VARIATIONAL INFERENCE AND MCMC

PROJECT REPORT

Dhruvil Sangani
Department of Mathematics
Indian Institute of Technology Kanpur
Kanpur, India 208016
dhruvil@iitk.ac.in

Nitin Garg
Department of Mathematics
Indian Institute of Technology Kanpur
Kanpur, India 208016
nitingrg@iitk.ac.in

August 19, 2020

ABSTRACT

Variational Inference is a very popular method to obtain the posterior for a probability distribution when it gets difficult to compute. We first put forward a family of densities and then obtain the member of that family which is closest to our true posterior. In this report, we measure the closeness using KL divergence.

1 Introduction

Variational Inference is widely used to optimize posterior densities for Bayesian models. Another method called Markov chain Monte Carlo (MCMC) is also used for this purpose. Variational inference runs faster and is easily scalable to huge data sets. So, keeping in mind these advantages we have applied variational inference to some datasets. The most general problem starts with considering the joint density of latent variables **z**=$\mathrm{z}_{1:m}$ and observations **x**=$\mathrm{x}_{1:n}$,

$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{z})p(\mathbf{x} \mid \mathbf{z}) \tag{1}$$

The idea behind variational inference is to turn the probelem at hand to an optimization problem. First we posit a family of approximate densities $\bar{2}$. Then we try to find a member of that family which minimizes the Kullback-Leibler(KL) divergence.

$$q^*(\mathbf{z}) = \arg\min KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) \tag{2}$$

2 Variational Inference

The main goal of variational inference is to approximate the conditional density of latent variables given the observed variables. We have a family of approximate densities and then we use KL divergence to see which member of the family most closely resembles the required density. The fitted density then serves as a proxy for the exact conditional density. We will focus our attention on Mean Field variational inference only which would involve approximating the posterior with a Bayesian Mixture of Gaussians.

2.1 The problem of approximate inference

We can write the conditional density as...

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{x})} \tag{3}$$

where p(x) is the marginal density of the observation, also called the *evidence*. We can calculate it by...

$$p(\mathbf{x}) = \int p(\mathbf{z},\mathbf{x})d\mathbf{z} \tag{4}$$

Bayesian mixture of Gaussians
The full hierarchical model is

$$\mu \sim \mathcal{N}(0, \sigma^2), \tag{5}$$

$$c_i \sim \text{Categorical}(1/K, ..., 1/K), \tag{6}$$

$$x_i|c_i, \mu \sim N(c_i^T \mu, 1) \tag{7}$$

For a sample n, the joint density of latent and observed variables is...

$$p(\mu, c, x) = p(\mu) \prod_{i=1}^{n} p(c_i) p(x_i|c_i, \mu) \tag{8}$$

here the latent variables are z={μ , c}, The K class means and n class assignments. Here, the evidence is

$$p(x) = \int p(\mu) \prod_{i=1}^{n} \sum_{c_i} p(c_i) p(x_i|c_i, \mu) \tag{9}$$

2.2 The mean-field variational family

Now we need to introduce variational family 2. A generic member of the mean-field variational family is given by...

$$q(z) = \Pi q_i(z_i)$$

here, each latent variable is governed by its variational factors $q_{i's}$ and these factors are used to maximize ELBO.

Consider again the Bayesian mixture of Gaussians, the mean fields variational densities are somewhat of the form

$$q(\mu, c) = \Pi q(\mu_k; m_k, s_k^2) \Pi q(c_i; \varphi) \tag{10}$$

2.3 Evidence Lower Bound

As we have seen above, we now need to solve the following optimization problem

$$q^*(z) = \arg\min KL(q(z)||p(z|x)) \tag{11}$$

Once we find q*(.), it would be the best approximation for the exact conditional density. However, this problem is not solvable easily as it requires computation of $\log p(x)$. Now, the KL divergence would be

$$KL(q(z)||p(z|x)) = \mathbb{E}[\log q(z)] - \mathbb{E}[\log p(z|x)] \tag{12}$$

Expanding the above equation, we get

$$KL(q(z)||p(z|x)) = \mathbb{E}[\log q(z)] - \mathbb{E}[\log p(z, x)] + \log p(x) \tag{13}$$

This shows the dependence on $\log p(x)$

To defy this problem we define a new variable named Evidence Lower Bound (ELBO) which is given by

$$ELBO(q) = \mathbb{E}[\log(p(z, x)] - \mathbb{E}[\log(q(z)] \tag{14}$$

We rewrite the ELBO equation as follows

$$ELBO(q) = \mathbb{E}[\log(p(z)] + \mathbb{E}[\log(p(x|z)] - \mathbb{E}(\log(q(z)) = \mathbb{E}(] \log(p(x|z)) - KL(q(z)||p(z)) \tag{15}$$

Also from above we know that

$$\log(p(x) = KL(q(z)||p(x|z)) + ELBO(q) \tag{16}$$

In particular, for our bayesian mixture of gaussians, consider K mixture components and n data points $x_{1:n}$. The latent variables are the K mean parameters $\mu = \mu_{1:K}$ and n latent-class assignments $c = c_{1:n}$. The assignment c_i indicates which latent cluster x_i comes from.
There is a fixed hyperparameter σ , the variance of the normal prior on the μk's. We vary it to observe different results.

