Mean-field Variational Inference for Bayesian Mixture of Gaussian

December 27, 2022

Abstract

Notes for the Bayesian Mixture of Gaussians section in Bishop's PRML. Compared to the book's treatment, this derivation more detailed. When choosing priors, we rely on Theorem 2.2 in Beal's PhD thesis, i.e., the priors should be conjugate to the complete-data log likelihood.

Table of contents

1 Complete-data log likelihood

Just write out the mixture of Gaussians assuming that the latent variables are observed:

$$\prod_{i=1}^{n} p(\boldsymbol{x}_i, \boldsymbol{z}_i | \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{c} \pi_k^{z_{ik}} \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{ik}}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}_{1:k}, \boldsymbol{\Lambda}_{1:k})$.

2 Prior for $\boldsymbol{\pi}$

Dirichlet:

$$\mathrm{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha}_0) \propto \prod_{k=1}^{c} \pi_k^{(\alpha_0)_k - 1}$$

Relevant part of the complete-data log likelihood:

$$\prod_{i=1}^{n} \prod_{k=1}^{c} \pi_k^{z_{ik}}$$

Proof that they are conjugate:

$$\left(\prod_{k=1}^{c} \pi_k^{(\alpha_0)_k - 1}\right)\left(\prod_{i=1}^{n}\prod_{k=1}^{c} \pi_k^{z_{ik}}\right) = \left(\prod_{k=1}^{c} \pi_k^{(\alpha_0)_k - 1}\right)\left(\prod_{k=1}^{c} \pi_k^{\sum_{i=1}^{n} z_{ik}}\right) = \prod_{k=1}^{c} \pi_k^{(\alpha_0)_k + \sum_{i=1}^{n} z_{ik} - 1}$$

3 Joint prior for $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$:

Gaussian-Wishart:

$$p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k | \, \boldsymbol{m}_0, \beta_0, v_0) = \mathcal{N}(\boldsymbol{\mu}_k | \, \boldsymbol{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) \, \mathrm{Wi}(\boldsymbol{\Lambda}_k | \, \boldsymbol{L}_0, v_0)$$

Relevant part of the complete-data log likelihood:

$$\prod_{i=1}^{n} \mathcal{N}(\boldsymbol{x}_i | \, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{ik}} = \prod_{i \text{ s.t. } z_{i,k=1}} \mathcal{N}(\boldsymbol{x}_i | \, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})$$

Proof that they are conjugate to each other: omitted, see Section 4.6.3.3 of Murphy.

4 Computing $q(Z)$: variational E-step

For notationaly clarity, we let $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}$ and $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_k\}$.

Starting from the CAVI update rule:

