Tutorial 8: Variational Inference on Gaussian Mixture Models using Python

DD2420 Probabilistic Graphical Models

Márk Antal Csizmadia
macs@kth.se

March 2021

1 Introduction

In this tutorial, the variational inference algorithm is implemented in the case of a Bayesian Multivariate Gaussian Mixture Model (GMM). The probabilistic graphical model is shown in 1, and the distributions of the random variables are listed below.

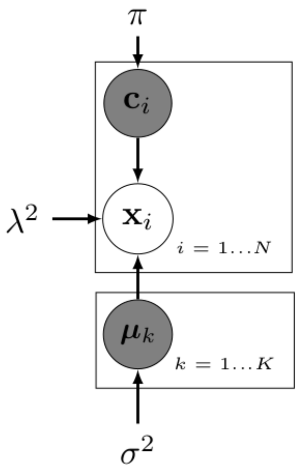

Figure 1: PGM of Bayesian Multivariate Gaussian Mixture Model with K components and N samples.

where

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\alpha}, \sigma^2\mathbf{I}) \qquad k = 1, ..., K \tag{1.1}$$

$$c_i \sim Categorical(\frac{1}{K}, ..., \frac{1}{K}) \qquad i = 1, ..., N \tag{1.2}$$

$$\mathbf{x}_i | c_i, \boldsymbol{\mu} \sim \mathcal{N}(c_i^T \boldsymbol{\mu}, \lambda^2\mathbf{I}) \qquad i = 1, ..., N \tag{1.3}$$

and the data are p-dimensional such that

$$\boldsymbol{\mu}_k \in \mathbb{R}^p, \mathbf{x}_i \in \mathbb{R}^p \tag{1.4}$$

To approximate the posterior over latent variables $\boldsymbol{\mu}, \mathbf{c}$ we can use mean-field approximation of the form

$$p(\boldsymbol{\mu}, \mathbf{c}) \approx q(\boldsymbol{\mu}, \mathbf{c}) = \prod_{k=1}^{K} q(\boldsymbol{\mu}_k) \prod_{i=1}^{N} q(c_i) \tag{1.5}$$

where

$$q(\boldsymbol{\mu}_k) = \mathcal{N}\left(\boldsymbol{\mu}_k | \mathbf{m}_k, s_k^2 \mathbf{I}\right) \tag{1.6}$$

and

$$q(c_i) = Categorical(\phi_i) \tag{1.7}$$

The variational parameters $\mathbf{m}_k, s_k^2$ and $\boldsymbol{\phi}_i = \{\phi_{i,1}, \phi_{i,2}, ..., \phi_{i,K}\}$ will be optimized with the mean-field coordinate ascend variational inference (CAVI) algorithm such that

$$q_j^*(z_j) \propto \exp\left(\mathbb{E}_{-j}\left[\log p\left(z_j, \mathbf{z}_{-j}, \mathbf{x}\right)\right]\right) \tag{1.8}$$

2 Assingment 1

Compute ELBO in closed form (by plugging in all distributions and taking expectations under the approximate distribution). This result will be later used to track the convergence of the model and compare different runs of optimization.

The ELBO of the Gaussian Mixture Model (GMM) is given as

$$\begin{aligned}
&\mathcal{L}(\mathbf{x}|\mathbf{m}, \mathbf{s}^2, \boldsymbol{\phi}) \\
&= \mathbb{E}_q\left[\log p\left(\mathbf{x}, \boldsymbol{\mu}, \mathbf{c}\right)\right] - \mathbb{E}_q\left[\log q\left(\boldsymbol{\mu}, \mathbf{c}\right)\right] \\
&= \mathbb{E}_q\left[\sum_{k=1}^{K} \log p(\boldsymbol{\mu}_k) + \sum_{i=1}^{N} \left(\log p(c_i) + \log p(\mathbf{x}_i|c_i, \boldsymbol{\mu})\right)\right] \\
&- \mathbb{E}_q\left[\sum_{k=1}^{K} \log q(\boldsymbol{\mu}_k) + \sum_{i=1}^{N} \log q(c_i)\right]
\end{aligned} \tag{2.1}$$

$$\begin{aligned}
&= \sum_{k=1}^{K} \mathbb{E}_q[\log p(\boldsymbol{\mu}_k)] + \sum_{i=1}^{N} \mathbb{E}_q[\log p(c_i)] + \sum_{i=1}^{N} \mathbb{E}_q[\log p(\mathbf{x}_i|c_i, \boldsymbol{\mu})] \\
&- \sum_{k=1}^{K} \mathbb{E}_q[\log q(\boldsymbol{\mu}_k)] - \sum_{i=1}^{N} \mathbb{E}_q[\log q(c_i)]
\end{aligned}$$

By plugging in all distributions and taking expectations under the approximate distribution, the closed form of the ELBO is computed.