The ELBO is a function of the variational parameters m, s^2, and ϕ,

$$\begin{aligned} ELBO(m, s^2, \varphi) = \sum_{k=1}^{K} \mathbb{E}[\log p(\mu_k); m_k, s_k^2] \\ + \sum_{i=1}^{n} (\mathbb{E}[\log p(c_i; \varphi_i)] + \mathbb{E}[\log p(x_i|c_i, \mu); \varphi_i, m, s^2]) \\ - \sum_{i=1}^{n} \mathbb{E}[\log q(c_i; \varphi_i) - \sum_{k=1}^{K} \mathbb{E}[\log q(\mu_k; m_k, s_k^2)] \end{aligned} \tag{17}$$

2.4 Optimal Updates for the variational parameters

We have to update the variational parameters in order to minimize the KL divergence-

2.4.1 The c update

We can write

$$p(x_i|c_i, \mu) = \prod_{k=1}^{K} p(x_i|\mu_k)^{c_{ik}} \tag{18}$$

Upon using this, we find that the update is

$$\varphi_{ik} \propto \exp\{\mathbb{E}[\mu_k; m_k, s_k^2]x_i - \mathbb{E}[\mu_k^2; m_k, s_k^2]/2\} \tag{19}$$

2.4.2 The m and the s update

Now φ_{ik} is the probability that the ith observation comes from the kth cluster so we see that $\varphi_{ik} = \mathbb{E}[c_i; \varphi_i]$. Upon using this, we obtain the updates for the mixture-component means as follows

$$m_k = \frac{\Sigma_i \varphi_{ik} x_i}{1/\sigma^2 + \Sigma_i \varphi_{ik}}, s_k^2 = \frac{1}{1/\sigma^2 + \Sigma_i \varphi_{ik}} \tag{20}$$

2.5 Co-ordinate ascent mean field variational inference

This algorithm goes on like this. First we consider the jth latent variable z_j. The complete conditional of z_j is its conditional density given all the other latent variables in the model and the observations, p(z_j|z_{-j},x). The optimal $q_j(z_j)$ is then proportional to

$$q_j^*(z_j) = \exp\{\mathbb{E}_{-j}[\log p(z_j|z_{-j}, x)]\} \tag{21}$$

Equivalently, the above equation is proportional to the exponentiated log of the joint

