Variational Inference for Gaussian Mixture Models

Kuljit S. Virk

March 4, 2024

Contents

1 Introduction

An observed sample $\mathbf{x}$ is a d-dimensional vector. We denote the set of N indepdnent observations as the data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. We let $p(\mathbf{x})$ be the probability of observing a sample $\mathbf{x}$. Then by independence of observations, the probability of observing the entire dataset $\mathbf{X}$ is,

$$p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x}_n). \tag{1}$$

This is called the *data likelihood.* The logarithm of this function is,

$$\log p(\mathbf{X}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n). \tag{2}$$

Another general quantity we will often make use of in the derivations is the *partition function*, which may be defined as the normalization constant for any probability distribution. Thus we write

$$\begin{aligned} \log p(x) &= f(x) - \log \mathcal{Z}, \\ \mathcal{Z} &= \sum_x e^{f(x)}. \end{aligned}$$

The Gaussian Mixtures Model (GMM) fits a parameterized function to the true probability $p(\mathbf{x})$ using a minimization principle. We denote the parameter set by

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}, \tag{3}$$

$$\boldsymbol{\theta}_k = (\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k), \tag{4}$$

where π_k are scalars that sum to 1, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$ are the mean and precision matrix parameters for the normal distribution. There are K such distributions and their weighted sum consitutes the GMM approximation to $p(\mathbf{x})$,

$$p(\boldsymbol{x}) \approx p(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}), \tag{5}$$

where $\mathcal{N}$ denotes the normal distribution.

In the most common and simplest realization, K is fixed, and $\boldsymbol{\theta}_k$ are determined by maximizing the log of the data-likelihood function,

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} \sum_{n=1}^{N} \log p(\mathbf{x}|\boldsymbol{\Theta}). \tag{6}$$

To reach the maximum, we must equate the derivatives of (2) to zero. It is convenient to first define *responsibilities,*

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1})}. \tag{7}$$

It follows directly from the definition that

$$\sum_k \gamma_{nk} = 1,$$

and we define the effective number of data points explained by component k as

$$N_k \equiv \sum_{n=1}^{N} \gamma_{nk}. \tag{8}$$

When maximizing with respect to π_k, we must use Lagrange multiplier to enforce the constraint on

their sum. Using the formulas proved in the Appendix, we obtain

$$0 = \frac{\partial \log p(\mathbf{X})}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{\Lambda}_k \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right), \tag{9}$$

$$0 = \frac{\partial \log p(\mathbf{X})}{\partial \boldsymbol{\Lambda}_k} = \sum_{n=1}^{N} \gamma_{nk} \left[\boldsymbol{\Lambda}_k^{-1} - \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T\right], \tag{10}$$

$$0 = \frac{\partial}{\partial \pi_k}\left[\log p(\mathbf{X}) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)\right] \tag{11}$$

$$= \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1})} - \lambda.$$

From the first two equations we get

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n, \tag{12}$$

$$\boldsymbol{\Lambda}_k^{-1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T. \tag{13}$$

Multiplying (11) by π_k and summing over k, we get $\lambda = -N$. Substiuting this back into (11), multiplying by π_k, and then substituing the definition (8), we get

$$\pi_k = \frac{N_k}{N}. \tag{14}$$

Equations (7), (12), (13), and (14) are solved self consistently. The most common algorithm for finding the solution is *Expectation Maximization*. According to this algortihm, we first initialize all $\boldsymbol{\theta}_k$ and compute γ_{nk} from (7). This is the expectation or E-step. We then fix γ_{nk} and update $\boldsymbol{\theta}_k$ according to (12)-(14), which are solutions for the maximum, and is thus called the M-step (or maximization step). The cycle continues until $\boldsymbol{\theta}_k$ stop changing.

2 Latent variable formulation of GMM

Latent variable formulation provides a bridge to new methods for modeling data using GMM. In this section, we show how to write the GMM in the latent variable formulation, and how it leads to the same solutions to parameters derived in the previous section.