$$\begin{aligned}
\mathcal{L}(\mathbf{x}|\mathbf{m}, \mathbf{s}^2, \boldsymbol{\phi}) =& \sum_{k=1}^{K} \mathbb{E}_q\left[\log \mathcal{N}(\boldsymbol{\alpha}, \sigma^2\mathbf{I}))\right] \\
&+ \sum_{i=1}^{N} \mathbb{E}_q\left[\log\, Categorical\left(\frac{1}{K}, ..., \frac{1}{K}\right)\right] \\
&+ \sum_{i=1}^{N} \mathbb{E}_q\left[\log \mathcal{N}(c_i^T\boldsymbol{\mu}, \lambda^2\mathbf{I})\right] \\
&- \sum_{k=1}^{K} \mathbb{E}_q[\log \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_k, s_k^2\mathbf{I}\right))] \\
&- \sum_{i=1}^{N} \mathbb{E}_q[\log\, Categorical(\phi_i)]
\end{aligned} \tag{2.2}$$

The closed form of the first term in Equation 2.2 on the right-hand side is as follows

$$
\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}_q \left[\log \mathcal{N}(\boldsymbol{\alpha}, \sigma^2 \mathbf{I}))\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_q \left[\log \left((2\pi)^{-\frac{p}{2}} \det\left(\sigma^2 \mathbf{I}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha}\right)^T \left(\sigma^2 \mathbf{I}\right)^{-1} \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha}\right)\right)\right)\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_q \left[-\frac{p}{2} \log\left(2\pi\sigma^2\right) - \frac{\sigma^{-2}}{2} \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha}\right)^T \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha}\right)\right] \\
&= \sum_{k=1}^{K} -\frac{p}{2} \log\left(2\pi\sigma^2\right) - \frac{\sigma^{-2}}{2} \mathbb{E}_q \left[\left(\boldsymbol{\mu}_k - \boldsymbol{\alpha}\right)^T \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha}\right)\right] \\
&= -\frac{1}{2} \sum_{k=1}^{K} p \log\left(2\pi\sigma^2\right) + \sigma^{-2} \mathbb{E}_q \left[\left(\boldsymbol{\mu}_k - \boldsymbol{\alpha}\right)^T \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha}\right)\right] \\
&= -\frac{1}{2} \left[pK \log\left(2\pi\sigma^2\right) + \sigma^{-2} \sum_{k=1}^{K} \left[p s_k^2 + \mathbf{m}_k^T \mathbf{m}_k - \mathbf{m}_k^T \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \mathbf{m}_k + \boldsymbol{\alpha}^T \boldsymbol{\alpha} \right] \right]
\end{aligned}
\tag{2.3}
$$

From the first to the second line, the probability density function of the multivariate normal distribution was plugged in. From the second to the third line, the log of the terms was taken and it was note that $\det\left(\sigma^2 \mathbf{I}\right)^{-\frac{1}{2}} = \sigma^{-p}$. From the third to the fourth line, the expectation was factored into the product, keeping in mind that the variational factor $q(\boldsymbol{\mu}_k)$ is only the function of $\boldsymbol{\mu}_k$ but any of the other parameters. From the fourth to the fifth line, the equation was contracted. From the fifth to the last line, the sum was factored in and the following was used

$$
\begin{aligned}
\mathbb{E}_q \left[\left(\boldsymbol{\mu}_k - \boldsymbol{\alpha}\right)^T \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha}\right)\right] &= \mathbb{E}_q \left[\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \boldsymbol{\mu}_k + \boldsymbol{\alpha}^T \boldsymbol{\alpha} \right] \\
&= \mathbb{E}_q \left[\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k \right] - \mathbb{E}_q \left[\boldsymbol{\mu}_k^T \boldsymbol{\alpha} \right] - \mathbb{E}_q \left[\boldsymbol{\alpha}^T \boldsymbol{\mu}_k \right] + \mathbb{E}_q \left[\boldsymbol{\alpha}^T \boldsymbol{\alpha} \right] \\
&= p s_k^2 + \mathbf{m}_k^T \mathbf{m}_k - \mathbf{m}_k^T \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \mathbf{m}_k + \boldsymbol{\alpha}^T \boldsymbol{\alpha}
\end{aligned}
\tag{2.4}
$$

where it is noted that $\mathbb{E}_q \left[\boldsymbol{\mu}_k\right] = \mathbf{m}_k$, and $\mathbb{E}_q \left[\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k\right] = p \ Var(\boldsymbol{\mu}_k) + \mathbb{E}\left[\boldsymbol{\mu}_k^T\right] \mathbb{E}\left[\boldsymbol{\mu}_k\right] = p s_k^2 + \mathbf{m}_k^T \mathbf{m}_k$ as $\boldsymbol{\mu}_k \in \mathbb{R}^p$.

The closed form of the second term in Equation 2.2 on the right-hand side is as follows

$$
\begin{aligned}
\sum_{i=1}^{N} \mathbb{E}_q \left[\log Categorical \left(\frac{1}{K}, ..., \frac{1}{K} \right) \right] &= \sum_{i=1}^{N} \mathbb{E}_q \left[\log \prod_{k=1}^{K} \left(\frac{1}{K} \right)^{c_{i,k}} \right] \\
&= \sum_{i=1}^{N} \mathbb{E}_q \left[\sum_{k=1}^{K} \log \left(\frac{1}{K} \right)^{c_{i,k}} \right] \\
&= -\sum_{i=1}^{N} \mathbb{E}_q \left[\sum_{k=1}^{K} c_{i,k} \log \left(K \right) \right] \\
&= -\sum_{i=1}^{N} \log \left(K \right) \sum_{k=1}^{K} \mathbb{E}_q \left[c_{i,k} \right] \\
&= -\sum_{i=1}^{N} \log \left(K \right) \sum_{k=1}^{K} \phi_{i,k} \\
&= -\sum_{i=1}^{N} \log \left(K \right) \\
&= -N \log \left(K \right)
\end{aligned}
\tag{2.5}
$$

where in the first line $c_{i,k}$ is a cluster assignment, $\mathbb{E}_q \left[c_{i,k} \right] = \phi_{i,k}$ is the expectation of the cluster assignment variable, and $\sum_{k=1}^{K} \phi_{i,k} = 1$ as $\phi_{i,k}$ defines a categorical distribution.

