CSED524 **Probabilistic Graphical Models**
Fall 2015
(Instructor: Prof. Seungjin Choi)

Homework#2: Variational Mixture of Gaussian

Yongkyu Cho (20130712)
jyg1124@postech.ac.kr

1. **Model description**

(a) **Graphical Model**

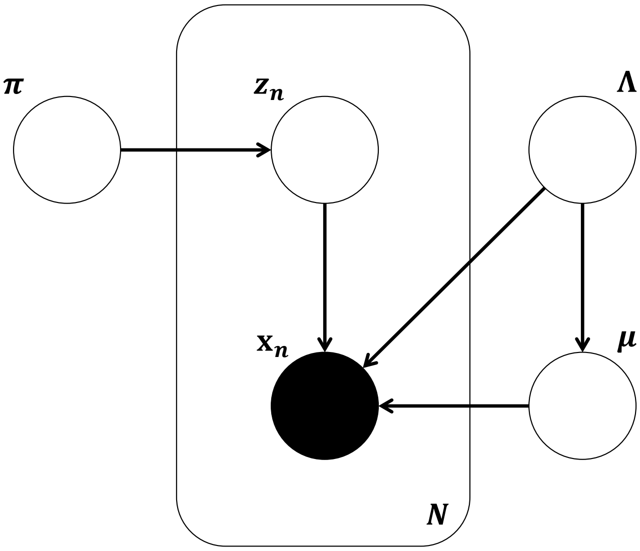

Figure 1: Directed acyclic graphical representation of Variational MoG

(b) **Notation**

$\mathbf{X} = \{\mathbf{x_1}, \cdots, \mathbf{x}_N\}$: Set of N data points, $\mathbf{x}_i$ is D-dim vector.
$\mathbf{Z} = \{\mathbf{z_1}, \cdots, \mathbf{z}_N\}$: Set of latent variables. Each $\mathbf{z}_i$ specifies the distribution of $\mathbf{x}_i$. $\mathbf{z}_i$ is K-dim vector.
$\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_K\}$: mixing coefficients, $\pi_k = \mathbb{P}(z_{nk} = 1)$
$\boldsymbol{\mu} = \{\mu_i\}_{i=1}^{K}, \mathbf{\Lambda} = \{\Lambda_i\}_{i=1}^{K}$: Multivariate Gaussian parameters corresponding to the mixing coefficients. (i.e.,$\pi_k \sim \mu_k, \Lambda_k^{-1}$)

(c) **Distributions for variables**

$\boldsymbol{\pi} \sim SymDir(K, \alpha_0)$: K-dimensional symmetric Dirichlet distribution with hyperparameter for each component set to α_0 (conjugate prior of the multinomial distribution)
$\boldsymbol{\Lambda} \sim \mathcal{W}(\mathbf{W}_0, \nu_0)$: Wishart distribution (conjugate prior of the precision matrix for a multivariate Gaussian distribution)
$\boldsymbol{\mu} \sim \mathcal{N}_K(\mu_0, (\beta_0 \Lambda_i)^{-1})$: K-dim multivariate Gaussian distribution
$\mathbf{z}_i \sim Multi(1, \boldsymbol{\pi})$: Multinomial distribution with sum-to-one constraint and parameter $\boldsymbol{\pi}$
$\mathbf{x}_i \sim \mathcal{N}_D(\mu_{z_i}, \Lambda_{z_i}^{-1})$: D-dim multivariate Gaussian distribution

2. **Derivation of variational distribution**

From the graphical model, the joint pdf is factorized by

$$p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) = p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})$$

The probability density functions for each factor are as follows.