$$q_j^*(z_j) = \exp\{\mathbb{E}_{-j}[\log p(z_j, z_{-j}, x)]\} \tag{22}$$

The CAVI algorithm can be coded as follows

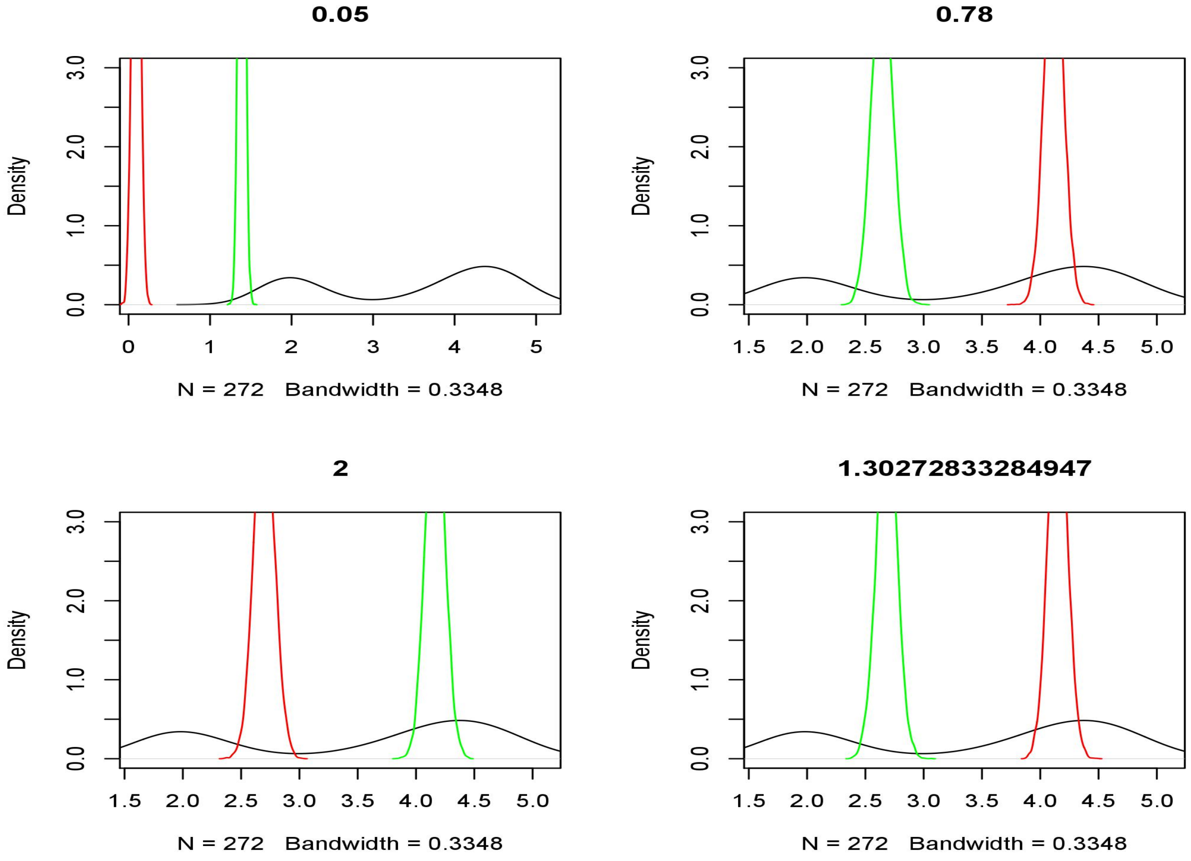
Algorithm 1 : CAVI for a Gaussian Mixture model

```
Input: Data $x_{1:n}$, number of components $K$, prior variance of component means $\sigma^2$
Output: Variational densities $q(\mu_k; m_k, s_k^2)$ (Gaussian) and $q(c_i; \varphi_i)$($K$-categorical)
Initialize: Variational parameters $m = m_{1:K}$, $s^2 = s_{1:K}^2$ and $\varphi = \varphi_{1:n}$
while the ELBO has not converged do
    for $i \in \{1, ..., n\}$ do
        Set $\varphi_{ik} \propto \exp\{\mathbb{E}[\mu_k; m_k, s_k^2]x_i - \mathbb{E}[\mu_k^2; m_k, s_k^2]/2\}$
    end
    for $k \in \{1, ..., K\}$ do
        Set $m_k \leftarrow \frac{\Sigma_i \varphi_{ik} x_i}{1/\sigma^2 + \Sigma_i \varphi_{ik}}$
        Set $s_k^2 \leftarrow \frac{1}{1/\sigma^2 + \Sigma_i \varphi_{ik}}$
    end
    Compute ELBO$(m, s^2, \varphi)$
end
return $q(m, s^2, \varphi)$
```

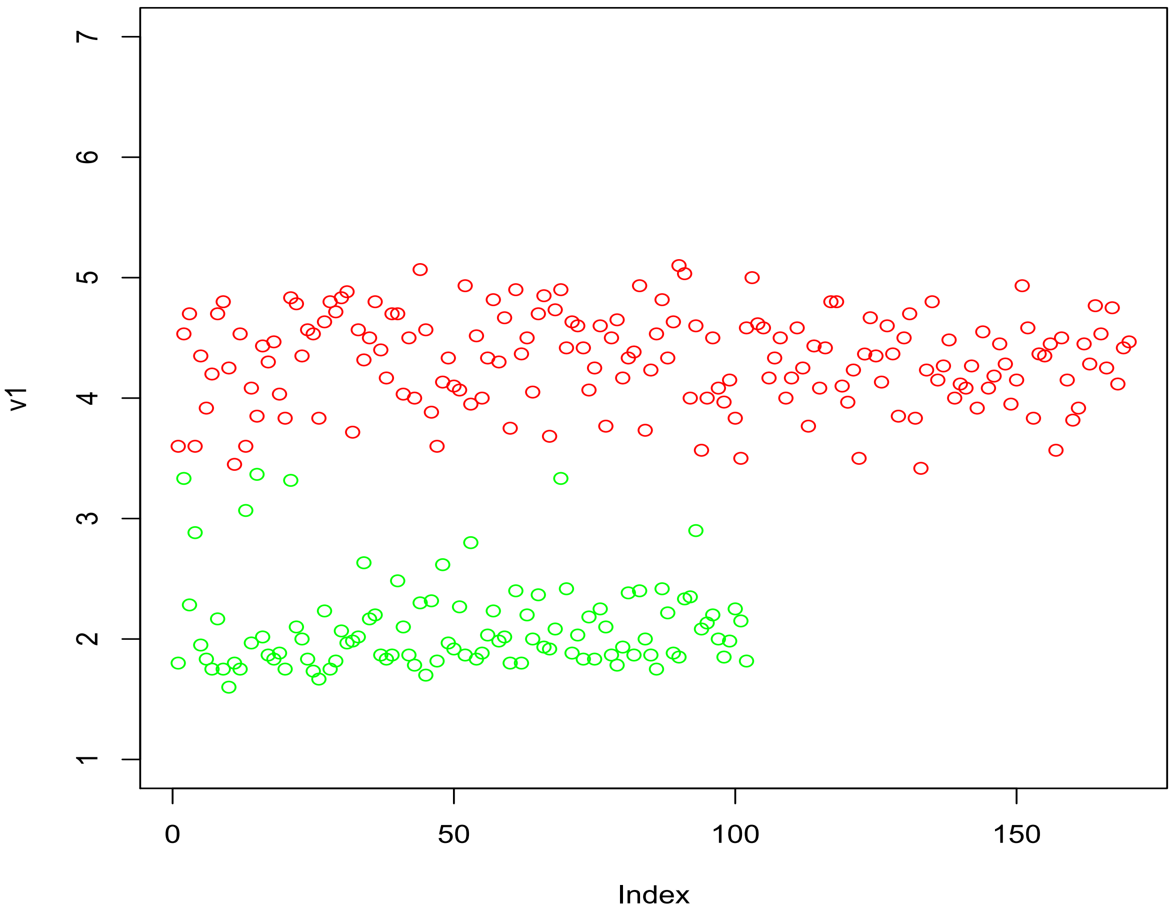
This algorithm takes elbo uphill till we arrive at a plane surface where elbo converges.

