EM algorithm and Variational Bayes algorithm

Daisuke Endo

January 22, 2019

1 ML estimation

1.1 About Gaussian mixtute model

The likelihood of gaussian mixture is given by

$$\begin{aligned} p(\mathbf{x};\theta) &= \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) \\ &= \sum_k \pi_k N(\mathbf{x}|\mu_{\mathbf{k}}, \mathbf{\Lambda}_k^{-1}) \end{aligned} \tag{1}$$

where θ is all parameters, that is $\theta = (\pi, \mu, \mathbf{\Lambda})$. The mixing coefficients π_k satisfy $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$. μ_k is the mean vector for cluster k. $\mathbf{\Lambda}_k$ is the precision matrix for cluster k.

$\mathbf{x}$ is an observation variable. $\mathbf{z}$ is a latent variable, and satisfy $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$. Then the probability of $\mathbf{z}$ is given by

$$\begin{aligned} p(\mathbf{z}) &= \prod_k p(z_k) \\ &= \prod_k \pi_k^{z_k} \end{aligned} \tag{2}$$

Given θ, the joint probability of $\mathbf{x}, \mathbf{z}$ is given by

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}|\theta) &= \prod_k p(z_k)p(\mathbf{x}|\mathbf{z};\theta) \\ &= \prod_k \pi_k^{z_k} N(\mathbf{x}|\mu_k, \mathbf{\Lambda}_k^{-1})^{z_k} \end{aligned} \tag{3}$$

Given $\mathbf{x}$ and θ, the probability of $\mathbf{z}$ is given by

$$p(z_k = 1|\mathbf{x};\theta) \propto p(z_k)p(\mathbf{x}|z_k = 1;\theta) \tag{4}$$

$$\begin{aligned} p(z_k = 1|\mathbf{x};\theta) &= \frac{\pi_k N(\mathbf{x}|\mu_k, \mathbf{\Lambda}_k)}{\sum_k \pi_k N(\mathbf{x}|\mu_k, \mathbf{\Lambda}_k^{-1})} \\ &= \gamma_k \end{aligned} \tag{5}$$

where γ_k is the "responsibility" that cluster k takes for 'explaining' the observation $\mathbf{x}$. Then, the probability of $\mathbf{z}$ is

$$p(\mathbf{z}|\mathbf{x};\theta) = \prod_k \gamma_k^{z_k} \tag{6}$$

1.2 EM Algorithm

From Jensen's ineqnality, the log likelihood is rewritten , as follows.

$$\begin{aligned}\log p(\mathbf{X};\theta) &= \log \int p(\mathbf{X},\mathbf{Z};\theta)d\mathbf{Z} \\ &= \log \int q(\mathbf{Z})\frac{p(\mathbf{X},\mathbf{Z};\theta)}{q(\mathbf{Z})}d\mathbf{Z} \\ &\geq \int q(\mathbf{Z})\log\frac{p(\mathbf{X},\mathbf{Z};\theta)}{q(\mathbf{Z})}d\mathbf{Z} \qquad (7)\end{aligned}$$

The eqality hold true when $q^*(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X};\theta)$. Here, the lower bound $L(\mathbf{X}|\theta)$ is written as $L(\mathbf{X}|\theta) = \int q(\mathbf{Z})\log\frac{p(\mathbf{X},\mathbf{Z};\theta)}{q(\mathbf{X})}d\mathbf{Z}$.

- In E step, we evaluate $q^*(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X};\theta)$. Thus we calculate a posterior of latent variables $\mathbf{Z}$.
- In M step, we substituted $q^*(\mathbf{Z})$ for the lower bound to maximize it when θ is fixed. Next, we estiamte parameters θ to maximize the lower bound.
- we evaluate the log likelihood $\log p(\mathbf{X}|\theta)$ and continue updating until the log likelihood converges.

1.3 E step of GMM

In E step of Gaussain mixture, from the above discussion,

$$\begin{aligned}q^*(\mathbf{Z}) &= p(\mathbf{Z}|\mathbf{X};\theta) \\ &= \prod_n p(\mathbf{z_n}|\mathbf{x_n};\theta_\mathbf{n}) \\ &= \prod_n\prod_k \gamma_{nk}^{z_{nk}} \qquad (8)\end{aligned}$$

where the n-th row of $\mathbf{X}$ is $\mathbf{x}_n^t$, the n-th row of $\mathbf{Z}$ is $\mathbf{z}_n^t$.

