Yuxuan Wang, December 2020, solution for final project Q1.

1 Variational Inference

1.1 Derivation of variational inference algorithm

First, supposing we partition the elements of $\mathbf{Z}$ into disjoint groups that we denote by $\mathbf{Z_i}$, we can get:

$$q(\mathbf{Z})=\prod_{i=1}^{M} q_i\left(\mathbf{Z}_i\right)$$

To minimize ELBO, we can reformulate it as:

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i\left\{\ln p(\mathbf{X}, \mathbf{Z})-\sum_i \ln q_i\right\} \mathrm{d}\mathbf{Z} \\
&= \int q_j\left\{\int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \ \mathrm{d}\mathbf{Z}_i\right\} \mathrm{d}\mathbf{Z}_j-\int q_j \ln q_j \ \mathrm{d}\mathbf{Z}_j+\text{const} \\
&= \int q_j \ln \tilde{p}\left(\mathbf{X}, \mathbf{Z}_j\right) \mathrm{d}\mathbf{Z}_j-\int q_j \ln q_j \ \mathrm{d}\mathbf{Z}_j+\text{const}
\end{aligned} \tag{1}$$

The maximum occurs when $q_j(\mathbf{Z_i})=\tilde{p}\left(\mathbf{X}, \mathbf{Z}_j\right)$. Then we obtain a general expression for the optimal solution for $q_j^*(\mathbf{Z_i})$:

$$\ln q_j^{\star}\left(\mathbf{Z}_j\right)=\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]+\text{ const.} \tag{2}$$

In order to formulate a variational treatment of this model, we next write down the joint distribution of all of the random variables, which is given by:

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\phi})=p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\phi}) p(\mathbf{Z} \mid \boldsymbol{\rho}) p(\boldsymbol{\rho}) p(\boldsymbol{\mu} \mid \boldsymbol{\phi}) p(\boldsymbol{\phi}) \tag{3}$$

Using the latent variables and the decomposition, then discarding any terms independent from $\mathbf{Z_i}$, (2) can be reformulated as:

$$\ln q_j^{\star}\left(\mathbf{Z}_j\right)=\mathbb{E}_{\boldsymbol{\rho}}[\ln p(\mathbf{Z}|\boldsymbol{\rho})]+\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\phi}}[\ln p(\mathbf{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\phi})]+\text{ const.} \tag{4}$$

$$\ln q^{\star}(\mathbf{Z})=\sum_{n=1}^N \sum_{k=1}^K z_{nk} \ln \sigma_{nk}+\text{ const} \tag{5}$$

where we define σ as:

$$\begin{aligned}
\ln \sigma_{nk} =& \mathbb{E}\left[\ln \rho_k\right]+\frac{1}{2} \mathbb{E}\left[\ln \left|\boldsymbol{\phi}_k\right|\right]-\frac{D}{2} \ln (2 \pi) \\
& -\frac{1}{2} \mathbb{E}_{\boldsymbol{\mu}_k, \boldsymbol{\phi}_k}\left[\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)^{\mathrm{T}} \boldsymbol{\phi}_k\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)\right]
\end{aligned}$$

Normalize σ:

$$r_{nk}=\frac{\sigma_{nk}}{\sum_{j=1}^K \sigma_{nk}}$$

In order to denote the symbols more convenient, define:

$$\begin{aligned}
N_k &= \sum_{n=1}^N r_{nk} \\
\overline{\mathrm{x}}_k &= \frac{1}{N_k} \sum_{n=1}^N r_{nk} \mathrm{x}_n \\
\mathrm{S}_k &= \frac{1}{N_k} \sum_{n=1}^N r_{nk}\left(\mathrm{x}_n-\overline{\mathrm{x}}_k\right)\left(\mathrm{x}_n-\overline{\mathrm{x}}_k\right)^{\mathrm{T}}
\end{aligned}$$

Using the fact that the model has conjugate priors, the functional form of the factors in the variational posterior distribution is known, namely discrete for $\mathbf{Z}$, Dirichlet for ρ, and Gaussian-Wishart for (μ_k, ϕ_k)[1].