2.6 An Implementation in R

We applied variational inference to a real data set viz. the Old Faithful Geyser Data. It contains the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. We implemented the algorithm in R and obtained the following results for different values of the hyperparameter σ-



For the case where $\sigma = 0.05$, we observe that as chosen variance is very small, the algorithm fails to predict very accurately as to what the posterior distribution is.
As σ is increased, the algorithm tends to perform much better but the result do not change much after a certain point.
In the last case, σ is chosen to be exactly equal to the variance of the input data set.



Above is a classification of the same data, which is obtained via the variational parameter φ. The data has been classified into two classes according to the clear gap that is present in the data.

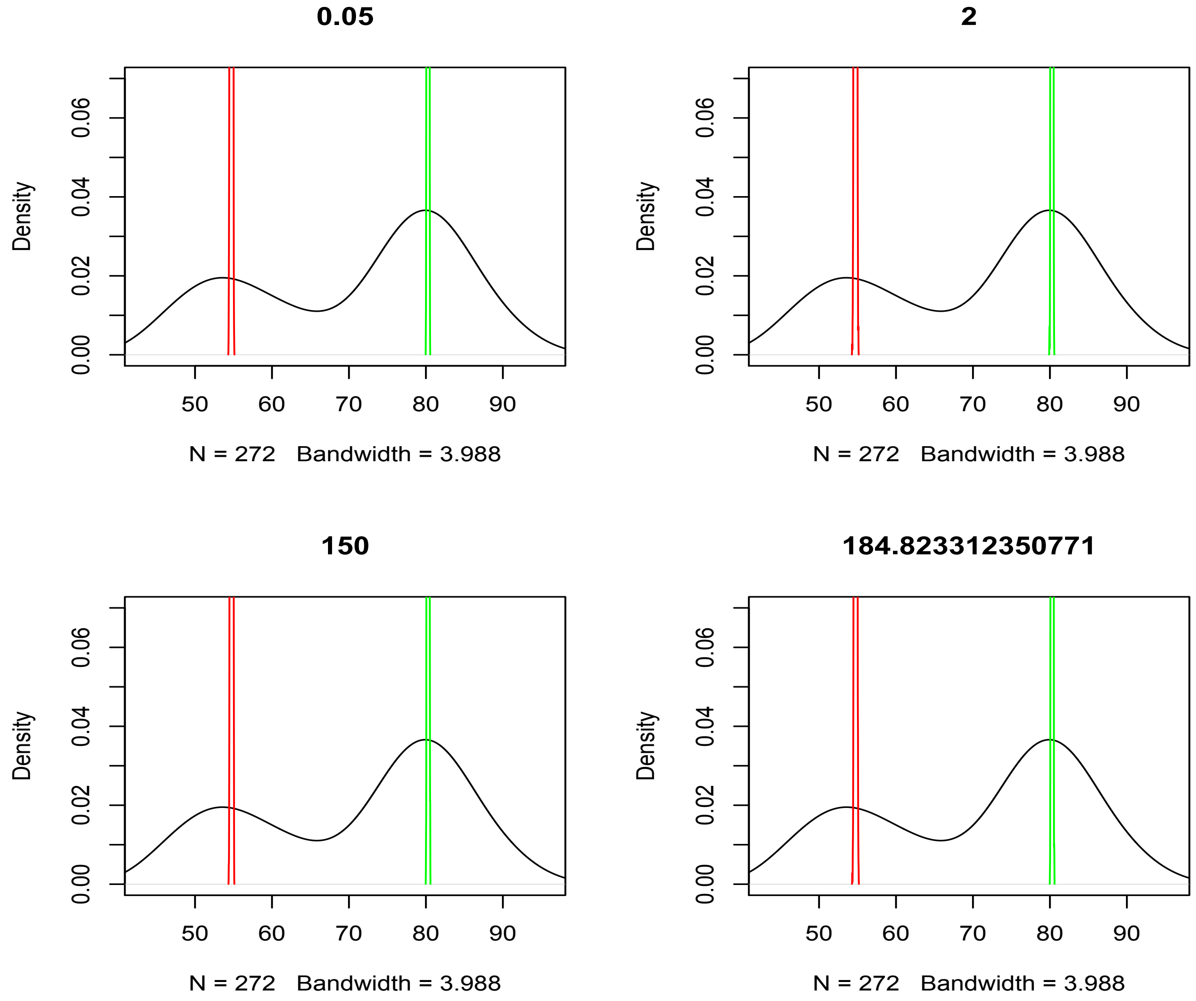
When we run the code on the waiting dataset, due to numerical instability it didn't give any relevant results.