To derive the closed form of the third term in Equation 2.2 on the right-hand side, the hint from Assignment 2 in this report is used, that is $p \left(\mathbf{x}_i | c_i, \boldsymbol{\mu} \right) = \prod_{k=1}^{K} p \left(\mathbf{x}_i | \boldsymbol{\mu}_k \right)^{c_{i,k}}$

$$
\begin{aligned}
\sum_{i=1}^{N} \mathbb{E}_q [\log p(\mathbf{x}_i | c_i, \boldsymbol{\mu})] &= \sum_{i=1}^{N} \mathbb{E}_q \left[\log \prod_{k=1}^{K} p(\mathbf{x}_i | \boldsymbol{\mu_k})^{c_{i,k}} \right] \\
&= \sum_{i=1}^{N} \mathbb{E}_q \left[\sum_{k=1}^{K} c_{i,k} \log p(\mathbf{x}_i | \boldsymbol{\mu_k}) \right] \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{E}_q \left[c_{i,k} \right] \mathbb{E}_q \left[\log p(\mathbf{x}_i | \boldsymbol{\mu_k}) \right] \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \phi_{i,k} \, \mathbb{E}_q \left[\log \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu_k}, \lambda^2 \mathbf{I}) \right]
\end{aligned}
\tag{2.6}
$$

where

$$
\begin{aligned}
&\mathbb{E}_q \left[\log \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu_k}, \lambda^2 \mathbf{I}) \right] \\
&= \mathbb{E}_q \left[\log \left((2\pi)^{-\frac{p}{2}} \det \left(\lambda^2 \mathbf{I} \right)^{-\frac{1}{2}} \exp \left(-\frac{1}{2} \left(\mathbf{x}_i - \boldsymbol{\mu}_k \right)^T \left(\lambda^2 \mathbf{I} \right)^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k \right) \right) \right) \right] \\
&= \mathbb{E}_q \left[-\frac{p}{2} \log \left(2\pi \right) - \frac{p}{2} \log \left(\lambda^2 \right) - \frac{\lambda^{-2}}{2} \left(\mathbf{x}_i - \boldsymbol{\mu}_k \right)^T \left(\mathbf{x}_i - \boldsymbol{\mu}_k \right) \right] \\
&= -\frac{p}{2} \log \left(2\pi\lambda^2 \right) - \frac{\lambda^{-2}}{2} \mathbb{E}_q \left[\left(\mathbf{x}_i - \boldsymbol{\mu}_k \right)^T \left(\mathbf{x}_i - \boldsymbol{\mu}_k \right) \right]
\end{aligned}
\tag{2.7}
$$

where it was noted that $\log\left(\det(\lambda^2\mathbf{I})^{-\frac{1}{2}}\right) = \log\left((\lambda^{2p})^{-\frac{1}{2}}\right) = \log(\lambda^{-p}) = -\frac{p}{2}\log(\lambda^2)$, and the last expectation term was derived as

$$\begin{aligned}\mathbb{E}_q\left[(\mathbf{x}_i-\boldsymbol{\mu}_k)^T(\mathbf{x}_i-\boldsymbol{\mu}_k)\right] &= \mathbb{E}_q\left[\mathbf{x}_i^T\mathbf{x}_i - \mathbf{x}_i^T\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T\mathbf{x}_i + \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right]\\ &= \mathbb{E}_q\left[\mathbf{x}_i^T\mathbf{x}_i\right] - \mathbb{E}_q\left[\mathbf{x}_i^T\boldsymbol{\mu}_k\right] - \mathbb{E}_q\left[\boldsymbol{\mu}_k^T\mathbf{x}_i\right] + \mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right]\\ &= \mathbf{x}_i^T\mathbf{x}_i - \mathbf{x}_i^T\mathbf{m}_k - \mathbf{m}_k^T\mathbf{x}_i + s_k^2p + \mathbf{m}_k^T\mathbf{m}_k\end{aligned} \tag{2.8}$$

where it is noted that $\mathbb{E}_q[\boldsymbol{\mu}_k] = \mathbf{m}_k$, and $\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right] = p\ Var(\boldsymbol{\mu}_k) + \mathbb{E}\left[\boldsymbol{\mu}_k^T\right]\mathbb{E}[\boldsymbol{\mu}_k] = ps_k^2 + \mathbf{m}_k^T\mathbf{m}_k$ as $\boldsymbol{\mu}_k \in \mathbb{R}^p$.