$$\begin{aligned}
p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) &= \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}(\mathbf{x}_n|\mu_k,\Lambda_k^{-1})^{z_{nk}}\\
p(\mathbf{Z}|\boldsymbol{\pi}) &= \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}\\
p(\boldsymbol{\pi}) &= \frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_0)^K}\prod_{k=1}^{K}\pi_k^{\alpha_0-1}\\
p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) &= \prod_{k=1}^{K}\mathcal{N}(\mu_k|\mu_0,(\beta_0\Lambda_k)^{-1})\\
p(\boldsymbol{\Lambda}) &= \prod_{k=1}^{K}\frac{|\mathbf{W}_0|^{1/2}|\Lambda_k|^{(\nu_0-D-1)/2}\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}(\mathbf{W_0^{-1}}\Lambda_k)}}{2^{\frac{\nu_0 D}{2}}\pi^{\frac{D(D-1)}{4}}\prod_{i=1}^{D}\Gamma(\frac{\nu_0+1-i}{2})}
\end{aligned}$$

where $\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$

Now, we want to use the mean field theory. That is, we assume $q(\mathbf{Z}) = \prod_{i=1}^{M}q_i(\mathbf{Z}_i)$. Then, we can get the optimal (i.e., maximizing the lower bound of complete data log-likelihood) marginal distribution of factorized variables using the calculus of variation as follows.

$$\begin{aligned}
\mathcal{L}(q(\mathbf{Z})) &= \int q(\mathbf{Z})\log\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\mathrm{d}\mathbf{Z}\\
&= \int\prod_i q_i\{\log p(\mathbf{X},\mathbf{Z}) - \sum_i \log q_i\}\mathrm{d}\mathbf{Z}\\
&= \int\left(\prod_{i\neq j}q_i\cdot q_j\right)\left\{\log p(\mathbf{X},\mathbf{Z}) - \sum_{i\neq j}\log q_i - \log q_j\right\}\mathrm{d}\mathbf{Z}\\
&= \int\left(\prod_{i\neq j}q_i\right)\left\{q_j\log p(\mathbf{X},\mathbf{Z}) - q_j\sum_{i\neq j}\log q_i - q_j\log q_j\right\}\mathrm{d}\mathbf{Z}\\
&= \int q_j\cdot\log p(\mathbf{X},\mathbf{Z})\cdot\prod_{i\neq j}q_i\mathrm{d}\mathbf{Z} - \int q(\mathbf{Z})\sum_{i\neq j}\log q_i + q(\mathbf{Z})\log q_j\mathrm{d}\mathbf{Z}\\
&= \int q_j\left(\int_{i\neq j}\log p(\mathbf{X},\mathbf{Z})\prod_{i\neq j}q_i\mathrm{d}\mathbf{Z}_i\right)\mathrm{d}\mathbf{Z}_j - \int q_j\log q_j\mathrm{d}\mathbf{Z}_j + \text{const}\\
&= \underbrace{\int q_j\log\tilde{p}(\mathbf{X},\mathbf{Z})\mathrm{d}\mathbf{Z}_j - \int q_j\log q_j\mathrm{d}\mathbf{Z}_j + \text{const}}_{\text{(negative KL divergence between } q_j \text{ and } \tilde{p})}\cdots\textcircled{A}
\end{aligned}$$

where $\log \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \int \log p(\mathbf{X}, \mathbf{Z}) \cdot \prod_{i \neq j} q_i \mathrm{d}\mathbf{Z}_i = \mathbb{E}_{i \neq j} [\log p(\mathbf{X}, \mathbf{Z})] + \text{const}$. Maximizing Ⓐ is equivalent to minimizing $KL(q_j || \tilde{p})$. This minimum is attained when $q_j \approx \tilde{p}$. Therefore,

$$\begin{aligned} \log q_j^*(\mathbf{Z}_j) &= \mathop{\mathbb{E}}_{i \neq j} [\log p(\mathbf{X}, \mathbf{Z})] + \text{const} \\ \implies q_j^*(\mathbf{Z}_j) &= \frac{\mathrm{e}^{\mathbb{E}_{i \neq j}[\log p(\mathbf{X}, \mathbf{Z})]}}{\int \mathrm{e}^{\mathbb{E}_{i \neq j}[\log p(\mathbf{X}, \mathbf{Z})]} \mathrm{d}\mathbf{Z}_j} \end{aligned}$$

