1)p(\pi_1|Z_{-i}, \alpha, \beta)d\pi_1 \\ &= \int \pi_1 \frac{\pi_1^{\alpha_n^{-i}-1}(1-\pi_1)^{\beta_n^{-i}-1}}{B(\alpha_n^{-i}, \beta_n^{-i})} d\pi_1 \\ &= \frac{B(\alpha_n^{-i}+1, \beta_n^{-i})}{B(\alpha_n^{-i}, \beta_n^{-i})} \\ &= \frac{n_1^{-i}+1}{n-1}, \end{aligned} \tag{2}$$

where $n_1^{-i} = \sum_{j \neq i} \mathbb{1}(z_j = 1)$, $\alpha_n^{-i} = \alpha + n_1^{-i}$ and $\beta_n^{-i} = \beta + n_2^{-i}$. Note that this result is the same as prediction rules for urn model with two states and similar result can be derived for $z_i = 2$.

Posterior predictive for x_i:

$$\begin{aligned} p(x_i|Z_{-i}, X_{-i}, z_i = 1) &= p(x_i|\{x_j|z_j = k, j \neq i\}) \\ &= \int p(x_i|\mu_1, \sigma^2)p(\mu_1|\{x_j|z_j = k, j \neq i\})d\mu_1 \\ &\sim N(x_i|\frac{\sigma^2}{n_1^{-i}\sigma_0^2 + \sigma^2}\mu_0 + \frac{n_1^{-i}\sigma_0^2}{n_1^{-i}\sigma_0^2 + \sigma^2}\bar{X_1}^{-i}, (n_1^{-i}/\sigma^2 + 1/\sigma_0^2)^{-1}), \end{aligned} \tag{3}$$

where $\bar{X_1}^{-i} = \frac{\sum_{\{j:z_j=1, j\neq i\}} x_j}{n_1^{-i}}$. The case when $z_i = 2$ can be derived in a similar way.

Collapsed Gibbs algorithm

With two predictive distribution, we summarize the collapsed Gibbs sampler for two component mixture model as follows:

- Input: data X, K and initilization of $\boldsymbol{Z}$.

- for m in {1,...,M}
  - Calculate n_1, n_2, $\bar{X_1}$ and $\bar{X_2}$ from previous iteration.
    - for each i in {1,...,n}, calculate
      - n_1^{-i}, n_2^{-i}, $\bar{X_1}^{-i}$ and $\bar{X_2}^{-i}$.
      - sample new z_i from $p(z_i|Z_{-i}, \boldsymbol{X})$
    - end inner loop
- end outer loop

Simulation study

We simulate 11 random dataset each of sample size 1000 from mixture distribution $0.2\times N(2,1)+0.8\times N(4,1)$ with the first 200 being from component with mean 2 and remaining 800 being from component with mean 4. Within each iteration from both sampler, we calculate the cluster allocation accuracy $A^{(s)} = \sum_{i=1}^{n} \mathbb{1}(z_i^{(s)} = z_i^{(true)})/n$, and compare the two samplers based on those A's. Agian, the MCMC output is affected by the lable swithcing problem.

Hyper parameters are fixed at $\mu_0 = 3$, $\sigma_0^2 = 0.5$, $\alpha = 1$ and $\beta = 1$. The variance parameter σ^2 for each component is set at 1.

Figure 1 is one plot for prediction accuracy from the MCMC output under two samplers. As the plot might suggest, in this MCMC sampling, the Gibbs sampler suffers from the lable switching problem from iteration around 6000 to 7000. This claim can be further checked by figure 2. The output of collapsed Gibbs sampler, however, seems to be more stable than the output from Gibbs sampler. This phenomenon can be just coincidence since the output is from one dataset and the MCMC algorithms only run for once. To have a more clear view of the phenomenon, we do this process 10 more times.