As a result, the third term in Equation 2.2 is

$$\begin{aligned}&\sum_{i=1}^{N}\mathbb{E}_q[\log p(\mathbf{x}_i|c_i,\boldsymbol{\mu})]\\ &= \sum_{i=1}^{N}\mathbb{E}_q\left[\log\mathcal{N}(c_i^T\boldsymbol{\mu},\lambda^2\mathbf{I})\right]\\ &= \sum_{i=1}^{N}\sum_{k=1}^{K}\phi_{i,k}\left(-\frac{p}{2}\log(2\pi\lambda^2) - \frac{\lambda^{-2}}{2}\left(\mathbf{x}_i^T\mathbf{x}_i - \mathbf{x}_i^T\mathbf{m}_k - \mathbf{m}_k^T\mathbf{x}_i + s_k^2p + \mathbf{m}_k^T\mathbf{m}_k\right)\right)\end{aligned} \tag{2.9}$$

The closed form of the fourth term in Equation 2.2 on the right-hand side is as follows

$$\begin{aligned}&\sum_{k=1}^{K}\mathbb{E}_q[\log\mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_k,s_k^2\mathbf{I})\right)]\\ &= \sum_{k=1}^{K}\mathbb{E}_q\left[\log\left((2\pi)^{-\frac{p}{2}}\det\left(s_k^2\mathbf{I}\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\mu}_k-\mathbf{m}_k)^T\left(s_k^2\mathbf{I}\right)^{-1}(\boldsymbol{\mu}_k-\mathbf{m}_k)\right)\right)\right]\\ &= \sum_{k=1}^{K}\mathbb{E}_q\left[-\frac{p}{2}\log(2\pi) - p\log(s_k) - \frac{s_k^{-2}}{2}(\boldsymbol{\mu}_k-\mathbf{m}_k)^T(\boldsymbol{\mu}_k-\mathbf{m}_k)\right]\\ &= \sum_{k=1}^{K}-\frac{p}{2}\log(2\pi) - p\log(s_k) - \frac{s_k^{-2}}{2}\mathbb{E}_q\left[(\boldsymbol{\mu}_k-\mathbf{m}_k)^T(\boldsymbol{\mu}_k-\mathbf{m}_k)\right]\end{aligned} \tag{2.10}$$

where it was noted that $\log\left(\det(s_k^2\mathbf{I})^{-\frac{1}{2}}\right) = \log\left(\left(s_k^{2p}\right)^{-\frac{1}{2}}\right) = \log\left(s_k^{-p}\right) = -p\log(s_k)$, and the last expectation term was derived as

$$\begin{aligned}\mathbb{E}_q\left[(\boldsymbol{\mu}_k-\mathbf{m}_k)^T(\boldsymbol{\mu}_k-\mathbf{m}_k)\right] &= \mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T\mathbf{m}_k - \mathbf{m}_k^T\boldsymbol{\mu}_k + \mathbf{m}_k^T\mathbf{m}_k\right]\\ &= \mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right] - \mathbb{E}_q\left[\boldsymbol{\mu}_k^T\mathbf{m}_k\right] - \mathbb{E}_q\left[\mathbf{m}_k^T\boldsymbol{\mu}_k\right] + \mathbb{E}_q\left[\mathbf{m}_k^T\mathbf{m}_k\right]\\ &= s_k^2p + \mathbf{m}_k^T\mathbf{m}_k - \mathbf{m}_k^T\mathbf{m}_k - \mathbf{m}_k^T\mathbf{m}_k + \mathbf{m}_k^T\mathbf{m}_k\\ &= s_k^2p\end{aligned} \tag{2.11}$$

where it is noted that $\mathbb{E}_q\left[\boldsymbol{\mu}_k\right] = \mathbf{m}_k$, and $\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right] = p\ Var(\boldsymbol{\mu}_k) + \mathbb{E}\left[\boldsymbol{\mu}_k^T\right]\mathbb{E}\left[\boldsymbol{\mu}_k\right] = ps_k^2 + \mathbf{m}_k^T\mathbf{m}_k$ as $\boldsymbol{\mu}_k \in \mathbb{R}^p$.

As a result, the fourth term in Equation 2.2 is

$$\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}_q[\log\mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_k, s_k^2\mathbf{I}\right))] \\
&= \sum_{k=1}^{K} -\frac{p}{2}\log\left(2\pi\right) - p\log\left(s_k\right) - \frac{s_k^{-2}}{2}\mathbb{E}_q\left[\left(\boldsymbol{\mu}_k - \mathbf{m}_k\right)^T\left(\boldsymbol{\mu}_k - \mathbf{m}_k\right)\right] \\
&= \sum_{k=1}^{K} -\frac{p}{2}\log\left(2\pi\right) - p\log\left(s_k\right) - \frac{s_k^{-2}}{2}s_k^2 p \qquad (2.12) \\
&= \sum_{k=1}^{K} -\frac{p}{2}\log\left(2\pi\right) - p\log\left(s_k\right) - \frac{p}{2} \\
&= -\frac{p}{2}\sum_{k=1}^{K}\left(\log\left(2\pi\right) + 2\log\left(s_k\right) + 1\right)
\end{aligned}$$

The closed form of the fifth and last term in Equation 2.2 on the right-hand side is as follows

$$\begin{aligned}
\sum_{i=1}^{N}\mathbb{E}_q[\log Categorical(\phi_i)] &= \sum_{i=1}^{N}\mathbb{E}_q\left[\log\prod_{k=1}^{K}\left(\phi_i^k\right)^{c_{i,k}}\right] \\
&= \sum_{i=1}^{N}\mathbb{E}_q\left[\sum_{k=1}^{K}\log\left(\phi_i^k\right)^{c_{i,k}}\right] \\
&= \sum_{i=1}^{N}\mathbb{E}_q\left[\sum_{k=1}^{K}c_{i,k}\log\left(\phi_i^k\right)\right] \qquad (2.13) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}_q\left[c_{i,k}\right]\log\left(\phi_i^k\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\phi_i^k\log\left(\phi_i^k\right)
\end{aligned}$$

where is was noted that $\mathbb{E}_q\left[c_{i,k}\right] = \phi_i^k$.