Applying this result to the factorized distribution, we can calculate the optimal functional form of the factors $q^*(\mathbf{Z})$ as followings.

$$\begin{aligned} \log q^*(\mathbf{Z}) &= \mathop{\mathbb{E}}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}} [\log (\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const} \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}} [\log p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) + \log p(\boldsymbol{Z} | \boldsymbol{\pi}) + \log p(\boldsymbol{\pi}) + \log p(\boldsymbol{\mu} | \boldsymbol{\Lambda}) + \log p(\boldsymbol{\Lambda})] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}} [\log p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \mathop{\mathbb{E}}_{\boldsymbol{\pi}} [\log p(\boldsymbol{Z} | \boldsymbol{\pi})] + \text{const} \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}} \left[\log \left\{ \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\frac{1}{(2\pi)^{D/2}} \cdot \frac{1}{|\Lambda_k^{-1}|^{1/2}} \cdot \mathrm{e}^{-\frac{1}{2}(\mathbf{x}_n - \mu_k)^{\mathrm{T}} \Lambda_k (\mathbf{x}_n - \mu_k)} \right)^{z_{nk}} \right\} \right] \\ &\quad + \mathop{\mathbb{E}}_{\boldsymbol{\pi}} \left[\log \left(\prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \right) \right] + \text{const} \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}} \left[\sum_n \sum_k z_{nk} \log \left(\frac{1}{(2\pi)^{D/2}} \cdot \frac{1}{|\Lambda_k^{-1}|^{1/2}} \cdot \mathrm{e}^{-\frac{1}{2}(\mathbf{x}_n - \mu_k)^{\mathrm{T}} \Lambda_k (\mathbf{x}_n - \mu_k)} \right) \right] \\ &\quad + \mathop{\mathbb{E}}_{\boldsymbol{\pi}} \left[\sum_n \sum_k z_{nk} \log \pi_k \right] + \text{const} \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}} \left[\sum_n \sum_k z_{nk} \left(\log \frac{1}{(2\pi)^{D/2}} + \frac{1}{2} \log |\Lambda_k| - \frac{1}{2} (\mathbf{x}_n - \mu_k)^{\mathrm{T}} \Lambda_k (\mathbf{x}_n - \mu_k) \right) \right] \\ &\quad + \sum_n \sum_k z_{nk} \, \mathbb{E} [\log \pi_k] + \text{const} \\ &= \sum_n \sum_k z_{nk} \left(-\frac{D}{2} \right) \log 2\pi + \frac{1}{2} \sum_n \sum_k z_{nk} \, \mathbb{E} [\log |\Lambda_k|] \\ &\quad - \frac{1}{2} \sum_n \sum_k \mathop{\mathbb{E}}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}} \left[(\mathbf{x}_n - \mu_k)^{\mathrm{T}} \Lambda_k (\mathbf{x}_n - \mu_k) \right] + \sum_n \sum_k z_{nk} \, \mathbb{E} [\log \pi_k] + \text{const} \end{aligned}$$

Defining $\log \rho_{nk} = -\frac{D}{2} \log 2\pi + \frac{1}{2} \mathbb{E} [\log |\Lambda_k|] - \frac{1}{2} \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}} \left[(\mathbf{x}_n - \mu_k)^{\mathrm{T}} \Lambda_k (\mathbf{x}_n - \mu_k) \right] + \mathbb{E} [\log \pi_k]$

gives us

$$\begin{aligned}
\log q^*(\boldsymbol{Z}) &= \sum_n \sum_k z_{nk} \log \rho_{nk} + \text{const} \\
\implies q^*(\boldsymbol{Z}) &= \mathrm{e}^{\sum_n \sum_k z_{nk} \log \rho_{nk} + \text{const}} \\
&= \prod_n \prod_k \rho_{nk}^{z_n k} \cdot e^{\text{const}} \\
\implies q^*(\boldsymbol{Z}) &\propto \prod_n \prod_k \rho_{nk}^{z_n k}
\end{aligned}$$

