Homework #3

Yongjin Shin
20090488
Industrial Management Engineering, POSTECH
dreimer@postech.ac.kr

November 26, 2018

Problem 1. Detailed derivation of variational MoG

Solution Description

[Introduction]

If we consider some Bayesian model, we have seen that there exist some observed variables(or can be denoted as evidence E), latent variables(or hidden H) and model parameters. This model represents the joint distribution of evidence and hidden variable such as $p(E, H)$. The main goal of our tasks is to find out the conditional distribution over the hidden variable H given our evidence E, $p(H|E)$, and the distribution of evidence, $p(E)$. To achieve our goal, we need to compute the following equation (but known as Bayes theorem. Also note that it is enough to show the discrete case WLOG):

$$p(H|E) = \frac{p(E|H) \times p(H)}{p(E)} = \frac{p(E|H) \times p(H)}{\sum_H p(E, H)} \tag{1}$$

We only have some evidence so that we need to compute all the possible choice of hidden variable to find $p(E)$. Unfortunately if hidden variable has very high dimension, how can we finish our job in a given time. Previously, EM was suggested to infer $p(H|E)$ however if there are more cases which we cannot find the posterior distribution. Thus there are other possible methods to solve this complicated problems, MCMC(Markov Chain Monte Carlo) and VI(Variational Inference). This homework will only focus on VI, but based on what I have acknowledged about MCMC, VI can handle bigger data with faster speed. After deriving the key concept of VI, the difference between VI and EM will be reviewed quickly and variational MoG will be covered.

[Variational Inference]

Thinking about ELBO(evidence lower bound) is a good starting point of variational inference. Like how we did in EM, $\ln p(E)$ is decomposed as follows (Note that the model parameter θ does not appear because the parameters are now stochastic variables and are absorbed into H On the other hand, λ is the variational parameter which comes with the variational distribution Q. From the perspective of calculus of variations, λ will be used to optimize the functional.):

$$\begin{aligned}
\ln p(E) &= \ln \sum_H p(H, E) = \ln \sum_H Q(H|E, \lambda) \frac{p(H, E)}{Q(H|E, \lambda)} \\
&\geq \sum_H Q(H|E, \lambda) \ln \frac{p(H, E)}{Q(H|E, \lambda)} \\
&= \sum_H Q(H|E, \lambda) \ln p(H, E) - Q(H|E, \lambda) \ln Q(H|E, \lambda) = \mathcal{L}(\lambda) \\
&= \mathbb{E}_Q[\ln p(E|H)] - KL[Q(H|E, \lambda)||p(H|E)]
\end{aligned} \tag{2}$$

We have already known that if we set $Q(H|E,\lambda) = p(H|E)$, KL divergence is zero so that ELBO($\mathcal{L}(\lambda)$) term can be same as $\ln p(E)$. As mentioned before, however, what if we cannot find the posterior distribution $p(H|E)$. Thus we need some assumption toward the distribution Q for inferring ELBO, which is mean field theory:

$$Q(H) = \prod_{i \le |H|} q_i(H_i|\lambda_i) \tag{3}$$

This means that Hidden variables with respect to Q don't have any dependency each others. It is a quite interesting fact since in our original model, there might be any dependency in terms of hidden variables (even though we will never know about). But the introduced distribution Q does not care about those dependency. This mean field variational approximation will make our life easier. Meanwhile, it should be empathized that Q is unknown so that it can be any distribution. Thus, any well known distribution can be introduced as far as we can find a good parameter λ to approximate the real distribution. This is the reason we will use conjugate prior. Then, let's look at ELBO closely keeping in mind about the mean field theorem:

$$\begin{aligned}
\mathcal{L}(\lambda) &= \sum_H \big[Q(H|E,\lambda)\ln p(H,E) - Q(H|E,\lambda)\ln Q(H|E,\lambda)\big] \\
&= \sum_H [\prod_{i \le |H|} q_i(H_i|E,\lambda_i)\ln p(H,E) - \prod_{i \le |H|} q_i(H_i|E,\lambda_i)\ln \prod_{i \le |H|} q_i(H_i|E,\lambda_i)] \\
&= \sum_H [\prod_{i \le |H|} q_i(H_i|E,\lambda_i)\ln p(H,E) - \prod_{i \le |H|} q_i(H_i|E,\lambda_i)\sum_{i \le |H|} \ln q_i(H_i|E,\lambda_i)]
\end{aligned} \tag{4}$$

Now we can optimize a single variational parameter λ_j by regarding others $\lambda_{\neg j}$ as a constant. In the end q_j will be a well approximated functional to the true distribution. Thus,

