A Brief Note on Variational Inference

Di Wang
280868861@qq.com
PhD Candidate of Xi'an Jiaotong University, China
04/28/2020

Interested readers are pointed to Chapter 10 of Pattern Recognition and Machine Learning (PRML). We present missing details in some equations. For instance, the updating formula of mean vector and precision matrix in a multi-variate Gaussian mixture model (GMM) is illustrated in details. For a beginner of variational inference (VI), the updating formula of parameters in a multi-variate GMM is formidable. Thus, we also presented a toy example of univariate GMM without mean vector, which is both conceptually and mathematically concise. Therefore, a beginner is encouraged to firstly understand the subsection titled "Variational Mixture of 1D GMM without Mean Vector" before delving into the details of VI.

Evident Lower Bound

Our purpose is to compute the posterior

$$p(\mathbf{z}\,|\,\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})}, \tag{1}$$

$\mathbf{x}$ denotes the observed variable and $\mathbf{z}$ denotes the latent variable. Typically, introducing latent variable $\mathbf{z}$ simplifies the probabilistic inference. For instance, $\mathbf{x}$ denotes the pixel values of an image, and $\mathbf{z}$ denotes the object category of the image such as cat/dog. $p(\mathbf{x},\mathbf{z})$ denotes the model or joint distribution. The joint distribution can be easily factored by utilizing the respective graph model. $p(\mathbf{x}) = \int p(\mathbf{x},\mathbf{z})d\mathbf{z}$ denotes the model evidence or marginal likelihood. Generally, the exact posterior $p(\mathbf{z}\,|\,\mathbf{x})$ is intractable. Thus we use a $q(\mathbf{z})$ to approximate the posterior as follows:

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z})} \mathrm{KL}\big[q(\mathbf{z}) \,\|\, p(\mathbf{z}\,|\,\mathbf{x})\big], \tag{2}$$

where

$$\mathrm{KL}\big[q(\mathbf{z}) \,\|\, p(\mathbf{z})\big] \triangleq \mathbb{E}_{\mathbf{z}}\left[\log \frac{q(\mathbf{z})}{p(\mathbf{z})}\right] = \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} \tag{3}$$

denotes the KL divergence between two probabilistic density function (PDF). Thus, Eq(2) can be re-written as follows:

$$\begin{aligned}\mathrm{KL}\left[q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x})\right] &= \mathbb{E}_{\mathbf{z}}\left[\log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}\right] = \mathbb{E}_{\mathbf{z}}\left[\log \frac{q(\mathbf{z})\, p(\mathbf{x})}{p(\mathbf{x},\mathbf{z})}\right] \\ &= \mathbb{E}_{\mathbf{z}}\left[\log q(\mathbf{z})\right] - \mathbb{E}_{\mathbf{z}}\left[\log p(\mathbf{x},\mathbf{z})\right] + \log p(\mathbf{x}),\end{aligned} \tag{4}$$

The readers should keep in mind that $\mathbf{z} \sim q(\mathbf{z})$. The ELBO is defined as

$$\mathrm{ELBO}\left[q(\mathbf{z})\right] \triangleq E_{\mathbf{z}}\left[\log p(\mathbf{x},\mathbf{z})\right] - E_{\mathbf{z}}\left[\log q(\mathbf{z})\right]. \tag{5}$$

Thus the log evidence is bounded by ELBO:

$$\log p(\mathbf{x}) = \mathrm{KL}\left[q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x})\right] + \mathrm{ELBO}\left[q(\mathbf{z})\right]. \tag{6}$$

The log evidence $\log p(\mathbf{x})$ is only related to the data, and $\mathrm{KL}\left[q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x})\right] \geq 0$.

To minimize the KL divergence is equivalent to maximize the ELBO as follows:

$$\begin{aligned}q^*(\mathbf{z}) &= \arg\min_{q(\mathbf{z})} \mathrm{KL}\left[q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x})\right] \\ &= \arg\min_{q(\mathbf{z})} \left\{\log p(\mathbf{x}) - \mathrm{ELBO}\left[q(\mathbf{z})\right]\right\} \\ &= \arg\max_{q(\mathbf{z})} \mathrm{ELBO}\left[q(\mathbf{z})\right].\end{aligned} \tag{7}$$

Notice that ELBO is only related to the model $p(\mathbf{x},\mathbf{z})$ and a variational family $q(\mathbf{z})$.