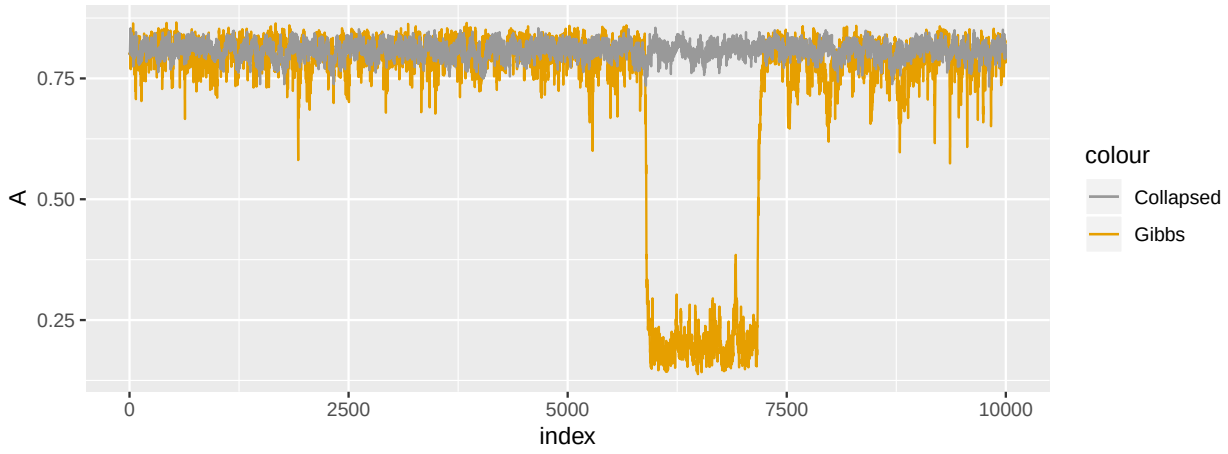

Figure 1: Trace plots for A

Figure 3 show the output of A from both Gibbs sampler and collapsed Gibbs sampler for 4 random dataset generated from mixture model $0.2 \times N(2,1) + 0.8 \times N(4,1)$. For these four dataset, we no longer observe the phenomenon in figure 1, but we can still see that the fluctuation of Gibbs sampler's trace plot tend to be bigger than that of collapsed Gibbs sampler. In terms of "prediction" acccuracy on average, both methods tend to be 80% accurate, which indicates that the means of allocation accuracy do not differ that much. In terms of convergence rate, both methods converge (or find the local mode in the posterior) pretty fast given reasonable initial values.