Now, to tackle this problem, the first thing that strikes our mind is to scale the data down. But due to scaling, we lose the true essence of the data. The hyperparameters also get changed.

So we use the log-sum-exp trick to compute φ.

$$\log(\sum_i \exp(x_i)) = m + \log(\sum_i (x_i - m)) \tag{23}$$

Upon implementing the improved code and running it on the waiting dataset, we get the following results for different values of the hyperparameter σ.



We can see that the estimates that we obtain for the means are very close to the actual means and the data given is thus much closer to a Gaussian distribution and we can use the obtained distribution as a proxy to the actual distribution without much error.

3 Monte Carlo Markov Chains

MCMC methods are used to approximate the posterior distribution over latent variables by random sampling in a probabilistic space. We will focus on a particular MCMC technique namely Gibbs Sampling where we approximate a distribution using a set of samples.

3.1 Our model

The model is essentially the same but for the sake of completeness, we have to to approximate a distribution with a mixture of gaussians. Once again, the posterior is intractable to compute because of the denominator i.e. $p(x)$ so we need approximate inference. To do this we will define a Markov chain whose state space are the latent variables and whose stationary distribution is the posterior that we want.

3.2 Gibbs Sampling

The Gibbs is an iterative algorithm. It maintains a value for each latent variable. In each iteration, it samples from each latent variable conditional on the other latent variables and the observations. Call each of these distributions a complete conditional. Now if we do this many times, then the resulting sample will be a sample from the exact posterior. Here is an algorithm for the same

Algorithm 2 Gibbs Sampler for mixture of Gaussians

```
Input: data x and a number of components K.
Initialize: mixture locations $\mu$ randomly
Maintain mixture locations $\mu$ and mixture assignments z.

repeat
    for each data point i do
        Sample $z_i$ | {$\mu$, x}
    end
    for each mixture component k do
        Sample $\mu_k$ | {z, x}
    end
until;
```

We sample z_i's according to the following rule

$$p(z_i|\mu, x_i) \propto \pi_{z_i}\phi(x_i; \mu_{z_i}, \sigma^2) \tag{24}$$

For simplicity we assume π_k to be 1/K.

The μ's are sampled as follows

$$\mu_k|z, x \sim \mathcal{N}(\hat{\mu}_k, \hat{\lambda}_k^2) \tag{25}$$

Define n_k as the no. of data points belonging to the Kth cluster i.e.

$$n_k = \sum_{i=1}^{n} z_i^k \tag{26}$$

where z_i is an indicator vector i.e. a K-vector with a single one.

Define $\bar{x_k}$ as the means of all the data points in the Kth cluster i.e.

$$\bar{x}_k = \frac{\sum_{i=1}^{n} z_i^k x_i}{n_k} \tag{27}$$

and

$$\hat{\mu}_k = (\frac{n_k/\sigma^2}{n_k/\sigma^2 + 1/\lambda^2})\bar{x_k} \tag{28}$$