Using the fact that for each value of n, the quatity z_{nk} are binaries and sum-to-1 over all values of k, we obtain

$$q^*(\boldsymbol{Z}) = \prod_n \prod_k r_{nk}^{z_{nk}}$$

where $r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nk}}$. We can regard this as product of single-observation multinomial distribution with parameter r_{nk} for $k = 1, \cdots, K$. Moreover, we know $\mathbb{E}\left[z_{nk}\right] = r_{nk}$. (mean of multinomial distribution)

Now, let's find another optimal factor $q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$. Applying the result we got above,

$$\begin{aligned}
\log q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \mathop{\mathbb{E}}_{\boldsymbol{Z}}\left[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] + \text{const} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{Z}}\left[\log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) + \log p(\mathbf{Z}|\boldsymbol{\pi}) + \log p(\boldsymbol{\pi}) + \log p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) + \log p(\boldsymbol{\Lambda})\right] + \text{const} \\
&= \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{Z}}\left[\log p(\boldsymbol{Z}|\boldsymbol{\pi})\right] + \log p(\boldsymbol{\pi}) +}_{\text{w.r.t. } \boldsymbol{\pi}} \underbrace{\log\left\{p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) p(\boldsymbol{\Lambda})\right\}}_{(=\log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}))\text{w.r.t. } \boldsymbol{\mu}, \boldsymbol{\Lambda}} + \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{Z}}\left[\log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right]}_{\text{w.r.t. } \boldsymbol{\mu}, \boldsymbol{\Lambda}} + \text{const} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{Z}}\left[\log \prod_n \prod_k \pi_k^{z_{nk}}\right] + \log\left(\frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_0)^K} \prod_k \pi_k^{\alpha_0 - 1}\right) \cdots \text{w.r.t. } \boldsymbol{\pi} \\
&\quad + \log\left(\prod_k \mathcal{N}(\mu_k|\mu_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k|\mathbf{W}_0, \nu_0)\right) \cdots \text{w.r.t. } \boldsymbol{\mu}, \boldsymbol{\Lambda} \\
&\quad + \mathop{\mathbb{E}}_{\mathbf{Z}}\left[\log\left(\prod_n \prod_k \mathcal{N}(\mathbf{x_n}|\mu_k, \Lambda_k^{-1})^{z_{nk}}\right)\right] + \text{const} \cdots \text{w.r.t. } \boldsymbol{\mu}, \boldsymbol{\Lambda} \\
&= \mathop{\mathbb{E}}_{\mathbf{Z}}\left[\log \prod_n \prod_k \pi_k^{z_{nk}}\right] + \log \frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_0)^K} + (\alpha_0 - 1) \sum_k \log \pi_k \\
&\quad + \sum_k \log\left\{\mathcal{N}(\mu_k|\mu_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k|\mathbf{W}_0, \nu_0)\right\} \\
&\quad + \sum_n \sum_k \mathbb{E}\left[z_{nk}\right] \log \mathcal{N}(\mathbf{x}_n|\mu_k, \Lambda_k^{-1}) + \text{const}
\end{aligned}$$

Therefore, we can notice that $q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ can be factorized into $q^*(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\mu_k, \Lambda_k)$ so that $\log q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \log q^*(\boldsymbol{\pi}) + \sum_k q^*(\mu_k, \Lambda_k)$. Moreover, identifying the terms depending on $\boldsymbol{\pi}$ yields us

$$\begin{aligned}
\log q^*(\boldsymbol{\pi}) &= \sum_n \sum_k \underbrace{\mathbb{E}\left[z_{nk}\right]}_{=r_{nk}} \log \pi_k + (\alpha_0 - 1) \sum_k \log \pi_k + \text{const} \\
&= \sum_k \underbrace{\sum_n r_{nk}}_{=:N_k} \log \pi_k + (\alpha_0 - 1) \sum_k \log \pi_k + \text{const} \\
\Longrightarrow q^*(\boldsymbol{\pi}) &= \mathrm{e}^{\sum_k N_k \log \pi_k} \cdot \mathrm{e}^{(\alpha_0 - 1)\sum_k \log \pi_k} \cdot \mathrm{e}^{\text{const}} \\
&= \prod_k \pi_k^{N_k + \alpha_0 - 1} \cdot \mathrm{e}^{\text{const}}
\end{aligned}$$

Then, we can recognize that $q^*(\boldsymbol{\pi}) \sim Dir(\boldsymbol{\pi}|\boldsymbol{\alpha})$ with parameter $\boldsymbol{\alpha}$ has components $\alpha_k = \alpha_0 + N_k$.

