

# 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.

Tips: vector- and matrix-based operations (integral and derivative, expectation over random variables) can be tedious and formidable. You can convert the operations into the respective scalar-base operations such as:

$$\begin{aligned} \mathbf{z} &\sim p(\mathbf{z} | \mathbf{x}) \\ \mathbb{E}_{\mathbf{z}} [\mathbf{z}\mathbf{z}^T] &= \int \int \begin{bmatrix} z_1 z_1 & z_1 z_2 \\ z_1 z_2 & z_2 z_2 \end{bmatrix} p(z_1, z_2 | \mathbf{x}) dz_1 dz_2. \end{aligned} \quad (1)$$

## Evident Lower Bound

Our purpose is to compute the posterior

$$p(\mathbf{z} | \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})}, \quad (2)$$

$\mathbf{x}$  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:

$$q^*(\mathbf{z}) = \arg \min_{q(\mathbf{z})} \text{KL}[q(\mathbf{z}) \| p(\mathbf{z} | \mathbf{x})], \quad (3)$$

where

$$\text{KL}[q(\mathbf{z}) \parallel p(\mathbf{z})] \triangleq \mathbb{E}_{\mathbf{z}} \left[ \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \right] = \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} \quad (4)$$

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

$$\begin{aligned} \text{KL}[q(\mathbf{z}) \parallel p(\mathbf{z} | \mathbf{x})] &= \mathbb{E}_{\mathbf{z}} \left[ \log \frac{q(\mathbf{z})}{p(\mathbf{z} | \mathbf{x})} \right] = \mathbb{E}_{\mathbf{z}} \left[ \log \frac{q(\mathbf{z}) p(\mathbf{x})}{p(\mathbf{x}, \mathbf{z})} \right] \\ &= \mathbb{E}_{\mathbf{z}} [\log q(\mathbf{z})] - \mathbb{E}_{\mathbf{z}} [\log p(\mathbf{x}, \mathbf{z})] + \log p(\mathbf{x}), \end{aligned} \quad (5)$$

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

$$\text{ELBO}[q(\mathbf{z})] \triangleq E_{\mathbf{z}} [\log p(\mathbf{x}, \mathbf{z})] - E_{\mathbf{z}} [\log q(\mathbf{z})]. \quad (6)$$

Thus the log evidence is bounded by ELBO:

$$\log p(\mathbf{x}) = \text{KL}[q(\mathbf{z}) \parallel p(\mathbf{z} | \mathbf{x})] + \text{ELBO}[q(\mathbf{z})]. \quad (7)$$

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

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

$$\begin{aligned} q^*(\mathbf{z}) &= \arg \min_{q(\mathbf{z})} \text{KL}[q(\mathbf{z}) \parallel p(\mathbf{z} | \mathbf{x})] \\ &= \arg \min_{q(\mathbf{z})} \{ \log p(\mathbf{x}) - \text{ELBO}[q(\mathbf{z})] \} \\ &= \arg \max_{q(\mathbf{z})} \text{ELBO}[q(\mathbf{z})]. \end{aligned} \quad (8)$$

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

$$q(\mathbf{z}) = \prod_{i=1}^3 q(z_i). \quad (9)$$

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

$$\begin{aligned} q^{t+1}(z_1) &= \max_{q(z_1)} \text{ELBO}[q(z_1) q^t(z_2) q^t(z_3)], \\ q^{t+1}(z_2) &= \max_{q(z_1)} \text{ELBO}[q^{t+1}(z_1) q(z_2) q^t(z_3)], \\ q^{t+1}(z_3) &= \max_{q(z_1)} \text{ELBO}[q^{t+1}(z_1) q^{t+1}(z_2) q^t(z_3)]. \end{aligned} \quad (10)$$

Given  $q'(z_i), i=1,2,3$ , the posterior  $q^{t+1}(z_i), i=1,2,3$  can be alternately updated.

Now we compute the ELBO:

$$\text{ELBO}[q(\mathbf{z})] = \mathbb{E}_{z_1 z_2 z_3} [\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_{z_1 z_2 z_3} [\log \{q(z_1)q(z_2)q(z_3)\}]. \quad (11)$$

The first term is:

$$\mathbb{E}_{\mathbf{z}} [\log p(\mathbf{x}, \mathbf{z})] = \iiint q(z_1)q(z_2)q(z_3) \log p(\mathbf{x}, \mathbf{z}) dz_1 dz_2 dz_3, \quad (12)$$

and the second term:

$$\begin{aligned} \mathbb{E}_{\mathbf{z}} [\log q(\mathbf{z})] &= \iiint q(z_1)q(z_2)q(z_3) \log \{q(z_1)q(z_2)q(z_3)\} dz_1 dz_2 dz_3 \\ &= \sum_{i=1}^3 \int q(z_i) \log q(z_i) dz_i. \end{aligned} \quad (13)$$