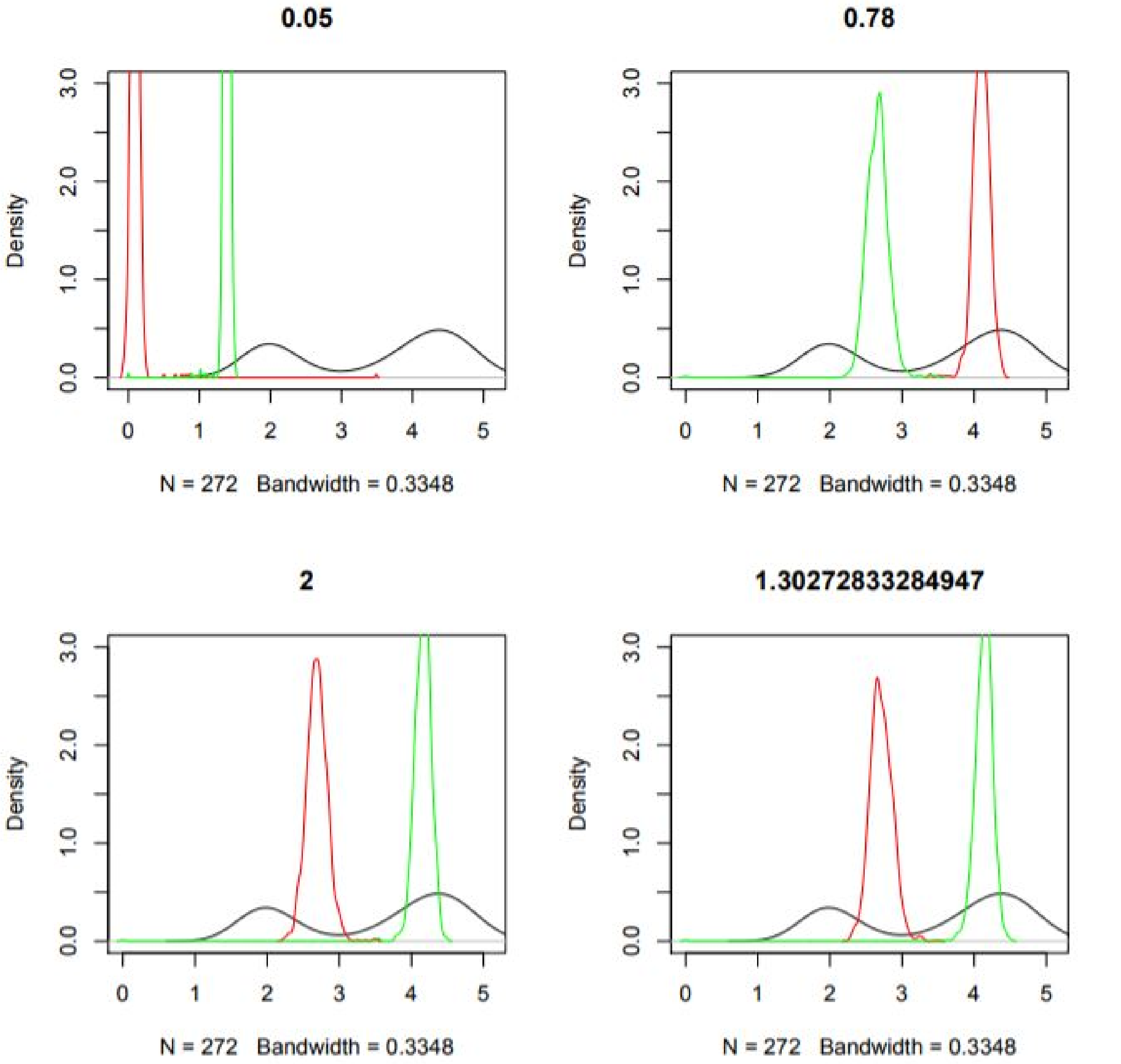
$$\hat{\lambda}_k^2 = (n_k/\sigma^2 + 1/\lambda^2)^{-1} \tag{29}$$

This is an approximate inference algorithm for estimating the posterior for a mixture of Gaussians. At each iteration, we first sample each mixture assignment from Equation 24 and then sample each mixture location from Equation 25.