$$\begin{aligned}
\mathcal{L}(\lambda_j) &= \sum_H [\prod_{i \le |H|} q_i(H_i|E,\lambda_i)\{\ln p(H,E) - \sum_{k \le |H|} \ln q_k(H_k|E,\lambda_k)\}] \\
&= \sum_{H_j}\sum_{H_{\neg j}} \big[q_j(H_j|E,\lambda_j) \prod_{i \le |H|, i \ne j} q_i(H_i|E,\lambda_i)\{\ln p(H,E) - \sum_{k \le |H|} \ln q_k(H_k|E,\lambda_k)\}\big] \\
&= \sum_{H_j}\sum_{H_{\neg j}} q_j(H_j|E,\lambda_j) \prod_{i \le |H|, i \ne j} q_i(H_i|E,\lambda_i)\ln p(H,E) \\
&\quad - \sum_{H_j}\sum_{H_{\neg j}} q_j(H_j|E,\lambda_j) \prod_{i \le |H|, i \ne j} q_i(H_i|E,\lambda_i)\sum_{k \le |H|} \ln q_k(H_k|E,\lambda_k) \\
&= \sum_{H_j}\sum_{H_{\neg j}} q_j(H_j|E,\lambda_j) \prod_{i \le |H|, i \ne j} q_i(H_i|E,\lambda_i)\ln p(H,E) \\
&\quad - \sum_{H_j}\sum_{H_{\neg j}} q_j(H_j|E,\lambda_j) \prod_{i \le |H|, i \ne j} q_i(H_i|E,\lambda_i)\{\sum_{k \le |H|, k \ne j} \ln q_k(H_k|E,\lambda_k) + \ln q_j(H_j|E,\lambda_j)\} \\
&= \sum_{H_j} q_j(H_j|E,\lambda_j)\big[\sum_{H_{\neg j}} \prod_{i \le |H|, i \ne j} q_i(H_i|E,\lambda_i)\ln p(H,E)\big] \\
&\quad - \sum_{H_j} q_j(H_j|E,\lambda_j)\big[\sum_{H_{\neg j}} \prod_{i \le |H|, i \ne j} q_i(H_i|E,\lambda_i)\{\sum_{k \le |H|, k \ne j} \ln q_k(H_k|E,\lambda_k) + \ln q_j(H_j|E,\lambda_j)\}\big]
\end{aligned} \tag{5}$$

We only concern about λ_j , however, $\sum_{H_{\neg j}} \prod_{i \le |H|, i \ne j} q_i(H_i|E,\lambda_i) \sum_{k \le |H|, k \ne j} \ln q_k(H_k|E,\lambda_k)$does not contain anything about λ_j. Therefore $\sum_{H_j}\sum_{H_{\neg j}} \prod_{i \le |H|, i \ne j} q_i(H_i|E,\lambda_i) \sum_{k \le |H|, k \ne j} \ln q_k(H_k|E,\lambda_k)$ has same value but it can be regarded as constant C. If we define a new distribution

$$\log \tilde{p}(H,E) = \sum_{H_{\neg j}} \prod_{i \le |H|, i \ne j} q_i(H_i|E,\lambda_i)\ln p(H,E) \tag{6}$$

, ELBO should be as follows:

$$
\begin{aligned}
\mathcal{L}(\lambda_j) &= \sum_{H_j} q_j(H_j|E,\lambda_j)\big[\sum_{H_{\neg j}} \prod_{i\leq|H|,i\neq j} q_i(H_i|E,\lambda_i)\ln p(H,E)\big] - \sum_{H_j} q_j(H_j|E,\lambda_j)\ln q_j(H_j|E,\lambda_j) + C \\
&= \sum_{H_j} q_j(H_j|E,\lambda_j)\ln \tilde{p}(H,E) - \sum_{H_j} q_j(H_j|E,\lambda_j)\ln q_j(H_j|E,\lambda_j) + C \\
&= \sum_{H_j} q_j(H_j|E,\lambda_j)\ln \frac{\tilde{p}(H,E)}{q_j(H_j|E,\lambda_j)} + C \\
&= C - KL[q_j(H_j|E,\lambda_j)||\tilde{p}(H,E)]
\end{aligned}
\tag{7}
$$

Therefore, we can maximize ELBO by let $q_j(H_j|E,\lambda_j)$ be same with $\tilde{p}(H,E)$. However when it comes to $\tilde{p}(H,E)$, it can be show that

$$
\begin{aligned}
\tilde{p}(H,E) &= \sum_{H_{\neg j}} \prod_{i\leq|H|,i\neq j} q_i(H_i|E,\lambda_i)\ln p(H,E) \\
&= \mathbb{E}_{q_{i\neq j}}[\ln p(H,E)] + C
\end{aligned}
\tag{8}
$$

which means an expectation with respect to the q distributions over all variables H_i for $i \neq j$. Therefore we obtain a general expression for the optimal solution:

$$
\ln q_j^*(H_j|E,\lambda_i) = \mathbb{E}_{q_{i\neq j}}[\ln p(H,E)] + C \tag{9}
$$

[Variational Inference vs EM algorithm]

Before deriving variational MoG, we need to compare what the differences are between VI and EM. Both methods want to infer the unknown variable with given evidence. Thus they introduce an arbitrary distribution Q to achieve their goal. The thing is that EM is finding the approximated distribution upon the posterior distribution of p (at E step) and try to optimize model parameter θ (at M step). Therefore, EM also can be a point estimator. On the other hand, VI assumes that we don't know about the posterior. Thus VI takes mean field theorem and regards all the hidden variables as random variables (so that it cannot but to attain variational parameters which doen't know appear explicitly in the model). Since random variables have their own distributions, each hidden variables have prior distributions. Conjugate prior distributions are commonly used for the simple calculations. In the end, VI give us posterior distributions of hidden variables so that it is not a point estimator.

