Bayesian Analysis of Gaussian Mixture Models

1 Markov Chain Monte Carlo

Validating the hyperparameters of our conjugate priors was facilitated by *A Compendium of Conjugate Priors* by Daniel Fink.

1.1 Derivation

Given our data for each cluster, k, is normally distributed: $\{x_1, ..., x_n\} \sim N(\mu_k, \sigma_k)$ and we assume a known σ, then our likelihood function can be represented as:

$$L(\mu|\{x_1, ..., x_n\}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \cdot exp\left(\frac{-\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right)$$

Using the identity:

$$\sum_{i=1}^{n}(x_i - \mu)^2 = n(\mu - \bar{x}) + \sum_{i=1}^{n}(x_i - \bar{x})^2$$

We can substitute into our original likelihood function:

$$L(\mu|\{x_1, ..., x_n\}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \cdot exp\left(-\frac{n(\mu - \bar{x})}{2\sigma^2} - \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2}\right)$$

where the variance $\sigma^2 = \frac{1}{\phi}$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

The mixture proportion, ρ, is a from a multinomial process, then we know our likelihood L is

$$L = \prod_{i=1}^{k} \rho_k^{n_k}$$

The assignment identifies this as proportional to a Dirichlet distribution which has parameters $(\delta_1, ..., \delta_k)$. The conjugate prior is also a Dirichlet distribution with a new set of

hyperparameters, $(\delta_1^*, ..., \delta_k^*)$ with density:

$$p(\rho_k) \propto \prod_{i=1}^{k} \rho_k^{\delta_k - 1}$$

The posterior is calculated using *Bayes Theorem* where the likelihood of the data is multiplied by the prior distribution. For proportionality, we only need to consider the numerator.

$$p(\rho_k|x, z) \propto p(x, z|\rho_k)p(\rho_k)$$

$$\propto \left(\prod_{i=1}^{k} \rho_k^{n_k}\right)\left(\prod_{i=1}^{k} \rho_k^{\delta_k - 1}\right)$$

$$\propto \prod_{i=1}^{k} \rho_k^{n_k + \delta_k - 1}$$

If we sum the parameters in the exponent, we find that δ_k^* is:

$$\Rightarrow \boxed{\delta_k^* = \delta_k + n_k}$$

Similarly, if we want to find the conjugate prior of the precision, we can say that the form of the likelihood function that depends on ϕ_k is:

$$L(\phi_k|\{x_1, ..., x_n\}) \propto \phi^{\frac{n}{2}} \cdot exp\left(-\frac{\phi}{2}\sum_{i:z_i=k}(x_i - \mu_k)^2\right)$$

As a function of ϕ, this is proportional to a Gamma distribution with hyper parameters $\alpha > 0$ and $\beta > 0$. The corresponding probability distribution is

$$\text{prior} = \frac{\beta^\alpha \phi^{\alpha-1} exp\,(-\beta\phi)}{\Gamma(\alpha)}$$

The posterior for $\phi_k|x, z$ $Gamma\left(\frac{a_k^*}{2}, \frac{b_k^*}{2}\right)$ is calculated as

$$p(\phi_k|x, z) \propto p(\phi_k)p(x, z|\phi_k)$$

$$\propto \left(\phi^{\alpha-1} exp\,(-\beta_k\phi)\right)\left(\phi^{\frac{n_k}{2}} exp\left(-\frac{\phi}{2}\sum_{i:z_i=k}(x_i - \mu_k)^2\right)\right)$$

$$\propto \left(\phi^{\frac{a}{2}-1} exp\left(-\frac{b}{2}\phi\right)\right)\left(\phi^{\frac{n_k}{2}} exp\left(-\frac{\phi}{2}\sum_{i:z_i=k}(x_i - \mu_k)^2\right)\right)$$

$$\propto \left(\phi^{\frac{a}{2}+\frac{n_k}{2}-1}\right)\left(exp\left(-\frac{b\phi}{2} - \frac{\phi}{2}\sum_{i:z_i=k}(x_i - \mu_k)^2\right)\right)$$

$$\propto \left(\phi^{\frac{a}{2}+\frac{n_k}{2}-1}\right)\left(exp\left(-\left(\frac{b}{2}+\frac{1}{2}\sum_{i:z_i=k}(x_i-\mu_k)^2\right)\phi\right)\right)$$