For each data sample $\mathbf{x}$, we introduce a binary vector $\mathbf{z}$ of dimension K such that $z_j = 1$ for only one $j \in \{1, \ldots, K\}$ and zero for all others, and with the probability

$$p(z_j = 1) = \pi_j. \tag{15}$$

Since $z_k \in \{0, 1\}$, we can write the probability of the full vector $\mathbf{z}$ as

$$p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_k}. \tag{16}$$

Due to the binary property of z_k, we can define,

$$p(\mathbf{x}|\mathbf{z},\boldsymbol{\Theta}) \equiv \prod_{k=1}^{K}\left[\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})\right]^{z_k}. \tag{17}$$

Finally, we see that the GMM as defined in (5) can be expressed as a marginal probability,

$$p(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z},\boldsymbol{\Theta})p(\mathbf{z}). \tag{18}$$

By fundamental laws of probability, the summand is equal to the joint probability distribution of $\mathbf{x}$ and $\mathbf{z}$, and thus we obtain the latent variable formulation as a marginalization of the probability distribution defined over the larger space $(\mathbf{x},\mathbf{z})$,

$$p(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z}|\boldsymbol{\Theta}). \tag{19}$$

For each $\mathbf{x}_n \in \mathbf{X}$, we define a latent vector $\mathbf{z}_n$, and denote the set of all latent vectors for the data points as $\mathbf{Z}$. The log-likelihood of the joint, or complete, data $\{\mathbf{X},\mathbf{Z}\}$ using the model (19), we obtain

$$\log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\Theta}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left[\log\pi_k + \log\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})\right]. \tag{20}$$

The binary variables z_{nk} select the data points that *belong* to component k. This association of a data point to a compnent occurs in the latent space and is not controlled by the observer. For given observed data $\mathbf{X}$, the expectation value of z_{nk} follows from Bayes' formula

$$p(\mathbf{Z}|\mathbf{X},\boldsymbol{\Theta}) = \Omega p(\mathbf{X}|\mathbf{Z},\boldsymbol{\Theta})p(\mathbf{Z}|\boldsymbol{\Theta}),$$

where Ω is a normaliation constant. Therefore,

$$\mathbb{E}_Z\left[z_{nk}\right] = \sum_{\mathbf{z}_n} z_{nk}\frac{\prod_{j=1}^{K}\left[\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Lambda}_j^{-1})\right]^{z_{nj}}}{\sum_{\mathbf{z}_n}\prod_{j=1}^{K}\left[\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Lambda}_j^{-1})\right]^{z_{nj}}}.$$

For fixed n, only one j contributes to the multiplication. The denominator in the summand is a sum over all binary vectors, or the vertices of K-dimensional cube. Thus for each term in the sum in the denominator the produce contributes exactly one factor. The only term that contributes in the numerator is the one for which $z_{nk}=1$ and since the sum is unconstrained, it is guarranteed to exist in the summation. Putting the two obeservations together, and using the definition (7) along with the fact that the only non-zero $z_{nk}=1$, we obtain,

$$\mathbb{E}_Z\left[z_{nk}\right] = \gamma_{nk}. \tag{21}$$

The expected value of (20) over the latent space is then

$$\mathbb{E}_Z\left[\log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\Theta})\right] = \sum_{n=1}^{N}\gamma_{nk}\left[\log\pi_k + \log\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})\right]. \tag{22}$$

Maximizing the likelihood of this expected value yields precisely the equations (12),(13), and (14) found above.

To conclude, in the latent variable formulation, we augment the observed space of $\mathbf{x}$ vectors with an unobserved space where $\mathbf{z}$ resides. However, the maximization principle is applied to the observed space by *averaging out* the latent variables. The averaging is performed using the posterior distribution of latent variables *given* the observed values.

3 Variational formulation

3.1 The Neal-Hinton representation

We begin with the following representation of $\log p(\mathbf{x})$, also often called the Neal-Hinton representation [2]. We take *any* probability distribution function $q(\mathbf{z})$ over *any* space, such that,

$$\sum_{\mathbf{z}} q(\mathbf{z}) \quad = \quad 1.$$

We then multiply $\log p$ by 1 on one one hand and the left hand side of this equation on the other, and using the relation $p(\mathbf{x}, \mathbf{z}, \boldsymbol{\Theta}) = p(\mathbf{z}, \boldsymbol{\Theta}|\mathbf{x})p(\mathbf{x})$,