Mean Field Approximation

Assuming the latent variable is composed of a set disjoint group $\mathbf{z} = \{z_i\}_{i=1}^3$, $z_i \in \mathbb{R}^3$, by utilizing the mean field theory we have

$$q(\mathbf{z}) = \prod_{i=1}^{3} q(z_i). \tag{8}$$

The maximization of ELBO can be accordingly is optimized as follows:

$$\begin{aligned}q^{t+1}(z_1) &= \max_{q(z_1)} \mathrm{ELBO}\left[q(z_1) q^t(z_2) q^t(z_3)\right], \\ q^{t+1}(z_2) &= \max_{q(z_1)} \mathrm{ELBO}\left[q^{t+1}(z_1) q(z_2) q^t(z_3)\right], \\ q^{t+1}(z_3) &= \max_{q(z_1)} \mathrm{ELBO}\left[q^{t+1}(z_1) q^{t+1}(z_2) q^t(z_3)\right].\end{aligned} \tag{9}$$

Given $q^t(z_i), i = 1,2,3$, the posterior $q^{t+1}(z_i), i = 1,2,3$ can be alternately updated.

Now we compute the ELBO:

$$\mathrm{ELBO}\left[q(\mathbf{z})\right] = \mathbb{E}_{z_1 z_2 z_3}\left[\log p(\mathbf{x},\mathbf{z})\right] - \mathbb{E}_{z_1 z_2 z_3}\left[\log\left\{q(z_1) q(z_2) q(z_3)\right\}\right]. \tag{10}$$

The first term is:

$$\mathbb{E}_{\mathbf{z}}\left[\log p(\mathbf{x},\mathbf{z})\right]=\iiint q(z_1)q(z_2)q(z_3)\log p(\mathbf{x},\mathbf{z})\,dz_1dz_2dz_3, \tag{11}$$

and the second term:

$$\begin{aligned}\mathbb{E}_{\mathbf{z}}\left[\log q(\mathbf{z})\right] &= \iiint q(z_1)q(z_2)q(z_3)\log\left\{q(z_1)q(z_2)q(z_3)\right\}dz_1dz_2dz_3 \\ &= \sum_{i=1}^{3}\int q(z_i)\log q(z_i)dz_i.\end{aligned} \tag{12}$$

If $q(z_2), q(z_3)$ is fixed, the ELBO is only related to $q(z_1)$. We compute the first term of ELBO as follows:

$$\begin{aligned}&\mathbb{E}_{z_1}\left[\log p(\mathbf{x},\mathbf{z})\right]=\iiint q(z_1)q(z_2)q(z_3)\log p(\mathbf{x},\mathbf{z})\,dz_1dz_2dz_3 \\ &=\int q(z_1)\left[\iint q(z_2)q(z_3)\log p(\mathbf{x},\mathbf{z})\,dz_2dz_3\right]dz_1 \\ &=\mathbb{E}_{z_1}\left[\mathbb{E}_{z_2z_3}\left[\log p(\mathbf{x},\mathbf{z})\right]\right].\end{aligned} \tag{13}$$

Notice the term $\mathbb{E}_{z_2z_3}\left[\log p(\mathbf{x},\mathbf{z})\right]$ only contains z_1 because z_2, z_3 have already eliminated by integral operator. The second term of ELBO is as follows:

$$\mathbb{E}_{z_1}\left[\log q(\mathbf{z})\right]=\int \log q(z_1)dz_1 + cst = \mathbb{E}_{z_1}\left[\log q(z_1)\right]. \tag{14}$$

Thus maximization of ELBO with respect to $q(z_1)$ is as follows

$$\begin{aligned}q^*(z_1) &= \max_{q(z_1)}\left\{\mathbb{E}_{z_1}\left[\mathbb{E}_{z_2z_3}\left[\log p(\mathbf{x},\mathbf{z})\right]\right]-\mathbb{E}_{z_1}\left[\log q(z_1)\right]\right\} \\ &= \max_{q(z_1)}\mathbb{E}_{z_1}\left[\log\frac{\exp\left\{\mathbb{E}_{z_2z_3}\left[\log p(\mathbf{x},\mathbf{z})\right]\right\}}{q(z_1)}\right] \\ &= \min_{q(z_1)}\mathrm{KL}\left[q(z_1)\,\|\,\exp\left\{\mathbb{E}_{z_2z_3}\left[\log p(\mathbf{x},\mathbf{z})\right]\right\}\right].\end{aligned} \tag{15}$$

Since KL divergence is always non-negative, the optimal posterior $q^*(z_1)$ is obtained when KL divergence equals to zero:

$$\begin{aligned}&q^*(z_1)\propto\exp\left\{\mathbb{E}_{z_2z_3}\left[\log p(\mathbf{x},\mathbf{z})\right]\right\}, \\ &\log q^*(z_1)=\mathbb{E}_{z_2z_3}\left[\log p(\mathbf{x},\mathbf{z})\right]+cst.\end{aligned} \tag{16}$$

Variational Mixture of Gaussians

Each observation $\mathbf{x}_n\in\mathbb{R}^D$ is associated with a latent variable $\mathbf{z}_n$, which is a 1-of-K binary vector with individual element $z_{nk}=\{0,1\}$. In some papers the $\mathbf{z}_n$ is said to be

a categorical distribution. To summary, the observed dataset denotes $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, and the latent variables denoted by $\mathbf{Z}=\{\mathbf{z}_1,\ldots,\mathbf{z}_N\}$.