$$\begin{aligned}
& \log q(\boldsymbol{Z}) \\
= \; & \mathbb{E}_{q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const} \\
= \; & \mathbb{E}_{q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})}[\log p(\boldsymbol{Z} | \, \boldsymbol{\pi}) + \log p(\boldsymbol{X} | \, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) + \log p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const} \\
= \; & \mathbb{E}_{q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})}[\log p(\boldsymbol{Z} | \, \boldsymbol{\pi}) + \log p(\boldsymbol{X} | \, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const} \\
= \; & \mathbb{E}_{q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})}[\log p(\boldsymbol{Z} | \, \boldsymbol{\pi})] + \mathbb{E}_{q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})}[\log p(\boldsymbol{X} | \, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const} \\
= \; & \mathbb{E}_{q(\boldsymbol{\pi})}[\log p(\boldsymbol{Z} | \, \boldsymbol{\pi})] + \mathbb{E}_{q(\boldsymbol{\mu},\boldsymbol{\Lambda})}[\log p(\boldsymbol{X} | \, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const} \\
= \; & \mathbb{E}_{q(\boldsymbol{\pi})}\left[\sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik} \log \pi_k\right] + \mathbb{E}_{q(\boldsymbol{\mu},\boldsymbol{\Lambda})}\left[\sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik} \log \mathcal{N}(\boldsymbol{x}_i | \, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})\right] + \text{const} \\
= \; & \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik} \, \mathbb{E}_{q(\boldsymbol{\pi})}[\log \pi_k] + \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik} \, \mathbb{E}_{q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)}\left[\frac{1}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{D}{2}\ln 2\pi - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right] + \text{const} \\
= \; & \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik} \, \mathbb{E}_{q(\boldsymbol{\pi})}[\log \pi_k] + \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik}\left[\frac{1}{2}\,\mathbb{E}_{q(\boldsymbol{\Lambda}_k)}[\ln|\boldsymbol{\Lambda}_k|] - \frac{D}{2}\ln 2\pi - \frac{1}{2}\,\mathbb{E}_{q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)}[(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_i - \right. \\
& \left. \boldsymbol{\mu}_k)]\right] + \text{const} \\
= \; & \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik}\left(\mathbb{E}_{q(\boldsymbol{\pi})}[\log \pi_k] + \frac{1}{2}\,\mathbb{E}_{q(\boldsymbol{\Lambda}_k)}[\ln|\boldsymbol{\Lambda}_k|] - \frac{D}{2}\ln 2\pi - \frac{1}{2}\,\mathbb{E}_{q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)}[(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_i - \boldsymbol{\mu}_k)]\right) + \text{const} \\
\triangleq \; & \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik} \log \rho_{ik}
\end{aligned}$$

Therefore

$$q(\boldsymbol{Z}) \;\propto\; \prod_{i=1}^{n}\prod_{k=1}^{c} \rho_{ik}^{z_{ik}}$$

Requring that the distributuon is normalized on each row:

$$q(\boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{c} r_{ik}^{z_{ik}}$$

where r_{ik} is the ρ_{ik} divided by row sum.

5 Computing $q(\boldsymbol{\pi}, \boldsymbol{\mu}_{1:C}, \boldsymbol{\Lambda}_{1:C})$: variational M-step

Starting from the CAVI update rule:

$$\begin{aligned}
& \ln q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\
= \; & \mathbb{E}_{q(\boldsymbol{Z})}[\log p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{Z}, \boldsymbol{X})] + \text{const} \\
= \; & \mathbb{E}_{q(\boldsymbol{Z})}[\log p(\boldsymbol{\pi}) + \log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) + \log p(\boldsymbol{Z}|\, \boldsymbol{\pi}) + \log p(\boldsymbol{X}|\, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const} \\
= \; & \log p(\boldsymbol{\pi}) + \log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) + \mathbb{E}_{q(\boldsymbol{Z})}[\log p(\boldsymbol{Z}|\, \boldsymbol{\pi})] + \mathbb{E}_{q(\boldsymbol{Z})}[\log p(\boldsymbol{X}|\, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const} \\
= \; & (\log p(\boldsymbol{\pi}) + \mathbb{E}_{q(\boldsymbol{Z})}[\log p(\boldsymbol{Z}|\, \boldsymbol{\pi})]) + \sum_{k=1}^{C} \left(p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + \sum_{i=1}^{N} r_{ik} \log \mathcal{N}(\boldsymbol{x}_i|\, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) \right) + \text{const}, \qquad (1)
\end{aligned}$$

which contains separate terms for $\boldsymbol{\pi}$ and each $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$. This implies that the approximate posterior further factors into

$$\begin{aligned}
\log q(\boldsymbol{\pi}, \boldsymbol{\mu}_1, \ldots, .\boldsymbol{\mu}_C, \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_C) &= \log q(\boldsymbol{\pi}) + \sum_{k=1}^{C} \ln q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \\
q(\boldsymbol{\pi}, \boldsymbol{\mu}_1, \ldots, .\boldsymbol{\mu}_C, \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_C) &= q(\boldsymbol{\pi}) \prod_{k=1}^{C} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)
\end{aligned}$$

so that we can optimize over each component distribution separately.