Now, we are going to get the optimal factor $q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. The logarithm of this factor can be factorized and expressed as

$$\begin{aligned}
\log q^*(\mu_k, \Lambda_k) &= \log q^*(\mu_k|\Lambda_k) + \log q^*(\Lambda_k) \\
&= \log \left\{\mathcal{N}(\mu_k|\mu_0, (\beta_0\Lambda_k)^{-1}) \mathcal{W}(\Lambda_k|\mathbf{W}_0, \nu_0)\right\} \\
&\quad + \sum_n \mathbb{E}\left[z_{nk}\right] \log \mathcal{N}(\mathbf{x}_n|\mu_k, \Lambda_k^{-1}) + \text{const} \\
&= -\frac{\beta_0}{2}(\mu_k - \mu_0)^{\mathrm{T}}\Lambda_k(\mu_k - \mu_0) + \frac{1}{2}\log|\Lambda_k| - \frac{1}{2}\mathrm{Tr}(\Lambda_k \mathbf{W}_0^{-1}) + \frac{\nu_0 - D - 1}{2}\log|\Lambda_k| \\
&\quad - \frac{1}{2}\sum_n \mathbb{E}\left[z_{nk}\right](\mathbf{x}_n - \mu_k)^{\mathrm{T}}\Lambda_k(\mathbf{x}_n - \mu_k) + \frac{1}{2}\log|\Lambda_k| \sum_n \mathbb{E}\left[z_{nk}\right] + \text{const}
\end{aligned}$$

To find the optimal functional form of μ_k, gathering the terms with respect to μ_k from above equation gives us.

$$\begin{aligned}
\log q^*(\mu_k|\Lambda_k) &= -\frac{\beta_0}{2}(\mu_k - \mu_0)^{\mathrm{T}}\Lambda_k(\mu_k - \mu_0) - \frac{1}{2}\sum_n \mathbb{E}\left[z_{nk}\right](\mathbf{x}_n - \mu_k)^{\mathrm{T}}\Lambda_k(\mathbf{x}_n - \mu_k) \\
&= -\frac{\beta_0}{2}\left(\mu_k^{\mathrm{T}}\Lambda_k\mu_k - \mu_k^{\mathrm{T}}\Lambda_k\mu_0 - \mu_0^{\mathrm{T}}\Lambda_k\mu_k + \cancelto{\text{const}}{\mu_0^{\mathrm{T}}\Lambda_k\mu_0}\right) \\
&\quad - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}\left[z_{nk}\right]\left(\cancelto{\text{const}}{\mathbf{x}_n^{\mathrm{T}}\Lambda_k\mu_k} - \mathbf{x}_n^{\mathrm{T}}\Lambda_k\mu_k - \mu_k^{\mathrm{T}}\Lambda_k\mathbf{x}_n + \mu_k^{\mathrm{T}}\Lambda_k\mu_k\right) \\
&= -\frac{1}{2}\mu_k^{\mathrm{T}}\left(\beta_0 + \sum_n \mathbb{E}\left[z_{nk}\right]\right)\Lambda_k\mu_k + \frac{\beta_0}{2}\left(\mu_k^{\mathrm{T}}\Lambda_k\mu_0 + \mu_0^{\mathrm{T}}\Lambda_k\mu_k\right) \\
&\quad + \frac{1}{2}\sum_n \mathbb{E}\left[z_{nk}\right]\left(\mathbf{x}_n^{\mathrm{T}}\Lambda_k\mu_k + \mu_k^{\mathrm{T}}\Lambda_k\mathbf{x}_n\right) + \text{const} \\
&= -\frac{1}{2}\mu_k^{\mathrm{T}}\left(\beta_0 + \underbrace{\sum_n \mathbb{E}\left[z_{nk}\right]}_{=:N_k}\right)\Lambda_k\mu_k + \mu_k^{\mathrm{T}}\Lambda_k\left(\beta_0\mu_0 + \underbrace{\sum_n \mathbb{E}\left[z_{nk}\right]\mathbf{x}_n}_{=:N_k\bar{\mathbf{x}}_k}\right) + \text{const} \\
&= -\frac{1}{2}\mu_k^{\mathrm{T}}\left(\beta_0 + N_k\right)\Lambda_k\mu_k + \mu_k^{\mathrm{T}}\Lambda_k\left(\beta_0\mu_0 + N_k\bar{\mathbf{x}}_k\right) + \text{const}
\end{aligned}$$