$$\begin{aligned}
\log p(\mathbf{x}) &= \sum_{\mathbf{z},\boldsymbol{\Theta}} q(\mathbf{z}, \boldsymbol{\Theta}) \log p(\mathbf{x}) \\
&= \sum_{\mathbf{z},\boldsymbol{\Theta}} q(\mathbf{z}, \boldsymbol{\Theta}) \log \left[\frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\Theta})}{p(\mathbf{z}, \boldsymbol{\Theta}|\mathbf{x})} \right] \\
&= \sum_{\mathbf{z},\boldsymbol{\Theta}} q(\mathbf{z}, \boldsymbol{\Theta}) \log \left[\frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\Theta})}{q(\mathbf{z}, \boldsymbol{\Theta})} \frac{q(\mathbf{z}, \boldsymbol{\Theta})}{p(\mathbf{z}, \boldsymbol{\Theta}|\mathbf{x})} \right] \\
&= \sum_{\mathbf{z},\boldsymbol{\Theta}} q(\mathbf{z}, \boldsymbol{\Theta}) \log \left[\frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\Theta})}{q(\mathbf{z}, \boldsymbol{\Theta})} \right] - \sum_{\mathbf{z},\boldsymbol{\Theta}} q(\mathbf{z}, \boldsymbol{\Theta}) \log \left[\frac{p(\mathbf{z}, \boldsymbol{\Theta}|\mathbf{x})}{q(\mathbf{z}, \boldsymbol{\Theta})} \right].
\end{aligned} \tag{23}$$

This is the Neal-Hinton transformation. It is an identity to re-write $\log p$. Note that the most general function $q(\mathbf{z})$ may be described by a several parameters, which will become dependent on $\boldsymbol{\Theta}$ and the observations $\mathbf{x}$ when we constrain q to minimize the second term as discussed below.

We define a functional for the first term in (23),

$$\mathcal{L}(q, \boldsymbol{\Theta}) \quad = \quad \mathbb{E}_q \left[\log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\Theta}) \right] + H(q), \tag{24}$$

where $H(q)$ is the entropy of the distribution q. We recognize the last term in (23) as the KL divergence,

$$\mathcal{D}(q(\mathbf{z}, \boldsymbol{\Theta}), p(\mathbf{z}, \boldsymbol{\Theta}|\mathbf{x})) \quad = \quad - \sum_{\mathbf{z},\boldsymbol{\Theta}} q(\mathbf{z}, \boldsymbol{\Theta}) \log \left[\frac{p(\mathbf{z}, \boldsymbol{\Theta}|\mathbf{x})}{q(\mathbf{z}, \boldsymbol{\Theta})} \right]. \tag{25}$$

Since $\mathcal{D}(.,.) \geq 0$ for any probability distributions, we see that for probability densities q, $\log p \geq \mathcal{L}(q, \boldsymbol{\Theta})$. Furthermore, $\mathcal{D} = 0$ only when its two input arguments are exactly the same functions. We conclude that,

$$\mathcal{L}(q, \boldsymbol{\Theta}) = \log p(\mathbf{x}) \quad \Longleftrightarrow \quad q(\mathbf{z}, \boldsymbol{\Theta}) = p(\mathbf{z}, \boldsymbol{\Theta}|\mathbf{x}).$$

Therefore, the solution $\boldsymbol{\Theta}^*$ that maximizes $\log p(\mathbf{x})$ also gives the globally maximum value of (24) as $\mathcal{L}(q, \boldsymbol{\Theta}^*)$.

To summarize this section, we introduced a parameterized space of functions q, and a functional $\mathcal{L}(q, \boldsymbol{\Theta})$ that achieves a maximum whenever $\log p(\mathbf{x})$ does. This allows us to introduce another variant of the EM algorithm, repeated until $\boldsymbol{\Theta}$ stops changing to within an acceptable level,

1. For a fixed $\boldsymbol{\Theta}$ vary the parameters of q to maximize $\mathcal{L}$.
2. For q fixed as in Step 1, vary $\boldsymbol{\Theta}$ to maximize $\mathcal{L}$.

3.2 Mean field solution to the posterior

Let us now introduce a factorization of the function q over disjoint sets of variables,

$$q(\boldsymbol{\xi}) \quad = \quad \prod_{i \in S} q_i(\boldsymbol{\xi}_i).$$

Then substitute this into (24),

$$\mathcal{L} \quad = \quad \sum_{\boldsymbol{\xi}_j} q_j(\boldsymbol{\xi}_j) \sum_{\{\boldsymbol{\xi}_{i \neq j}\}} \prod_{i \neq j} q_i(\boldsymbol{\xi}_i) \log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\Theta}) - \sum_{\boldsymbol{\xi}_j} q_j(\boldsymbol{\xi}_j) \log q_j(\boldsymbol{\xi}_j) - \sum_{\{\boldsymbol{\xi}_{i \neq j}\}} q_i(\boldsymbol{\xi}_i) \log q_i(\boldsymbol{\xi}_i).$$