5.1 $q(\boldsymbol{\pi})$

Isolating the terms containing $\boldsymbol{\pi}$ from Equation 1, we see that

$$\begin{aligned}
\log q^*(\boldsymbol{\pi}) &= \log p(\boldsymbol{\pi}) + \mathbb{E}_{q(\boldsymbol{Z})}[\log p(\boldsymbol{Z}|\, \boldsymbol{\pi})] + \text{const} \\
&= \sum_{k=1}^{C} (\alpha_0 - 1) \ln \pi_k + \sum_{k=1}^{C} \left(\sum_{i=1}^{N} r_{ik} \right) \ln \pi_k + \text{const} \\
&= \sum_{k=1}^{C} \left(\alpha_0 + \sum_{i=1}^{N} r_{ik} - 1 \right) \ln \pi_k + \sum_{k=1}^{C} \ln \pi_k + \text{const} \\
&\Rightarrow \\
q^*(\boldsymbol{\pi}) &= \text{Dir}(\boldsymbol{\pi}|\, \boldsymbol{\alpha}) \quad \text{with} \quad \alpha_k = \alpha_0 + N_k
\end{aligned}$$

5.2 $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$

Isolating the terms containing $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ from Equation 1, we see that

$$\log q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \log p(\boldsymbol{\mu}_k|\, \boldsymbol{\Lambda}_k) + \log p(\boldsymbol{\Lambda}_k) + \sum_{i=1}^{N} r_{ik} \log \mathcal{N}(\boldsymbol{x}_i|\, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) + \text{const}$$

By the Gaussian-Wishart prior we specified earlier, we have

$$\begin{aligned} p(\boldsymbol{\mu}_k|\,\boldsymbol{\Lambda}_k) &= \mathcal{N}(\boldsymbol{\mu}_k|\,\boldsymbol{m}_0,(\beta_0\boldsymbol{\Lambda}_k)^{-1}) \\ &\propto |\beta_0\boldsymbol{\Lambda}_k|^{1/2}\exp\!\Big(-\frac{1}{2}(\boldsymbol{\mu}_k-\boldsymbol{m}_0)^T(\beta_0\boldsymbol{\Lambda}_k)(\boldsymbol{\mu}_k-\boldsymbol{m}_0)\Big) \\ &\propto |\boldsymbol{\Lambda}_k|^{1/2}\exp\!\Big(-\frac{1}{2}(\boldsymbol{\mu}_k-\boldsymbol{m}_0)^T(\beta_0\boldsymbol{\Lambda}_k)(\boldsymbol{\mu}_k-\boldsymbol{m}_0)\Big) \\ &\Rightarrow \\ \log p(\boldsymbol{\mu}_k|\,\boldsymbol{\Lambda}_k) &= \frac{1}{2}\log|\boldsymbol{\Lambda}_k|-\frac{1}{2}(\boldsymbol{\mu}_k-\boldsymbol{m}_0)^T(\beta_0\boldsymbol{\Lambda}_k)(\boldsymbol{\mu}_k-\boldsymbol{m}_0)+\text{const} \end{aligned}$$

and

$$\begin{aligned} p(\boldsymbol{\Lambda}_k) &= \mathcal{W}(\boldsymbol{\Lambda}_k|\,\boldsymbol{W}_0,\upsilon_0) \\ &\propto |\boldsymbol{\Lambda}_k|^{(\upsilon_0-C-1)/2}\exp\!\Big(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{W}_0^{-1}\boldsymbol{\Lambda}_k)\Big) \\ &\Rightarrow \\ \log p(\boldsymbol{\Lambda}_k) &= \frac{(\upsilon_0-C-1)}{2}\log|\boldsymbol{\Lambda}_k|-\frac{1}{2}\,\mathrm{Tr}(\boldsymbol{W}_0^{-1}\boldsymbol{\Lambda}_k)+\text{const.} \end{aligned}$$