Summing the arguments in the exponent of ϕ and the exponential function respectively and knowing that the posterior parameters are of the form $Gamma\left(\frac{a_k^*}{2}, \frac{b_k^*}{2}\right)$, we find the posterior parameters to be:

$$\Rightarrow \boxed{a_k^* = a + n_k}$$

$$\Rightarrow \boxed{b_k^* = b + \sum_{i:z_i=k}(x_i-\mu_k)^2}$$

For the mean μ_k, the only part that of the likelihood function that depends on μ is

$$L(\mu_k|\{x_1,...,x_n\}) \propto exp\left(-\frac{n(\mu-\bar{x})}{2\sigma^2}\right)$$

We see that this proportional to a normal distribution, so our prior must also be of the normal family. To find the hyperparameters of the prior, lets call them m and v respectively as Fink does. However for our case, $v = \frac{1}{\alpha_k\phi_k}$.

$$prior(\mu|m,v) = \frac{1}{\sqrt{2\pi v}}\, exp\left(\left(-\frac{(\mu-m)^2}{2v}\right)\right)$$

The posterior for $\mu_k|x,z,\phi_k$ $\mathcal{N}(m_k^*, v_k^*)$ is calculated with Bayes Theorem and simplified by completing the square on the variable μ for the sum of the arguments of the exponential functions.

$$p(\mu_k|x,z,\phi_k) \propto p(\mu_k)p(x,z,\phi_k|\mu_k)$$

$$\propto exp\left(-\frac{n(\mu-\bar{x})}{2\sigma^2}\right)exp\left(-\frac{(\mu-m)^2)}{2v}\right)$$

$$\propto exp\left(-\frac{\left(\mu-\frac{\sigma^2 m+vn\bar{x}}{\sigma^2+vn}\right)^2}{\frac{2v\sigma^2}{\sigma^2+vn}}\right)$$

We see that the unknown parameters for the posterior are:

$$m_k^* = \frac{\sigma_k^2 m_k + v_k n_k \bar{x}_k}{\sigma_k^2 + v_k n_k}$$

$$v_k^* = \frac{v_k\sigma_k^2}{\sigma_k^2 + v_k n_k}$$

Substituting in the given values of $\sigma_k = \frac{1}{\phi}$ and $v_k = \frac{1}{\alpha_k\phi_k}$ and simplifying we get can find the posterior parameters.

For α_k^*:

$$v_k^* = \frac{1}{\alpha_k^*\phi_k} = \frac{v_k\sigma_k^2}{\sigma_k^2 + v_k n_k}$$

$$= \frac{\frac{1}{\alpha_k \phi_k} \frac{1}{\phi_k}}{\frac{1}{\phi_k} + \frac{n_k}{\alpha_k \phi_k}}$$

$$= \frac{\frac{1}{\alpha_k \phi_k^2}}{\frac{\alpha_k \phi_k + n_k \phi_k}{\alpha_k \phi_k^2}}$$

$$\frac{1}{(\alpha_k + n_k)\phi_k}$$

$$\Rightarrow \boxed{\alpha_k^* = \alpha_k + n_k}$$

For m_k^*:

$$m_k^* = \frac{\sigma_k^2 m_k + v_k n_k \bar{x}_k}{\sigma_k^2 + v_k n_k}$$

$$= \frac{\frac{m_k}{\phi_k} + \frac{n_k \bar{x}_k}{\alpha_k \phi_k}}{\frac{1}{\phi_k} + \frac{n_k}{\alpha_k \phi_k}}$$