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:

$$\begin{aligned} \mathbb{E}_{z_1} [\log p(\mathbf{x}, \mathbf{z})] &= \iint q(z_2)q(z_3) \log p(\mathbf{x}, \mathbf{z}) dz_2 dz_3 \\ &= \int q(z_1) \left[ \iint q(z_2)q(z_3) \log p(\mathbf{x}, \mathbf{z}) dz_2 dz_3 \right] dz_1 \\ &= \mathbb{E}_{z_1} [\mathbb{E}_{z_2 z_3} [\log p(\mathbf{x}, \mathbf{z})]]. \end{aligned} \quad (14)$$

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

$$\mathbb{E}_{z_1} [\log q(\mathbf{z})] = \int \log q(z_1) dz_1 + cst = \mathbb{E}_{z_1} [\log q(z_1)]. \quad (15)$$

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

$$\begin{aligned} q^*(z_1) &= \max_{q(z_1)} \left\{ \mathbb{E}_{z_1} [\mathbb{E}_{z_2 z_3} [\log p(\mathbf{x}, \mathbf{z})]] - \mathbb{E}_{z_1} [\log q(z_1)] \right\} \\ &= \max_{q(z_1)} \mathbb{E}_{z_1} \left[ \log \frac{\exp \{ \mathbb{E}_{z_2 z_3} [\log p(\mathbf{x}, \mathbf{z})] \}}{q(z_1)} \right] \\ &= \min_{q(z_1)} \text{KL} [q(z_1) \parallel \exp \{ \mathbb{E}_{z_2 z_3} [\log p(\mathbf{x}, \mathbf{z})] \}]. \end{aligned} \quad (16)$$

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

$$\begin{aligned} q^*(z_1) &\propto \exp \{ \mathbb{E}_{z_2 z_3} [\log p(\mathbf{x}, \mathbf{z})] \}, \\ \log q^*(z_1) &= \mathbb{E}_{z_2 z_3} [\log p(\mathbf{x}, \mathbf{z})] + cst. \end{aligned} \quad (17)$$

## 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  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , and the latent variables denoted by  $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ .

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:

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^K \pi_k^{\alpha_0 - 1}. \quad (18)$$

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

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu} | \boldsymbol{\Lambda}) p(\boldsymbol{\Lambda}) = \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0) \quad (19)$$

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

$$p(\mu, \lambda) = \mathcal{N}(\mu | \mu_0, (\beta \lambda)^{-1}) \text{Gam}(\lambda | a, b). \quad (20)$$

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

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\mathbf{Z} | \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu} | \boldsymbol{\Lambda}) p(\boldsymbol{\Lambda}). \quad (21)$$

Clarify each conditional distribution and respective logarithm:

$$\begin{aligned}
p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\mu}, \Lambda) &= \prod_{n=1}^N \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \Lambda_k^{-1})^{z_{nk}} \\
\log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\mu}, \Lambda) &= -\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K z_{nk} \left\{ (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \Lambda_k (\mathbf{x}_n - \boldsymbol{\mu}_k) - \log |\Lambda_k| \right\} \\
p(\mathbf{Z} | \boldsymbol{\pi}) &= \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \Leftrightarrow \log p(\mathbf{Z} | \boldsymbol{\pi}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \log \pi_k \\
p(\boldsymbol{\pi}) &= \prod_{k=1}^K \pi_k^{\alpha_0 - 1} \Leftrightarrow \log p(\boldsymbol{\pi}) = \sum_{k=1}^K (\alpha_0 - 1) \log \pi_k \\
p(\boldsymbol{\mu} | \Lambda) &= \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_0, (\beta_0 \Lambda_k)^{-1}) \\
\log p(\boldsymbol{\mu} | \Lambda) &= -\frac{1}{2} \sum_{k=1}^K \beta_0 (\boldsymbol{\mu}_k - \mathbf{m}_0)^T \Lambda_k (\boldsymbol{\mu}_k - \mathbf{m}_0) - \log |\Lambda_k| \\
p(\Lambda) &= \prod_{k=1}^K \mathcal{W}(\Lambda_k | \mathbf{W}_0, \nu_0) \Leftrightarrow \log p(\Lambda) = \sum_{k=1}^K \frac{\nu_0 - K - 1}{2} \log |\Lambda_k| - \frac{1}{2} \text{Tr}(\mathbf{W}_0^{-1} \Lambda_k).
\end{aligned} \tag{22}$$

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

## Assignment Factor

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

$$\begin{aligned}
\log q^*(\mathbf{Z}) &= \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda} [\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda)] \\
&= \mathbb{E}_{\boldsymbol{\mu}, \Lambda} [\log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\mu}, \Lambda)] + \mathbb{E}_{\boldsymbol{\pi}} [\log p(\mathbf{Z} | \boldsymbol{\pi})].
\end{aligned} \tag{23}$$

The first term:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\mu}, \Lambda} [\log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\mu}, \Lambda)] \\
&= -\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K z_{nk} \iint \left\{ (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \Lambda_k (\mathbf{x}_n - \boldsymbol{\mu}_k) - \log |\Lambda_k| \right\} d\boldsymbol{\mu}_k d\Lambda_k.
\end{aligned} \tag{24}$$

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

The second term:

$$\mathbb{E}_{\boldsymbol{\pi}} [\log p(\mathbf{Z} | \boldsymbol{\pi})] = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \int \log \pi_k d\pi_k. \tag{25}$$