1.4 M step of GMM

Next we maximize the lower bound with respect to the pareameters θ. From the above $q^*(\mathbf{Z})$, the lower bound $L(\mathbf{X}|\theta)$ is given by

$$\begin{aligned}L(\mathbf{X}|\theta) &= \sum_\mathbf{Z} q^*(\mathbf{Z})\log p(\mathbf{X},\mathbf{Z};\theta) - \sum_\mathbf{Z} q^*(\mathbf{Z})\log q^*(\mathbf{Z}) \\ &= \sum_\mathbf{Z} q^*(\mathbf{Z})\sum_n \log p(\mathbf{x}_n,\mathbf{z}_n;\theta) - \sum_\mathbf{Z} q^*(\mathbf{Z})\log q^*(\mathbf{Z}) \\ &= \sum_\mathbf{Z} q^*(\mathbf{Z})\sum_n\sum_k z_{nk}(\log\pi_k + \log N(\mathbf{x_n}|\mu_k,\mathbf{\Lambda}_k^{-1}) \\ &\quad -\sum_\mathbf{Z} q^*(\mathbf{Z})\log q^*(\mathbf{Z}) \\ &= \sum_n\sum_k\sum_\mathbf{Z} q^*(\mathbf{Z})z_{nk}(\log\pi_k + \log N(\mathbf{x_n}|\mu_k,\mathbf{\Lambda}_k^{-1}) \\ &\quad -\sum_\mathbf{Z} q^*(\mathbf{Z})\log q^*(\mathbf{Z}) \qquad (9)\end{aligned}$$

At first, we calculate $\int q^*(\mathbf{Z}) z_{nk} d\mathbf{Z}$.

$$\begin{aligned}\sum_{\mathbf{Z}} q^*(\mathbf{Z}) z_{nk} &= \sum_{\mathbf{Z}} \prod_{n'} \prod_{k'} \gamma_{n'k'}^{z'_N k'} z_{nk} \\ &= \gamma_{nk} \end{aligned} \tag{10}$$

Then, we pay attention a part of $L(\mathbf{X}|\theta)$ depending on θ

$$\begin{aligned} L(\mathbf{X}|\theta) &= \sum_n \sum_k \gamma_{nk} (\log \pi_k + \log N(\mathbf{x_n}|\mu_k, \mathbf{\Lambda}_k^{-1}) + const \\ &= \sum_n \sum_k \gamma_{nk} (\log \pi_k - \frac{1}{2}(\mathbf{x}_n - \mu_k)^T \mathbf{\Lambda}_k (\mathbf{x}_n - \mu_k) + \frac{1}{2} \log |\mathbf{\Lambda}_k|) \\ & \quad +const \end{aligned} \tag{11}$$

Next, we estimate new parameters $\pi, \mu, \mathbf{\Lambda}$ to maximize the lower bound $L(\mathbf{X}|\theta)$.

- estimate new π
  By the constraint of π, we maximize the lower bound bound with respect π using Lagrange multiplier , as follows.

$$\tilde{L}(\mathbf{X}|\theta) = L(\mathbf{X}|\theta) + \lambda(\sum_k \pi_k - 1) \tag{12}$$

$$\frac{\partial \tilde{L}(\mathbf{X}|\theta)}{\partial \pi_k} = 0 \quad (\text{for all } k) \tag{13}$$

$$\frac{\partial \tilde{L}(\mathbf{X}|\theta)}{\partial \lambda} = 0 \tag{14}$$

  we compute the above eqations.

$$\begin{aligned} \frac{\partial}{\partial \pi_k} \{\sum_n \sum_{k'} \gamma_{nk'} \log \pi_{k'} + \lambda \sum_{k'} \pi_{k'}\} &= 0 \\ \pi_k &= \frac{\sum_n \gamma_{nk}}{\lambda} \\ \frac{\partial}{\partial \lambda} \lambda(\sum_k \pi_k - 1) &= 0 \\ \sum_k \pi_k - 1 &= 0 \end{aligned}$$

  Then,

$$\begin{aligned} \sum_k \pi_k &= \sum_k \frac{\sum_n \gamma_{nk}}{\lambda} \\ \lambda &= \sum_k \sum_n \gamma_{nk} \end{aligned}$$

  Therefore, new π_k is give by

$$\pi_k^* = \frac{\sum_n \gamma_{nk}}{\sum_k \sum_n \gamma_{nk}} \tag{15}$$

- estimate new μ
  we maximize the lower bound with respect to μ.

$$\begin{aligned}
\frac{\partial L(\mathbf{X}|\theta)}{\partial \mu_k} &= 0 \qquad (16)\\
\sum_n \gamma_{nk} \frac{\partial}{\partial \mu_k} \frac{1}{2}(\mathbf{x}_n - \mu_k)^T \mathbf{\Lambda}_k (\mathbf{x}_n - \mu_k) &= 0\\
\sum_n \gamma_{nk} (\mathbf{x}_n - \mu_{\mathbf{k}}) &= 0
\end{aligned}$$