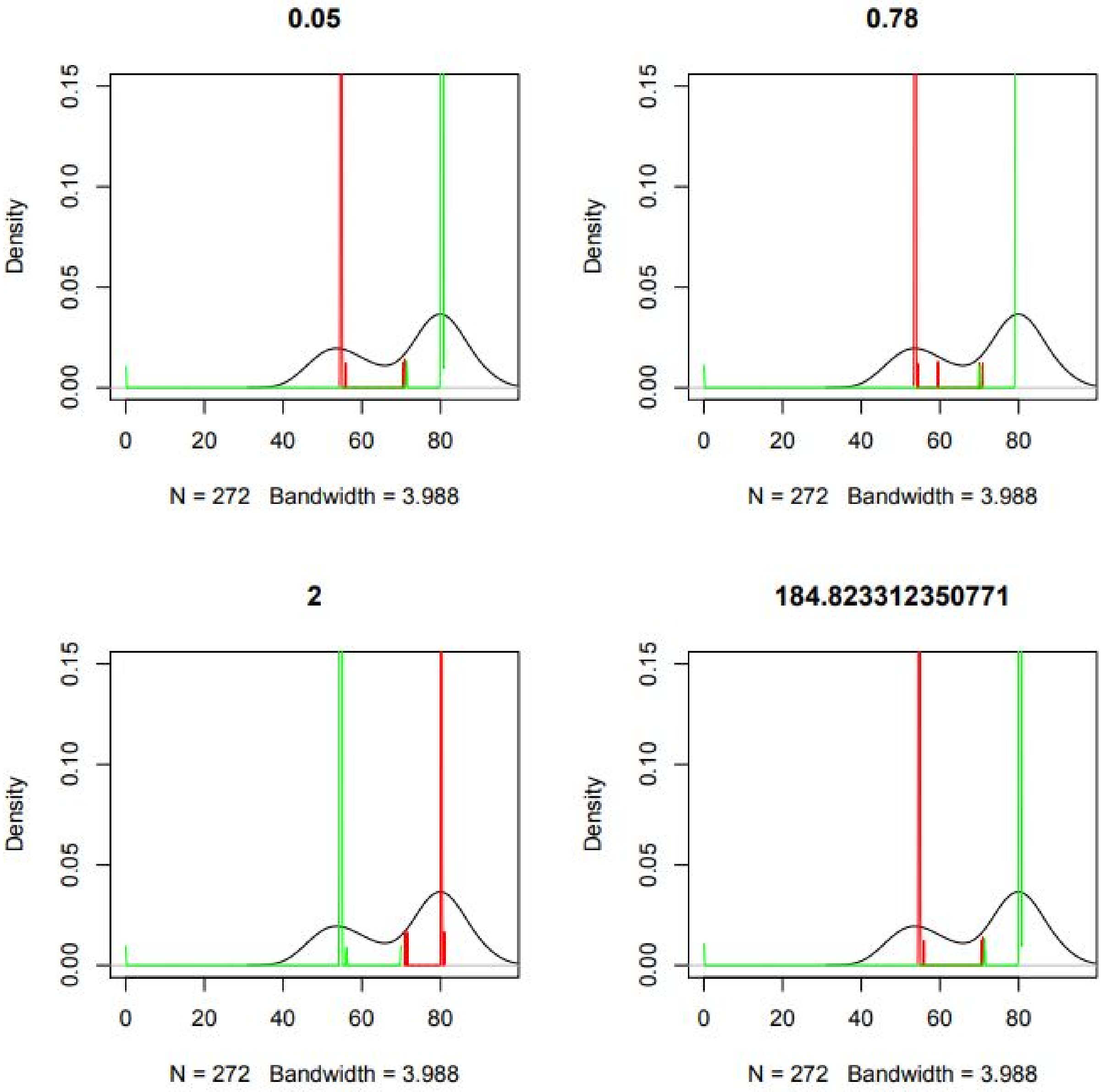
3.3 Implementation

We implemented the Gibbs sampling algorithm for the Old Faithful Geyser Data like we did for variational inference.

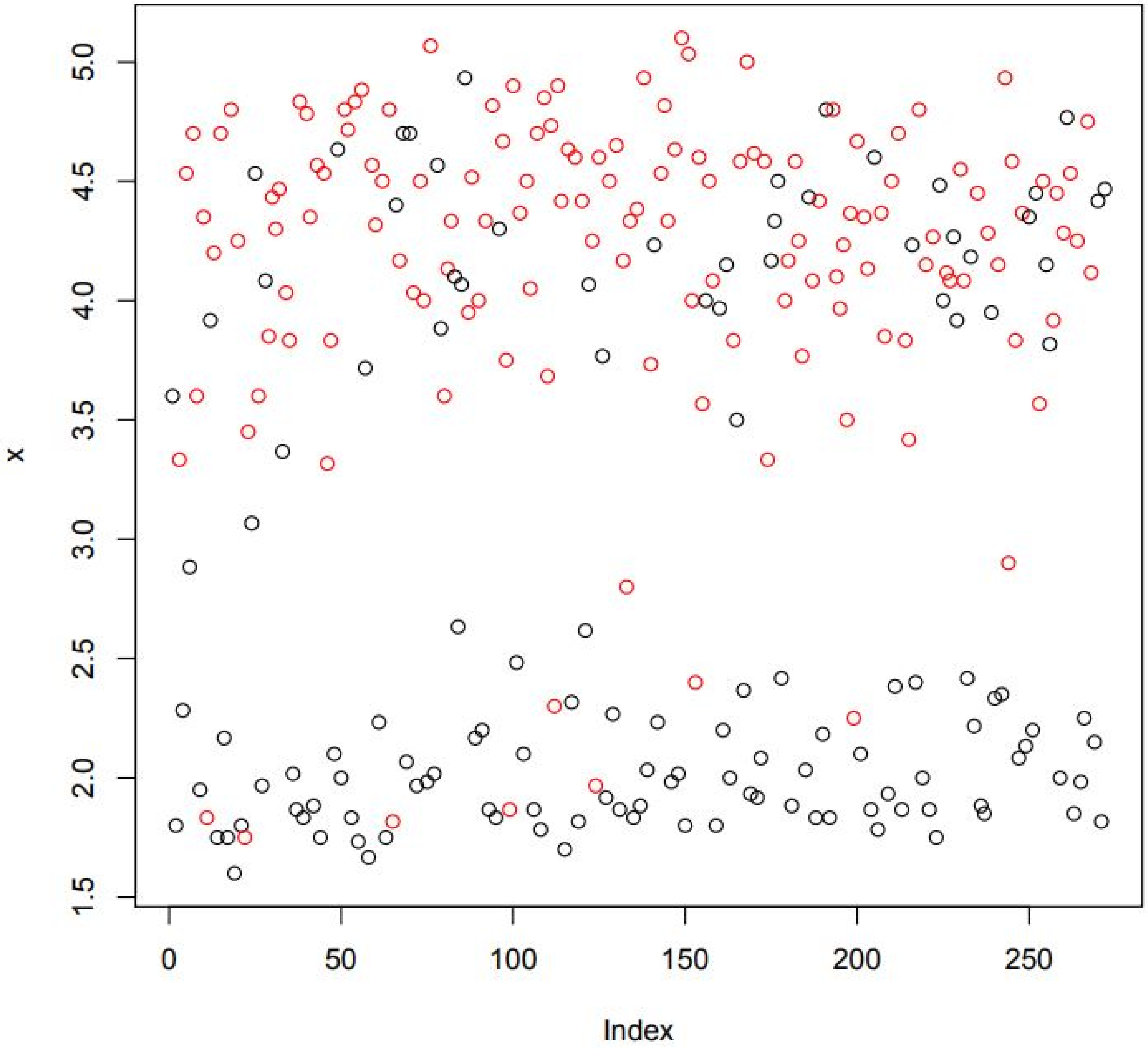
We obtained the following results for different values of the hyperparameter σ.