Thus

$$\begin{aligned}
\log q^*(\mathbf{Z}) &= \sum_{n=1}^N \sum_{k=1}^K z_{nk} \rho_{nk} \\
\rho_{nk} &= -\frac{1}{2} \mathbb{E}_{\boldsymbol{\mu}_k, \Lambda_k} \left[ (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \Lambda_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \right] + \frac{1}{2} \mathbb{E}_{\Lambda_k} [\log |\Lambda_k|] + \mathbb{E}_{\pi_k} [\log \pi_k],
\end{aligned} \tag{26}$$

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

$$r_{nk} = \frac{\exp(\tilde{\rho}_{nk})}{\sum_{k=1}^K \exp(\tilde{\rho}_{nk})} \quad (27)$$

$$\tilde{\rho}_{nk} = \rho_{nk} - \max_k \rho_{nk}.$$

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

$$q^*(\mathbf{Z}) = \prod_{n=1}^N \prod_{k=1}^K r_{nk}^{z_{nk}}. \quad (28)$$

The term  $\mathbb{E}_{\mu_k, \Lambda_k} \left[ (\mathbf{x}_n - \mu_k)^T \Lambda_k (\mathbf{x}_n - \mu_k) \right]$  is difficult to be computed. Thus we need do some transformations as follows

$$\begin{aligned} & (\mathbf{x}_n - \mu_k)^T \Lambda_k (\mathbf{x}_n - \mu_k) \\ &= (\mathbf{y} - \mathbf{d})^T \Lambda_k (\mathbf{y} - \mathbf{d}) \\ & \mathbf{y} \triangleq \mu_k - \mathbf{m}_k \sim \mathcal{N}(\mathbf{0}, (\beta_k \Lambda_k)^{-1}) \\ & \mathbf{d} \triangleq \mathbf{x}_n - \mathbf{m}_k. \end{aligned} \quad (29)$$

The integral can be transformed accordingly:

$$\begin{aligned} & (\mathbf{y}, \Lambda_k) \sim \mathcal{N}(\mathbf{y} | \mathbf{0}, (\beta_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | \mathbf{W}_0, \nu_0) \\ & \mathbb{E}_{\mu_k, \Lambda_k} \left[ (\mathbf{x}_n - \mu_k)^T \Lambda_k (\mathbf{x}_n - \mu_k) \right] \\ &= \mathbb{E}_{\mathbf{y}, \Lambda_k} \left[ (\mathbf{y} - \mathbf{d})^T \Lambda_k (\mathbf{y} - \mathbf{d}) \right] \\ &= \mathbb{E} \left[ \mathbf{y}^T \Lambda_k \mathbf{y} \right] - 2\mathbf{d}^T \mathbb{E}[\Lambda_k \mathbf{y}] + \mathbf{d}^T \mathbb{E}[\Lambda_k] \mathbf{d}. \end{aligned} \quad (30)$$

The second term equals to zeros:

$$\begin{aligned} & \mathbb{E}[\Lambda_k \mathbf{y}] \\ &= \int \left\{ \int \mathcal{N}(\mathbf{y} | \mathbf{0}, (\beta_k \Lambda_k)^{-1}) \Lambda_k \mathbf{y} d\mathbf{y} \right\} \mathcal{W}(\Lambda_k | \mathbf{W}_0, \nu_0) d\Lambda_k \\ & \int \mathcal{N}(\mathbf{y} | \mathbf{0}, (\beta_k \Lambda_k)^{-1}) \Lambda_k \mathbf{y} d\mathbf{y} = \Lambda_k \mathbb{E}(\mathbf{y}) = \mathbf{0} \\ & \Rightarrow \mathbb{E}[\Lambda_k \mathbf{y}] = \mathbf{0}. \end{aligned} \quad (31)$$

The first term:

$$\begin{aligned} & \mathbb{E}[\mathbf{y}^T \Lambda_k \mathbf{y}] \\ &= \int \int \mathcal{N}(\mathbf{y} | \mathbf{0}, (\beta_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | \mathbf{W}_0, \nu_0) \mathbf{y}^T \Lambda_k \mathbf{y} d\mathbf{y} d\Lambda_k \\ &= \int \left\{ \int \mathcal{N}(\mathbf{y} | \mathbf{0}, (\beta_k \Lambda_k)^{-1}) \mathbf{y}^T \Lambda_k \mathbf{y} d\mathbf{y} \right\} \mathcal{W}(\Lambda_k | \mathbf{W}_0, \nu_0) d\Lambda_k. \end{aligned} \quad (32)$$

To compute the inner integral, we need a linear transformation based on Cholesky decomposition of precision matrix.

$$\mathbf{\Lambda}_k = \mathbf{L}\mathbf{L}^T \Rightarrow |\mathbf{L}| = |\mathbf{L}^T| = \sqrt{|\mathbf{\Lambda}_k|}, \quad (33)$$

Where  $\mathbf{L}$  denotes a lower triangular matrix.