[Variational MoG]

For deploying our knowledge about Variational inference, we will derive variational MoG step by step. Firstly, we are going to check our model (we have discussed the initial model in EM-MoG) and then secondly, prior distributions will be discussed. And then we will use the result of variational inference (equation 9) for achieving posterior distribution of each hidden variables.

For each observation x_n we have a corresponding latent variable z_n comprising a 1-of K binary vector with elements z_{nk} for k = 1, 2, ... , K. As before we denote the observed data set by $X = \{x_1, x_2, ..., x_N\}$, and similarly we denote the latent variables by $Z = \{z_1, z_2, ..., z_N\}$. We can write down the conditional distribution of Z, give the mixing coefficient π, in the form

$$
p(Z|\pi) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{10}
$$

Similarly, we can write down the conditional distribution of the observed data vectors, given the latent variables and the component parameters

$$
p(X|Z,\mu,\Lambda) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}(x_n|\mu_k,\Lambda_k^{-1})^{z_{nk}} \tag{11}
$$

where $\mu = \{\mu_k\}$ and $\Lambda = \{\Lambda_k\}$. Our random variables are $\{Z, \mu, \Lambda\, \pi\}$. Next we should introduce priors over $\{\mu, \Lambda, \pi\}$.

The analysis is considerably simplified if we use conjugate prior distributions. We therefore choose a Dirichlet distribution over the mixing coefficients π as the likelihood $p(Z|\pi)$ is multinomial distribution. Thus,

$$p(\pi) = Dir(\pi|\alpha_0) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} \tag{12}$$

where by symmetry we have chosen the same parameter α_0 for each of the components, and $C(\alpha_0)$ is the normalization constant for the Dirichlet distribution. Since $(\pi_1, \pi_2, ..., \pi_K)|(x_1, x_2, ..., x_K) \sim Dir(a_0 + x_1, a_0 + x_2, ..., a_0 + x_K)$, if the value of α_0 is small, then the posterior distribution will be influenced primarily by the data rather than the prior.
Similarly, we introduce an independent Gaussian-Wishart prior governing the mean and precision of each Gaussian component, given by

$$\begin{aligned} p(\mu, \Lambda) &= \prod_{k=1}^{K} \mathcal{N}(\mu_k|m_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k|W_0, \nu_0) \\ &= p(\mu|\Lambda)p(\nu) \end{aligned} \tag{13}$$

because this represents the conjugate prior distribution when both the mean and precision are unknown. Typically we would choose $m_0 = 0$ by symmetry. Therefore, our model introduce variational parameters $\{\alpha_0, m_0, \beta_0, W_0, \mu_0\}$which will be used to optimize variational distribution.

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

$$p(X, Z, \pi, \mu, \Lambda) = p(X|Z, \mu, \Lambda)p(Z|\pi)p(\pi)p(\mu|\Lambda)p(\Lambda) \tag{14}$$

We now consider a variational distribution which can be factorized as follows:

$$q(Z, \pi, \mu, \Lambda) = q(Z)q(\pi, \mu, \Lambda) \tag{15}$$

This is the assumption that we need to make in order to obtain a tractable practical solution to our Bayesian mixture model. The functional form of the factors $q(Z)$ and $q(\pi, \mu, \Lambda)$ will be determined automatically by optimization of the variational distribution. The corresponding sequential update equations for these factors can be easily derived by making use of *equation 9*.

(1)$\log q^*(Z)$

$$\begin{aligned} \ln q^*(Z) &= \mathbb{E}_{\pi,\mu,\Lambda}[\ln p(X, Z, \pi, \mu, \Lambda)] + const. \\ &= \mathbb{E}_{\pi,\mu,\Lambda}[\ln p(X|Z, \mu, \Lambda) + \ln p(Z|\pi) + \ln p(\pi) + \ln p(\mu|\Lambda) + \ln p(\Lambda)] + const. \\ &= \mathbb{E}_{\pi,\mu,\Lambda}[\ln p(X|Z, \mu, \Lambda)] + \mathbb{E}_{\pi,\mu,\Lambda}[\ln p(Z|\pi)] + const. \\ &= \mathbb{E}_{\mu,\Lambda}[\ln p(X|Z, \mu, \Lambda)] + \mathbb{E}_{\pi}[\ln p(Z|\pi)] + const. \end{aligned} \tag{16}$$

Substituting for the two conditional distribution on the right hand side, and absorbing any terms that are independent of Z into the additive constant, we have

$$\begin{aligned} \ln q^*(Z) &= \mathbb{E}_{\mu,\Lambda}[\ln p(X|Z, \mu, \Lambda)] + \mathbb{E}_{\pi}[\ln p(Z|\pi)] + const. \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[\mathbb{E}_{\pi_k}[\ln \pi_k] - \frac{1}{2} \mathbb{E}_{\Lambda_k}[\ln \frac{1}{|\Lambda_k|}] - \frac{D}{2} \ln 2\pi - \frac{1}{2} \mathbb{E}_{\mu_k, \Lambda_k}[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] \right] + consts. \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + const. \end{aligned} \tag{17}$$

Therefore, taking the exponential of both sides of the equation above

$$q^*(Z) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}} \tag{18}$$

From the result, it seems like multi-nomial distribution. Since there is no restriction to the distribution q , we can consider any posterior distribution. Thus as far as it has sufficient conditions as a probability distribution, we can have the multi-nomial distribution as our posterior distribution. If we introduce

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}} \tag{19}$$

, then we can see that as ρ_{nk} is given by the exponential of a real quantity, the quantitnies r_{nk} will be nonnegative and $\sum_k r_{nk} = 1$. Since $z_{nk} \in \{0, 1\}$, $\sum_k z_{nk} = 1$. Therefore we can perfectly make multinomial distribution with $q^*(Z)$ and r_{nk}. Thus,

$$q(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}} \tag{20}$$

We might need the statistic value of $q(Z)$. For the multinomial distribution, $\mathbb{E}_Z[z_{nk}]$ is known as r_{nk}.