$$= \frac{\frac{\alpha_k \phi_k m_k + \phi_k n_k \bar{x}_k}{\alpha_k \phi_k^2}}{\frac{\alpha_k \phi_k + n_k \phi_k}{\alpha_k \phi_k^2}}$$

$$= \frac{(\alpha_k m_k + n_k \bar{x}_k)\phi_k}{(\alpha_k + n_k)\phi_k}$$

$$\Rightarrow \boxed{m_k^* = \frac{(\alpha_k m_k + n_k \bar{x}_k)}{(\alpha_k + n_k)}}$$

where

$$\bar{x}_k = \frac{1}{n_k} \sum_{i:z_i=k} x_i$$

1.2 Implementation

The MCMC algorithm is implemented with $k = 2$ on the toy dataset. The algorithm was run for 20,000 iterations, with the first 10,000 iterations discarded as a *burn-in* phase. The mean of the unknown parameters, μ, ϕ, and ρ, is calculated using the remaining 10,000 samples. Using the mean of the unknown parameters, μ, ϕ, and ρ, 1000 samples of the data are generated.

1.3 Results

The posterior mean of the unknown parameters are given in Table 1. The posterior distribution of the unknown parameters μ, ϕ, and ρ are shown in Figures 1, 2, 3. The histogram of the original data set and the generated data set are shown in Figures 4 and 6.

k	μ_k	ϕ_k	ρ_k
0	0.328	108.079	0.516
1	0.587	91.867	0.484

Table 1: The posterior mean of the unknown parameters.

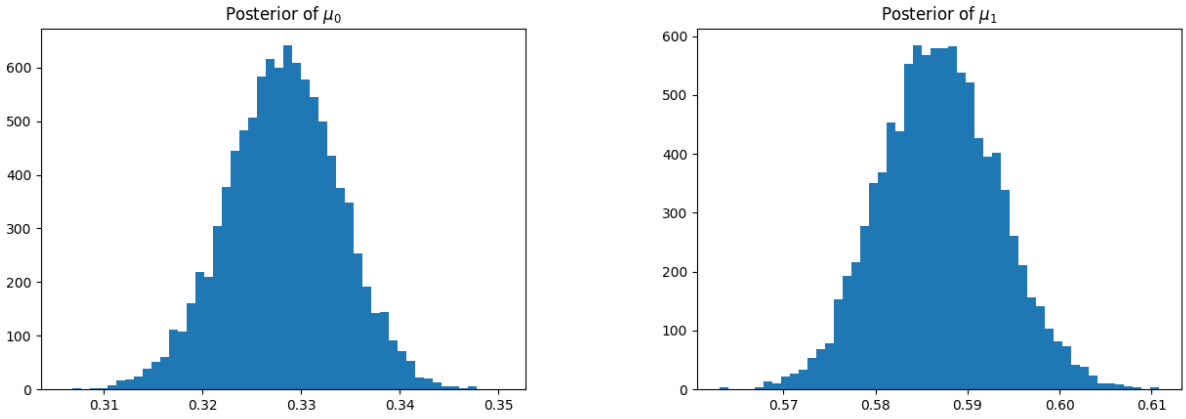

Figure 1: Distributions of μ.