$$\begin{aligned} \mathbf{u} &= \mathbf{L}^T \mathbf{y} \Leftrightarrow \mathbf{u}^T \mathbf{u} = \mathbf{y}^T \mathbf{L}\mathbf{L}^T \mathbf{y} = \mathbf{y}^T \mathbf{\Lambda}_k \mathbf{y} \Leftrightarrow \mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\beta_k} \mathbf{I}\right) \\ \mathbf{u} &\sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\beta_k} \mathbf{I}\right) \\ \Rightarrow \text{cov}[u_i] &= \sqrt{\frac{\beta_k}{2\pi}} \int \exp\left(-\frac{\beta_k}{2} u_i^2\right) u_i^2 du_i = \frac{1}{\beta_k}, i=1, 2, \dots, D. \end{aligned} \quad (34)$$

Thus the inner integral is computed as follows:

$$\begin{aligned} &\int \mathcal{N}(\mathbf{y} | \mathbf{0}, (\beta_k \mathbf{\Lambda}_k)^{-1}) \mathbf{y}^T \mathbf{\Lambda}_k \mathbf{y} d\mathbf{y} \\ &= \sqrt{\frac{|\beta_k \mathbf{\Lambda}_k|}{(2\pi)^D}} \int \exp\left(-\frac{\beta_k}{2} \mathbf{y}^T \mathbf{\Lambda}_k \mathbf{y}\right) \mathbf{y}^T \mathbf{\Lambda}_k \mathbf{y} d\mathbf{y} \\ &= \sqrt{\frac{|\beta_k \mathbf{\Lambda}_k|}{(2\pi)^D}} \int \exp\left(-\frac{\beta_k}{2} \mathbf{u}^T \mathbf{u}\right) \mathbf{u}^T \mathbf{u} \frac{1}{|\mathbf{L}^T|} d\mathbf{u} \\ &= \sqrt{\frac{\beta_k^D}{(2\pi)^D}} \int \exp\left(-\frac{\beta_k}{2} \mathbf{u}^T \mathbf{u}\right) \mathbf{u}^T \mathbf{u} d\mathbf{u} \\ &= \sum_{i=1}^D \left\{ \sqrt{\frac{\beta_k}{2\pi}} \int u_i^2 \exp\left(-\frac{\beta_k}{2} u_i^2\right) du_i \right\} \\ &= \sum_{i=1}^D \text{cov}[u_i] = \frac{D}{\beta_k} \Rightarrow \mathbb{E}[\mathbf{y}^T \mathbf{\Lambda}_k \mathbf{y}] = \frac{D}{\beta_k}. \end{aligned} \quad (35)$$

Thus the whole integral is as follows

$$\begin{aligned} &\mathbb{E}_{\mathbf{\mu}_k, \mathbf{\Lambda}_k} \left[ (\mathbf{x}_n - \mathbf{\mu}_k)^T \mathbf{\Lambda}_k (\mathbf{x}_n - \mathbf{\mu}_k) \right] \\ &= \mathbb{E}[\mathbf{y}^T \mathbf{\Lambda}_k \mathbf{y}] - 2\mathbf{d}^T \mathbb{E}[\mathbf{\Lambda}_k \mathbf{y}] + \mathbf{d}^T \mathbb{E}[\mathbf{\Lambda}_k] \mathbf{d} \\ &= \frac{D}{\beta_k} + v_k (\mathbf{x}_n - \mathbf{m}_k)^T \mathbf{W}_k (\mathbf{x}_n - \mathbf{m}_k)^T. \end{aligned} \quad (36)$$

Which is same with Eq(10.64) in PRML.

## Other Factors

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

$$\begin{aligned} &\log q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ &= \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{Z} | \boldsymbol{\pi})] + \log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) + \log p(\boldsymbol{\pi}). \end{aligned} \quad (37)$$

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

$$q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) \prod_{k=1}^K q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k). \quad (38)$$

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

## Weight Factor

$$\begin{aligned} \log q^*(\boldsymbol{\pi}) &= \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{Z} | \boldsymbol{\pi})] + \log p(\boldsymbol{\pi}) \\ &= \sum_{n=1}^N \sum_{k=1}^K r_{nk} \log \pi_k + \sum_{k=1}^K (\alpha_0 - 1) \log \pi_k \\ &= \sum_{k=1}^K (N_k + \alpha_0 - 1) \log \pi_k, \end{aligned} \quad (39)$$

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

$$\begin{aligned} q^*(\boldsymbol{\pi}) &= \text{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha}) \\ \alpha_k &= \alpha_0 + N_k. \end{aligned} \quad (40)$$