For GMM model, the parameter set of mean vectors, precision matrices and weights $\boldsymbol{\mu}\triangleq\{\boldsymbol{\mu}_k\},\boldsymbol{\Lambda}\triangleq\{\boldsymbol{\Lambda}_k\},\boldsymbol{\pi}\triangleq\{\pi_k\}$ is what we want. From the Bayesian point of view, these are random variables instead of deterministic parameters, thus their priors should be specified.

For weight vector, the Dirichlet distribution is specified:

$$p(\boldsymbol{\pi})=\mathrm{Dir}(\boldsymbol{\pi}\mid\alpha_0)=C(\alpha_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1}. \tag{17}$$

For mean vector and precision matrices, the Gaussian-Wishart distribution is specified:

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda})=p(\boldsymbol{\mu}\mid\boldsymbol{\Lambda})\,p(\boldsymbol{\Lambda})=\prod_{k=1}^{K}\mathcal{N}\left(\boldsymbol{\mu}_k\mid\mathbf{m}_0,(\beta_0\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k\mid\mathbf{W}_0,\nu_0) \tag{18}$$

Notice Gaussian-Gamma distribution is 1D Gaussian-Wishart distribution

$$p(\mu,\lambda)=\mathcal{N}\left(\mu\mid\mu_0,(\beta\lambda)^{-1}\right)\mathrm{Gam}(\lambda\mid a,b). \tag{19}$$

Based on the graph model in PRML, the joint distribution over all random variables is:

$$p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})=p(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})\,p(\mathbf{Z}\mid\boldsymbol{\pi})\,p(\boldsymbol{\pi})\,p(\boldsymbol{\mu}\mid\boldsymbol{\Lambda})\,p(\boldsymbol{\Lambda}). \tag{20}$$

Clarify each conditional distribution and respective logarithm:

$$\begin{aligned}
&p(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})=\prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}\left(\mathbf{x}_n\mid\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}\\
&\log p(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})=-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\left\{(\mathbf{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)-\log|\boldsymbol{\Lambda}_k|\right\}\\
&p(\mathbf{Z}\mid\boldsymbol{\pi})=\prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}\Leftrightarrow\log p(\mathbf{Z}\mid\boldsymbol{\pi})=\sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\log\pi_k\\
&p(\boldsymbol{\pi})=\prod_{k=1}^{K}\pi_k^{\alpha_0-1}\Leftrightarrow\log p(\boldsymbol{\pi})=\sum_{k=1}^{K}(\alpha_0-1)\log\pi_k\\
&p(\boldsymbol{\mu}\mid\boldsymbol{\Lambda})=\prod_{k=1}^{K}\mathcal{N}\left(\boldsymbol{\mu}_k\mid\mathbf{m}_0,(\beta_0\boldsymbol{\Lambda}_k)^{-1}\right)\\
&\log p(\boldsymbol{\mu}\mid\boldsymbol{\Lambda})=-\frac{1}{2}\sum_{k=1}^{K}\beta_0(\boldsymbol{\mu}_k-\mathbf{m}_0)^T\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k-\mathbf{m}_0)-\log|\boldsymbol{\Lambda}_k|\\
&p(\boldsymbol{\Lambda})=\prod_{k=1}^{K}\mathcal{W}(\boldsymbol{\Lambda}_k\mid\mathbf{W}_0,\nu_0)\Leftrightarrow\log p(\boldsymbol{\Lambda})=\sum_{k=1}^{K}\frac{\nu_0-K-1}{2}\log|\boldsymbol{\Lambda}|-\frac{1}{2}\mathrm{Tr}\left(\mathbf{W}_0^{-1}\boldsymbol{\Lambda}\right).
\end{aligned} \tag{21}$$

Notice that we drop some constant values and only retain sufficient statistics.

Assignment Factor

The posterior of indicator variable $\mathbf{Z}$ is computed as:

$$\begin{aligned}\log q^*(\mathbf{Z}) &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})\right] \\ &= \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log p(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})\right]+\mathbb{E}_{\boldsymbol{\pi}}\left[\log p(\mathbf{Z}\mid\boldsymbol{\pi})\right].\end{aligned} \tag{22}$$

The first term:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log p(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})\right] \\ &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\iint\left\{(\mathbf{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)-\log|\boldsymbol{\Lambda}_k|\right\}d\boldsymbol{\mu}_k d\boldsymbol{\Lambda}_k.\end{aligned} \tag{23}$$

We drop the constant value which is unrelated to $\mathbf{Z}$.
The second term:

$$\mathbb{E}_{\boldsymbol{\pi}}\left[\log p(\mathbf{Z}\mid\boldsymbol{\pi})\right]=\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\int\log\pi_k d\pi_k. \tag{24}$$