The normal likelihoods for all the observations can be written as

$$\sum_{i=1}^{N} r_{ik}\log\mathcal{N}(\boldsymbol{x}_i|\,\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})=\sum_{i=1}^{N} r_{ik}\left(\frac{1}{2}\log|\boldsymbol{\Lambda}_k|-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right)+\text{const.}$$

Now, we can assemble all terms together:

$$\begin{aligned} \log q^*(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) &= \frac{1}{2}\log|\boldsymbol{\Lambda}_k|-\frac{1}{2}(\boldsymbol{\mu}_k-\boldsymbol{m}_0)^T(\beta_0\boldsymbol{\Lambda}_k)(\boldsymbol{\mu}_k-\boldsymbol{m}_0) \\ &\quad+\frac{(\upsilon_0-C-1)}{2}\log|\boldsymbol{\Lambda}_k|-\frac{1}{2}\,\mathrm{Tr}(\boldsymbol{W}_0^{-1}\boldsymbol{\Lambda}_k) \\ &\quad+\sum_{i=1}^{N} r_{ik}\left(\frac{1}{2}\log|\boldsymbol{\Lambda}_k|-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right)+\text{const.} \end{aligned}$$

It may not be immediate clear what we should do here, so it's good to remind ourselves that by Theorem 2 of Beal tells us that the approximate posterior over $(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)$ would also be Gaussian-Wishart.

Firstly, we see that

$$\frac{(\upsilon_0-C-1)}{2}\log|\boldsymbol{\Lambda}_k|+\frac{(\sum_{i=1}^{N} r_{ik})}{2}\log|\boldsymbol{\Lambda}_k|=\frac{(\upsilon_0+N_k-C-1)}{2}\log|\boldsymbol{\Lambda}_k|,$$

which means that the Wishart posterior distribution on $\boldsymbol{\Lambda}_k$ have DOF $\upsilon_k=\upsilon_0+N-C$.

Secondly, we need manipulate these terms

$$-\frac{1}{2}(\boldsymbol{\mu}_k-\boldsymbol{m}_0)^T(\beta_0\boldsymbol{\Lambda}_k)(\boldsymbol{\mu}_k-\boldsymbol{m}_0) \quad -\frac{1}{2}\,\mathrm{Tr}(\boldsymbol{W}_0^{-1}\boldsymbol{\Lambda}_k) \qquad \sum_{i=1}^{N} r_{ik}\Big(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\Big)$$

into two terms of the form

$$-\frac{1}{2}(\boldsymbol{\mu}_k-?_1)^T(?\boldsymbol{\Lambda}_k)(\boldsymbol{\mu}_k-?_1) \quad -\frac{1}{2}\,\mathrm{Tr}(?_2\boldsymbol{\Lambda}_k)$$

where the $?_2$ in the second term shouldn't contain $\boldsymbol{\mu}_k$ because it would be a parameter of the Wishart posterior. For simplicity, we will drop the shared $-1/2$ term shortly.

To begin, we do an expansion with a trick:

$$
\begin{aligned}
& \sum_{i=1}^{N} r_{ik}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \\
= & \sum_{i=1}^{N} r_{ik} \left((\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k) - (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k)\right)^T \boldsymbol{\Lambda}_k \left((\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k) - (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k)\right) \\
= & \sum_{i=1}^{N} r_{ik} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k) \boldsymbol{\Lambda}_k (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)^T - 2 (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k)^T \boldsymbol{\Lambda}_k \underbrace{\sum_{i=1}^{N} r_{ik} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)}_{0} + N_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k) \\
= & N_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k) + \sum_{i=1}^{N} r_{ik} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k) \\
= & N_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k) + \mathrm{Tr}\left[\left(\sum_{i=1}^{N} r_{ik} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)^T \right) \boldsymbol{\Lambda}_k \right],
\end{aligned}
$$

which is great because the first term here contains $\boldsymbol{\mu}_k$ while the second term here doesn't and can be directly merged with $\mathrm{Tr}(\boldsymbol{W}_0^{-1} \boldsymbol{\Lambda}_k)$.