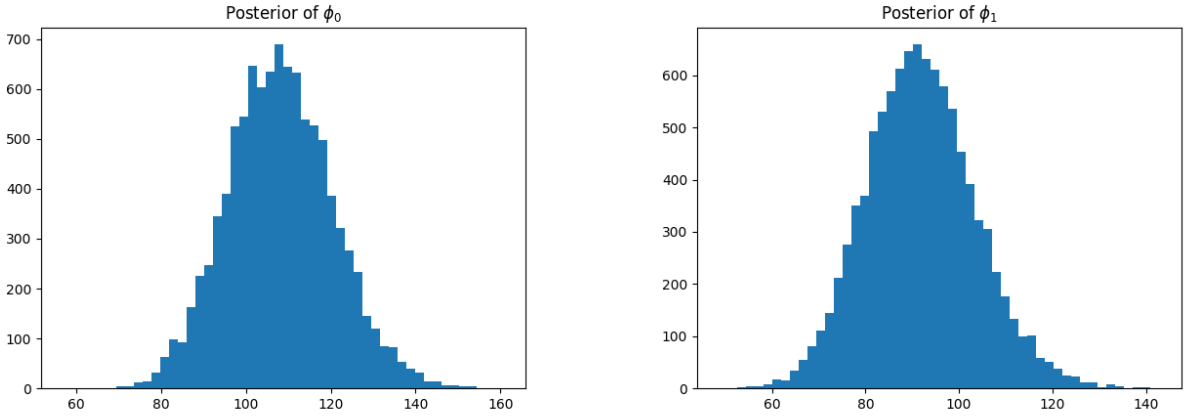

Figure 2: Distributions of ϕ.

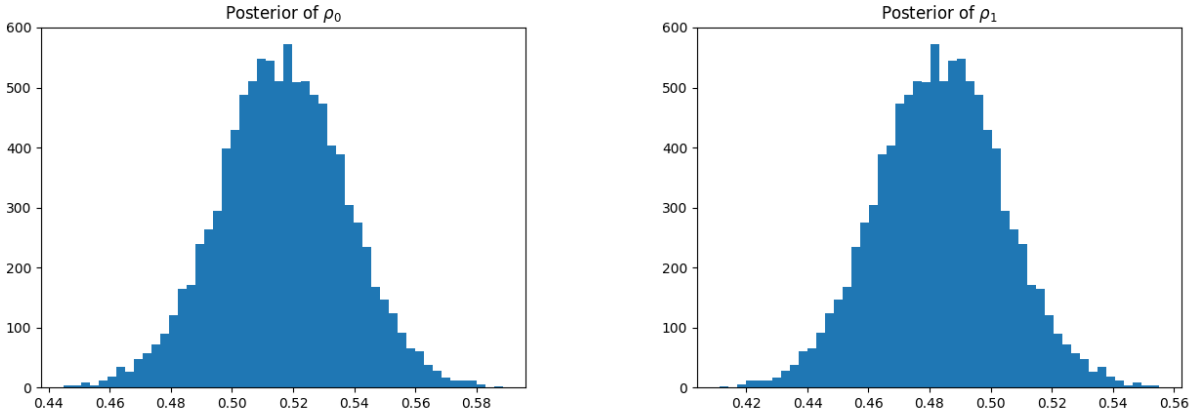

Figure 3: Distributions of ρ.

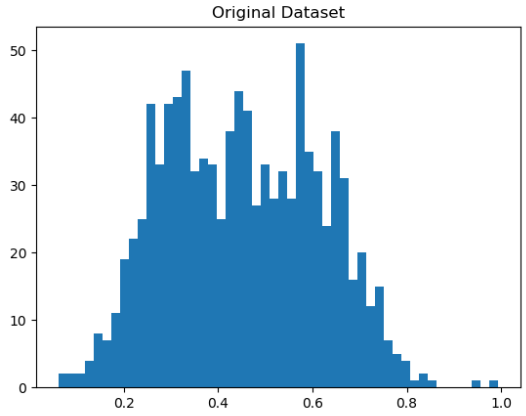

Figure 4: Histogram of the Original Dataset.

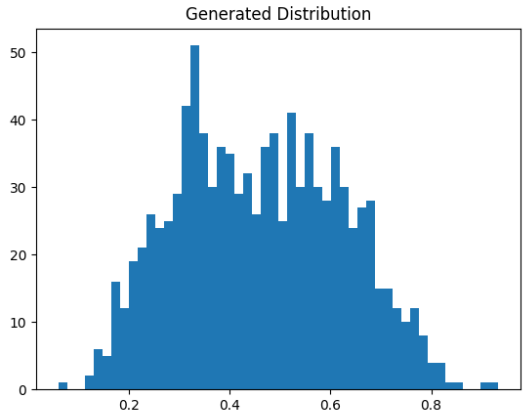

Figure 5: Histogram of the Generated Dataset.