  Therefore, new μ_k is give by

$$\mu_k^* = \frac{\sum_n \gamma_{nk} \mathbf{x}_n}{\sum_n \gamma_{nk}} \qquad (17)$$

- estimate new $\mathbf{\Lambda}$
  we maximize the lower bound with respect to λ.

$$\begin{aligned}
\frac{\partial L(\mathbf{X}|\theta)}{\partial \mathbf{\Lambda}_k} &= 0 \qquad (18)\\
\sum_n \gamma_{nk} \frac{\partial}{\partial \mathbf{\Lambda}_k} \{-\frac{1}{2}(\mathbf{x}_n - \mu_k)^T \mathbf{\Lambda}_k (\mathbf{x}_n - \mu_k) + \frac{1}{2} \log |\mathbf{\Lambda}_k|\} &= 0\\
\sum_n \gamma_{nk} \frac{\partial}{\partial \mathbf{\Lambda}_k} \{-tr[(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \mathbf{\Lambda}_k] + \log |\mathbf{\Lambda}_k|\} &= 0\\
\sum_n \gamma_{nk} [(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T]^T - \sum_n \gamma_{nk} \mathbf{\Lambda}^{-1\,T} &= 0
\end{aligned}$$

  Therefore, new $\mathbf{\Lambda_k}$ is given by

$$\mathbf{\Lambda}_k^{*\,-1} = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \qquad (19)$$

  After μ_k is updated such as

$$\mu_k = \frac{\sum_n \gamma_{nk} \mathbf{x}_n}{\sum_n \gamma_{nk}}$$

  , therefore

$$\mathbf{\Lambda}_k^{*\,-1} = \frac{\sum_n \gamma_{nk} \mathbf{x}_n \mathbf{x}_n^T}{\sum_n \gamma_{nk}} - \mu_k \mu_k^T \qquad (20)$$

2 Implement of EM algorithm

I assumed 4 classes of Gaussian Mixtures, that is $K = 4$. I gave the initial paramters, such as the mixing coefficients $\pi_{\mathbf{k}}$ are uniform, the mean vectors of gaussians $\mu_{\mathbf{k}}$ are random, and the covariance matrices $\mathbf{\Sigma_k}$ are identity matrices.

I iterated E-step and M-step until the log likelihood $\log P(\mathbf{X}; \theta)$ converges at $O(10^{-5})$. I showed the log likelihood at each iterations.

```
iter	log-likelihood
0	-94560.039710
1	-77988.729995
2	-74171.989247
3	-73319.776206
4	-72904.415632
5	-72873.098154
6	-72873.012123
7	-72873.009706
8	-72873.009562
9	-72873.009553
```

Then, last estimated parameters was recored in "params.dat", and the probability of being each class for each data was recored in "z.csv". I used python for this implement. The code of this implement was written in "EM_gmm.py".

I showed the data classified by coloring each data points.

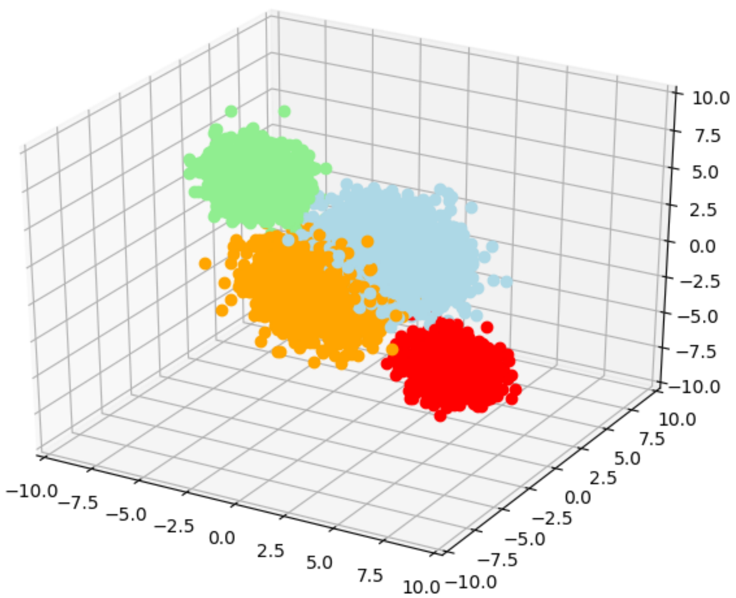

Figure 1: EM algorithm for x.csv

3 Bayesian estimation

3.1 Variational Bayes

We think a fully Bayesian model, so we suppose all parameters are given prior distributions. The model have latent variables as well as parameters, and let us define $\mathbf{Z}$ as the set of all latent variables and parameters. Similarly, we define $\mathbf{X}$

as the set of all observed variables. From Jensen's ineqnality, the log likelihood of $\mathbf{X}$ is such as

$$\begin{aligned}\log p(\mathbf{X}) &= \log \int p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} \\ &= \log \int q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\ &\geq \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} = L(q) \end{aligned} \tag{21}$$