We can merge the first term here with

$$
\begin{aligned}
& \beta_0 (\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_0) + N_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k) \\
= & \mathrm{Tr}[(\beta_0 (\boldsymbol{\mu}_k - \boldsymbol{m}_0)(\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T + N_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k)(\boldsymbol{\mu}_k - \bar{\boldsymbol{x}}_k)^T) \boldsymbol{\Lambda}_k] \\
= & \mathrm{Tr}[(\beta_0 (\boldsymbol{\mu}_k \boldsymbol{\mu}_k^T - \boldsymbol{\mu}_k \boldsymbol{m}_0^T - \boldsymbol{m}_0 \boldsymbol{\mu}_k^T + \boldsymbol{m}_0 \boldsymbol{m}_0^T) + N_k (\boldsymbol{\mu}_k \boldsymbol{\mu}_k^T - \boldsymbol{\mu}_k \bar{\boldsymbol{x}}_k^T - \bar{\boldsymbol{x}}_k \boldsymbol{\mu}_k^T + \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T)) \boldsymbol{\Lambda}_k] \\
= & \mathrm{Tr}[((\beta_0 + N_k) \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T - \boldsymbol{\mu}_k (\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)^T - (\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k) \boldsymbol{\mu}_k^T + \beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T + N_k \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T) \boldsymbol{\Lambda}_k] \\
= & \mathrm{Tr}\Bigg[\Bigg((\beta_0 + N_k) \Bigg(\boldsymbol{\mu}_k \boldsymbol{\mu}_k^T - \boldsymbol{\mu}_k \frac{(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)^T}{\beta_0 + N_k} - \frac{(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)^T}{\beta_0 + N_k} \boldsymbol{\mu}_k^T \Bigg) + \beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T + \\
& N_k \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T \Bigg) \boldsymbol{\Lambda}_k \Bigg] \\
= & \mathrm{Tr}\Bigg[\Bigg((\beta_0 + N_k) \Bigg(\boldsymbol{\mu}_k - \frac{(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)}{\beta_0 + N_k} \Bigg) \Bigg(\boldsymbol{\mu}_k - \frac{(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)}{\beta_0 + N_k} \Bigg)^T - \\
& \frac{(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)^T}{\beta_0 + N_k} + \beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T + N_k \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T \Bigg) \boldsymbol{\Lambda}_k \Bigg] \text{(complete the squre)} \\
= & \mathrm{Tr}\Bigg[\Bigg(\cdots - \frac{(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)^T}{\beta_0 + N_k} + \beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T + N_k \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T \Bigg) \boldsymbol{\Lambda}_k \Bigg] \\
= & \mathrm{Tr}\Bigg[\Bigg(\cdots - \frac{\beta_0^2 \boldsymbol{m}_0 \boldsymbol{m}_0^T}{\beta_0 + N_k} - \frac{\beta_0 N_k \boldsymbol{m}_0 \bar{\boldsymbol{x}}_k^T}{\beta_0 + N_k} - \frac{\beta_0 N_k \bar{\boldsymbol{x}}_k \boldsymbol{m}_0^T}{\beta_0 + N_k} - \frac{N_k^2 \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T}{\beta_0 + N_k} + \beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T + N_k \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T \Bigg) \boldsymbol{\Lambda}_k \Bigg] \\