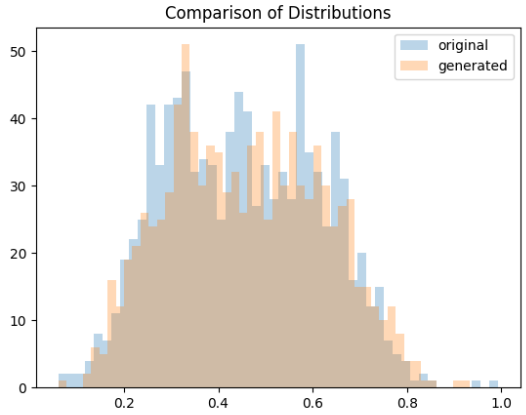

Figure 6: Comparison of the Original and the Generated Datasets.

The comparison of the histogram of the original dataset to the generated distribution shows that the MCMC algorithm has successfully learned the unknown parameters of this generative model. The generated dataset appears to closely match what was given originally.

2 Variational Inference

2.1 Variational Inference Algorithm

Variational inference estimates the posterior distribution $p(\rho, \phi, \mu|x)$ by creating a variational distribution q defined as $q(Z, \rho, \phi, \mu)$. $q(Z, \rho, \phi, \mu)$ can be factored as $q(Z)q(\rho, \phi, \mu)$

The goal of variation inference is to choose the parameters Z, ρ, ϕ, μ that give the tightest bound on the marginal probability of x. This can be expressed as maximizing the ELBO, or Evidence Lower Bound on the log of the marginal probability of x. The ELBO is:

$$log\, p(X) \geq E_q[log\, p(x, Z)] - E_q[log\, q(Z)]$$

Maximizing the ELBO is equivalent to minimizing the KL divergence between between q(z) and p(Z—x).

In general, the q that makes the ELBO the largest is

$$ln\, q_i^*(Z_i) = E_i[ln\, p(x, z)] + constant$$

The distribution of the data can be factorized as follow:

$$p(x, z, \rho, \mu, \phi) = p(x|Z, \mu, \phi)p(z|\rho)p(\mu|\phi)p(\phi)$$

We will use a Dirichlet prior for ρ, and a Gaussian-Wishart prior for μ and ϕ. The factored conjugate prior distribution is:

$$p(\rho, \mu, \phi) = Dir(\rho|\alpha_0)\Pi_k N(\mu_k|m_0, (\beta_0\phi_k)^{-1})Wi(\phi_k|L_0, \nu_0)$$

The variational distribution $q(z, \phi, \rho, \mu)$ is as follows:

$$q(z|\mu, \phi, \rho)\, q(\mu, \phi, \rho) = Dir(\rho|\alpha)\Pi_k N(\mu_k|m_k, (\beta_k\phi_k)^{-1})Wi(\phi_k|L_k, \nu_k)$$

The variational inference algorithm is similar to the EM algorithm. It has an "E like" step and an "M like" step. The steps are summarized below:

Variational E Step

The E Step focuses on calculating the expectation of log q(z) with respect to the hidden variables ϕ, μ, and ρ, summarized in the below derivation as θ.

$$log\, q(z) =_{q(\theta)} [log\, p(x, z, \theta)] + const$$

$$= \sum_i \sum_i z_{ik} log\, \eta_{ik} + const$$

where

$$log\, \eta_{ik} =_{q(\theta)} [log\, \rho_k] + \tfrac{1}{2}{}_{q(\theta)}[log\, |\phi_k|] - \tfrac{D}{2}\, log(2\pi) - \tfrac{1}{2}{}_{q(\theta)}[(x_i - \mu_k)\, \phi_k\, (x_i - \mu_k)]$$

Because $q(\pi)$ is Dirichlet distributed, its expectation is:

$$log\, \tilde{\rho} = [log\, \rho_k] = \psi(\alpha_k) - \psi(\textstyle\sum_i \alpha_i)$$

Where $\psi()$ is the digamma function.