Since $\log q^*(\mu_k|\Lambda_k)$ depends on μ_k quadratically, it follows a multivariate Gaussian distibution with parameters $\beta_k = \beta_0 + N_k$ and $\mathbf{m}_k = \frac{1}{\beta_k}(\beta_0\mu_0 + N_k\bar{\mathbf{x}}_k)$. That is,

$$q^*(\mu_k|\Lambda_k) \sim \mathcal{N}(\mu_k|\mathbf{m}_k, \beta_k\Lambda_k).$$

To find the optimal functional form of Λ_k, we are going to use the following equality.

$$\log q^*(\Lambda_k) = \log q^*(\mu_k, \Lambda_k) - \log q^*(\mu_k|\Lambda_k)$$

Plugging the results from above into the right hand side of the equality and gathering the terms with respect to Λ_k gives us

$$\begin{aligned}
\log q^*(\Lambda_k) = &-\frac{\beta_0}{2}(\mu_k - \mu_0)^{\mathrm{T}}\Lambda_k(\mu_k - \mu_0) + \frac{1}{2}\log|\Lambda_k| - \frac{1}{2}\mathrm{Tr}(\Lambda_k\mathbf{W}_0^{-1}) \\
&+ \frac{\nu_0 - D - 1}{2}\log\Lambda_k - \frac{1}{2}\sum_n \mathbb{E}\left[z_{nk}\right](\mathbf{x}_n - \mu_k)^{\mathrm{T}}\Lambda_k(\mathbf{x}_n - \mu_k) \\
&+ \frac{1}{2}\left(\sum_n \mathbb{E}\left[z_{nk}\right]\right)\log|\Lambda_k| + \frac{\beta_k}{2}(\mu_k - \mathbf{m}_k)^{\mathrm{T}}\Lambda_k(\mu_k - \mathbf{m}_k) - \frac{1}{2}\log|\Lambda_k| + \text{const} \\
= &\frac{\nu_k - D - 1}{2}\log|\Lambda_k - \frac{1}{2}\mathrm{Tr}(\Lambda_k\mathbf{W}_k^{-1}) + \text{const}
\end{aligned}$$

where we have defined

$$\begin{aligned}
\mathbf{S}_k &= \frac{1}{N_k}\sum_n r_{nk}(\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^{\mathrm{T}} \\
\mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + \beta_0(\mu_k - \mu_0)(\mu_k - \mu_0)^{\mathrm{T}} + \sum_n \mathbb{E}\left[z_{nk}\right](\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^{\mathrm{T}} \\
&\quad - \beta_k(\mu_k - \mathbf{m}_k)(\mu_k - \mathbf{m}_k)^{\mathrm{T}} \\
&= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\mathbf{x}}_k - \mu_0)(\bar{\mathbf{x}_k} - \mu_0)^{\mathrm{T}} \\
\nu_k &= \nu_0 + N_k
\end{aligned}$$