Thus

$$\begin{aligned}&\log q^*(\mathbf{Z})=\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log\rho_{nk} \\ &\log\rho_{nk}=-\frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)\right]+\frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}_k}\left[\log|\boldsymbol{\Lambda}_k|\right]+\mathbb{E}_{\pi_k}\left[\log\pi_k\right]\end{aligned} \tag{25}$$

Where $\log\rho_{nk}$ is irrespective with z_{nk}, and ρ_{nk} should be normalized in order to be a valid PDF:

$$\begin{aligned}r_{nk}&=\frac{\exp(\tilde{\rho}_{nk})}{\sum_{k=1}^{K}\exp(\tilde{\rho}_{nk})} \\ \tilde{\rho}_{nk}&=\rho_{nk}-\max_k\rho_{nk}\end{aligned} \tag{26}$$

Thus, the posterior of $\mathbf{Z}$ consists of a set of independent categorical distributions

$$q^*(\mathbf{Z})\propto\prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}} \tag{27}$$

Other Factors

Next we consider $\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}$:

$$\begin{aligned}&\log q^*(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) \\ &=\mathbb{E}_{\mathbf{Z}}\left[\log p(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})\right]+\mathbb{E}_{\mathbf{Z}}\left[\log p(\mathbf{Z}\mid\boldsymbol{\pi})\right]+\log p(\boldsymbol{\mu},\boldsymbol{\Lambda})+\log p(\boldsymbol{\pi})\end{aligned} \tag{28}$$

By observing the above equation, $\boldsymbol{\pi}$ and $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ are decoupled as follows:

$$q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \tag{29}$$

This indicates that we can separately estimate $\boldsymbol{\pi}$ and a pair of $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$

Weight Factor

$$\begin{aligned} &\log q^*(\boldsymbol{\pi}) = \mathbb{E}_{\mathbf{Z}}\left[\log p(\mathbf{Z} \mid \boldsymbol{\pi})\right] + \log p(\boldsymbol{\pi}) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \log \pi_k + \sum_{k=1}^{K} (\alpha_0 - 1) \log \pi_k \\ &= \sum_{k=1}^{K} (N_k + \alpha_0 - 1) \log \pi_k \end{aligned} \tag{30}$$

where $N_k \triangleq \sum_{n=1}^{N} r_{nk}$ denotes the effective number of k^{th} component of GMM. Since the prior $p(\boldsymbol{\pi})$ is Dirichlet distribution, its variational factor has the same distribution as follows

$$\begin{aligned} q^*(\boldsymbol{\pi}) &= Dir(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \\ \alpha_k &= \alpha_0 + N_k \end{aligned} \tag{31}$$

Mean and Precision Factors

This part is sort of difficult owing to the coupling between mean vector and precision matrix. The prior of mean vector and precision matrix are as follows:

$$\begin{aligned} &\log p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \\ &= -\frac{1}{2}\beta_0 (\boldsymbol{\mu}_k - \mathbf{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_0) - \log|\boldsymbol{\Lambda}_k| + \frac{\nu_0 - K - 1}{2} \log|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left(\mathbf{W}_0^{-1} \boldsymbol{\Lambda}_k\right) \\ &= -\frac{1}{2}\beta_0 (\boldsymbol{\mu}_k - \mathbf{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_0) + \frac{\nu_0 - K - 3}{2} \log|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left(\mathbf{W}_0^{-1} \boldsymbol{\Lambda}_k\right) \\ &= -\frac{\beta_0}{2}\left(\boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k - 2\boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \mathbf{m}_0\right) + \frac{\nu_0 - K - 3}{2} \log|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left(\mathbf{W}_0^{-1} \boldsymbol{\Lambda}_k\right) - \frac{\beta_0}{2}\mathbf{m}_0^T \boldsymbol{\Lambda}_k \mathbf{m}_0 \\ &= -\frac{\beta_0}{2}\left(\boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k - 2\boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \mathbf{m}_0\right) + \frac{\nu_0 - K - 3}{2} \log|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left\{\left(\mathbf{W}_0^{-1} + \beta_0 \mathbf{m}_0 \mathbf{m}_0^T\right) \boldsymbol{\Lambda}_k\right\} \end{aligned} \tag{32}$$

Where we use the cyclic property of trace operator:

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA}) \tag{33}$$

For the mean and precision matrices,

$$\log q^*(\boldsymbol{\mu},\boldsymbol{\Lambda}) = \mathbb{E}_{\mathbf{Z}}\left[\log p(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})\right] + \log p(\boldsymbol{\mu},\boldsymbol{\Lambda}) = \sum_{k=1}^{K} t_k \tag{34}$$