(2)$\log q^*(\pi, \mu, \Lambda)$

$$\begin{aligned}
\ln q^*(\pi, \mu, \Lambda) &= \mathbb{E}_Z[\ln p(X, Z, \pi, \mu, \Lambda)] + const. \\
&= \mathbb{E}_Z[\ln p(X|Z, \mu, \Lambda) + \ln p(Z|\pi) + \ln p(\pi) + \ln p(\mu|\Lambda) + \ln p(\Lambda)] + const. \\
&= \big[\ln p(\pi) + \mathbb{E}_Z[\ln p(Z|\pi)]\big] + \big[\sum_{k=1}^{K} \ln p(\mu_k, \Lambda_k) + \sum_{k=1}^{K} \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}] \ln \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})\big] + const.
\end{aligned} \tag{21}$$

From the equation above, $q^*(\pi)$ and $q^*(\mu, \Lambda)$ are explicitly divided into two terms. Thus,

$$\begin{aligned}
\ln q^*(\pi) &= \ln p(\pi) + \mathbb{E}_Z[\ln p(Z|\pi)] + const. \\
&= (\alpha_0 - 1) \sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K} \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}] \ln \pi_k + const. \\
&= (\alpha_0 - 1) \sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \ln \pi_k + const.
\end{aligned} \tag{22}$$

Obviously, the posterior distribution will be Dirichlet distribution since we used conjugate prior distribution. From the equation we can easily check this after taking the exponential of both sides.

$$\begin{aligned}
q^*(\pi) &= const \times \prod_{k=1}^{K} \pi_k^{(a_0 - 1) + (\sum_{n=1}^{N} r_{nk})} \\
&= const \times \prod_{k=1}^{K} \pi_k^{a_0 + N_k - 1} \\
&= Dir(\pi|\alpha)
\end{aligned} \tag{23}$$

where α has components α_k given by $\alpha_k = \alpha_0 + N_k$ and $N_k = \sum_{n=1}^{N} r_{nk}$.

Finally, the variational posterior distribution $q^*(\mu_k, \Lambda_k)$ does not factorize into the product of the marginals, but we can always use the product rule to write it in the form $q^*(\mu_k, \Lambda_k) = q^*(\mu_k|\Lambda_k)q^*(\Lambda_k)$.

Thus,

$$\begin{aligned}
\ln q^*(\mu_k, \Lambda_k) &= \ln q^*(\mu_k|\Lambda_k) q^*(\Lambda_k) \\
&= \ln q^*(\mu_k|\Lambda_k) + \ln q^*(\Lambda_k) \\
&= \ln p(\mu_k, \Lambda_k) + \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}] \ln \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1}) + const. \\
&= \ln \mathcal{N}(\mu_k|m_0, \beta_0\Lambda_k) + \ln \mathcal{W}(\Lambda_k|W_0, \nu_0) + \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}] \ln \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1}) + const. \\
&= -\frac{\beta_0}{2}(\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) + \frac{1}{2}\ln|\Lambda_k| - \frac{1}{2} Tr(\Lambda_k W_0^{-1}) \\
&\quad + \frac{\nu_0 - D - 1}{2} \ln|\Lambda_k| - \frac{1}{2}\sum_{n=1}^{N} \mathbb{E}[z_{nk}](x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) + \frac{1}{2}(\sum_{n=1}^{N} \mathbb{E}[z_{nk}]) \ln|\Lambda_k| + const.
\end{aligned} \tag{24}$$

First of all, to identify the distribution for μ_k, we need only consider terms on the right hand side of the above equation depending on μ_k;

$$\begin{aligned}
\ln q^*(\mu_k|\Lambda_k) &= -\frac{1}{2}\mu_k^T \big[\beta_0 + \sum_{n=1}^{N} \mathbb{E}[z_{nk}]\big] \Lambda_k \mu_k + \mu_k^T \Lambda_k \big[\beta_0 m_0 + \sum_{n=1}^{N} \mathbb{E}[z_{nk}] x_n\big] + const. \\
&= -\frac{1}{2}\mu_k^T [\beta_0 + N_k] \Lambda_k \mu_k + \mu_k^T \Lambda_k [\beta_0 m_0 + N_k \bar{x}_k] + const.
\end{aligned} \tag{25}$$

where $\bar{x}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk} x_n$. Thus we see that $\ln q^*(\mu_k|\Lambda_k)$ depends quadratically on μ_k and hence $q^*(\mu_k|\Lambda_k)$ can be assumed a Gaussian distribution. If we compared the log likelihood of arbitrary multivariate normal distribution $\mathcal{N}(\mu|m, \Lambda^{-1})$,