## 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:

$$\begin{aligned} \log p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) &= -\frac{1}{2} \beta_0 (\boldsymbol{\mu}_k - \mathbf{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_0) - \log |\boldsymbol{\Lambda}_k| + \frac{v_0 - K - 1}{2} \log |\boldsymbol{\Lambda}_k| - \frac{1}{2} \text{Tr}(\mathbf{W}_0^{-1} \boldsymbol{\Lambda}_k) \\ &= -\frac{1}{2} \beta_0 (\boldsymbol{\mu}_k - \mathbf{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_0) + \frac{v_0 - K - 3}{2} \log |\boldsymbol{\Lambda}_k| - \frac{1}{2} \text{Tr}(\mathbf{W}_0^{-1} \boldsymbol{\Lambda}_k) \\ &= -\frac{\beta_0}{2} (\boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k - 2 \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \mathbf{m}_0) + \frac{v_0 - K - 3}{2} \log |\boldsymbol{\Lambda}_k| - \frac{1}{2} \text{Tr}(\mathbf{W}_0^{-1} \boldsymbol{\Lambda}_k) - \frac{\beta_0}{2} \mathbf{m}_0^T \boldsymbol{\Lambda}_k \mathbf{m}_0 \\ &= -\frac{\beta_0}{2} (\boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k - 2 \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \mathbf{m}_0) + \frac{v_0 - K - 3}{2} \log |\boldsymbol{\Lambda}_k| - \frac{1}{2} \text{Tr} \left\{ (\mathbf{W}_0^{-1} + \beta_0 \mathbf{m}_0 \mathbf{m}_0^T) \boldsymbol{\Lambda}_k \right\}, \end{aligned} \quad (41)$$

Where we use the cyclic property of trace operator:

$$\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA}). \quad (42)$$

For the mean and precision matrices,



$$\log q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_{k=1}^K t_k. \quad (43)$$

The first term:

$$\begin{aligned} & \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \\ &= -\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K r_{nk} \left\{ (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) - \log |\boldsymbol{\Lambda}_k| \right\} \\ &= -\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \left\{ r_{nk} \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k - 2r_{nk} \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \mathbf{x}_n + r_{nk} \mathbf{x}_n^T \boldsymbol{\Lambda}_k \mathbf{x}_n - r_{nk} \log |\boldsymbol{\Lambda}_k| \right\} \\ &= \sum_{k=1}^K \left\{ -\frac{N_k}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k + N_k \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \bar{\mathbf{x}}_k + \frac{1}{2} N_k \log |\boldsymbol{\Lambda}_k| - \frac{1}{2} \text{Tr} \left\{ \left( \sum_{n=1}^N r_{nk} \mathbf{x}_n \mathbf{x}_n^T \right) \boldsymbol{\Lambda}_k \right\} \right\}, \end{aligned} \quad (44)$$

where we use the cyclic property of trace operator.

Thus we combine the likelihood the prior:

$$\begin{aligned} t_k &= -\frac{N_k}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k + N_k \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \bar{\mathbf{x}}_k + \frac{1}{2} N_k \log |\boldsymbol{\Lambda}_k| - \frac{1}{2} \text{Tr} \left\{ \left( \sum_{n=1}^N r_{nk} \mathbf{x}_n \mathbf{x}_n^T \right) \boldsymbol{\Lambda}_k \right\} \\ &\quad - \frac{1}{2} \beta_0 \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k + \beta_0 \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \mathbf{m}_0 - \log |\boldsymbol{\Lambda}_k| + \frac{v_0 - K - 3}{2} \log |\boldsymbol{\Lambda}_k| - \frac{1}{2} \text{Tr} \left\{ (\mathbf{W}_0^{-1} + \beta_0 \mathbf{m}_0 \mathbf{m}_0^T) \boldsymbol{\Lambda}_k \right\} \\ &= -\frac{1}{2} \left\{ (N_k + \beta_0) \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k - 2\boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k (N_k \bar{\mathbf{x}}_k + \beta_0 \mathbf{m}_0) \right\} + \frac{v_0 - K - 3 + N_k}{2} \log |\boldsymbol{\Lambda}_k| \\ &\quad - \frac{1}{2} \text{Tr} \left\{ (\mathbf{C}_k + \mathbf{W}_0^{-1} + \beta_0 \mathbf{m}_0 \mathbf{m}_0^T) \boldsymbol{\Lambda}_k \right\} \\ \mathbf{C}_k &\triangleq \sum_{n=1}^N r_{nk} \mathbf{x}_n \mathbf{x}_n^T. \end{aligned} \quad (45)$$

Comparing with prior, we have:

$$\begin{aligned} \beta_k &= N_k + \beta_0 \\ \mathbf{m}_k &= \frac{1}{\beta_k} (N_k \bar{\mathbf{x}}_k + \beta_0 \mathbf{m}_0) \\ \mathbf{W}_k^{-1} &= \mathbf{C}_k + \mathbf{W}_0^{-1} + \beta_0 \mathbf{m}_0 \mathbf{m}_0^T - \beta_k \mathbf{m}_k \mathbf{m}_k^T \\ v_k &= v_0 + N_k. \end{aligned} \quad (46)$$

Notice that the updating formula of  $\mathbf{W}_k$  or Eq(10.62) in PRML is as follows:

$$\begin{aligned} \mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + N_k \mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\mathbf{x}}_k - \mathbf{m}_0)(\bar{\mathbf{x}}_k - \mathbf{m}_0)^T \\ \mathbf{S}_k &\triangleq \frac{1}{N_k} \sum_{n=1}^N r_{nk} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^T. \end{aligned} \quad (47)$$

