Homework 2
Hengxu Lin (hl3541)
October 29, 2022

1 Problem: Gibbs Sampling

1.1 Brief Recap of the Data

- **Description**. The Expedia dataset is a private dataset obtained from Wharton Customer Analytics. The data is comprised of a search transaction dataset and a clickstream dataset. I use the search transaction dataset in this project, which includes user search record per hotel on Expedia.
- **Selected Features.** To cut down the running time of the algorithm, I manually select 3 numerical variables which is related to consumer clicks (Hotel Price, Room Capacity and Traveler Review Counts), and yield 10,000 observations. The original features include searching criteria, search results, price, rating, promotion, hotel characteristics and the dependent variable is whether a hotel is clicked. I choose these three features since it is nearly normal distributed and fit the Mixture Gaussian assumption.

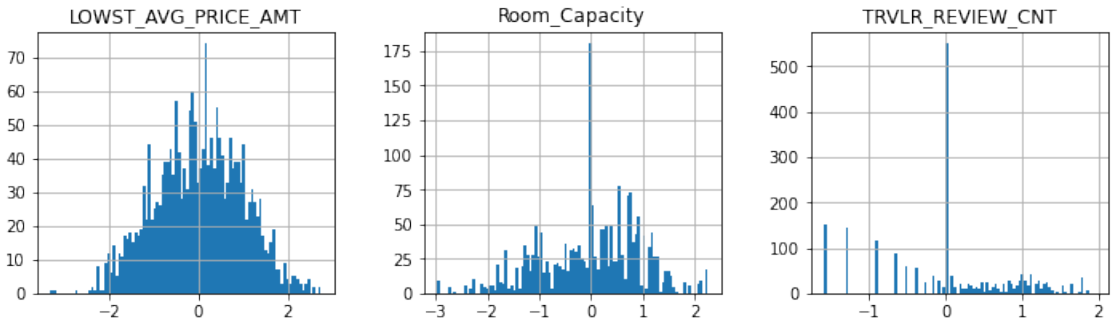

Figure 1: Selected Click-stream Features

1.2 Log Joint Probability

I use a Mixture Gaussian to represent my data. The proportion distribution is a multinomial distribution with the conjugate prior as Dirichlet distribution.

I implement one Gibbs Sampling on my data set, and record the log joint to check the convergence. Although I only have one sweep of Gibbs sampling, the algorithm converge pretty well. It should be good to continue to maximize the posterior probability.

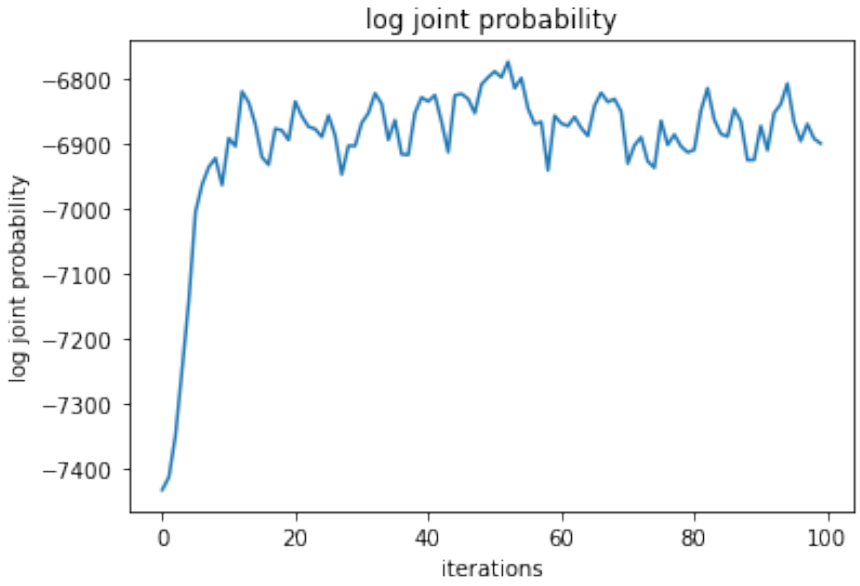

Figure 2: Log Joint Probability of the Posterior

To check how the groups are identified by Mixture Gaussian, I plot the group data in two dimensional space. It seems with one Gibbs sampling the data was not divided pretty well. Hence I further create a synthetic data set to check my Gibbs sampler.

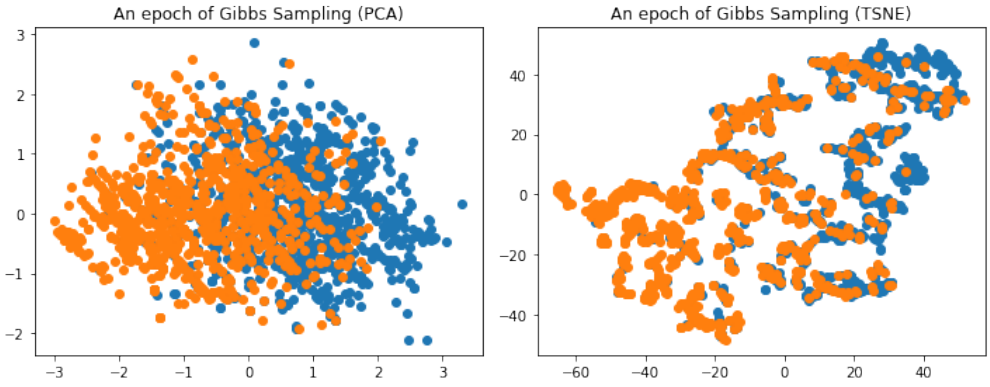

Figure 3: Visualization of the Group Data

The synthetic data has salient two groups, I manually select half of the data in the population with a higher Price.