= & \mathrm{Tr}\Bigg[\Bigg(\cdots - \frac{\beta_0^2 \boldsymbol{m}_0 \boldsymbol{m}_0^T}{\beta_0 + N_k} - \frac{\beta_0 N_k \boldsymbol{m}_0 \bar{\boldsymbol{x}}_k^T}{\beta_0 + N_k} - \frac{\beta_0 N_k \bar{\boldsymbol{x}}_k \boldsymbol{m}_0^T}{\beta_0 + N_k} - \frac{N_k^2 \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T}{\beta_0 + N_k} + \frac{\beta_0^2 + \beta_0 N_k}{\beta_0 + N_k} \boldsymbol{m}_0 \boldsymbol{m}_0^T + \\
& \frac{\beta_0 N_k + N_k^2}{\beta_0 + N_k} \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T \Bigg) \boldsymbol{\Lambda}_k \Bigg] \\
= & \mathrm{Tr}\Bigg[\Bigg(\cdots - \frac{\beta_0 N_k \boldsymbol{m}_0 \bar{\boldsymbol{x}}_k^T}{\beta_0 + N_k} - \frac{\beta_0 N_k \bar{\boldsymbol{x}}_k \boldsymbol{m}_0^T}{\beta_0 + N_k} + \frac{\beta_0 N_k}{\beta_0 + N_k} \boldsymbol{m}_0 \boldsymbol{m}_0^T + \frac{\beta_0 N_k}{\beta_0 + N_k} \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T \Bigg) \boldsymbol{\Lambda}_k \Bigg] \\
= & \mathrm{Tr}\Bigg[\Bigg(\cdots + \Bigg(\frac{\beta_0 N_k}{\beta_0 + N_k} \Bigg) (\boldsymbol{m}_0 \boldsymbol{m}_0^T - \boldsymbol{m}_0 \bar{\boldsymbol{x}}_k^T - \bar{\boldsymbol{x}}_k \boldsymbol{m}_0^T + \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T) \Bigg) \boldsymbol{\Lambda}_k \Bigg] \\
= & \mathrm{Tr}\Bigg[\Bigg(\cdots + \Bigg(\frac{\beta_0 N_k}{\beta_0 + N_k} \Bigg) (\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)^T \Bigg) \boldsymbol{\Lambda}_k \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathrm{Tr}\Bigg[\Bigg(\cdots + \left(\frac{\beta_0 N_k}{\beta_0+N_k}\right)(\bar{\boldsymbol{x}}_k-\boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k-\boldsymbol{m}_0)^T\Bigg)\boldsymbol{\Lambda}_k\Bigg] \\
&= \mathrm{Tr}\Bigg[\Bigg((\beta_0+N_k)\Bigg(\boldsymbol{\mu}_k-\frac{(\beta_0\boldsymbol{m}_0+N_k\bar{\boldsymbol{x}}_k)}{\beta_0+N_k}\Bigg)\Bigg(\boldsymbol{\mu}_k-\frac{(\beta_0\boldsymbol{m}_0+N_k\bar{\boldsymbol{x}}_k)}{\beta_0+N_k}\Bigg)^T + \left(\frac{\beta_0 N_k}{\beta_0+N_k}\right)(\bar{\boldsymbol{x}}_k-\boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k- \\
&\quad \boldsymbol{m}_0)^T\Bigg)\boldsymbol{\Lambda}_k\Bigg] \\
&= \Bigg(\boldsymbol{\mu}_k-\frac{(\beta_0\boldsymbol{m}_0+N_k\bar{\boldsymbol{x}}_k)}{\beta_0+N_k}\Bigg)^T((\beta_0+N_k)\boldsymbol{\Lambda}_k)\Bigg(\boldsymbol{\mu}_k-\frac{(\beta_0\boldsymbol{m}_0+N_k\bar{\boldsymbol{x}}_k)}{\beta_0+N_k}\Bigg)^T + \mathrm{Tr}\Bigg[\left(\frac{\beta_0 N_k}{\beta_0+N_k}\right)(\bar{\boldsymbol{x}}_k- \\
&\quad \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k-\boldsymbol{m}_0)^T\boldsymbol{\Lambda}_k\Bigg],
\end{aligned}
$$

where the first term here has the required form of $(\boldsymbol{\mu}_k - ?_1)^T(?\boldsymbol{\Lambda}_k)(\boldsymbol{\mu}_k - ?_1)$ and the second term can be directly merged with $\mathrm{Tr}(\boldsymbol{W}_0^{-1}\boldsymbol{\Lambda}_k)$.