Let us define,

$$\log \overline{p}(\mathbf{x}, \mathbf{z}_j, \boldsymbol{\Theta}_j) \quad = \quad \mathbb{E}_{i \neq j} \left[\log p \left(\mathbf{x}, \mathbf{z}, \boldsymbol{\Theta} \right) \right].$$

Then,

$$\mathcal{L} \quad = \quad \sum_{\mathbf{z}_j, \boldsymbol{\Theta}_j} q_j(\mathbf{z}_j, \boldsymbol{\Theta}_j) \log \overline{p}(\mathbf{x}, \mathbf{z}_j, \boldsymbol{\Theta}_j) - \sum_{\boldsymbol{\xi}_j} q_j(\boldsymbol{\xi}_j) \log q_j(\boldsymbol{\xi}_j) + \ldots,$$

where the omitted terms do not depend on $\boldsymbol{\xi}_j$. Suppose we keep all q_i fixed for $i \neq j$, and carry out the first step in the Neal-Hinton variant of the EM. We get

$$\log q_j^*(\mathbf{z}_j, \boldsymbol{\Theta}_j) \quad = \quad \mathbb{E}_{i \neq j} \left[\log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\Theta}) \right] + c, \tag{26}$$

where c can be determined by normalizing q_j. It is thus the *partition function*, and the normalized distribution is

$$q_j^*(\mathbf{z}_j) \quad = \quad \frac{\exp \left[\mathbb{E}_{i \neq j} \left[\log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\Theta}) \right] \right]}{\sum_j \exp \left[\mathbb{E}_{i \neq j} \left[\log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\Theta}) \right] \right]}. \tag{27}$$

This solution is called the *mean-field* solution. The right hand side of this expression is a joint probability distribution of the observed variables, latent variables, and model parameters. We do not have a model for that. Instead we have (20), and the chain rule of probability gives $p(\mathbf{x}, \mathbf{z}, \boldsymbol{\Theta}) =$

$p(\mathbf{x}, \mathbf{z}|\mathbf{\Theta})p(\mathbf{\Theta})$. The mean field solution can then be stated in terms of the likelihood of the complete data $\{\mathbf{x}, \mathbf{z}\}$ and the prior for the model parameters,

$$q_j^*(\mathbf{z}_j) = \frac{\exp\left[\mathbb{E}_{i\neq j}\left[\log p(\mathbf{x}, \mathbf{z}|\mathbf{\Theta}) + \log p(\mathbf{\Theta})\right]\right]}{\sum_j \exp\left[\mathbb{E}_{i\neq j}\left[\log p(\mathbf{x}, \mathbf{z}|\mathbf{\Theta}) + \log p(\mathbf{\Theta})\right]\right]}.$$

Furthermore, the prior distributions of $\boldsymbol{\theta}_k$ are all independent, and thus

$$p(\mathbf{\Theta}) = \prod_{k=1}^{K} p(\pi_k)p(\boldsymbol{\mu}_k, \mathbf{\Lambda}_k). \tag{28}$$

To conclude, we now have the equation for updating in step 1 of the Neal-Hinton variant of EM. In fact, there is no loss of generality in amalmagating $\mathbf{\Theta}$ into the latent space itself and thus reducing both steps to a single step of replacing each q_j by the mean-field solution interatively until convergence.

3.3 Mean field solution to GMM with conjugate priors

In this section, we turn to deriving the equations for the variational posterior based on the mean field solution given in (27). Any function for q that is normalized as a probability density function is admissible. Optimization over this unconstrained set of functions will select the true posterior. However, that is still a class too large for numerical calculation (why?). Instead, we limit ourselves to the functions q that take the form of conjugate priors. This allows us to convert the problem from a search in function space to a search of numerical parameters.

In the conventional approach taken here, one postulates that the variational posterior factorizes into a function over latent variables and a function over the model parameters.