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

$$\begin{aligned}
& \sum_{n=1}^N r_{nk} \mathbf{x}_n \mathbf{x}_n^T + \beta_0 \mathbf{m}_0 \mathbf{m}_0^T - \beta_k \mathbf{m}_k \mathbf{m}_k^T \\
&= \sum_{n=1}^N r_{nk} (\mathbf{x}_n - \bar{\mathbf{x}}_k) (\mathbf{x}_n - \bar{\mathbf{x}}_k)^T + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\mathbf{x}}_k - \mathbf{m}_0) (\bar{\mathbf{x}}_k - \mathbf{m}_0)^T.
\end{aligned} \tag{48}$$

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

$$\begin{aligned}
& \sum_{n=1}^N r_{nk} (\mathbf{x}_n - \bar{\mathbf{x}}_k) (\mathbf{x}_n - \bar{\mathbf{x}}_k)^T = \sum_{n=1}^N \left\{ r_{nk} \mathbf{x}_n \mathbf{x}_n^T + r_{nk} \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T - r_{nk} \mathbf{x}_n \bar{\mathbf{x}}_k^T \right\} \\
&= \sum_{n=1}^N r_{nk} \mathbf{x}_n \mathbf{x}_n^T + \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T \sum_{n=1}^N r_{nk} - \bar{\mathbf{x}}_k \sum_{n=1}^N r_{nk} \mathbf{x}_n^T - \left( \sum_{n=1}^N r_{nk} \mathbf{x}_n \right) \bar{\mathbf{x}}_k^T \\
&= \sum_{n=1}^N r_{nk} \mathbf{x}_n \mathbf{x}_n^T + \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T N_k - \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T N_k - N_k \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T \\
&= \sum_{n=1}^N r_{nk} \mathbf{x}_n \mathbf{x}_n^T - N_k \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T.
\end{aligned} \tag{49}$$

Thus the following equation should be hold:

$$\begin{aligned}
& \beta_0 \mathbf{m}_0 \mathbf{m}_0^T - \beta_k \mathbf{m}_k \mathbf{m}_k^T = -N_k \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\mathbf{x}}_k - \mathbf{m}_0) (\bar{\mathbf{x}}_k - \mathbf{m}_0)^T \\
& \Downarrow \\
& \beta_0 (\beta_0 + N_k) \mathbf{m}_0 \mathbf{m}_0^T - (N_k \bar{\mathbf{x}}_k + \beta_0 \mathbf{m}_0)^2 \\
&= -N_k (\beta_0 + N_k) \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T + \beta_0 N_k (\bar{\mathbf{x}}_k - \mathbf{m}_0) (\bar{\mathbf{x}}_k - \mathbf{m}_0)^T.
\end{aligned} \tag{50}$$

The left term is

$$\begin{aligned}
& \beta_0 (N_k + \beta_0) \mathbf{m}_0 \mathbf{m}_0^T - (N_k \bar{\mathbf{x}}_k + \beta_0 \mathbf{m}_0) (N_k \bar{\mathbf{x}}_k^T + \beta_0 \mathbf{m}_0^T) \\
&= \beta_0 N_k \mathbf{m}_0 \mathbf{m}_0^T + \beta_0^2 \mathbf{m}_0 \mathbf{m}_0^T - \left\{ N_k^2 \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T + \beta_0^2 \mathbf{m}_0 \mathbf{m}_0^T + \beta_0 N_k \mathbf{m}_0 \bar{\mathbf{x}}_k^T + \beta_0 N_k \bar{\mathbf{x}}_k \mathbf{m}_0^T \right\} \\
&= \beta_0 N_k \mathbf{m}_0 \mathbf{m}_0^T - \left\{ N_k^2 \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T + \beta_0 N_k \mathbf{m}_0 \bar{\mathbf{x}}_k^T + N_k \beta_0 \bar{\mathbf{x}}_k \mathbf{m}_0^T \right\}.
\end{aligned} \tag{51}$$

Thus we should prove

$$\begin{aligned}
& \beta_0 \mathbf{m}_0 \mathbf{m}_0^T - \left\{ N_k \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T + \beta_0 \mathbf{m}_0 \bar{\mathbf{x}}_k^T + \beta_0 \bar{\mathbf{x}}_k \mathbf{m}_0^T \right\} + (\beta_0 + N_k) \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T \\
&= \beta_0 (\bar{\mathbf{x}}_k - \mathbf{m}_0) (\bar{\mathbf{x}}_k - \mathbf{m}_0)^T.
\end{aligned} \tag{52}$$

The left term is:

$$\begin{aligned}
& \beta_0 \mathbf{m}_0 \mathbf{m}_0^T - N_k \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T - \beta_0 \mathbf{m}_0 \bar{\mathbf{x}}_k^T - \beta_0 \bar{\mathbf{x}}_k \mathbf{m}_0^T + \beta_0 \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T + N_k \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T \\
&= \beta_0 \mathbf{m}_0 \mathbf{m}_0^T - \beta_0 \mathbf{m}_0 \bar{\mathbf{x}}_k^T - \beta_0 \bar{\mathbf{x}}_k \mathbf{m}_0^T + \beta_0 \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T \\
&= \beta_0 (\mathbf{m}_0 \mathbf{m}_0^T - \mathbf{m}_0 \bar{\mathbf{x}}_k^T - \bar{\mathbf{x}}_k \mathbf{m}_0^T + \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T) \\
&= \beta_0 (\bar{\mathbf{x}}_k - \mathbf{m}_0) (\bar{\mathbf{x}}_k - \mathbf{m}_0)^T.
\end{aligned} \tag{53}$$