I repeat the Gibbs sampler and obtain two salient groups. The log joint probability shows a more stable convergence in this data. In my manual-crafted data set, the distribution is easier to identify hence only one Gibbs sampling epoch splits the data very well.

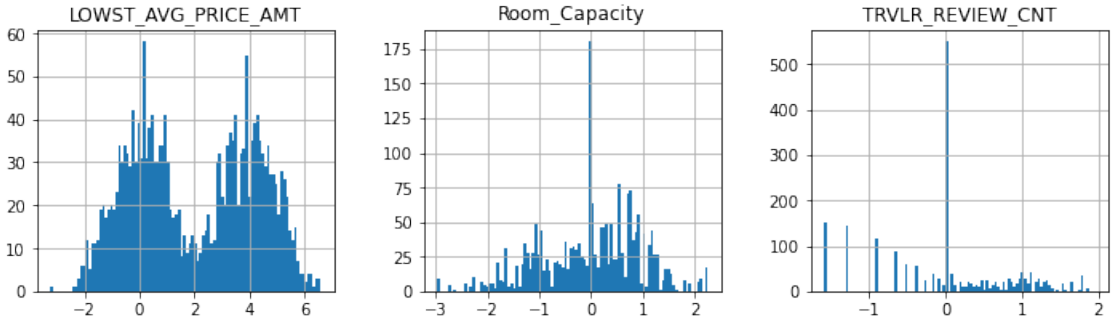

Figure 4: Price distribution with two groups in the data

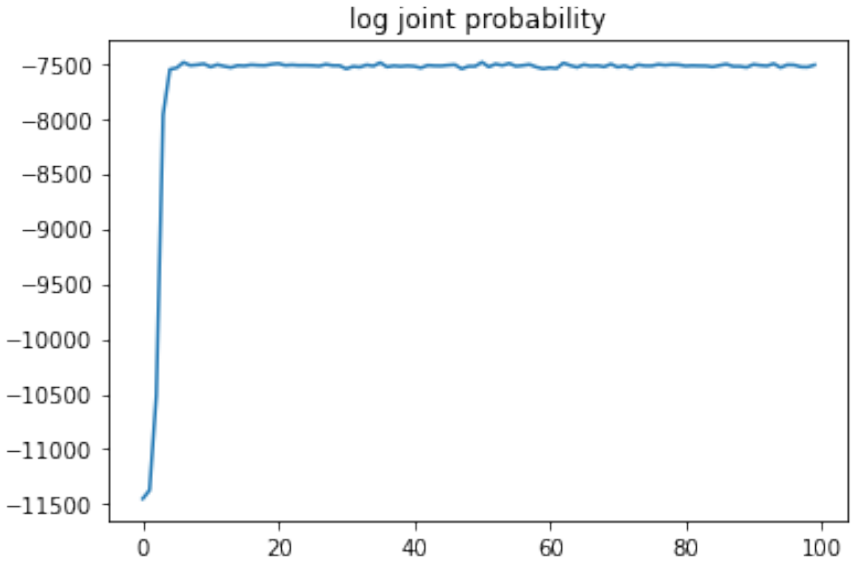

Figure 5: Log Joint Probability of the Posterior

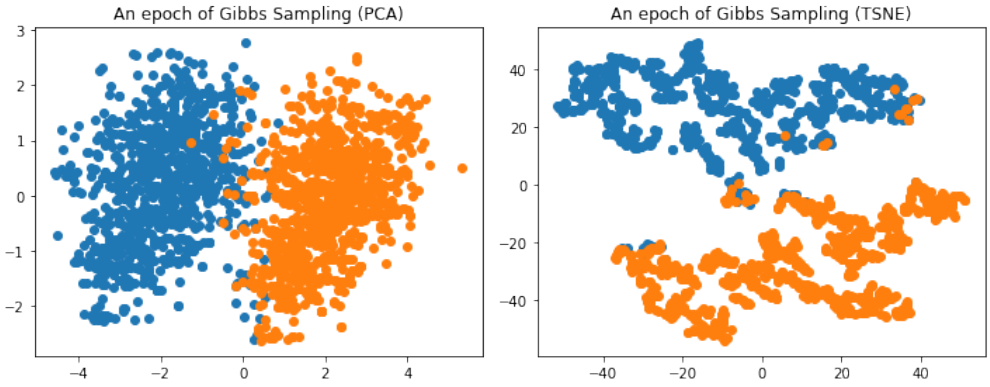

Figure 6: Visualization of the Group Data