To summarize, using the closed-form derivations, the ELBO is

$$
\begin{aligned}
&\mathcal{L}(\mathbf{x}|\mathbf{m}, \mathrm{s}^2, \boldsymbol{\phi}) \\
&= -\frac{1}{2}\left[pK \log\left(2\pi\sigma^2\right) + \sigma^{-2}\sum_{k=1}^{K}\left[ps_k^2 + \mathbf{m}_k^T\mathbf{m}_k - \mathbf{m}_k^T\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\mathbf{m}_k + \boldsymbol{\alpha}^T\boldsymbol{\alpha}\right]\right] \\
&- N\log\left(K\right) \\
&+ \sum_{i=1}^{N}\sum_{k=1}^{K}\phi_{i,k}\left(-\frac{p}{2}\log\left(2\pi\lambda^2\right) - \frac{\lambda^{-2}}{2}\left(\mathbf{x}_i^T\mathbf{x}_i - \mathbf{x}_i^T\mathbf{m}_k - \mathbf{m}_k^T\mathbf{x}_i + s_k^2 p + \mathbf{m}_k^T\mathbf{m}_k\right)\right) \qquad (2.14) \\
&+ \frac{p}{2}\sum_{k=1}^{K}\left(\log\left(2\pi\right) + 2\log\left(s_k\right) + 1\right) \\
&- \sum_{i=1}^{N}\sum_{k=1}^{K}\phi_{i,k}\log\left(\phi_{i,k}\right)
\end{aligned}
$$

The mathematical derivations were partly based on [1], [2], and [3].

3 Assingment 2

Show that the variational update for i-th cluster assignment is

$$\phi_{i,k} \propto \exp\left(\frac{\mathbf{x}_i^T\mathbb{E}\left[\boldsymbol{\mu}_k\right]}{\lambda^2} - \frac{\mathbb{E}\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right]}{2\lambda^2}\right) \qquad (3.1)$$

Hint. Use the fact that

$$p\left(\mathbf{x}_i|c_i, \boldsymbol{\mu}\right) = \prod_{k=1}^{K} p\left(\mathbf{x}_i|\boldsymbol{\mu}_k\right)^{c_{i,k}} \qquad (3.2)$$

The variational distribution parameter updates result from minimizing the KL divergence between the mean-field approximation and the true posterior, or equivalently, maximizing the ELBO. The functional form of each variational factor $q_j(\theta_j)$ is the ELBO, derived in Assignment 1. Taking the functional derivative of the ELBO with respect to a variational factor $q_j(\theta_j)$ is analogous to taking the expectation of the ELBO over the variational factor $q_j(\theta_j)$ while all of the other variational factors $q_{-j}(\theta_{-j})$ are kept fixed. To do this, the hint was already used previously in solving Assignment 1.

Taking the derivative of the closed-form ELBO in Equation 2.14 with respect to $\phi_{i,k}$, it is noted that only the third and the fifth terms are functions of $\phi_{i,k}$, and therefore only these terms have non-zero derivatives with respect to $\phi_{i,k}$.

$$
\begin{aligned}
&\frac{\partial \mathcal{L}(\mathbf{x}|\mathbf{m},\mathbf{s}^2,\boldsymbol{\phi})}{\partial \phi_{i,k}} \\
&= \frac{\partial \left(\sum_{i=1}^{N}\sum_{k=1}^{K}\phi_{i,k}\left(-\frac{p}{2}\log\left(2\pi\lambda^2\right)-\frac{\lambda^{-2}}{2}\left(\mathbf{x}_i^T\mathbf{x}_i-\mathbf{x}_i^T\mathbf{m}_k-\mathbf{m}_k^T\mathbf{x}_i+s_k^2p+\mathbf{m}_k^T\mathbf{m}_k\right)\right)\right)}{\partial \phi_{i,k}} \\
&+ \frac{\partial\left(-\sum_{i=1}^{N}\sum_{k=1}^{K}\phi_{i,k}\log\left(\phi_{i,k}\right)\right)}{\partial \phi_{i,k}} \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\left(-\frac{p}{2}\log\left(2\pi\lambda^2\right)-\frac{\lambda^{-2}}{2}\left(\mathbf{x}_i^T\mathbf{x}_i-\mathbf{x}_i^T\mathbf{m}_k-\mathbf{m}_k^T\mathbf{x}_i+s_k^2p+\mathbf{m}_k^T\mathbf{m}_k\right)\right) \\
&- \sum_{i=1}^{N}\sum_{k=1}^{K}\left(\log\left(\phi_{i,k}\right)+1\right)
\end{aligned}
\tag{3.3}
$$