The first term:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z}}\left[\log p(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})\right] \\
&= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\{(\mathbf{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k) - \log|\boldsymbol{\Lambda}_k|\right\} \\
&= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\left\{r_{nk}\boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k - 2r_{nk}\boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k\mathbf{x}_n + r_{nk}\mathbf{x}_n^T\boldsymbol{\Lambda}_k\mathbf{x}_n - r_{nk}\log|\boldsymbol{\Lambda}_k|\right\} \\
&= \sum_{k=1}^{K}\left\{-\frac{N_k}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k + N_k\boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k\bar{\mathbf{x}}_k + \frac{1}{2}N_k\log|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left\{\left(\sum_{n=1}^{N} r_{nk}\mathbf{x}_n\mathbf{x}_n^T\right)\boldsymbol{\Lambda}_k\right\}\right\}
\end{aligned} \tag{35}$$

Where we use the cyclic property of trace operator.
Thus we combine the likelihood the prior:

$$\begin{aligned}
t_k &= -\frac{N_k}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k + N_k\boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k\bar{\mathbf{x}}_k + \frac{1}{2}N_k\log|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left\{\left(\sum_{n=1}^{N} r_{nk}\mathbf{x}_n\mathbf{x}_n^T\right)\boldsymbol{\Lambda}_k\right\} \\
&-\frac{1}{2}\beta_0\boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k + \beta_0\boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k\mathbf{m}_0 - \log|\boldsymbol{\Lambda}_k| + \frac{\nu_0-K-3}{2}\log|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left\{\left(\mathbf{W}_0^{-1}+\beta_0\mathbf{m}_0\mathbf{m}_0^T\right)\boldsymbol{\Lambda}_k\right\} \\
&= -\frac{1}{2}\left\{(N_k+\beta_0)\boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k - 2\boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k(N_k\bar{\mathbf{x}}_k+\beta_0\mathbf{m}_0)\right\} + \frac{\nu_0-K-3+N_k}{2}\log|\boldsymbol{\Lambda}_k| \\
&-\frac{1}{2}\mathrm{Tr}\left\{\left(\mathbf{C}_k+\mathbf{W}_0^{-1}+\beta_0\mathbf{m}_0\mathbf{m}_0^T\right)\boldsymbol{\Lambda}_k\right\} \\
\mathbf{C}_k &\triangleq \sum_{n=1}^{N} r_{nk}\mathbf{x}_n\mathbf{x}_n^T
\end{aligned} \tag{36}$$

Comparing with prior, we have:

$$\begin{aligned}
\beta_k &= N_k + \beta_0 \\
\mathbf{m}_k &= \frac{1}{\beta_k}(N_k\bar{\mathbf{x}}_k + \beta_0\mathbf{m}_0) \\
\mathbf{W}_k^{-1} &= \mathbf{C}_k + \mathbf{W}_0^{-1} + \beta_0\mathbf{m}_0\mathbf{m}_0^T - \beta_k\mathbf{m}_k\mathbf{m}_k^T \\
\nu_k &= \nu_0 + N_k
\end{aligned} \tag{37}$$

Notice that the updating formula of $\mathbf{W}_k$ in PRML is as follows:

$$\begin{aligned}
\mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0+N_k}(\bar{\mathbf{x}}_k-\mathbf{m}_0)(\bar{\mathbf{x}}_k-\mathbf{m}_0)^T \\
\mathbf{S}_k &\triangleq \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n-\bar{\mathbf{x}}_k)(\mathbf{x}_n-\bar{\mathbf{x}}_k)^T
\end{aligned} \tag{38}$$

We will prove that the updating formula in this manuscript is same with the above formula, namely,

$$\begin{aligned}&\sum_{n=1}^{N} r_{nk}\mathbf{x}_n\mathbf{x}_n^T + \beta_0\mathbf{m}_0\mathbf{m}_0^T - \beta_k\mathbf{m}_k\mathbf{m}_k^T \\ &= \sum_{n=1}^{N} r_{nk}\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)^T + \frac{\beta_0 N_k}{\beta_0 + N_k}\left(\bar{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\bar{\mathbf{x}}_k - \mathbf{m}_0\right)^T\end{aligned} \tag{39}$$

We firstly simplify the term $\sum_{n=1}^{N} r_{nk}\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)^T$ as follows

$$\begin{aligned}&\sum_{n=1}^{N} r_{nk}\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)^T = \sum_{n=1}^{N}\left\{r_{nk}\mathbf{x}_n\mathbf{x}_n^T + r_{nk}\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T - r_{nk}\mathbf{x}_n\bar{\mathbf{x}}_k^T\right\} \\ &= \sum_{n=1}^{N} r_{nk}\mathbf{x}_n\mathbf{x}_n^T + \bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T\sum_{n=1}^{N} r_{nk} - \bar{\mathbf{x}}_k\sum_{n=1}^{N} r_{nk}\mathbf{x}_n^T - \left(\sum_{n=1}^{n} r_{nk}\mathbf{x}_n\right)\bar{\mathbf{x}}_k^T \\ &= \sum_{n=1}^{N} r_{nk}\mathbf{x}_n\mathbf{x}_n^T + \bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T N_k - \bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T N_k - N_k\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T \\ &= \sum_{n=1}^{N} r_{nk}\mathbf{x}_n\mathbf{x}_n^T - N_k\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T\end{aligned} \tag{40}$$