Now we can notice that

$$q^*(\Lambda_k) \sim \mathcal{W}(\Lambda_k|\mathbf{W}_k, \nu_k).$$

Therefore,

$$q^*(\mu_k, \Lambda_k) = \mathcal{N}\left(\mu_k|\mathbf{m}_k, (\beta_k\Lambda_k)^{-1}\right)\mathcal{W}(\Lambda_k|\mathbf{W}_k, \nu_k). \quad \text{(Gaussian-Wishart distribution)}$$

3. **Variational EM algorithm**

   (a) **Variational E-Step**: Using the variational distributions we obtained above, we can evaluate the expectations over the current model parameters.

      i. Evaluate the followings.

$$\underset{\mu_k,\Lambda_k}{\mathbb{E}}\left[(\mathbf{x}_n-\mu_k)^{\mathrm{T}}\Lambda_k(\mathbf{x}_n-\mu_k)\right]=D\beta_k^{-1}+\nu_k(\mathbf{x}_n-\mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\mathbf{x}_n-\mathbf{m}_k)$$

$$\log\tilde{\Lambda}_k:=\mathbb{E}\left[\log|\Lambda_k|\right]=\sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}+D\log 2+\log|\mathbf{W}_k|\right)$$

$$\log\tilde{\pi}_k:=\mathbb{E}\left[\log\pi_k\right]=\psi(\alpha_k)-\psi\left(\sum_k\alpha_k\right)$$

      where $\psi(x)=\frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$.

      ii. Evaluate the following using the result from i.

$$\log\rho_{nk}=-\frac{D}{2}\log 2\pi+\frac{1}{2}\mathbb{E}\left[\log|\Lambda_k|\right]-\frac{1}{2}\underset{\mu_k,\Lambda_k}{\mathbb{E}}\left[(\mathbf{x}_n-\mu_k)^{\mathrm{T}}\Lambda_k(\mathbf{x}_n-\mu_k)\right]+\mathbb{E}\left[\log\pi_k\right]$$

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

      iii. Evaluate

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

   (b) **Variational M-Step**: Compute the optimized variational distribution over model parameters using the current distributions over latent variables. Fix r_{nk}.

      i. Compute the following quantities.

$$N_k=\sum_{n=1}^{N}r_{nk}$$

$$\bar{\mathbf{x}}_k=\frac{1}{N_k}\sum_{n=1}^{N}r_{nk}\mathbf{x}_n$$

$$\mathbf{S}_k=\frac{1}{N_k}\sum_{n}r_{nk}(\mathbf{x}_n-\bar{\mathbf{x}}_k)(\mathbf{x}_n-\bar{\mathbf{x}}_k)^{\mathrm{T}}$$

$$\beta_k=\beta_0+N_k$$

$$\mathbf{m}_k=\frac{1}{\beta_k}(\beta_0\mu_0+N_k\bar{\mathbf{x}}_k)$$

$$\nu_k=\nu_0+N_k$$

$$\mathbf{W}_k^{-1}=\mathbf{W}_0^{-1}+N_k\mathbf{S}_k+\frac{\beta_0N_k}{\beta_0+N_k}(\bar{\mathbf{x}}_k-\mu_0)(\bar{\mathbf{x}_k}-\mu_0)^{\mathrm{T}}$$

ii. Compute the following optimized variational distirubiton using the quantities above.

$$q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda})$$

$$= \prod_{k=1}^{K} \frac{1}{(2\pi)^{D/2}} |\beta_k \Lambda_k|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(\mu_k - \mathbf{m}_k)^T \beta_k \Lambda_k (\mu_k - \mathbf{m}_k)} \frac{|\mathbf{W}_k|^{1/2} |\Lambda_k|^{(\nu_k - D - 1)/2} \mathrm{e}^{-\frac{1}{2}\mathrm{Tr}(\mathbf{W_k^{-1}} \Lambda_k)}}{2^{\frac{\nu_k D}{2}} \pi^{\frac{D(D-1)}{4}} \prod_{i=1}^{D} \Gamma(\frac{\nu_k + 1 - i}{2})}$$