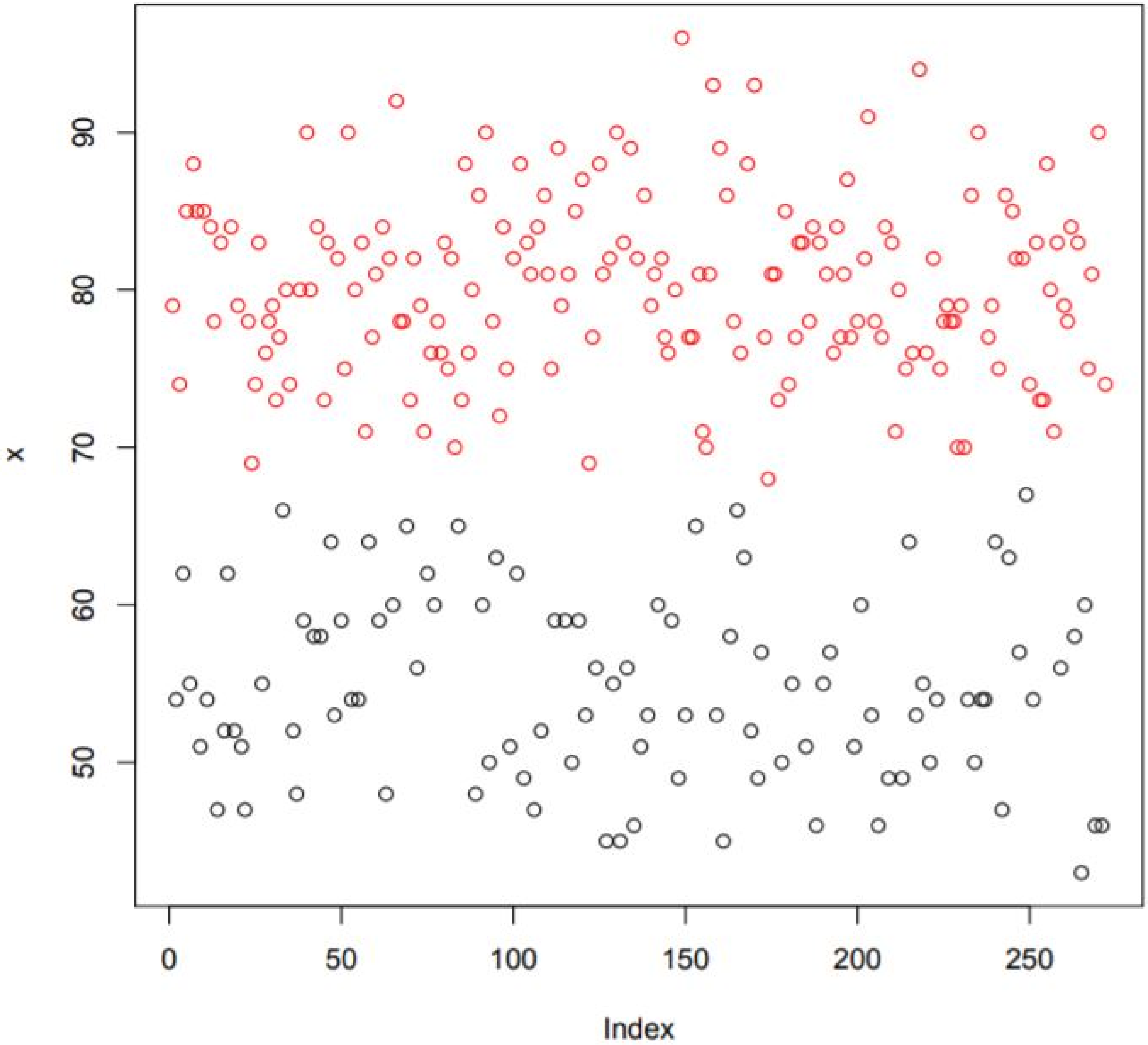
Plots for MCMC results with different sigmas on eruptions dataset.



Plots for MCMC results with different sigmas on waiting dataset.



MCMC Eruptions dataset Distribution



MCMC Waiting Dataset Distribution

References

[1] David M. Blei, Alp Kucukelbir, Jon D. McAuliffe. *Variational Inference: A Review for Statisticians*. arXiv:1601.00670 [stat.CO].

[2] David M. Blei. *Bayesian Mixture Models and the Gibbs Sampler*.