Thus the following equation should be hold:

$$\begin{aligned}&\beta_0\mathbf{m}_0\mathbf{m}_0^T - \beta_k\mathbf{m}_k\mathbf{m}_k^T = -N_k\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T + \frac{\beta_0 N_k}{\beta_0 + N_k}\left(\bar{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\bar{\mathbf{x}}_k - \mathbf{m}_0\right)^T \\ &\Updownarrow \\ &\beta_0\left(\beta_0 + N_k\right)\mathbf{m}_0\mathbf{m}_0^T - \left(N_k\bar{\mathbf{x}}_k + \beta_0\mathbf{m}_0\right)^2 \\ &= -N_k\left(\beta_0 + N_k\right)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T + \beta_0 N_k\left(\bar{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\bar{\mathbf{x}}_k - \mathbf{m}_0\right)^T\end{aligned} \tag{41}$$

The left term is

$$\begin{aligned}&\beta_0\left(N_k + \beta_0\right)\mathbf{m}_0\mathbf{m}_0^T - \left(N_k\bar{\mathbf{x}}_k + \beta_0\mathbf{m}_0\right)\left(N_k\bar{\mathbf{x}}_k^T + \beta_0\mathbf{m}_0^T\right) \\ &= \beta_0 N_k\mathbf{m}_0\mathbf{m}_0^T + \beta_0^2\mathbf{m}_0\mathbf{m}_0^T - \left\{N_k^2\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T + \beta_0^2\mathbf{m}_0\mathbf{m}_0^T + \beta_0 N_k\mathbf{m}_0\bar{\mathbf{x}}_k^T + \beta_0 N_k\bar{\mathbf{x}}_k\mathbf{m}_0^T\right\} \\ &= \beta_0 N_k\mathbf{m}_0\mathbf{m}_0^T - \left\{N_k^2\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T + \beta_0 N_k\mathbf{m}_0\bar{\mathbf{x}}_k^T + N_k\beta_0\bar{\mathbf{x}}_k\mathbf{m}_0^T\right\}\end{aligned} \tag{42}$$

Thus we should prove

$$\begin{aligned}&\beta_0\mathbf{m}_0\mathbf{m}_0^T - \left\{N_k\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T + \beta_0\mathbf{m}_0\bar{\mathbf{x}}_k^T + \beta_0\bar{\mathbf{x}}_k\mathbf{m}_0^T\right\} + \left(\beta_0 + N_k\right)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T \\ &= \beta_0\left(\bar{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\bar{\mathbf{x}}_k - \mathbf{m}_0\right)^T\end{aligned} \tag{43}$$

The left term is:

$$\begin{aligned}&\beta_0\mathbf{m}_0\mathbf{m}_0^T - N_k\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T - \beta_0\mathbf{m}_0\bar{\mathbf{x}}_k^T - \beta_0\bar{\mathbf{x}}_k\mathbf{m}_0^T + \beta_0\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T + N_k\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T \\ &= \beta_0\mathbf{m}_0\mathbf{m}_0^T - \beta_0\mathbf{m}_0\bar{\mathbf{x}}_k^T - \beta_0\bar{\mathbf{x}}_k\mathbf{m}_0^T + \beta_0\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T \\ &= \beta_0\left(\mathbf{m}_0\mathbf{m}_0^T - \mathbf{m}_0\bar{\mathbf{x}}_k^T - \bar{\mathbf{x}}_k\mathbf{m}_0^T + \bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T\right) \\ &= \beta_0\left(\bar{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\bar{\mathbf{x}}_k - \mathbf{m}_0\right)^T\end{aligned} \tag{44}$$

To summary, our posterior updating formula of mean vector and precision matrix is same with PRML.

Variational Linear Regression

The model

$$
\begin{aligned}
&p(\mathbf{t}\mid\mathbf{w})=\prod_{n=1}^{N}\mathcal{N}\left(t_n\mid\mathbf{w}^T\phi_n,\beta^{-1}\right)\\
&\log p(\mathbf{t}\mid\mathbf{w})=\log\beta+\frac{1}{2}\log 2\pi-\frac{1}{2}\beta\left(t_n-\mathbf{w}^T\phi_n\right)\\
&p(\mathbf{w}\mid\alpha)=\mathcal{N}\left(\mathbf{w}\mid\mathbf{0},\alpha^{-1}\mathbf{I}\right)\\
&\log p(\mathbf{w}\mid\alpha)=\frac{M}{2}\log\alpha-\frac{M}{2}\log 2\pi-\frac{1}{2}\mathbf{w}^T\mathbf{w}\alpha\\
&p(\alpha)=\mathrm{Gam}(\alpha\mid a_0,b_0)\\
&\log p(\alpha)=a_0\log b_0+(a_0-1)\log\alpha-b_0\alpha
\end{aligned}
\tag{45}
$$

Where $\mathrm{Gam}(\tau\mid a,b)=\frac{1}{\Gamma(a)}b^a\tau^{a-1}e^{-b\tau}$.