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda}) = q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda}). \tag{29}$$

We apply (27) to calculate the posterior on the partition containing the model parameters. In this case, the expectation value is needed only over the latent variables $\mathbf{Z}$,

$$\begin{aligned}\log q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda}) &= c + \sum_{k=1}^{K} \log p(\boldsymbol{\pi}_k) + \log p(\boldsymbol{\mu}_k, \mathbf{\Lambda}_k) \\ &\quad + \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}\left(z_{nk}\right)\left[\log \pi_k + \log \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{\Lambda}_k^{-1})\right],\end{aligned} \tag{30}$$

where c is the normalization or the partition function. Since the right hand side of(30) is a sum of two terms, one independent of π_k, we can write the function as the product $q(\boldsymbol{\pi})q(\boldsymbol{\mu}, \mathbf{\Lambda})$. The function $q(\boldsymbol{\pi})$ follows easily after exponentiation of the first term, which takes the form of the Dirichlet distribution [cite]. Since we are modeling the priors to be conjugate to the posterior, we set it to a Dirichlet distribution with parameter α_0. The result for the prior and the posterior are respectively,

$$p(\pi_k) = C(\alpha_0)\prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}, \tag{31}$$

$$q^*(\boldsymbol{\pi}) = C(\boldsymbol{\alpha})\prod_{k=1}^{K} \pi_k^{\alpha_k - 1}, \text{ where, } \alpha_k = \alpha_0 + N_k. \tag{32}$$

To obtain the factor $q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, we set it equal to the terms independent of π_k in (30),

$$\begin{aligned}\log q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \sum_{k=1}^{K}\left[\frac{N_k}{2}\log|\boldsymbol{\Lambda}_k| - \frac{d}{2}\log 2\pi\right] - \sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\gamma_{nk}}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k) \\ &\quad + \sum_{k=1}^{K}\log p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + c. \end{aligned} \tag{33}$$

The sum of logarithms shows that the posterior will also be a product of the posteriors for each component k,

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

where

$$\begin{aligned}\log q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) &= c + \log p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + \frac{N_k}{2}\log|\boldsymbol{\Lambda}_k| - \frac{d}{2}\log 2\pi \\ &\quad - \sum_{n=1}^{N}\frac{\gamma_{nk}}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k). \end{aligned} \tag{34}$$

We choose the prior to be conjugate to the posterior and thus write its logarithm in the same functional form with different parameters. The algebraic manipulations detailed in section §B show that the functional form is a normal-Wishart distribution. The results for the prior are shown in (39) with parameters $(N_0, \mathbf{m}_0, \mathbf{V}_0)$. Note that setting $N_0 = 0$ leads to the uninformative prior which may often be the choice while modeling data. These parameters enter into the self-consistent equations for the parameters for the posterior as shown in (41) for $q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ and (44)-(47) for the equations for the parameters.

We now turn to the calculation of $q(\mathbf{Z})$. Equation (27), gives the result,

$$\log q^*(\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log\rho_{nk} + c, \tag{35}$$

where c is the logarithm of the partition function, and

$$\begin{aligned}\log\rho_{nk} &= -\frac{d}{2}\log 2\pi + \mathbb{E}_{\pi_k}[\log\pi_k] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}_k}[\log|\boldsymbol{\Lambda}_k|] \\ &\quad -\frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right]. \end{aligned} \tag{36}$$

From the properties of the Dirichlet distribution [cite], $\mathbb{E}_{\pi_k}[\log\pi_k] = \psi(N_k + 1) - \psi(N + K)$. From the properties of the Wishart distribution, with $\nu_k = N_0 + N_k$ effective degrees of freedom,

$$\mathbb{E}_{\boldsymbol{\Lambda}_k}[\log|\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{d}\psi\left(\frac{\nu_k + 1 - i}{2}\right) + d\log 2 + \log\left|\overline{\mathbf{W}}_k\right|. \tag{37}$$

The last term of (36) is the expectation value of the Mahalanobis distance squared between a data point and a component mean. To evaluat this term, we re-write that term as follows, where we have used $\langle\cdot\rangle$ to indicate the expectation value over $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ for brevity,