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

## Variational Linear Regression

The model

$$\begin{aligned}
 p(\mathbf{t} | \mathbf{w}) &= \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \phi_n, \beta^{-1}) \\
 \log p(\mathbf{t} | \mathbf{w}) &= \log \beta + \frac{1}{2} \log 2\pi - \frac{1}{2} \beta (t_n - \mathbf{w}^T \phi_n)^2 \\
 p(\mathbf{w} | \alpha) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}) \\
 \log p(\mathbf{w} | \alpha) &= \frac{M}{2} \log \alpha - \frac{M}{2} \log 2\pi - \frac{1}{2} \mathbf{w}^T \mathbf{w} \alpha \\
 p(\alpha) &= \text{Gam}(\alpha | a_0, b_0) \\
 \log p(\alpha) &= a_0 \log b_0 + (a_0 - 1) \log \alpha - b_0 \alpha
 \end{aligned} \tag{54}$$

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

The model is

$$p(\mathbf{t}, \mathbf{w}, \alpha) = p(\mathbf{t} | \mathbf{w}) p(\mathbf{w} | \alpha) p(\alpha) \tag{55}$$

Compute Alpha:

$$\begin{aligned}
 &p(\mathbf{t} | \mathbf{w}) p(\mathbf{w} | \alpha) p(\alpha) \\
 &\log q^*(\alpha) \\
 &= \mathbb{E}[\log p(\mathbf{t}, \mathbf{w}, \alpha)] \\
 &= \mathbb{E}_{\mathbf{w}}[\log p(\mathbf{w} | \alpha)] + \log p(\alpha) \\
 &= \frac{M}{2} \log \alpha - \frac{1}{2} \mathbb{E}_{\mathbf{w}}[\mathbf{w}^T \mathbf{w}] \alpha + (a_0 - 1) \log \alpha - b_0 \alpha + \text{const.} \\
 &= \left( \frac{M}{2} + a_0 - 1 \right) \log \alpha - \left( \frac{1}{2} \mathbb{E}_{\mathbf{w}}[\mathbf{w}^T \mathbf{w}] + b_0 \right) \alpha + \text{const.}
 \end{aligned} \tag{56}$$

Compute the weight vector:

$$\begin{aligned}
 &\log p(\mathbf{t} | \mathbf{w}) = \log \beta + \frac{1}{2} \log 2\pi - \frac{1}{2} \beta (t_n - \mathbf{w}^T \phi_n)^2 \\
 &p(\mathbf{t} | \mathbf{w}) p(\mathbf{w} | \alpha) p(\alpha) \\
 &\log q^*(\mathbf{w}) \\
 &= \mathbb{E}_{\alpha}[\log p(\mathbf{w} | \alpha)] + \log p(\mathbf{t} | \mathbf{w}) \\
 &= -\frac{1}{2} E_{\alpha}(\alpha) \mathbf{w}^T \mathbf{w} - \frac{1}{2} \beta \sum_{n=1}^N (t_n - \mathbf{w}^T \phi_n)^2 + \text{const.} \\
 &= -\frac{1}{2} \left\{ E_{\alpha}(\alpha) \mathbf{w}^T \mathbf{w} + \beta \mathbf{w}^T \left( \sum_{n=1}^N \phi_n \phi_n^T \right) \mathbf{w} + 2\beta \mathbf{w}^T \sum_{n=1}^N \phi_n t_n \right\} + \text{const.}
 \end{aligned} \tag{57}$$

## 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  $\mathbf{X} = \{x_1, \dots, x_N\}$ , and the latent variables denoted by  $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ .

**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:

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^K \pi_k^{\alpha_0 - 1} \quad (58)$$

For precision, the Gamma distribution is specified:

$$p(\boldsymbol{\tau}) = \prod_{k=1}^K \text{Gam}(\tau_k | a_k, b_k) \quad (59)$$

Therefore, the model is

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}) = p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\tau}) p(\mathbf{Z} | \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\tau}) \quad (60)$$

Clarify each term:

$$\begin{aligned} p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\tau}) &= \prod_{n=1}^N \prod_{k=1}^K \mathcal{N}(x_n | 0, \tau_k^{-1})^{z_{nk}} \Leftrightarrow \log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\tau}) = -\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K z_{nk} (\tau_k x_n^2 - \log \tau_k) \\ p(\mathbf{Z} | \boldsymbol{\pi}) &= \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \Leftrightarrow \log p(\mathbf{Z} | \boldsymbol{\pi}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \log \pi_k \\ p(\boldsymbol{\pi}) &= \prod_{k=1}^K \pi_k^{\alpha_0 - 1} \Leftrightarrow \log p(\boldsymbol{\pi}) = \sum_{k=1}^K (\alpha_0 - 1) \log \pi_k \\ p(\boldsymbol{\tau}) &= \prod_{k=1}^K \text{Gam}(\tau_k | a_k, b_k) \Leftrightarrow \log p(\boldsymbol{\tau}) = \sum_{k=1}^K (a_k - 1) \log \tau_k - b_k \tau_k \end{aligned} \quad (61)$$