4. Experiment

(a) Data

Data for experiment is generated synthetically. Four 2-dimensional multivariate Gaussians are used. The synthetic data is described in the figure below. Each Gaussian component is

$$\mu_1 = \begin{bmatrix} 2 & 6 \end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 3 \end{bmatrix}$$

$$\mu_2 = \begin{bmatrix} 7 & 9 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$$

$$\mu_3 = \begin{bmatrix} 9 & 3 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

$$\mu_4 = \begin{bmatrix} 5 & 5 \end{bmatrix} \qquad \Sigma_4 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}$$

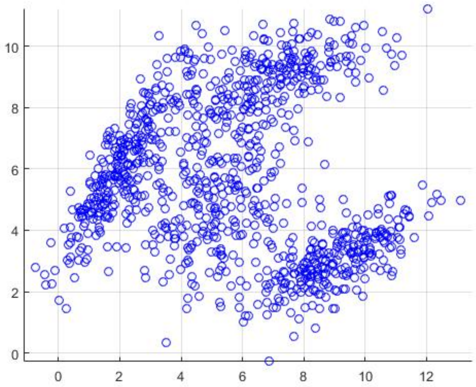

Figure 2: Synthetic Data

(b) **Result: Standard MoG (EM)**
Standard MoG often shows different results even though experemental setting is not changed.

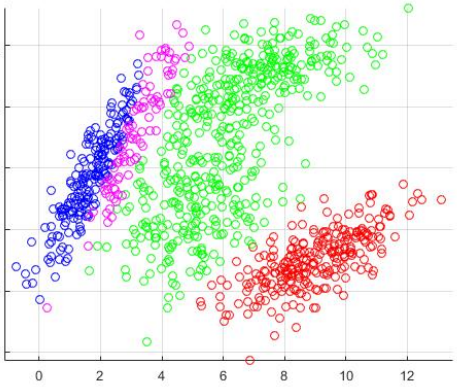

Figure 3: Standard MoG 1 (K=4)

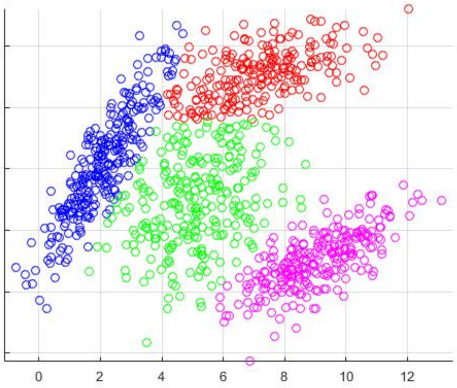

Figure 4: Standard MoG 2 (K=4)

(c) **Result: Variational MoG (VBEM)**

VBMoG shows robust and nice performance whenever the number of components is chosen sufficiently largely. However, the larger the number of component is chosen, the longer time it takes until convergence.

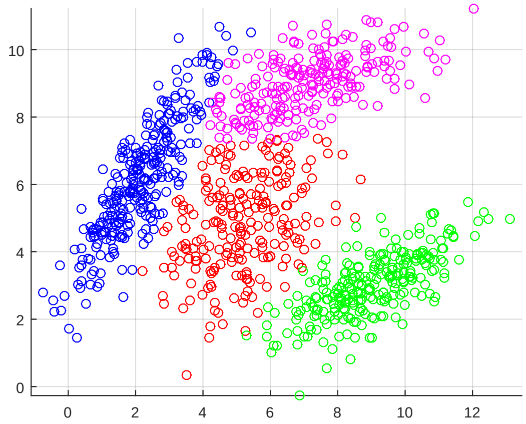

Figure 5: VBMoG (initial K=10)

5. **Comparison with standard MoG**

- VBMoG is robuster than MoG. MoG often shows different result even under the same experemental setting whereas VBMoG always shows consistently nice performance.
- VBMoG does not allow the singular solutions often arising in the machine learning approach where a Gaussian component becomes responsible for a single data point.
- VBMoG can directly determine the optimal number of components without resorting to methods such as cross-validation.
- There is no over-fitting if we choose a large number K of components in the mixture.