$$
\begin{aligned}
& \left\langle (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \right\rangle_{\boldsymbol{\Lambda}} \\
= \; & \mathrm{Tr} \left\langle \boldsymbol{\Lambda}_k (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \right\rangle_{\boldsymbol{\Lambda}} \\
= \; & \mathrm{Tr} \left\langle \boldsymbol{\Lambda}_k (\mathbf{x}_n - \mathbf{m}_k + \mathbf{m}_k - \boldsymbol{\mu}_k)(\mathbf{x}_n - \mathbf{m}_k + \mathbf{m}_k - \boldsymbol{\mu}_k)^T \right\rangle_{\boldsymbol{\Lambda}} \\
= \; & \mathrm{Tr} \left\langle \boldsymbol{\Lambda}_k (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T \right\rangle_{\boldsymbol{\Lambda}} + \mathrm{Tr} \left\langle \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_k)(\boldsymbol{\mu}_k - \mathbf{m}_k)^T \right\rangle_{\boldsymbol{\Lambda}} + 2\mathrm{Tr} \left\langle \boldsymbol{\Lambda}_k (\mathbf{x}_n - \mathbf{m}_k)(\boldsymbol{\mu}_k - \mathbf{m}_k)^T \right\rangle_{\boldsymbol{\Lambda}} \\
= \; & \mathrm{Tr} \left\langle \boldsymbol{\Lambda}_k (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T \right\rangle_{\boldsymbol{\Lambda}} + \mathrm{Tr} \left\langle \boldsymbol{\Lambda}_k \left\langle (\boldsymbol{\mu}_k - \mathbf{m}_k)(\boldsymbol{\mu}_k - \mathbf{m}_k)^T \right\rangle_{\boldsymbol{\mu}|\boldsymbol{\Lambda}} \right\rangle_{\boldsymbol{\Lambda}}, \\
= \; & \nu_k \mathrm{Tr} \left[\mathbf{W}_k (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T \right] + \frac{d}{N_0 + N_k} \\
= \; & \nu_k (\mathbf{x}_n - \mathbf{m}_k)^T \mathbf{W}_k (\mathbf{x}_n - \mathbf{m}_k) + \frac{d}{N_0 + N_k}
\end{aligned}
$$

Note that the cross term vanishes by (32). The first term is the square of the Mahalanobis distance between data point $\mathbf{x}_n$ and the *expectation value* of the mean parameter$\langle \boldsymbol{\mu}_k \rangle = \mathbf{m}_k$ with the expected value for the precision matrix equal to $\mathbf{W}_k$. It is thus the Mahalanobis distance to the center of mass of the component k. The prefactor ν_k accounts for the population that enters into evidence for this component.

The second term above is nothing but a scaled Mahalanobis distaance of a data point from the component mean. Here the expectation values are over the probability distributions for θ_k. Including the normalization, the final form of $q^*(\mathbf{Z})$ is given by (40), and (42) below.

To summarize this section, we found explicit expression for the mean-field solution (27) to the variational computation of the posterior distribution of the latent variables and model parameters. We did not make assumptions on the functional form of the mean field solutions, but only that the solution factorizes between the latent variables and the model parameters. The solution itself produced the classic functions of Dirichlet distribution for the occupancies π_k, and normal-Wishart solution the mean and variance parameters of GMM. The prior distributions are

$$
p(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}, \tag{38}
$$

$$
p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \left[\mathcal{N}\left(\boldsymbol{\mu}_k | \boldsymbol{m}_0, N_0^{-1} \boldsymbol{\Lambda}_0^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0\right) \right]. \tag{39}
$$

The posterior distributions are,

$$
q^*(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \gamma_{nk}^{z_{nk}}, \tag{40}
$$

$$
q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \left[\mathcal{N}\left(\boldsymbol{\mu}_k | \mathbf{m}_k, \left(N_k + N_0\right)^{-1} \boldsymbol{\Lambda}_k^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}_k | \mathbf{W}_k, \nu_0 + N_k + 1\right) \right]. \tag{41}
$$

The update equations to be solved self-consistently are,

$$\gamma_{nk} = \frac{\rho_{nk}}{\sum_j \rho_{nj}}, \tag{42}$$

$$N_k = \sum_{n=1}^{N} \gamma_{nk}, \tag{43}$$

$$\overline{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n, \tag{44}$$

$$\mathbf{m}_k = \frac{N_0 \boldsymbol{m}_0 + N_k \overline{\mathbf{x}}_k}{N_0 + N_k}, \tag{45}$$

$$\overline{\mathbf{V}}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \overline{\mathbf{x}}_k)(\mathbf{x}_n - \overline{\mathbf{x}}_k)^T, \tag{46}$$

$$\mathbf{W}_k^{-1} = \mathbf{W}_0^{-1} + N_k \overline{\mathbf{V}}_k + \frac{N_0 N_k}{N_0 + N_k} (\overline{\mathbf{x}}_k - \boldsymbol{m}_0)(\overline{\mathbf{x}}_k - \boldsymbol{m}_0)^T. \tag{47}$$

