# GMMs by Mean Field Variational Inference

Jiajun He

In Bayesian framework, we want to find a set of parameters

$$\boldsymbol{\theta} = \{\boldsymbol{\pi}, z_1, z_2, ..., z_N, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K\}$$

, which maximizes the posterior probability $P(\boldsymbol{\theta}|\boldsymbol{X})$. But the form of the posterior distribution is relatively complicated in GMMs, so we use $q(\boldsymbol{\theta})$ to approximate it.

According to Bayes Rule, the probability of observing $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ is

$$P(\boldsymbol{X}) = \frac{P(\boldsymbol{\theta}, \boldsymbol{X})}{P(\boldsymbol{\theta}|\boldsymbol{X})} = \frac{P(\boldsymbol{\theta}, \boldsymbol{X})}{q(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta})}{P(\boldsymbol{\theta}|\boldsymbol{X})}$$

Take logarithm and expectation on both sides, we get

$$\int q(\boldsymbol{\theta}) \log P(\boldsymbol{X}) \mathrm{d}\boldsymbol{\theta} = \int q(\boldsymbol{\theta}) \log \frac{P(\boldsymbol{\theta}, \boldsymbol{X})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{P(\boldsymbol{\theta}|\boldsymbol{X})} \mathrm{d}\boldsymbol{\theta}$$

i.e.,

$$\log P(\boldsymbol{X}) = \int q(\boldsymbol{\theta}) \log \frac{P(\boldsymbol{\theta}, \boldsymbol{X})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} + D_{KL}[q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{X})]$$

The tern on the left-hand side is a constant, so minimizing KL-divergence between $q(\boldsymbol{\theta})$ and real posterior distribution is equivalent to maximizing the first term on the right-hand side, which is named as *ELBO*.

Under mean field assumption, i.e., $q(\boldsymbol{\theta}) = \prod_{\theta_i \in \boldsymbol{\theta}} q(\theta_i)$ , each variable can be optimized successively:
Specifically, for each $\theta_i \in \boldsymbol{\theta}$,

$$\begin{aligned}
ELBO &= \int q(\boldsymbol{\theta}) \log(\frac{P(\boldsymbol{\theta}, \boldsymbol{X})}{q(\boldsymbol{\theta})}) \mathrm{d}\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log P(\boldsymbol{\theta}, \boldsymbol{X}) \mathrm{d}\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&= \mathbb{E}_{\theta_i}[\mathbb{E}_{\boldsymbol{\theta}-\{\theta_i\}}[\log P(\boldsymbol{\theta}, \boldsymbol{X})]] - \mathbb{E}_{\boldsymbol{\theta}}[\log q(\theta_i)] - \sum_{j \neq i} \mathbb{E}_{\boldsymbol{\theta}}[\log q(\theta_j)] \\
&= \int q(\theta_i) \mathbb{E}_{\boldsymbol{\theta}-\{\theta_i\}}[\log P(\boldsymbol{\theta}, \boldsymbol{X})] \mathrm{d}\theta_i - \int q(\theta_i) \log q(\theta_i) \mathrm{d}\theta_i - \sum_{j \neq i} \int q(\theta_j) \log q(\theta_j) \mathrm{d}\theta_j
\end{aligned}$$

The last term is a constant for $\theta_i$, i.e.,

$$ELBO = \int q(\theta_i) \mathbb{E}_{\boldsymbol{\theta}-\{\theta_i\}}[\log P(\boldsymbol{\theta}, \boldsymbol{X})] \mathrm{d}\theta_i - \int q(\theta_i) \log q(\theta_i) \mathrm{d}\theta_i + const$$

As the expectation term is positive, we can view it as an log-probability $\log \tilde{P}$, then

$$ELBO = -D_{KL}[q(\theta_i)||\tilde{P}] + const$$

To maximum $ELBO$ is equivalent to maximum the KL-divergence between each $q(\theta_i)$ and corresponding $\tilde{P}$.

Here, our task is just to find an MAP assignment, so to furthermore simplify the calculation, I set all $q$ to be one-point distribution, in which situation, $ELBO$ is maximized when

$$q(\theta_i) = \mathbb{I}\Big\{\theta_i = \text{mode}\{\tilde{P}\}\Big\}$$

Hereinafter, to simplify the notation, I will use $\theta_i = a$ to represent $q(\theta_i) = \mathbb{I}\{\theta_i = a\}$.

In our Gaussian Mixture Model, in each iteration step, I successively optimize $\boldsymbol{\pi}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K, z_1, z_2, ..., z_N$as follows:

**Optimize $\boldsymbol{\pi}$:**

$$\begin{aligned}\log q(\boldsymbol{\pi}) = \int & q(z_1)q(z_2)...q(z_N)q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)...q(\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K) \\ & \log\Big\{P(\boldsymbol{\pi})\prod_{i=1}^{k}P(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\prod_{j=1}^{N}[P(z_j|\boldsymbol{\pi})P(\boldsymbol{x}_j|z_j, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)]\Big\} \\ & \mathrm{d}z_1\mathrm{d}z_2...\mathrm{d}z_N\mathrm{d}\boldsymbol{\mu}_1\mathrm{d}\boldsymbol{\Sigma}_1\mathrm{d}\boldsymbol{\mu}_2\mathrm{d}\boldsymbol{\Sigma}_2...\mathrm{d}\boldsymbol{\mu}_K\mathrm{d}\boldsymbol{\Sigma}_K\end{aligned}$$