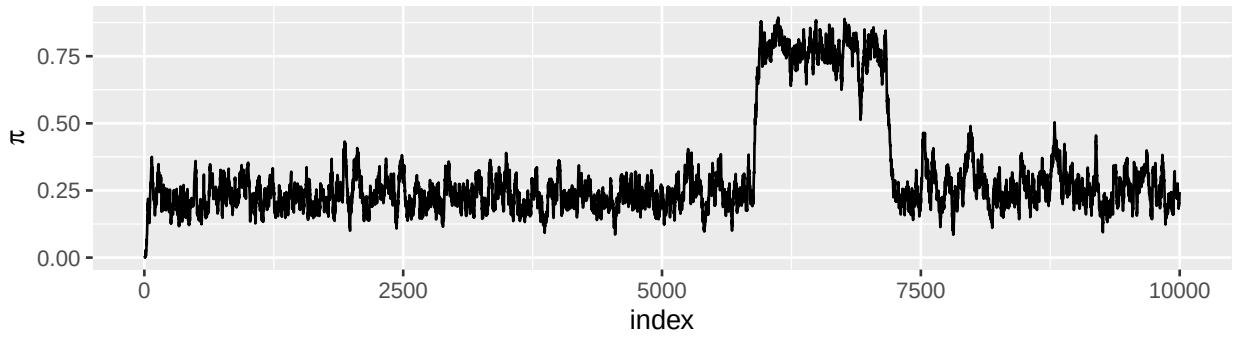

Figure 2: Trace plots for pi

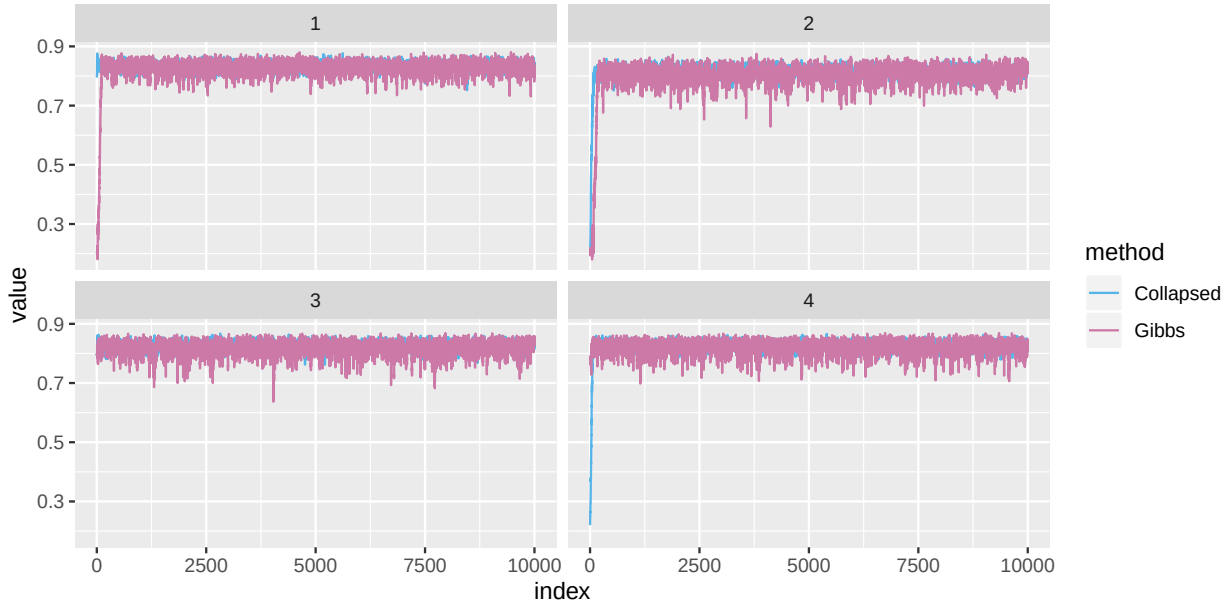

Figure 3: Trace plots for A for different dataset

Figure 4 describes the mean allocation accuracy and its uncertainty. The dot within the error bar represents the mean value of allocation accuracy and the error bar is given by plus and minus one standard deviation of the allocation accuracy's MCMC output. From the plot, we can see clearly that collapsed gibbs sampler's results have less uncertainty and tend to be more stable in assigning "good" allocation. However, in dataset 6 and 11, Gibbs sampler's results are relatively unstable. Especially in dataset 6, Gibbs sampler's results behave more or less like a random guessing.

We also do the same experiment under different dataset and observe similar pattern of the performance of those two different MCMC method. As is always the case, the performance (allocation accuracy) of two samplers is not only affected by lable switching, but also heavily predetermined by the intrinsic level of difficulty of the clustering problem. This is equivalent to say that if two clusters' mean parameters are really close, we do not expect the two MCMC algorithm to give "good" clustering results. in which two method are both inefficient or redundant.

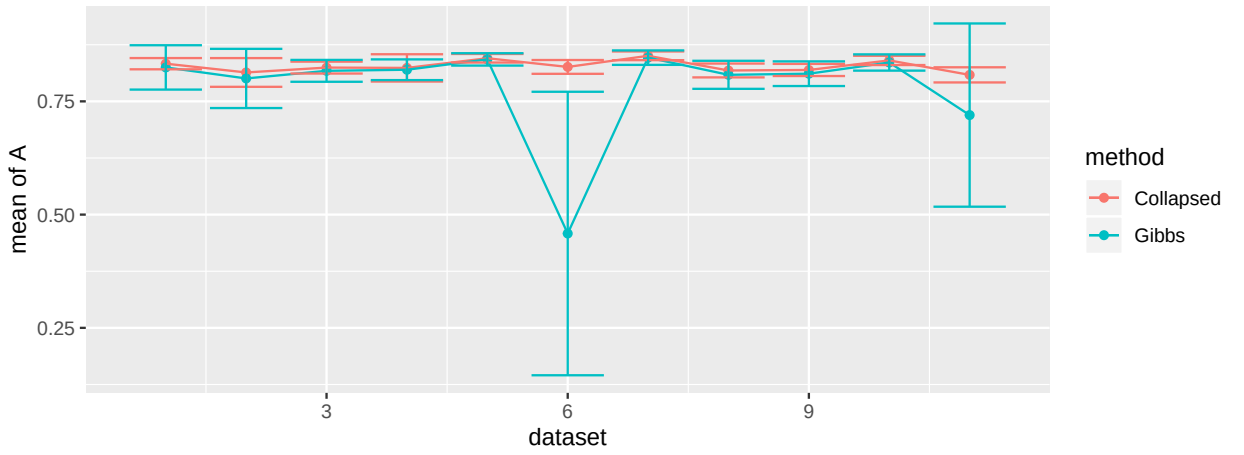

Figure 4: Error bar plots for A

Conclusion and Problem

In this final project, we compare Gibbs sampler and collapsed Gibbs sampler under the simplest two componet normal mixture model. The comparison criterion is based on the allocation accuracy from the MCMC output. Our findings are that collapsed Gibbs sampler yeild better MCMC result in the sense of the allocation accuracy's fluctuation is smaller. However, as we always mentioned in this project, lable switching should not be ignored in this clustering problem. But from our empirical resluts under our simulation dataset, collapsed Gibbs sampler seems to be less affected by lable swithcing.

Limitations for our project are also obvious. We choose number-of-component parameter $K = 2$ only because, to some extent, we can ameliorate the lable switching problem since there are only two models with its labels switched that yeild the same likelihood and the models are mutually exclusive. With K increases, more advanced techinques dealing with lable switching should be done for making reliable inference about the mixture component.

The future work is to compare those two samplers under nonparametric mixture model where the K, the number of mixture component, is no longer fixed but a random variable. Both Gibbs sampler and collapsed Gibbs sampler can be implemented for making inference about K. Also notice that this random quantity K is not affected by lable switching and thus the nonparametric model is a more appropriate setting for comparing or exploring the differncenes of Gibbs sampler and collapsed Gibbs sampler. Chances are that with nonparametric model, collapsed Gibbs sampler will yield faster inference about K in terms of convergence.

Appendix

Codes available at: https://github.com/HongxuanZhai/BDA_final_project

Reference

1. http://sap.ist.i.kyoto-u.ac.jp/members/yoshii/lectures/pattern_recognition/2017/20170606-npb-gmm.pdf
2. http://www.cs.columbia.edu/~blei/fogm/2015F/notes/mixtures-and-gibbs.pdf
3. https://dp.tdhopper.com/collapsed-gibbs/