$$\begin{aligned}
q^*(\boldsymbol{\rho}) &= \operatorname{Dir}(\boldsymbol{\rho}|\boldsymbol{\alpha}) \\
q^{\star}\left(\boldsymbol{\mu}_k, \boldsymbol{\phi}_k\right) &= \mathcal{N}\left(\boldsymbol{\mu}_k \mid \mathbf{m}_k,\left(\beta_k \boldsymbol{\phi}_k\right)^{-1}\right) \mathcal{W}\left(\boldsymbol{\phi}_k \mid \mathbf{W}_k, \nu_k\right)
\end{aligned} \tag{6}$$

At last, we introduce definitions of $\widetilde{\phi}_k$ and $\tilde{\rho}_k$:

$$\begin{aligned}\ln \widetilde{\phi}_k \equiv \mathbb{E}\left[\ln \left|\boldsymbol{\phi}_k\right|\right]=\sum_{i=1}^D \psi\left(\frac{\nu_k+1-i}{2}\right)+D \ln 2+\ln \left|\mathbf{W}_k\right| \\ \ln \tilde{\rho}_k \equiv \mathbb{E}\left[\ln \rho_k\right]=\psi\left(\alpha_k\right)-\psi(\widehat{\alpha})\end{aligned} \tag{7}$$

$\widehat{\alpha}=\sum_i \alpha_i$. We then reformulated (1) as:

$$\begin{aligned}\mathcal{L}= & \sum_{\mathbf{Z}} \iiint q(\mathbf{Z}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\phi}) \ln \left\{\frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\phi})}{q(\mathbf{Z}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\phi})}\right\} \mathrm{d} \boldsymbol{\rho} \mathrm{d} \boldsymbol{\mu} \mathrm{d} \boldsymbol{\phi} \\ = & \mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\phi})]-\mathbb{E}[\ln q(\mathbf{Z}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\phi})] \\ = & \mathbb{E}[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\phi})]+\mathbb{E}[\ln p(\mathbf{Z} \mid \boldsymbol{\rho})]+\mathbb{E}[\ln p(\boldsymbol{\rho})]+\mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\phi})] \\ & -\mathbb{E}[\ln q(\mathbf{Z})]-\mathbb{E}[\ln q(\boldsymbol{\rho})]-\mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\phi})]\end{aligned} \tag{8}$$

$$\begin{aligned}\mathbb{E}[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\phi})]= & \frac{1}{2} \sum_{k=1}^K N_k\left\{\ln \widetilde{\phi}_k-D \beta_k^{-1}-\nu_k \operatorname{Tr}\left(\mathbf{S}_k \mathbf{W}_k\right)\right. \\ & \left.-\nu_k\left(\overline{\mathbf{x}}_k-\mathbf{m}_k\right)^{\mathrm{T}} \mathbf{W}_k\left(\overline{\mathbf{x}}_k-\mathbf{m}_k\right)-D \ln (2 \rho)\right\} \\ \mathbb{E}[\ln p(\mathbf{Z} \mid \boldsymbol{\rho})]= & \sum_{n=1}^N \sum_{k=1}^K r_{n k} \ln \tilde{\rho}_k \\ \mathbb{E}[\ln p(\boldsymbol{\rho})]= & \ln C\left(\boldsymbol{\alpha}_0\right)+\left(\alpha_0-1\right) \sum_{k=1}^K \ln \tilde{\rho}_k \\ \mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\phi})]= & \frac{1}{2} \sum_{k=1}^K\left\{D \ln \left(\beta_0 / 2 \rho\right)+\ln \widetilde{\phi}_k-\frac{D \beta_0}{\beta_k}\right. \\ & \left.-\beta_0 \nu_k\left(\mathbf{m}_k-\mathbf{m}_0\right)^{\mathrm{T}} \mathbf{W}_k\left(\mathbf{m}_k-\mathbf{m}_0\right)\right\}+K \ln B\left(\mathbf{W}_0, \nu_0\right) \\ \mathbb{E}[\ln q(\mathbf{Z})]= & \sum_{n=1}^N \sum_{k=1}^K r_{n k} \ln r_{n k} \\ \mathbb{E}[\ln q(\boldsymbol{\rho})]= & \sum_{k=1}^K\left(\alpha_k-1\right) \ln \tilde{\rho}_k+\ln C(\boldsymbol{\alpha}) \\ \mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\phi})]= & \sum_{k=1}^K\left\{\frac{1}{2} \ln \widetilde{\phi}_k+\frac{D}{2} \ln \left(\frac{\beta_k}{2 \rho}\right)-\frac{D}{2}-\mathrm{H}\left[q\left(\boldsymbol{\phi}_k\right)\right]\right\}\end{aligned}$$