Finally, we have

$$
\begin{aligned}
&\log q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \\
&= -\frac{1}{2}\Bigg(\boldsymbol{\mu}_k-\frac{\beta_0\boldsymbol{m}_0+N_k\bar{\boldsymbol{x}}_k}{\beta_0+N}\Bigg)^T((\beta_0+N_k)\boldsymbol{\Lambda}_k)\Bigg(\boldsymbol{\mu}_k-\frac{\beta_0\boldsymbol{m}_0+N_k\bar{\boldsymbol{x}}_k}{\beta_0+N}\Bigg) + \\
&\quad \frac{(v_0+\sum_{i=1}^N r_{ik}-1)}{2}\log|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\Bigg(\Bigg(\boldsymbol{W}_0^{-1}+N_k(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_k)^T+\frac{\beta_0 N_k}{\beta_0+N_k}(\bar{\boldsymbol{x}}_k-\boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k- \\
&\quad \boldsymbol{m}_0)^T\Bigg)\boldsymbol{\Lambda}_k\Bigg) + \text{const} \\
&= \log\mathcal{N}\Bigg(\boldsymbol{\mu}_k\Bigg|\frac{\beta_0\boldsymbol{m}_0+N_k\bar{\boldsymbol{x}}_k}{\beta_0+N_k}, ((\beta_0+N_k)\boldsymbol{\Lambda}_k)^{-1}\Bigg) + \log\mathcal{W}\Bigg(\boldsymbol{\Lambda}_k\Bigg|\Bigg(\boldsymbol{W}_0^{-1}+N_k(\boldsymbol{x}_n-\bar{\boldsymbol{x}}_k)(\boldsymbol{x}_n-\bar{\boldsymbol{x}}_k)^T+ \\
&\quad \frac{\beta_0 N_k}{\beta_0+N_k}(\bar{\boldsymbol{x}}_k-\boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k-\boldsymbol{m}_0)^T\Bigg)^{-1}, v_0+N_k\Bigg).
\end{aligned}
$$

Or we can define

$$
\begin{aligned}
\boldsymbol{m}_k &= \frac{\beta_0\boldsymbol{m}_0+N_k\bar{\boldsymbol{x}}_k}{\beta_0+N_k} \\
\beta_k &= \beta_0+N \\
\boldsymbol{W}_k^{-1} &= \boldsymbol{W}_0^{-1}+N_k(\boldsymbol{x}_n-\bar{\boldsymbol{x}}_k)(\boldsymbol{x}_n-\bar{\boldsymbol{x}}_k)^T+\frac{\beta_0 N_k}{\beta_0+N_k}(\bar{\boldsymbol{x}}_k-\boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k-\boldsymbol{m}_0)^T \\
v_k &= v_0+N_k
\end{aligned}
$$

This results turn out to be very similar to the standard Gaussian-Wishart posterior without soft assignment. The approach here should be directly applicable to that, too.

6 Quantities required for computing $q(Z)$

Standard results:

$$
\rho_{ik} = \mathbb{E}_{q(\boldsymbol{\pi})}[\log\pi_k] + \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\Lambda}_k)}[\ln|\boldsymbol{\Lambda}_k|] - \frac{D}{2}\ln 2\pi - \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)}[(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)]
$$

Please refer to Bishop's PRML from computing these expectations.

7 Quantities required for computing $q(\boldsymbol{\pi})$ and $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$

Please refer to Bishop's PRML from computing the associated expectations.