We assume partition the elements of $\mathbf{Z}$ into disjoint groups that we denote by $\mathbf{Z_i}$, and decompose distribution over $\mathbf{Z}$ as follows

$$q(\mathbf{Z}) = \prod_i q_i(\mathbf{Z_i})$$

Then, we rewrite the lower bound $L(q)$ such as

$$\begin{aligned} L(q) &= \sum_i \{ \int q_i(\mathbf{Z_i}) (\int q_{-i}(\mathbf{Z_{-i}}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z_{-i}}) d\mathbf{Z_i} - \int q_i(\mathbf{Z_i}) \log q_i(\mathbf{Z_i}) d\mathbf{Z_i} \} \\ &= \sum_i \{ \int q_i(\mathbf{Z_i}) < \log p(\mathbf{X}, \mathbf{Z}) >_{q(\mathbf{Z_{-i}})} d\mathbf{Z_i} - \int q_i(\mathbf{Z_i}) \log q_i(\mathbf{Z_i}) d\mathbf{Z_i} \} \end{aligned}$$

where $\mathbf{Z_{-i}}$ is what $\mathbf{Z_i}$ is removed from $\mathbf{Z}$. From the above and Jensen's ineqnality, the lower bound is maximized when

$$\log q^*(\mathbf{Z_i}) =< \log p(\mathbf{X}, \mathbf{Z}) >_{q(\mathbf{Z_{-i}})} + const.$$

and we update $\log q^*(\mathbf{Z_i})$ for all i until the lower bound converges.

3.2 VB for GMM

3.2.1 About GMM

we consider about the likelihood functions for the Gaussian mixture model.We can write down the conditional distribution of the observed data vectors $\mathbf{X}$ given $\mathbf{Z}, \mu$ and $\mathbf{\Lambda}$, and the conditional distribution of the latent variables $\mathbf{Z}$ given π, in the form

$$p(\mathbf{X}|\mathbf{Z}, \mu, \mathbf{\Lambda}) = \prod_n \prod_k N(\mathbf{x_n}|\mu_{\mathbf{k}}, \mathbf{\Lambda_k^{-1}})^{z_{nk}} \tag{22}$$

$$p(\mathbf{Z}|\pi) = \prod_n \prod_k z_k^{nk} \tag{23}$$

where n represents the index of n-th data, k represents the index of k-th cluster. Next, we introduce prior distributions over the parameters μ, Λ and π.

$$p(\pi) = \frac{\Gamma(\sum_k \alpha_{0k})}{\prod_k \Gamma(\alpha_{0k})} \prod_k \pi_k^{\alpha_{0k}-1} \tag{24}$$

$$p(\mu, \mathbf{\Lambda}) = \prod_k N(\mu_{\mathbf{k}}|\mathbf{m_0}, (\beta_{\mathbf{0}}\mathbf{\Lambda_k})^{-1}) W(\mathbf{\Lambda_k}|\mathbf{W_0}, \nu_{\mathbf{0}}) \tag{25}$$

We regard updating variational posterior over latent variables $\mathbf{Z}$ and parameters in variational Bayes as EM algorithm of the maximum likelihood. Then

- VB-E step

$$\log q^*(\mathbf{Z}) = < \log p(\mathbf{X}|\mathbf{Z}, \mu, \mathbf{\Lambda}) + \log p(\mathbf{Z}|\pi) >_{q(\pi,\mu,\mathbf{\Lambda})} + const. \tag{26}$$

- VB-M step

$$\begin{aligned} \log q^*(\pi) &= < \log p(\mathbf{Z}|\pi)p(\pi) >_{q(\mathbf{Z})} + const. && (27) \\ \log q^*(\mu, \mathbf{\Lambda}) &= < \log p(\mathbf{X}|\mathbf{Z}, \mu, \mathbf{\Lambda})p(\mu, \mathbf{\Lambda}) >_{q(\mathbf{Z})} + const. && (28) \end{aligned}$$