$$\ln MND \sim -\mu^T \Lambda \mu + \mu^T \Lambda m + m^T \Lambda \mu - m^T \Lambda m \tag{26}$$

we can have

$$q^*(\mu_k|\Lambda_k) = \mathcal{N}(\mu_k|m_k, (\beta_k \Lambda_k)^{-1}) \tag{27}$$

where $\beta_k = \beta_0 + N_k$, $m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k \bar{x}_k)$.
For Λ_k, by using the fact that $\ln q^*(\Lambda_k) = \ln q^*(\mu_k, \Lambda_k) - \ln q^*(\mu_k|\Lambda_k)$ we can substitute *equation 25* from *equation 27*;

$$\begin{aligned}
\ln q^*(\Lambda_k) &= \ln q^*(\mu_k, \Lambda_k) - \ln q^*(\mu_k|\Lambda_k) \\
&= -\frac{\beta_0}{2}(\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) + \frac{1}{2}\ln|\Lambda_k| - \frac{1}{2} Tr(\Lambda_k W_0^{-1}) \\
&\quad + \frac{\nu_0 - D - 1}{2} \ln|\Lambda_k| - \frac{1}{2}\sum_{n=1}^{N} \mathbb{E}[z_{nk}](x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) + \frac{1}{2}(\sum_{n=1}^{N} \mathbb{E}[z_{nk}]) \ln|\Lambda_k| \\
&\quad + \frac{1}{2}\beta_k(\mu_k - m_k)^T \Lambda_k (\mu_k - m_k) - \frac{1}{2}\ln|\Lambda_k| + const. \\
&= -\frac{1}{2}\big[Tr(\Lambda_k W_0^{-1}) + \Lambda_k \beta_0 (\mu_k - m_0)(\mu_k - m_0)^T + \sum_{n=1}^{N} \mathbb{E}[z_{nk}] \Lambda_k (x_n - \mu_k)(x_n - \mu_k)^T \\
&\quad - \beta_k \Lambda_k (\mu_k - m_k)(\mu_k - m_k)^T\big] + \ln|\Lambda_k| (\frac{\nu_0 - D - 1}{2} + \frac{\sum_{n=1}^{N} \mathbb{E}[z_{nk}]}{2}) \\
&= -\frac{1}{2} Tr(\Lambda_k \big[W_0^{-1} + \beta_0 (\mu_k - m_0)(\mu_k - m_0)^T + \sum_{n=1}^{N} \mathbb{E}[z_{nk}](x_n - \mu_k)(x_n - \mu_k)^T \\
&\quad - \beta_k (\mu_k - m_k)(\mu_k - m_k)^T\big] + \ln|\Lambda_k| \frac{\nu_0 + N_k - D - 1}{2} \\
&= -\frac{1}{2} Tr(\Lambda_k W_k^{-1}) + \frac{\nu_k - D - 1}{2} \ln|\Lambda_k| + const.
\end{aligned} \tag{28}$$

where $\nu_k = \nu_0 + N_k$ and,

$$\begin{aligned}
W_k^{-1} &= W_0^{-1} + \beta_0(\mu_k - m_0)(\mu_k - m_0)^T - \beta_k(\mu_k - m_k)(\mu_k - m_k)^T + \sum_{n=1}^{N} \mathbb{E}[z_{nk}](x_n - \mu_k)(x_n - \mu_k)^T \\
&= W_0^{-1} + (\beta_0 \mu_k \mu_k^T - \beta_0 m_0(\mu_k + \mu_k^T) + \beta_0 m_0^2) - (\beta_0 \mu_k \mu_k^T + N_k \mu_k \mu_k^T - \beta_k m_k(\mu_k + \mu_k^T) + \beta_k m_k^2) \\
&\quad + \sum_{n=1}^{N} \mathbb{E}[z_{nk}](x_n - \mu_k)(x_n - \mu_k)^T \\
&= W_0^{-1} - N_k \mu_k \mu_k^T + (\beta_k m_k - \beta_0 m_0)(\mu_k + \mu_k^T) + \beta_0 m_0^2 - \beta_k m_k^2) \\
&\quad + \sum_{n=1}^{N} \mathbb{E}[z_{nk}](x_n - \mu_k)(x_n - \mu_k)^T \\
&= W_0^{-1} - N_k \mu_k \mu_k^T + (\beta_0 m_0 + N_k \bar{x}_k - \beta_0 m_0)(\mu_k + \mu_k^T) + \beta_0 m_0^2 - \beta_k m_k^2 \\
&\quad + (\sum_{n=1}^{N} \mathbb{E}[z_{nk}] x_n x_n^T - \sum_{n=1}^{N} \mathbb{E}[z_{nk}] x_n(\mu_k + \mu_k^T) + \sum_{n=1}^{N} \mathbb{E}[z_{nk}] \mu_k \mu_k^T) \\
&= W_0^{-1} - N_k \mu_k \mu_k^T + N_k \bar{x}_k(\mu_k + \mu_k^T) + \beta_0 m_0^2 - \beta_k m_k^2 + \sum_{n=1}^{N} \mathbb{E}[z_{nk}] x_n x_n^T \\
&\quad - N_k \bar{x}_k(\mu_k + \mu_k^T) + N_k \mu_k \mu_k^T \\
&= W_0^{-1} + \beta_0 m_0^2 - \beta_k m_k^2 + \sum_{n=1}^{N} \mathbb{E}[z_{nk}] x_n x_n^T
\end{aligned} \tag{29}$$