The model is

$$
p(\mathbf{t},\mathbf{w},\alpha)=p(\mathbf{t}\mid\mathbf{w})\,p(\mathbf{w}\mid\alpha)\,p(\alpha) \tag{46}
$$

Compute Alpha:

$$
\begin{aligned}
&p(\mathbf{t}\mid\mathbf{w})\,p(\mathbf{w}\mid\alpha)\,p(\alpha)\\
&\log q^*(\alpha)\\
&=\mathbb{E}\left[\log p(\mathbf{t},\mathbf{w},\alpha)\right]\\
&=\mathbb{E}_{\mathbf{w}}\left[\log p(\mathbf{w}\mid\alpha)\right]+\log p(\alpha)\\
&=\frac{M}{2}\log\alpha-\frac{1}{2}\mathbb{E}_{\mathbf{w}}\left[\mathbf{w}^T\mathbf{w}\right]\alpha+(a_0-1)\log\alpha-b_0\alpha+const.\\
&=\left(\frac{M}{2}+a_0-1\right)\log\alpha-\left(\frac{1}{2}\mathbb{E}_{\mathbf{w}}\left[\mathbf{w}^T\mathbf{w}\right]+b_0\right)\alpha+const.
\end{aligned}
\tag{47}
$$

Compute the weight vector:

$$
\begin{aligned}
&\log p(\mathbf{t}\mid\mathbf{w})=\log\beta+\frac{1}{2}\log 2\pi-\frac{1}{2}\beta\left(t_n-\mathbf{w}^T\phi_n\right)^2\\
&p(\mathbf{t}\mid\mathbf{w})\,p(\mathbf{w}\mid\alpha)\,p(\alpha)\\
&\log q^*(\mathbf{w})\\
&=\mathbb{E}_{\alpha}\left[\log p(\mathbf{w}\mid\alpha)\right]+p(\mathbf{t}\mid\mathbf{w})\\
&=-\frac{1}{2}E_{\alpha}(\alpha)\mathbf{w}^T\mathbf{w}-\frac{1}{2}\beta\sum_{n=1}^{N}\left(t_n-\mathbf{w}^T\phi_n\right)^2+const.\\
&=-\frac{1}{2}\left\{E_{\alpha}(\alpha)\mathbf{w}^T\mathbf{w}+\beta\mathbf{w}^T\left(\sum_{n=1}^{N}\phi_n\phi_n^T\right)\mathbf{w}+2\beta\mathbf{w}^T\sum_{n=1}^{N}\phi_n\right\}+const.
\end{aligned}
\tag{48}
$$

Variational Mixture of 1D GMM without Mean Vector

Each observation $x_n \in \mathbb{R}$ is associated with a latent variable $\mathbf{z}_n$, which is a 1-of-K binary vector with individual element z_{nk}. To summary, the observed dataset denotes $\mathbf{X} = \{x_1, \ldots, x_N\}$, and the latent variables denoted by $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$.

We will see that variational inference without the mean vector is much easier than aforementioned GMM. Only the parameter set of precision and weight $\boldsymbol{\tau} \triangleq \{\tau_k\}, \boldsymbol{\pi} \triangleq \{\pi_k\}$ is what we want. For weight parameter, the Dirichlet distribution is specified:

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \alpha_0) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} \tag{49}$$

For precision, the Gamma distribution is specified:

$$p(\boldsymbol{\tau}) = \prod_{k=1}^{K} \mathrm{Gam}(\tau_k \mid a_k, b_k) \tag{50}$$

Therefore, the model is

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}) = p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\tau})\, p(\mathbf{Z} \mid \boldsymbol{\pi})\, p(\boldsymbol{\pi})\, p(\boldsymbol{\tau}) \tag{51}$$

Clarify each term:

$$\begin{aligned}
&p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\tau}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}\left(x_n \mid 0, \tau_k^{-1}\right)^{z_{nk}} \Leftrightarrow \log p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\tau}) = -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left(\tau_k x_n^2 - \log \tau_k\right) \\
&p(\mathbf{Z} \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \Leftrightarrow \log p(\mathbf{Z} \mid \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log \pi_k \\
&p(\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} \Leftrightarrow \log p(\boldsymbol{\pi}) = \sum_{k=1}^{K} (\alpha_0 - 1) \log \pi_k \\
&p(\boldsymbol{\tau}) = \prod_{k=1}^{K} \mathrm{Gam}(\tau_k \mid a_k, b_k) \Leftrightarrow \log p(\boldsymbol{\tau}) = \sum_{k=1}^{K} (a_k - 1) \log \tau_k - b_k \tau_k
\end{aligned} \tag{52}$$