## Assignment Factor

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

$$\begin{aligned} \log q^*(\mathbf{Z}) &= E_{\boldsymbol{\pi}, \boldsymbol{\tau}} [\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\tau})] \\ &= \mathbb{E}_{\boldsymbol{\tau}} [\log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\tau})] + \mathbb{E}_{\boldsymbol{\pi}} [\log p(\mathbf{Z} | \boldsymbol{\pi})] \end{aligned} \quad (62)$$

The first term:

$$\mathbb{E}_{\boldsymbol{\tau}} \left[ \log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\tau}) \right] = -\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K z_{nk} \int (x_n^2 \tau_k + \log 2\pi - \log \tau_k) d\tau_k \quad (63)$$

The second term:

$$\mathbb{E}_{\boldsymbol{\pi}} \left[ \log p(\mathbf{Z} | \boldsymbol{\pi}) \right] = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \int \log \pi_k d\pi_k \quad (64)$$

Thus

$$\log q^*(\mathbf{Z}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \rho_{nk}, \quad (65)$$

where

$$\rho_{nk} = -\frac{1}{2} x_n^2 \mathbb{E}(\tau_k) + \frac{1}{2} \mathbb{E}(\log \tau_k) - \frac{1}{2} \log 2\pi + \mathbb{E}(\log \pi_k) \quad (66)$$

Similarly,

$$\begin{aligned} r_{nk} &= \frac{\exp(\tilde{\rho}_{nk})}{\sum_{k=1}^K \exp(\tilde{\rho}_{nk})} \\ \tilde{\rho}_{nk} &= \rho_{nk} - \max_k \rho_{nk} \\ q^*(\mathbf{Z}) &= \prod_{n=1}^N \prod_{k=1}^K r_{nk}^{\tilde{z}_{nk}}. \end{aligned} \quad (67)$$

## Other Factors

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

$$\log q^*(\boldsymbol{\pi}, \boldsymbol{\tau}) = \mathbb{E}_{\mathbf{Z}} \left[ \log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\tau}) \right] + \mathbb{E}_{\mathbf{Z}} \left[ \log p(\mathbf{Z} | \boldsymbol{\pi}) \right] + \log p(\boldsymbol{\tau}) + \log p(\boldsymbol{\pi}) \quad (68)$$

By observing the above equation,

$$q(\boldsymbol{\pi}, \boldsymbol{\tau}) = q(\boldsymbol{\pi}) \prod_{k=1}^K q(\tau_k) \quad (69)$$

## Weight Factor

$$\begin{aligned}
& \log q^*(\boldsymbol{\pi}) \\
&= \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{Z} | \boldsymbol{\pi})] + \log p(\boldsymbol{\pi}) \\
&= \sum_{n=1}^N \sum_{k=1}^K r_{nk} \log \pi_k + \sum_{k=1}^K (\alpha_0 - 1) \log \pi_k \\
&= \sum_{k=1}^K (N_k + \alpha_0 - 1) \log \pi_k
\end{aligned} \tag{70}$$

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.

$$\begin{aligned}
q^*(\boldsymbol{\pi}) &= \text{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha}) \\
\log q^*(\boldsymbol{\pi}) &= \sum_{k=1}^K (\alpha_k - 1) \log \pi_k \\
&\Rightarrow \alpha_k = \alpha_0 + N_k
\end{aligned} \tag{71}$$

## Precision Factors

The prior is:

$$\log p(\tau_k) = (a_k - 1) \log \tau_k - b_k \tau_k \tag{72}$$

For the precision matrix,

$$\begin{aligned}
& \log q^*(\boldsymbol{\tau}) = \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\tau})] + \log p(\boldsymbol{\tau}) \\
&= -\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K r_{nk} (\tau_k x_n^2 + \log 2\pi - \log \tau_k) + \sum_{k=1}^K (a_0 - 1) \log \tau_k - b_0 \tau_k \\
&= \sum_{k=1}^K \left\{ -\left( \frac{1}{2} \sum_{n=1}^N r_{nk} x_n^2 \right) \tau_k + \frac{1}{2} N_k \log \tau_k + (a_0 - 1) \log \tau_k - b_0 \tau_k \right\} \\
&= \left( \frac{1}{2} N_k + a_0 - 1 \right) \log \tau_k - \left( \frac{1}{2} c_k + b_0 \right) \tau_k + cst. \\
&\Rightarrow \begin{cases} a_k = a_0 + \frac{1}{2} N_k \\ b_k = b_0 + \frac{1}{2} c_k \end{cases} \\
&c_k \triangleq \sum_{n=1}^N r_{nk} x_n^2
\end{aligned} \tag{73}$$