Define $S_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk}(x_n - \bar{x}_k)(x_n - \bar{x}_k)^T$. Then,

$$\begin{aligned}
N_k S_k + N_k \bar{x}_k \bar{x}_k^T &= \sum_{n=1}^{N} \mathbb{E}[z_{nk}](x_n - \bar{x}_k)(x_n - \bar{x}_k)^T + N_k \bar{x}_k \bar{x}_k^T \\
&= (\sum_{n=1}^{N} \mathbb{E}[z_{nk}] x_n x_n^T - \frac{2}{N_k} \sum_{n=1}^{N} r_{nk}^2 x_n x_n^T + \frac{1}{N_k} \sum_{n=1}^{N} r_{nk}^2 x_n x_n^T) + \frac{1}{N_k} \sum_{n=1}^{N} r_{nk}^2 x_n x_n^T \\
&= \sum_{n=1}^{N} \mathbb{E}[z_{nk}] x_n x_n^T
\end{aligned} \tag{30}$$

Therefore,

$$\begin{aligned}
W_k^{-1} &= W_0^{-1} + \beta_0 m_0^2 - \beta_k m_k^2 + \sum_{n=1}^{N} \mathbb{E}[z_{nk}] x_n x_n^T \\
&= W_0^{-1} + \beta_0 m_0^2 - \beta_k m_k^2 + N_k S_k + N_k \bar{x}_k \bar{x}_k^T \\
&= W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)^T
\end{aligned} \tag{31}$$

Thus, we see that $q^*(\Lambda_k)$ is a Wishart distribution of the form

$$q^*(\Lambda_k) = \mathcal{W}(\Lambda_k | W_k, \nu_k) \tag{32}$$

(3)Update $\mathbb{E}[\ln \Lambda_k], \mathbb{E}[\ln pi_k], \mathbb{E}[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)]$

These update equations are analogous to the M-step equations of the EM algorithm for the maximum likelihood solution of the mixture of Gaussian. In order to perform this variational M step, we need the expectations $\mathbb{E}[z_{nk}] = r_{nk}$ representing the responsibilities. These are obtained by normalizing the ρ_{nk} that are given by *equation 19.* For $\ln\tilde{\Lambda_k} = \mathbb{E}[\ln\Lambda_k]$ and $\ln\tilde{\pi_k} = \mathbb{E}[ln_{pi_k}]$ are easily found from the *reference [1] B.21 and B.81.*

$$\ln\tilde{\Lambda_k} = \mathbb{E}[\ln\Lambda_k] = \sum_{i=1}^{D}\psi(\frac{\nu_k+1-i}{2}) + D\ln 2 + \ln|W_k| \tag{33}$$

$$\ln\tilde{\pi_k} = \mathbb{E}[lnpi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}) = \psi(\alpha_k) - \psi(\sum_k \alpha_k) \tag{34}$$

where $\psi(.)$ is the digamma function defined in the *reference [1] B.25.* Another $\mathbb{E}_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)]$ can be derived as follows:

$$\begin{aligned}\mathbb{E}_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)] &= \mathbb{E}_{\mu_k,\Lambda_k}[Tr(\Lambda_k(x_n-\mu_k(x_n-\mu_k)^T)] + \mathbb{E}_{\Lambda_k}[(x_n-m_k)^T\Lambda_k(x_n-m_k)]\\ &= \mathbb{E}_{\mu_k,\Lambda_k}[Tr(\Lambda_k var(x_n-\mu_k)] + \mathbb{E}_{\Lambda_k}[(x_n-m_k)^T\Lambda_k(x_n-m_k)]\\ &= \mathbb{E}_{\Lambda_k}[Tr(\Lambda_k(\beta_k\Lambda_k)^{-1})] + (x_n-m_k)^T\nu_k W_k(x_n-m_k)\\ &= \mathbb{E}_{\Lambda_k}[Tr(\Lambda_k(\beta_k\Lambda_k)^{-1}] + (x_n-m_k)^T\nu_k W_k(x_n-m_k)\\ &= D\beta_k^{-1} + \nu_k(x_n-m_k)^T W_k(x_n-m_k)\end{aligned} \tag{35}$$

The optimization of the variational posterior distribution involves cycling between two stages analogous to the E and M steps of the maximum likelihood EM algorithm. In the variational equivalent of the E step, we use the current distributions over $\{\pi, \mu, \Lambda\}$ to evaluate the moment in equation (33), (34), (35) and hence evaluate $\mathbb{E}[z_{nk}] = r_{nk}$. Then in the subsequent variational equivalent of the M step, we keep these responsibilities fixed and use them to re-compute the variational distribution over the parameters using *equation (27) and (32).*

[Final Notes]

From the derivation, we have seen that variational inference is quite similar with EM since it also has two steps like Estep and Mstep. In fact if we have magnificent number of samples, then the Bayesian treatment will converge to the maximum likelihood EM algorithm.

For the Bayesian mixture of Gaussian, a large computation is preceded in Estep, which is about updating r_{nk}. So we can say that there is little computational overhead in using this Bayesian approach as compared to the ML method.