All terms without $\boldsymbol{\pi}$ in it can be viewed as constants, and integrate out variables that do not occur in log term, we can get

$$\begin{aligned}\log q(\boldsymbol{\pi}) &= \int q(z_1)q(z_2)...q(z_N)\log P(\boldsymbol{\pi})P(z_1|\boldsymbol{\pi})P(z_2|\boldsymbol{\pi})...P(z_N|\boldsymbol{\pi})\mathrm{d}z_1\mathrm{d}z_2...\mathrm{d}z_N + const \\ &= \log P(\boldsymbol{\pi}) + \log P(z_1, z_2, ..., z_N|\boldsymbol{\pi}) + const\end{aligned}$$

The last step is correct as $q(z_i)$ is one-point distribution according to my assumption.

Considering that $\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\alpha})$ and $P(z_1, z_2, ..., z_N|\boldsymbol{\pi}) \sim \text{Multi}(\boldsymbol{\pi})$, the right-hand side is just the posterior Dirichlet distribution. Therefore, $\boldsymbol{\pi} = \text{mode}\Big\{\text{Dir}(\alpha_1 + N_1, \alpha_2 + N_2..., \alpha_K + N_K)\Big\}$ where $N_k$ is the number of observations in cluster $k$.

**Optimize $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$:**

When optimizing $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$, given that the entanglement of them does not bring us any inconvenience because we have a well-formed distribution over $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$, I do not entangle them into two $q$. Similar to $\boldsymbol{\pi}$, we can cancel out irrelevant variables and only get

$$\log q(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \log P(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + \sum_{z_j=i}\log P(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + const$$

The right-hand side is just the posterior Normal-Inverse-Wisart distribution. Thus, $\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j = \text{mode}\Big\{\mathcal{NIW}(\frac{\lambda\boldsymbol{\mu}_0 + N_j\bar{\boldsymbol{x}}}{\lambda + N_j}, \lambda + N_j, \nu + N_j, \boldsymbol{\Psi} + \mathrm{C} + \frac{\lambda N_j}{\lambda + N_j}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T)\Big\}$, where $\mathbf{C} = \sum_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ and $\bar{\boldsymbol{x}}$ is the mean of samples in class $j$;

**Optimize $z_1, z_2, ..., z_N$:**

Via similar calculation, it is easy to find that each $z_i$ can be optimized as $z_i = \max_{j=1}^{k}\{\boldsymbol{\pi}_j P(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\}$.

**Predict $z$ for new $\boldsymbol{x}$:**

After finding the MAP assignments for all parameters (i.e., one-point posterior distribution), we can predict class labels for new input $\boldsymbol{x}$.

$$
\begin{aligned}
&P(z|\boldsymbol{x}, \boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N}) \\
=&\int P(z|\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K, z_1, z_2, ..., z_N) \\
&\qquad P(\boldsymbol{\pi}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K, z_1, z_2, ..., z_N|\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N}) \\
&\qquad\qquad \mathrm{d}\boldsymbol{\pi}\mathrm{d}\boldsymbol{\mu}_1\mathrm{d}\boldsymbol{\Sigma}_1\mathrm{d}\boldsymbol{\mu}_2\mathrm{d}\boldsymbol{\Sigma}_2...\mathrm{d}\boldsymbol{\mu}_K\mathrm{d}\boldsymbol{\Sigma}_K\mathrm{d}z_1\mathrm{d}z_2...\mathrm{d}z_N \\
=&\int P(z|\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K) \\
&\qquad P(\boldsymbol{\pi}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K, z_1, z_2, ..., z_N|\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N}) \\
&\qquad\qquad \mathrm{d}\boldsymbol{\pi}\mathrm{d}\boldsymbol{\mu}_1\mathrm{d}\boldsymbol{\Sigma}_1\mathrm{d}\boldsymbol{\mu}_2\mathrm{d}\boldsymbol{\Sigma}_2...\mathrm{d}\boldsymbol{\mu}_K\mathrm{d}\boldsymbol{\Sigma}_K\mathrm{d}z_1\mathrm{d}z_2...\mathrm{d}z_N \\
=&P(z|\boldsymbol{x}, \boldsymbol{\pi}^{MAP}, \boldsymbol{\mu}_1^{MAP}, \boldsymbol{\Sigma}_1^{MAP}, \boldsymbol{\mu}_2^{MAP}, \boldsymbol{\Sigma}_2^{MAP}, ..., \boldsymbol{\mu}_K^{MAP}, \boldsymbol{\Sigma}_K^{MAP}) \\
\propto&P(\boldsymbol{x}|z, \boldsymbol{\mu}_1^{MAP}, \boldsymbol{\Sigma}_1^{MAP}, \boldsymbol{\mu}_2^{MAP}, \boldsymbol{\Sigma}_2^{MAP}, ..., \boldsymbol{\mu}_K^{MAP}, \boldsymbol{\Sigma}_K^{MAP})P(z|\boldsymbol{\pi}^{MAP})
\end{aligned}
$$