D is the the dimensionality of x, $\mathrm{H}[q(\phi_k)]$ is the entropy of the Wishart distribution, and the coefficients $C(\alpha)$ and $B(\boldsymbol{W}, v)$ are given in the appendix. $\beta_0, m_0, W_0, v_0, \alpha_0$ is the initial parameters we need to generalized. In order to maximize (8), we get the partial of the latent parameters and make them equal to 0, finally we get:

$$\begin{aligned}\beta_k & =\beta_0+N_k \\ \mathbf{m}_k & =\frac{1}{\beta_k}\left(\beta_0 \mathbf{m}_0+N_k \overline{\mathbf{x}}_k\right) \\ \mathbf{W}_k^{-1} & =\mathbf{W}_0^{-1}+N_k \mathbf{S}_k+\frac{\beta_0 N_k}{\beta_0+N_k}\left(\overline{\mathbf{x}}_k-\mathbf{m}_0\right)\left(\overline{\mathbf{x}}_k-\mathbf{m}_0\right)^{\mathrm{T}} \\ \nu_k & =\nu_0+N_k \\ \alpha_k & =\alpha_0+N_k \\ r_{n k} & \propto \widetilde{\rho}_k \widetilde{\phi}_k^{1 / 2} \exp \left\{-\frac{D}{2 \beta_k}-\frac{1}{2} \nu_k\left(\mathbf{x}_n-\mathbf{m}_k\right)^{\mathrm{T}} \mathbf{W}_k\left(\mathbf{x}_n-\mathbf{m}_k\right)\right\}\end{aligned} \tag{9}$$

Using (9),we can implement iteration until all parameters converge. Then, using the laten variable, we can solve the distribution of $\mathbf{x}$.

1.2 Implement the variational inference approximation

1.2.1 Data generation and visualization

Using 9 to iterate 100 times, the algorithm converges and latent variables have been solved. Using value of parameters and the conclusion in 6, the plot for distribution of ρ, ϕ and μ is given as follows.

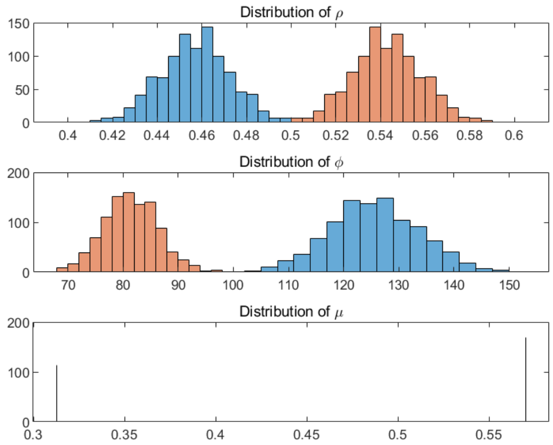


Figure 1: Posterior distribution of latent variables

Generating 1000 times ρ, ϕ and μ and use their average as the final solution, new data has been generated and compared with the original data.

$$\rho_1 = 0.4581 \quad \rho_2 = 0.5418$$
$$\phi_1 = 125.0265 \quad \phi_2 = 81.3975$$
$$\mu_1 = 0.3127 \quad \mu_2 = 0.5699$$