Note that there is no singularity, a component converges into a single sample only. Alas, if we simply introduce a prior and then use a MAP estimate instead of ML, these singularities are removed. Another to mention is that we can infer the number of K by the Bayesian method. We will see this in the problem 2.

Problem 2. Implement the algorithm and evaluate it on 'old faithful data'

Solution Description

Algorithm 1: VBGMM

```
input  : Dataset $\mathcal{D} = \{x_1, ..., x_N\}$
output: $\pi_k, \mu_k, \Lambda_k, r_{n,k}, N_k$ s.t $k = \{1, ..., K\}$
while $NOT\ converging\ OR\ iteration < max\_iteration$ do
    E-like-step
    $\ln \tilde{\Lambda_k} = \mathbb{E}[\ln \Lambda_k] = \sum_{i=1}^{D} \psi(\frac{\nu_k+1-i}{2}) + D \ln 2 + \ln |W_k|$
    $\ln \tilde{\pi_k} = \mathbb{E}[lnpi_k] = \psi(\alpha_k) - \psi(\sum_k \alpha_k)$
    $\mathbb{E}_{\mu_k, \Lambda_k}[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] = D\beta_k^{-1} + \nu_k (x_n - m_k)^T W_k (x_n - m_k)$

    M-like-step
    $\alpha_k = \alpha_0 + N_k$
    $\beta_k = \beta_0 + N_k$
    $m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k \bar{x}_k)$
    $W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)^T$
    $\nu_k = \nu_0 + N_k$
end
Return
```

[Dataset and Method]

Old faithful data is composed of 273 samples and its samples have two dimensions. Thus it is nice to visualize the process. Data was handled by Bayesian Mixture of Gaussian. In the end, it will converge a certain number of K, but six mixture components was the starting point. Total number of iteration was 60. and every 10th iteration, two plot was taken. One for the bar plot of responsibility (N_{nk}) and the other is for one-standard deviation density contours of each component. Each component has colored depends on the responsibility.

Note that in order to find r_{nk}, we should compute $log \sum \exp \rho$. Since it can be so easy to fall in the underflow, *log-sum trick*[3] was used in this case.

[Results]

From *figure 1 and 2*, we can see that after 60th iterations, the model only has two component which have dominant responsibility. Others are numerically distinguishable from their prior values. This can be regarded as trade-off between fitting the data and the complexity of the model. If we think of LASSO, when the model gets complex, the penalty terms increase highly. Likewise, the complexity from components will let its parameters pushed away so that the model only has simple as possible though it still evaluates the data well.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning.*