where it was noted that $\frac{\partial\left(-\sum_{i=1}^{N}\sum_{k=1}^{K}\phi_{i,k}\log\left(\phi_{i,k}\right)\right)}{\partial \phi_{i,k}} = -\sum_{i=1}^{N}\sum_{k=1}^{K}\left(1\times\log\left(\phi_{i,k}\right)+\phi_{i,k}\times\frac{1}{\phi_{i,k}}\right)$
As mentioned, minimizing the KL divergence is equivalent to maximizing the ELBO, that in turn implies that we are seeking the parameter $\phi_{i,k}$ that sets the derivative to zero. Therefore, the argument of the double summation has to be equal to zero, that is

$$
\begin{aligned}
&-\frac{p}{2}\log\left(2\pi\lambda^2\right)-\frac{\lambda^{-2}}{2}\left(\mathbf{x}_i^T\mathbf{x}_i-\mathbf{x}_i^T\mathbf{m}_k-\mathbf{m}_k^T\mathbf{x}_i+s_k^2p+\mathbf{m}_k^T\mathbf{m}_k\right) \\
&-\left(\log\left(\phi_{i,k}\right)+1\right)=0
\end{aligned}
\tag{3.4}
$$

Rearranging to solve for $\log\left(\phi_{i,k}\right)$ yields

$$
\begin{aligned}
&\log\left(\phi_{i,k}\right) \\
&=-\frac{p}{2}\log\left(2\pi\lambda^2\right)-\frac{\lambda^{-2}}{2}\left(\mathbf{x}_i^T\mathbf{x}_i-\mathbf{x}_i^T\mathbf{m}_k-\mathbf{m}_k^T\mathbf{x}_i+s_k^2p+\mathbf{m}_k^T\mathbf{m}_k\right)-1
\end{aligned}
\tag{3.5}
$$

Extending and regrouping the terms, and using that $\mathbb{E}_q\left[\boldsymbol{\mu}_k\right]=\mathbf{m}_k$, $\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right]=p\,Var(\boldsymbol{\mu}_k)+\mathbb{E}\left[\boldsymbol{\mu}_k^T\right]\mathbb{E}\left[\boldsymbol{\mu}_k\right]=ps_k^2+\mathbf{m}_k^T\mathbf{m}_k$ as $\boldsymbol{\mu}_k\in\mathbb{R}^p$, and $\mathbf{x}_i^T\mathbf{m}_k=\mathbf{m}_k^T\mathbf{x}_i$ yields

$$
\begin{aligned}
\log\left(\phi_{i,k}\right)&=-\frac{p}{2}\log\left(2\pi\lambda^2\right)-1-\frac{\lambda^{-2}}{2}\mathbf{x}_i^T\mathbf{x}_i+2\frac{\lambda^{-2}}{2}\mathbf{x}_i^T\mathbb{E}_q\left[\boldsymbol{\mu}_k\right]-\frac{\lambda^{-2}}{2}\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right] \\
&=-\frac{p}{2}\log\left(2\pi\lambda^2\right)-1-\frac{\lambda^{-2}}{2}\mathbf{x}_i^T\mathbf{x}_i+\lambda^{-2}\mathbf{x}_i^T\mathbb{E}_q\left[\boldsymbol{\mu}_k\right]-\frac{\lambda^{-2}}{2}\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right]
\end{aligned}
\tag{3.6}
$$

Taking the exponential of both sides then yields

$$
\phi_{i,k}=\exp\left(-\frac{p}{2}\log\left(2\pi\lambda^2\right)-1-\frac{\lambda^{-2}}{2}\mathbf{x}_i^T\mathbf{x}_i\right)\times\exp\left(\lambda^{-2}\mathbf{x}_i^T\mathbb{E}_q\left[\boldsymbol{\mu}_k\right]-\frac{\lambda^{-2}}{2}\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right]\right)
\tag{3.7}
$$

That shows that

$$\phi_{i,k} \propto \exp\left(\lambda^{-2}\mathbf{x}_i^T\mathbb{E}_q\left[\boldsymbol{\mu}_k\right] - \frac{\lambda^{-2}}{2}\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right]\right) \tag{3.8}$$

up to a multiplicative constant that is equal to $\exp\left(-\frac{p}{2}\log\left(2\pi\lambda^2\right) - 1 - \frac{\lambda^{-2}}{2}\mathbf{x}_i^T\mathbf{x}_i\right)$.

The mathematical derivation above, and in particular, taking the derivative of the lower bound, was based on [2].

4 Assingment 3

Complete the square to find the parameters of the optimal Gaussian $\boldsymbol{\mu}_k \sim \mathcal{N}\left(\mathbf{m}_k, s_k^2\mathbf{I}\right)$. Those parameters will be used for variational updates of the posterior of the mixture component means.

Based on Equation 1.8, we have that

$$q^*\left(\boldsymbol{\mu}_k\right) \propto \exp\left(\log p\left(\boldsymbol{\mu}_k\right) + \sum_{i=1}^{N}\mathbb{E}_{q\left(\boldsymbol{\mu}_{-k},c_i\right)}\left[\log p\left(\mathbf{x}_i|c_i,\boldsymbol{\mu}\right)\right]\right) \tag{4.1}$$

where $q^*\left(\boldsymbol{\mu}_k\right)$ itself is a Gaussian. Therefore, comparing the terms in the exponents of the CAVI expression above and the Gaussian varitaional factor distribution, the parameter updates are found.