The responsibilites are calculated as,

$$\begin{aligned}
\alpha_k &= \alpha_0 + N_k, \\
\alpha &= \sum_{k=1}^{K} \alpha_k = N + K\alpha_0, \\
\nu_k &= \nu_0 + N_k, \\
\log \rho_{nk} &= \psi(\alpha_k) - \psi(N + \alpha_0 K) + \frac{1}{2} \mathbb{E}\left[\log |\mathbf{\Lambda}_k|\right] \\
&\quad - \frac{d}{2} \log 2\pi - \frac{\nu_k}{2} D_{nk} - \frac{1}{2} \frac{d}{N_0 + N_k}, \\
\mathbb{E}\left[\log |\mathbf{\Lambda}_k|\right] &= \sum_{i=1}^{d} \psi\left(\frac{N_0 + N_k + 1 - i}{2}\right) + d \log 2 + \log \left|\overline{\mathbf{W}}_k\right|, \\
D_{nk} &= (\mathbf{x}_n - \mathbf{m}_k)^T \mathbf{W}_k (\mathbf{x}_n - \mathbf{m}_k).
\end{aligned}$$

References

[1] Bishop, Christopher. Pattern Recognition and Machine Learning. Springer, New York (2006)

[2] Neal, R.M. and Hinton, G.E. A view of the EM algorithm that justifies incremental, sparse, and other variants. Learning in Graphical Models, 355-368 (1998). Kluwer Academic Press, Norwell, MA.

A Matrix derivatives

B Derivation of variational posterior

We define two expectation values,

$$\overline{\mathbf{x}}_k \equiv \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}\mathbf{x}_n, \tag{48}$$

$$\overline{\mathbf{V}}_k \equiv \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}(\mathbf{x}_n-\overline{\mathbf{x}}_k)(\mathbf{x}_n-\overline{\mathbf{x}}_k)^T. \tag{49}$$

We first re-write the sum over data samples in second term on the right hand side of (34) as,

$$\begin{aligned}
&= \sum_{n=1}^{N}\gamma_{nk}(\mathbf{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)\\
&= \sum_{n=1}^{N}\gamma_{nk}(\mathbf{x}_n-\overline{\mathbf{x}}_k-(\boldsymbol{\mu}_k-\overline{\mathbf{x}}_k))^T\boldsymbol{\Lambda}_k(\mathbf{x}_n-\overline{\mathbf{x}}_k-(\boldsymbol{\mu}_k-\overline{\mathbf{x}}_k))\\
&= \sum_{n=1}^{N}\gamma_{nk}(\mathbf{x}_n-\overline{\mathbf{x}}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x}_n-\overline{\mathbf{x}}_k)+N_k(\boldsymbol{\mu}_k-\overline{\mathbf{x}}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k-\overline{\mathbf{x}}_k)\\
&\quad -2\sum_{n=1}^{N}\gamma_{nk}(\boldsymbol{\mu}_k-\overline{\mathbf{x}}_k)^T\boldsymbol{\Lambda}_k\left(\mathbf{x}_n-\overline{\mathbf{x}}_k\right).
\end{aligned}$$

The last term vanishes, while the inner product in the first can be written as $\mathrm{Tr}\left[N_k\overline{\mathbf{V}}_k\boldsymbol{\Lambda}_k\right]$. Therefore, we get,

$$\begin{aligned}
\log q^*(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) &= c+\log p(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)+\frac{1}{2}\mathrm{Tr}\left[N_k\overline{\mathbf{V}}_k\boldsymbol{\Lambda}_k\right]\\
&\quad \frac{1}{2}\left[-dN_k\log 2\pi+N_k\log|\boldsymbol{\Lambda}_k|-(\boldsymbol{\mu}_k-\overline{\mathbf{x}}_k)^T\left(N_k\boldsymbol{\Lambda}_k\right)(\boldsymbol{\mu}_k-\overline{\mathbf{x}}_k)\right].
\end{aligned}$$

In the summand second line, we employed the relation $|\mathbf{A}^n|=|\mathbf{A}|^n$. The summand in the second term in the equation above is the logarithm of a Gaussian distribution. The summand in the first term is the exponent term in the Wishart distribution, where its normalization factors are absorbed into c. Again, modeling the prior to be the conjugate, we write,

$$\begin{aligned}
\log q^*(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) &= c-\frac{1}{2}\mathrm{Tr}\left[N_0\mathbf{V}_0\boldsymbol{\Lambda}_k+N_k\overline{\mathbf{V}}_k\boldsymbol{\Lambda}_k\right]+\frac{N_0+N_k}{2}\left[-d\log 2\pi+\log|\boldsymbol{\Lambda}_k|\right]\\
&\quad -\frac{1}{2}\left[N_0(\boldsymbol{\mu}_k-\boldsymbol{m}_0)^T\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k-\boldsymbol{m}_0)+N_k(\boldsymbol{\mu}_k-\overline{\mathbf{x}}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k-\overline{\mathbf{x}}_k)\right].
\end{aligned}$$