[2] Seungjin Choi. *Lecture Note.*

[3] https://www.xarg.org/2016/06/the-log-sum-exp-trick-in-machine-learning/

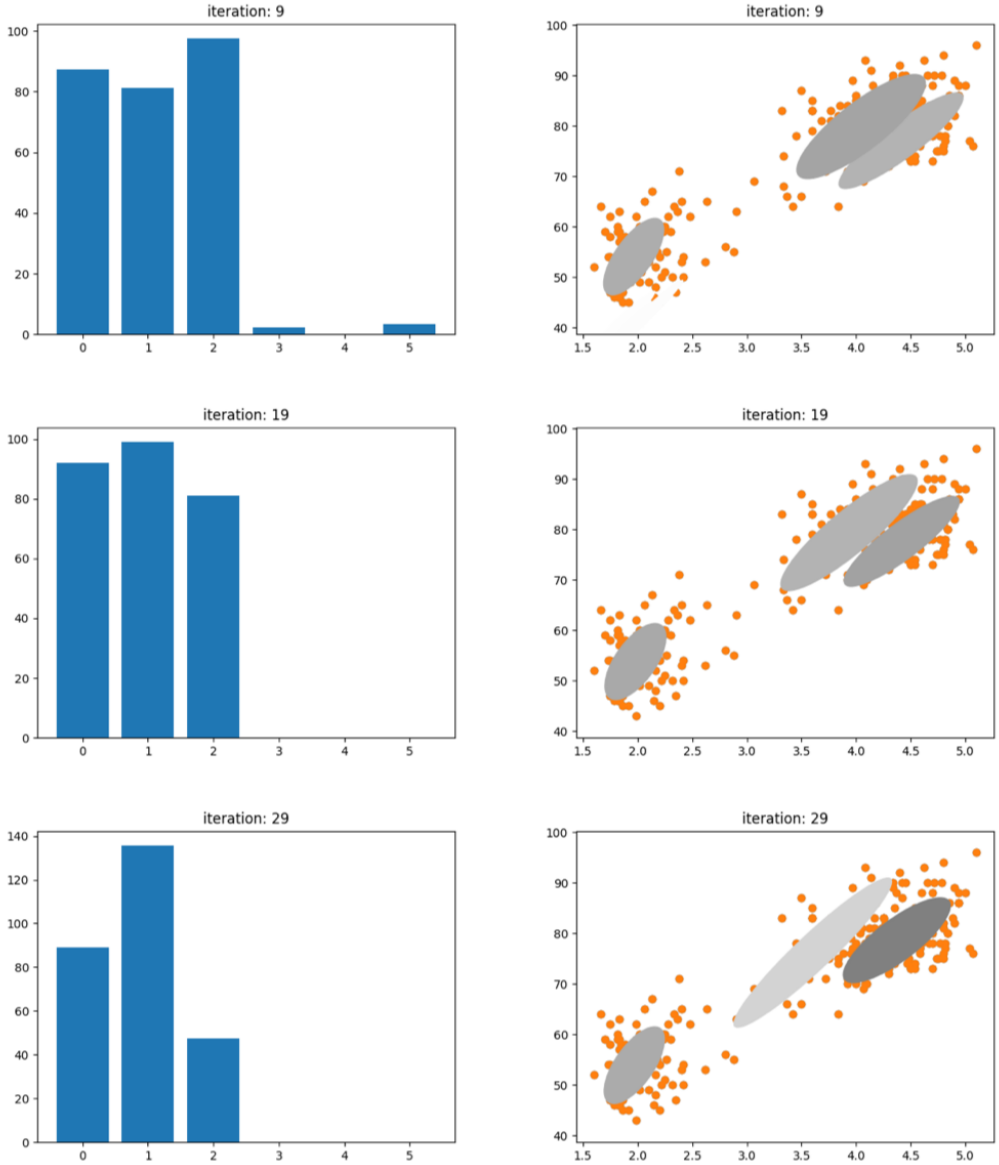

Figure 1: Result 1

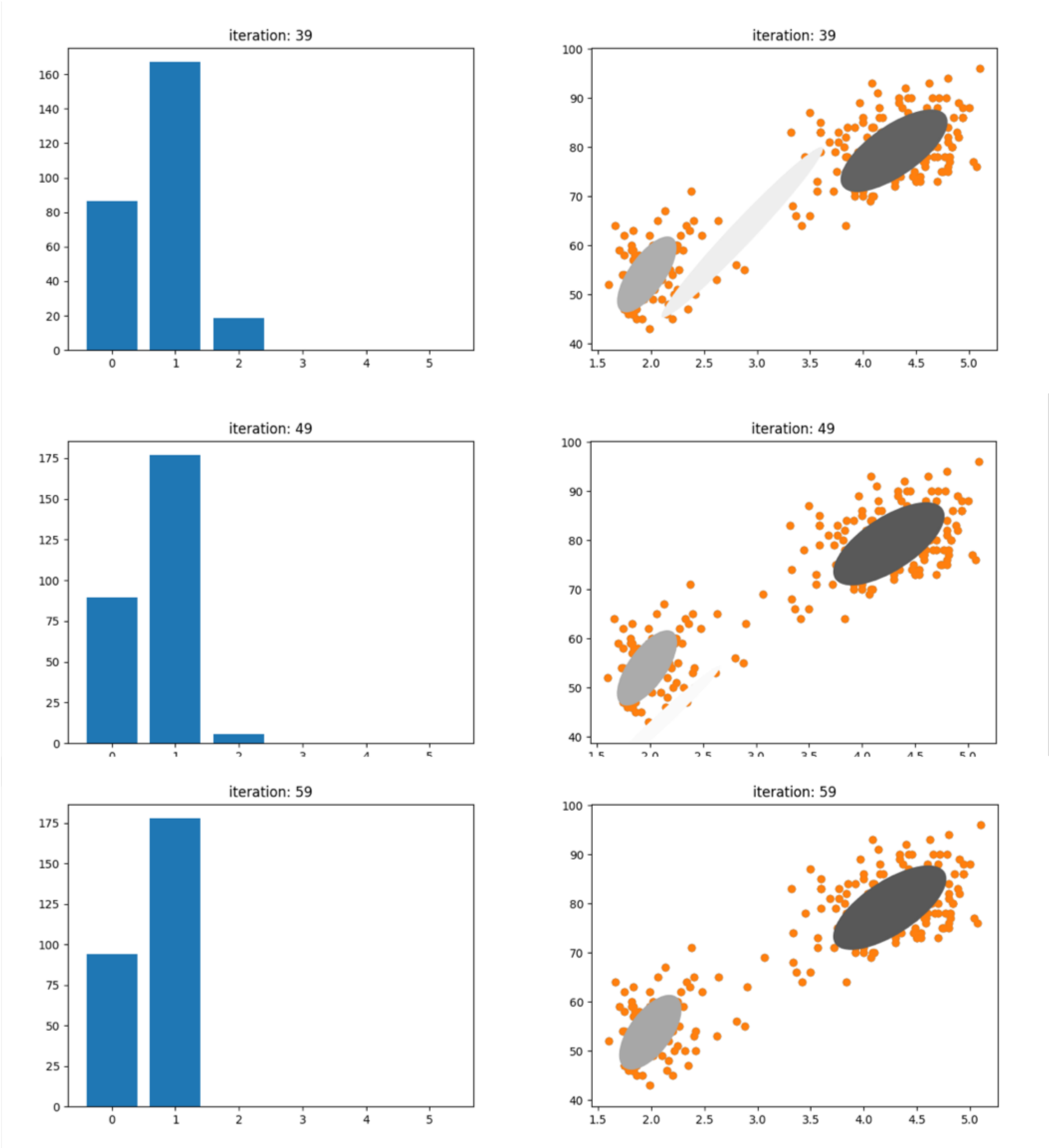

Figure 2: Result 2