Because the variational distribution factorizes into

$$q(\mu_k, \phi_k) = N(\mu_k|m_k, (\beta\, \phi_k)^{-1}Wi(\phi_k|L_k, \nu_k)$$

Because ϕ_k is Wishart distributed, its log expectation is:

$$log\, \tilde{\phi} = [log|\phi_k|] = \psi(\tfrac{\nu_k}{2}) + log\, 2 + log|\phi_k|$$

Then, for the expectation of the quadratic:

$$[(x_i - \mu_k)\, \phi_k\, (x_i - \mu_k)] = \beta_k^{-1} + \nu_k\, (x_i - m_k)\, \phi_k\, (x_i - m_k)$$

Combining these expectations into one, we get the responsibility r_{ik} of cluster k for data point i to be proportional to the following:

$$r_{ik} \propto \tilde{\pi_k} \tilde{\phi_k} exp(-\frac{1}{2\beta_k} - \frac{\nu_k}{2} (x_i - m_k) \phi_k (x_i - m_k)$$

The values of r_{ik} must be normalized so that the sum of responsibilities for each data point equals 1.

r_{ik} is then used in the M step to update the remaining variables.

Variational M Step

Using mean field factorization:

$$log\, q(\theta) = \\ log\, p(\rho) + \sum_k log\, p(\mu_k, \phi_k + \sum_i {}_{q(z)}[log\, p(z_i|\rho)] + \sum_i \sum_k {}_{q(z)}[z_{ik}] log\, N(x_i|\mu_k, \phi_k^{-1}) + const$$

This factorizes into:

$$q(\theta) = q(\rho) \Pi_k q(\mu_k, \phi_k)$$

For π

$$log\, q(\rho) = (\alpha_0 - 1) \sum_k log\, \rho_k + \sum_k \sum_i r_{ik} log \rho_k + const$$

This exponentiated is the Dirichlet distribution

$$q(\rho) = Dir(\rho|\alpha) \\ \alpha_k = \alpha_0 + N_k \\ N_k = \sum_i r_{ik}$$

The updates for μ_k and ϕ_k are as follows

$$q(\mu_k, \phi_k) = N(\mu_k|m_k, (\beta_k \phi_k)^{-1}) Wi(\phi_k|L_k, \nu_k)$$
$$\beta_k = \beta_0 + N_k$$
$$m_k = (\beta_0\, m_0 + N_k\, \bar{x}_k)/\beta_k$$
$$\nu_k = \nu_0 + N_k + 1$$
$$\bar{x}_k = \frac{1}{N_k} \sum_i r_{ik} x_i$$
$$S_k = \frac{1}{N_k} \sum_i r_{ik} (x_i - \bar{x}_k)^2$$
$$L_k^{-1} = L_0^{-1} + N_k\, S_k + \frac{\beta_0\, N_k}{\beta_0 + N_k} (\bar{x} - m_0)^2$$

These steps are alternated iteratively until the variational distribution becomes very close to the posterior distribution. Then, the values of m_k, L_k^{-1}, and $\frac{N_k}{N_{total}}$ are used as estimates of the posterior μ_k, ϕ_k, and ρ_k respectively.

2.2 Variational Inference Implementation

The variational inference algorithm was implemented on a Gaussian mixture model and run for 750 iterations. The resulting estimates for the posterior means, precisions, and mixing coefficients are summarized in the table below.

k	μ_k	ϕ_k	ρ_k
0	0.30	137.02	0.41
1	0.55	69.74	0.59

Because the variational inference algorithm does not sample from the posterior, we do not have sample distributions for the posterior. However, we generated 1000 data points based on our posterior estimates. A histogram of these datapoints is shown below.