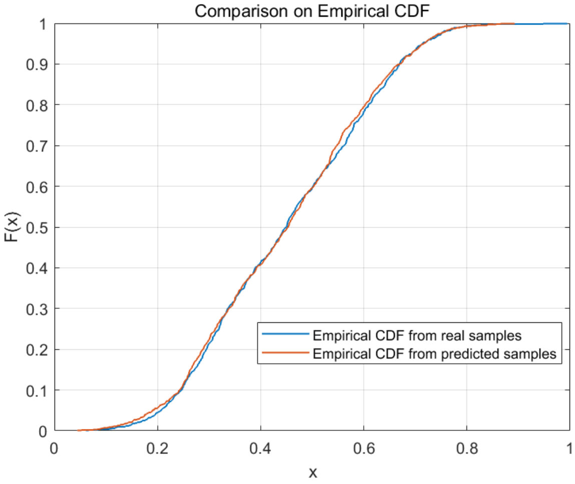


Figure 2: Empirical CDF of the original dataset and prediction data

From the observation of the comparison of the generated data and samples from real distribution, we find that they are pretty close to the true data samples.

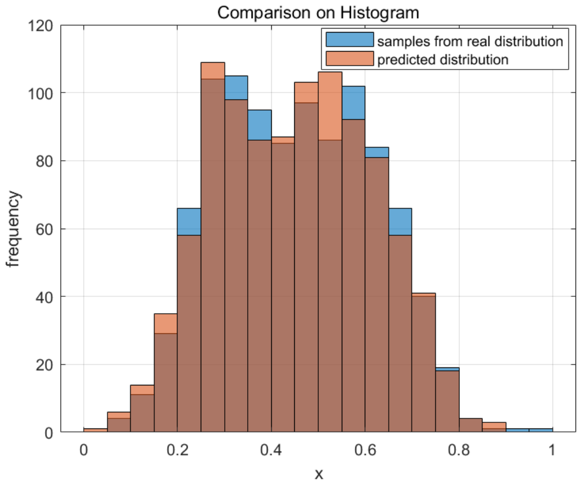


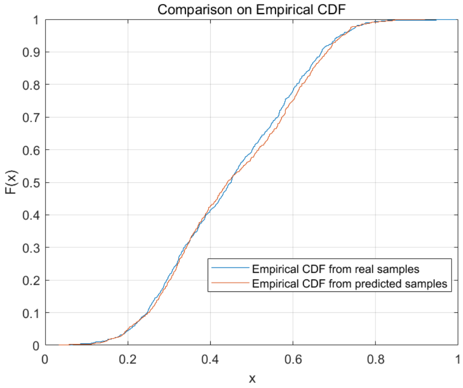
Figure 3: Histogram of the original dataset and prediction data

1.2.2 Comparison with MCMC

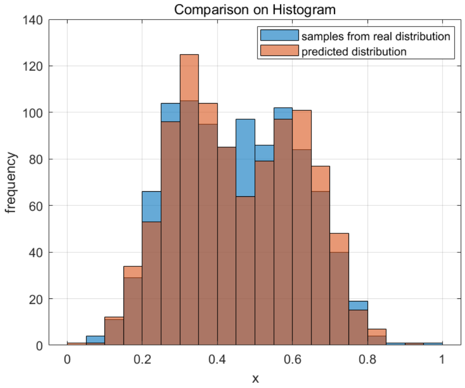
Using MCMC, the parameters we have is:

$$\rho_1 = 0.4206 \quad \rho_2 = 0.5794$$
$$\phi_1 = 195.8116 \quad \phi_2 = 86.6682$$
$$\mu_1 = 0.3366 \quad \mu_2 = 0.6168$$

The generated results are given by 1.2.2.



(a) CDF of MCMC



(b) Histogram of MCMC

Variational inference is an algorithm performs better efficiency than MCMC. The sampling process of MCMC approaches is pretty heavy but has no bias. MCMC are preferred when accurate results are expected, without regards to the time it takes. From the two histograms, it can be observed that the new data samples generated from variational inference are close to the original data set. Besides, the method of using variational inference only takes 100 iterations to converge, while MCMC need 500 times or even more to attain its final converge points.

Appendix

$C(\boldsymbol{\alpha}) = \frac{\Gamma(\widehat{\alpha})}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}$

$B(\mathbf{W}, \nu) \equiv |\mathbf{W}|^{-\nu/2} \left(2^{\nu D/2} \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}$