The exponent of the CAVI equation is as follows

$$\begin{aligned}
&\log p\left(\boldsymbol{\mu}_k\right) + \sum_{i=1}^{N}\mathbb{E}_{q\left(\boldsymbol{\mu}_{-k},c_i\right)}\left[\log p\left(\mathbf{x}_i|c_i,\boldsymbol{\mu}\right)\right] \\
&= -\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2\sigma^2} + \frac{\boldsymbol{\mu}_k^T\boldsymbol{\alpha}}{\sigma^2} + \sum_{i=1}^{N}\phi_{i,k}\left(-\frac{1}{2\lambda^2}\left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T\left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)\right) + \text{const} \\
&= -\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2\sigma^2} + \frac{\boldsymbol{\mu}_k^T\boldsymbol{\alpha}}{\sigma^2} - \frac{1}{2\lambda^2}\sum_{i=1}^{N}\phi_{i,k}\left(\mathbf{x}_i^T\mathbf{x}_i - \mathbf{x}_i^T\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T\mathbf{x}_i + \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right) + \text{const} \\
&= -\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2\sigma^2} + \frac{\boldsymbol{\mu}_k^T\boldsymbol{\alpha}}{\sigma^2} - \frac{1}{2\lambda^2}\sum_{i=1}^{N}\phi_{i,k}\left(\mathbf{x}_i^T\mathbf{x}_i - 2\boldsymbol{\mu}_k^T\mathbf{x}_i + \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right) + \text{const} \\
&= -\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2\sigma^2} + \frac{\boldsymbol{\mu}_k^T\boldsymbol{\alpha}}{\sigma^2} - \frac{1}{2\lambda^2}\sum_{i=1}^{N}\phi_{i,k}\mathbf{x}_i^T\mathbf{x}_i + \boldsymbol{\mu}_k^T\frac{1}{\lambda^2}\sum_{i=1}^{N}\phi_{i,k}\mathbf{x}_i - \frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2}\frac{1}{\lambda^2}\sum_{i=1}^{N}\phi_{i,k} + \text{const} \\
&= -\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\lambda^2}\sum_{i=1}^{N}\phi_{i,k}\right) + \boldsymbol{\mu}_k^T\left(\frac{\boldsymbol{\alpha}}{\sigma^2} + \frac{1}{\lambda^2}\sum_{i=1}^{N}\phi_{i,k}\mathbf{x}_i\right) - \frac{1}{2\lambda^2}\sum_{i=1}^{N}\phi_{i,k}\mathbf{x}_i^T\mathbf{x}_i + \text{const} \\
&= -\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\lambda^2}\sum_{i=1}^{N}\phi_{i,k}\right) + \boldsymbol{\mu}_k^T\left(\frac{\boldsymbol{\alpha}}{\sigma^2} + \frac{1}{\lambda^2}\sum_{i=1}^{N}\phi_{i,k}\mathbf{x}_i\right) + \text{const}
\end{aligned} \tag{4.2}$$

where it was noted that $\mathbf{x}_i^T \boldsymbol{\mu}_k = \boldsymbol{\mu}_k^T \mathbf{x}_i$, and the term $\frac{\boldsymbol{\mu}_k^T \boldsymbol{\alpha}}{\sigma^2}$ was added as a correction to the tutorial guide. The correction was based on the fact that this terms also depends on $\boldsymbol{\mu}_k$ based on Equations 2.3 and 2.4, and therefore has to be included. It is important as this term contributes to the mean update of the mean posterior $\mathbf{m}_k$.

The log of the variational factor distribution is

$$\begin{aligned}
\log q^* \left(\boldsymbol{\mu}_k\right) &= \log \mathcal{N}\left(\boldsymbol{\mu}_k | \mathbf{m}_k, s_k^2 \mathbf{I}\right) \\
&= \log \left((2\pi)^{-\frac{p}{2}} \det \left(s_k^2 \mathbf{I}\right)^{-\frac{1}{2}} \exp \left(-\frac{1}{2} \left(\boldsymbol{\mu}_k - \mathbf{m}_k\right)^T \left(s_k^2 \mathbf{I}\right)^{-1} \left(\boldsymbol{\mu}_k - \mathbf{m}_k\right) \right) \right) \\
&= -\frac{p}{2} \log\left(2\pi\right) - p \log\left(s_k\right) - \frac{s_k^{-2}}{2} \left(\boldsymbol{\mu}_k - \mathbf{m}_k\right)^T \left(\boldsymbol{\mu}_k - \mathbf{m}_k\right) \\
&= -\frac{p}{2} \log\left(2\pi\right) - p \log\left(s_k\right) - \frac{s_k^{-2}}{2} \left(\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T \mathbf{m}_k - \mathbf{m}_k^T \boldsymbol{\mu}_k + \mathbf{m}_k^T \mathbf{m}_k\right) \\
&= -\frac{p}{2} \log\left(2\pi\right) - p \log\left(s_k\right) - \frac{s_k^{-2}}{2} \boldsymbol{\mu}_k^T \boldsymbol{\mu}_k + s_k^{-2} \boldsymbol{\mu}_k^T \mathbf{m}_k - \frac{s_k^{-2}}{2} \mathbf{m}_k^T \mathbf{m}_k \\
&= -\frac{\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k}{2} s_k^{-2} + \boldsymbol{\mu}_k^T s_k^{-2} \mathbf{m}_k + \text{const}
\end{aligned} \tag{4.3}$$

where it was noted that $\boldsymbol{\mu}_k^T \mathbf{m}_k = \mathbf{m}_k^T \boldsymbol{\mu}_k$.