Assignment Factor

The indicator variable $\mathbf{Z}$ is computed as:

$$\begin{aligned}
\log q^*(\mathbf{Z}) &= E_{\boldsymbol{\pi}, \boldsymbol{\tau}} \left[\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\tau})\right] \\
&= \mathbb{E}_{\boldsymbol{\tau}} \left[\log p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\tau})\right] + \mathbb{E}_{\boldsymbol{\pi}} \left[\log p(\mathbf{Z} \mid \boldsymbol{\pi})\right]
\end{aligned} \tag{53}$$

The first term:

$$\mathbb{E}_{\boldsymbol{\tau}} \left[\log p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\tau})\right] = -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \int \left(x_n^2 \tau_k + \log 2\pi - \log \tau_k\right) d\tau_k \tag{54}$$

The second term:

$$\mathbb{E}_{\boldsymbol{\pi}}\left[\log p(\mathbf{Z}\mid\boldsymbol{\pi})\right]=\sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\int\log\pi_k d\pi_k \tag{55}$$

Thus

$$\log q^*(\mathbf{Z})=\sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\log\rho_{nk}, \tag{56}$$

where

$$\log\rho_{nk}=-\frac{1}{2}x_n^2\mathbb{E}(\tau_k)+\frac{1}{2}\mathbb{E}(\log\tau_k)-\frac{1}{2}\log 2\pi+\mathbb{E}(\log\pi_k) \tag{57}$$

Similarly,

$$\begin{aligned} r_{nk}&=\frac{\exp(\tilde{\rho}_{nk})}{\sum_{k=1}^{K}\exp(\tilde{\rho}_{nk})}\\ \tilde{\rho}_{nk}&=\rho_{nk}-\max_k\rho_{nk}\\ q^*(\mathbf{Z})&=\prod_{n=1}^{N}\prod_{k=1}^{K}r_{nk}^{z_{nk}}. \end{aligned} \tag{58}$$

Other Factors

Next we consider $\boldsymbol{\pi},\boldsymbol{\tau}$:

$$\log q^*(\boldsymbol{\pi},\boldsymbol{\tau})=\mathbb{E}_{\mathbf{Z}}\left[\log p(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\tau})\right]+\mathbb{E}_{\mathbf{Z}}\left[\log p(\mathbf{Z}\mid\boldsymbol{\pi})\right]+\log p(\boldsymbol{\tau})+\log p(\boldsymbol{\pi}) \tag{59}$$

By observing the above equation,

$$q(\boldsymbol{\pi},\boldsymbol{\tau})=q(\boldsymbol{\pi})\prod_{k=1}^{K}q(\tau_k) \tag{60}$$

Weight Factor

$$\begin{aligned} &\log q^*(\boldsymbol{\pi})\\ &=\mathbb{E}_{\mathbf{Z}}\left[\log p(\mathbf{Z}\mid\boldsymbol{\pi})\right]+\log p(\boldsymbol{\pi})\\ &=\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\log\pi_k+\sum_{k=1}^{K}(\alpha_0-1)\log\pi_k\\ &=\sum_{k=1}^{K}(N_k+\alpha_0-1)\log\pi_k \end{aligned} \tag{61}$$

where $N_k \triangleq \sum_{n=1}^{N} r_{nk}$ denotes the effective number. Since the prior $p(\boldsymbol{\pi})$ is Dirichlet distribution, its variational factor has the same distribution.

$$
\begin{aligned}
& q^*(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \\
& \log q^*(\boldsymbol{\pi}) = \sum_{k=1}^{K} (\alpha_k - 1) \log \pi_k \\
& \Rightarrow \alpha_k = \alpha_0 + N_k
\end{aligned}
\tag{62}
$$

Precision Factors

The prior is:

$$
\log p(\tau_k) = (a_k - 1) \log \tau_k - b_k \tau_k \tag{63}
$$

For the precision matrix,

$$
\begin{aligned}
& \log q^*(\boldsymbol{\tau}) = \mathbb{E}_{\mathbf{Z}} \left[\log p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\tau}) \right] + \log p(\boldsymbol{\tau}) \\
& = -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left(\tau_k x_n^2 + \log 2\pi - \log \tau_k \right) + \sum_{k=1}^{K} (a_0 - 1) \log \tau_k - b_0 \tau_k \\
& = \sum_{k=1}^{K} \left\{ -\left(\frac{1}{2} \sum_{n=1}^{N} r_{nk} x_n^2 \right) \tau_k + \frac{1}{2} N_k \log \tau_k + (a_0 - 1) \log \tau_k - b_0 \tau_k \right\} \\
& = \left(\frac{1}{2} N_k + a_0 - 1 \right) \log \tau_k - \left(\frac{1}{2} c_k + b_0 \right) \tau_k + cst. \\
& \Rightarrow \begin{cases} a_k = a_0 + \frac{1}{2} N_k \\ b_k = b_0 + \frac{1}{2} c_k \end{cases} \\
& c_k \triangleq \sum_{n=1}^{N} r_{nk} x_n^2
\end{aligned}
\tag{64}
$$