We define

$$\beta_k = N_0+N_k, \tag{50}$$

$$\mathbf{m}_k = \frac{N_0\boldsymbol{m}_0+N_k\overline{\mathbf{x}}_k}{N_0+N_k}, \tag{51}$$

and re-write the quadratic term as

$$\begin{aligned}
& N_0(\boldsymbol{\mu}_k - \mathbf{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_0) + N_k (\boldsymbol{\mu}_k - \overline{\mathbf{x}}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \overline{\mathbf{x}}_k) \\
= \; & (N_0 + N_k) \left[\boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k + \frac{N_0}{N_0 + N_k} \mathbf{m}_0^T \boldsymbol{\Lambda}_k \mathbf{m}_0 + \frac{N_k}{N_0 + N_k} \overline{\mathbf{x}}_k^T \boldsymbol{\Lambda}_k \overline{\mathbf{x}}_k - 2 \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \mathbf{m}_k \right] \\
= \; & (N_0 + N_k) (\boldsymbol{\mu}_k - \mathbf{m}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_k) \\
+ \; & N_0 \mathbf{m}_0^T \boldsymbol{\Lambda}_k \mathbf{m}_0 + N_k \overline{\mathbf{x}}_k^T \boldsymbol{\Lambda}_k \overline{\mathbf{x}}_k - (N_0 + N_k) \boldsymbol{m}_k^T \boldsymbol{\Lambda}_k \boldsymbol{m}_k .
\end{aligned}$$

The last term is,

$$\begin{aligned}
& N_0 \mathbf{m}_0^T \boldsymbol{\Lambda}_k \mathbf{m}_0 + N_k \overline{\mathbf{x}}_k^T \boldsymbol{\Lambda}_k \overline{\mathbf{x}}_k - (N_0 + N_k) \boldsymbol{m}_k^T \boldsymbol{\Lambda}_k \boldsymbol{m}_k \\
= \; & \left(N_0 - \frac{N_0^2}{N_0 + N_k} \right) \boldsymbol{m}_0^T \boldsymbol{\Lambda}_k \boldsymbol{m}_0 + \left(N_k - \frac{N_k^2}{N_0 + N_k} \right) \overline{\mathbf{x}}_k^T \boldsymbol{\Lambda}_k \overline{\mathbf{x}}_k - 2 \frac{N_0 N_k}{N_0 + N_k} \boldsymbol{m}_0^T \boldsymbol{\Lambda}_k \overline{\mathbf{x}}_k \\
= \; & \left(\frac{N_0 N_k}{N_0 + N_k} \right) \left[\boldsymbol{m}_0^T \boldsymbol{\Lambda}_k \boldsymbol{m}_0 + \overline{\mathbf{x}}_k^T \boldsymbol{\Lambda}_k \overline{\mathbf{x}}_k - 2 \boldsymbol{m}_0^T \boldsymbol{\Lambda}_k \overline{\mathbf{x}}_k \right] \\
= \; & \left(\frac{N_0 N_k}{N_0 + N_k} \right) (\boldsymbol{m}_0 - \overline{\mathbf{x}}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{m}_0 - \overline{\mathbf{x}}_k) .
\end{aligned}$$

Note that by the cyclic property of the trace, $\mathbf{a}^T \boldsymbol{\Lambda}_k \mathbf{a} = \mathrm{Tr}\left[\mathbf{a}\mathbf{a}^T \boldsymbol{\Lambda}_k \right]$. Putting all these results together, we get,

$$\begin{aligned}
\log q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \; = \; & c - \frac{1}{2} \mathrm{Tr} \left[\left\{ N_0 \mathbf{V}_0 + N_k \overline{\mathbf{V}}_k - \frac{N_0 N_k}{2\beta_k} (\boldsymbol{m}_0 - \overline{\mathbf{x}}_k) (\boldsymbol{m}_0 - \overline{\mathbf{x}}_k) \right\} \boldsymbol{\Lambda}_k \right] \\
& + \frac{\beta_k}{2} \left[-d \log 2\pi + \log |\boldsymbol{\Lambda}_k| \right] - \frac{1}{2} (\boldsymbol{\mu}_k - \mathbf{m}_k)^T (\beta_k \boldsymbol{\Lambda}_k) (\boldsymbol{\mu}_k - \mathbf{m}_k) .
\end{aligned}$$