Comparing the coefficients in the equations above, it is concluded that the optimal parameter updates are

$$s_k^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\lambda^2} \sum_{i=1}^{N} \phi_{i,k} \right)^{-1} \tag{4.4}$$

and

$$\mathbf{m}_k = s_k^2 \left(\frac{\boldsymbol{\alpha}}{\sigma^2} + \frac{1}{\lambda^2} \sum_{i=1}^{N} \phi_{i,k} \mathbf{x}_i \right) \tag{4.5}$$

5 Assignment 4

Using results of the previous assignments implement missing parts of the algorithm in the provided python code.

Please, refer to the code for the programmatic implementation. By running the code, the results shown below were obtained. Figure 2 shows the true data with the true cluster assignments (colours) and the true cluster means (red dots), Figure 3 shows the posterior with with the estimated cluster assignments (colours) and the estimated cluster means (red dots), and Figure 4 shows the ELBO over iterations until convergence.

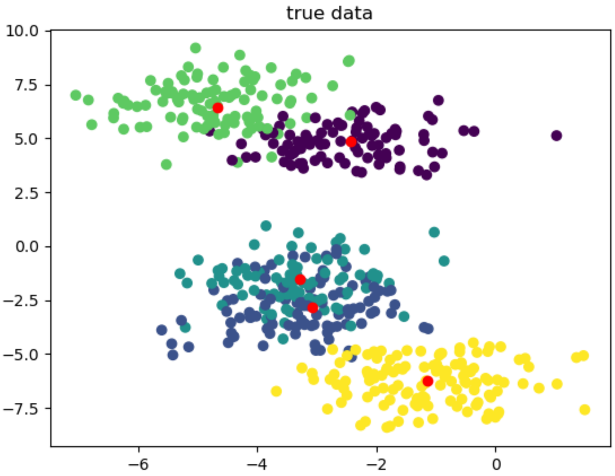

Figure 2: True data of Multivariate Gaussian Mixture Model

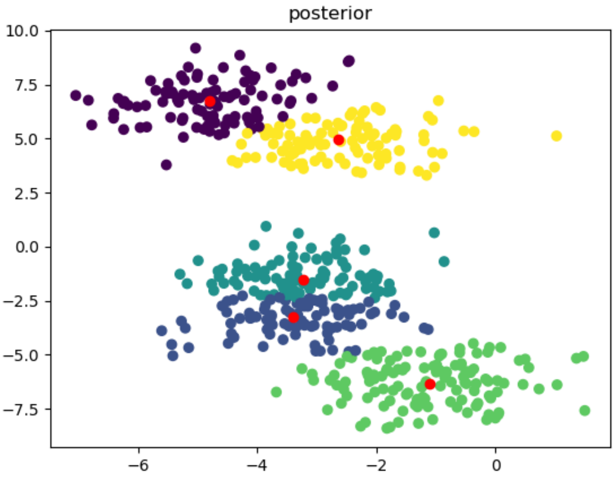

Figure 3: Posterior estimates of Multivariate Gaussian Mixture Model

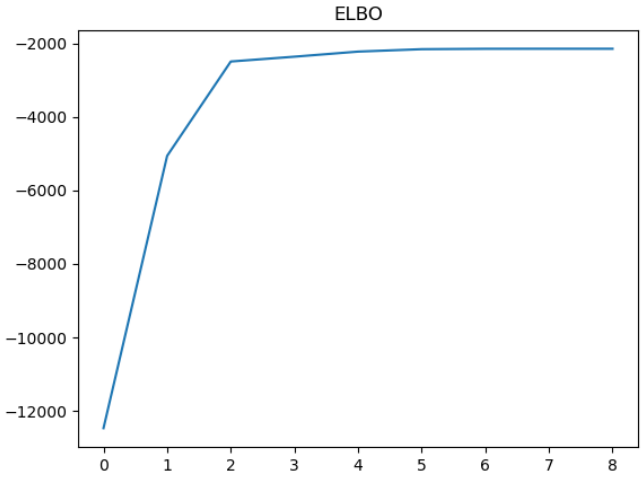

Figure 4: ELBO per iteration until convergence

References

[1] K. Yoshii, *Learning Algorithms for Gaussian Mixture Models*. Kyoto University, 2017. [Online]. Available: `http://sap.ist.i.kyoto-u.ac.jp/members/yoshii/lectures/pattern_recognition/2017/20170606-npb-gmm.pdf`.

[2] B. Keng, *Variational Bayes and The Mean-Field Approximation*. Github.io, 2017. [Online]. Available: `https://bjlkeng.github.io/posts/variational-bayes-and-the-mean-field-approximation/`.

[3] Wikipedia, *Variational Bayesian methods*. Wikipedia, 2021. [Online]. Available: `https://en.wikipedia.org/wiki/Variational_Bayesian_methods`.