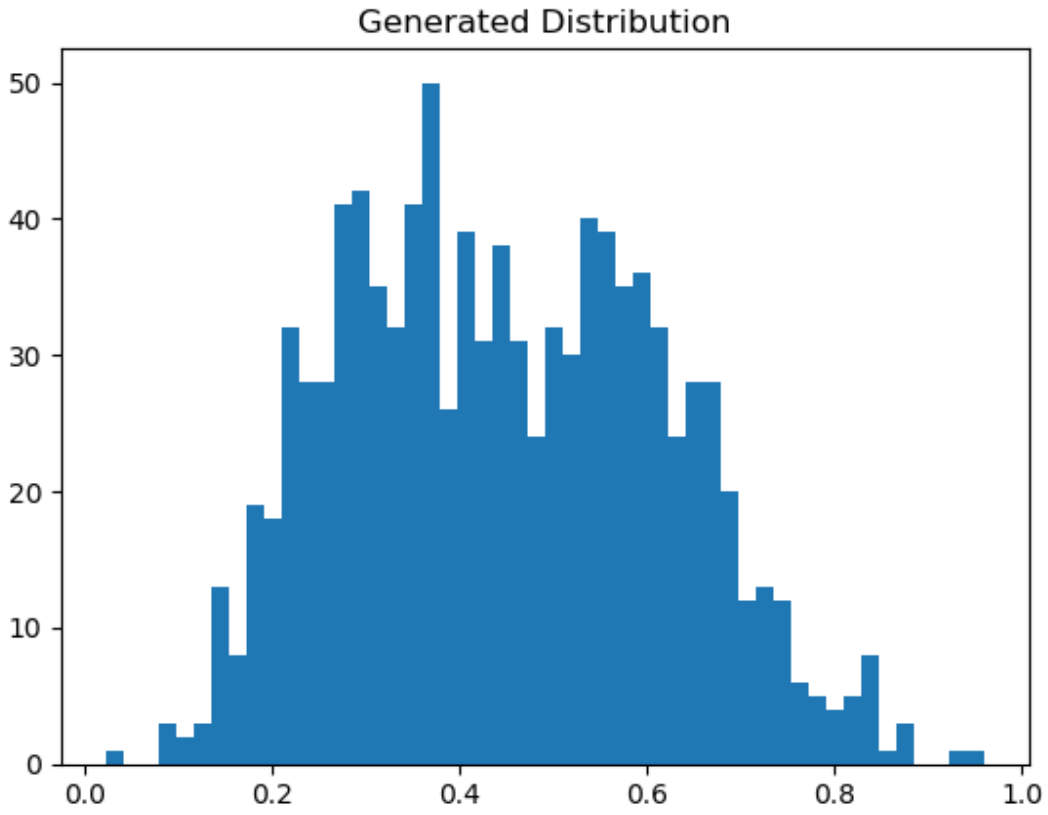

Figure 7: Generated Distribution

This algorithm converges in fewer iterations than the MCMC algorithm. To obtain accurate posterior estimates in MCMC, around 10,000 iterations were run. Variational inference required only 750 iterations. Without knowing the true μ, ϕ, and ρ values of the original distribution, it is difficult to asses the accuracy of the two methods based on how close their estimates are to the true values. It appears that the generated distributions using the estimated posteriors of both MCMC and variational inference appear to have similar accuracy in replication the original data. However, the variational inference algorithm was noticably more sensitive to initial values. For example, in the implementation, the algorithm struggled to converge on accurate posterior estimates when Ł$_0$ was set to 1. Setting L_0 to 0.1 led to much better performance of variational inference. Because of the sensitivity to initial values, MCMC might be more accurate unless a sensible method for determining initial values is used.