And, the variational lower bound is given as

$$\begin{aligned} L &= < \log p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \mathbf{\Lambda}) >_{q(\mathbf{Z},\pi,\mu,\mathbf{\Lambda})} - < \log q(\mathbf{Z}, \pi, \mu, \mathbf{\Lambda}) >_{q(\mathbf{Z},\pi,\mu,\mathbf{\Lambda})} \\ &= < \log p(\mathbf{X}|\mathbf{Z}, \mu, \mathbf{\Lambda}) >_{q(\mathbf{Z},\mu,\mathbf{\Lambda})} + < \log p(\mathbf{Z}|\pi) >_{q(\mathbf{Z},\pi)} + < \log p(\pi) >_{q(\pi)} \\ &\quad + < \log p(\mu, \mathbf{\Lambda}) >_{q(\mu,\mathbf{\Lambda})} - < \log q(\mathbf{Z}) >_{q(\mathbf{Z})} - < \log q(\pi) >_{q(\pi)} \\ &\quad - < \log q(\mu, \mathbf{\Lambda}) >_{q(\mu,\mathbf{\Lambda})} \end{aligned}$$

3.2.2 VB-E step for GMM

$$\begin{aligned} \log q^*(\mathbf{Z}) &= \sum_n \sum_k z_{nk} \{ < \log \pi_k >_{q(\pi)} \\ &\quad + < \log N(\mathbf{x_n}|\mu_\mathbf{k}, \mathbf{\Lambda_k^{-1}}) >_{q(\mu_\mathbf{k},\mathbf{\Lambda_k})} \} \\ &= \sum_n \sum_k z_{nk} \log \rho_{nk} + const. && (29) \end{aligned}$$

where we define D as the dimension of $\mathbf{x_n}$, and

$$\begin{aligned} \log \rho_{nk} &= < \log \pi_k >_{q(\pi)} + \frac{1}{2} < \log |\mathbf{\Lambda_k}| >_{q(\mathbf{\Lambda_k})} \\ &\quad - \frac{D}{2} \log(2\pi) - \frac{1}{2} < (\mathbf{x_n} - \mu_\mathbf{k})^\mathbf{T} \mathbf{\Lambda_k} (\mathbf{x_n} - \mu_\mathbf{k}) >_{q(\mu_\mathbf{k},\mathbf{\Lambda_k})} \end{aligned}$$

Each term is computed such as

$$\begin{aligned} < \log \pi_k >_{q(\pi)} &= \psi(\alpha_k) - \psi(\sum_k \alpha_k) \\ < \log |\mathbf{\Lambda_k}| >_{q(\mathbf{\Lambda_k})} &= \sum_{i=1}^{D} \psi(\frac{\nu_k + 1 - i}{2}) + D \log 2 + \log |\mathbf{W}_k| \\ < (\mathbf{x_n} - \mu_\mathbf{k})^\mathbf{T} \mathbf{\Lambda_k} (\mathbf{x_n} - \mu_\mathbf{k}) >_{q(\mu_\mathbf{k},\mathbf{\Lambda_k})} &= D\beta_k^{-1} + \nu_k (\mathbf{x_n} - \mathbf{m_k})^T \mathbf{W_k} (\mathbf{x_n} - \mathbf{m_k}) \end{aligned}$$

where ψ is a digamma function and α_k, β_k, $\mathbf{W_k}$ are hyperparameters for π_k, $\mu_\mathbf{k}$, $\mathbf{\Lambda}_k$. From the above discussion,

$$q^*(\mathbf{Z}) = \prod_n \prod_k \gamma_{nk}^{z_{nk}} \tag{30}$$

where

$$\gamma_{nk} = \frac{\rho_{nk}}{\sum_j \rho_{nj}} \tag{31}$$

3.2.3 VB-M step for GMM

- about $q^*(\pi)$

$$\log q^*(\pi) = \log p(\pi) + < \log p(\mathbf{Z}|\pi) >_{q(\mathbf{Z})} + const. \tag{32}$$

Here, we look at each term.

$$\begin{aligned}\log p(\pi) &= \sum_k (\alpha_{0k} - 1)\log \pi_k + const. \\ < \log p(\mathbf{Z}|\pi) >_{q(\mathbf{Z})} &= \sum_n \sum_k < z_{nk} >_{q(\mathbf{Z})} \log \pi_k\end{aligned}$$

First, we compute $< z_{nk} >_{q(\mathbf{Z})}$ such as

$$\begin{aligned}< z_{nk} >_{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} z_{nk} q(\mathbf{Z}) \\ &= \sum_{\mathbf{Z}} z_{nk} \prod_{n'} \prod_{k'} \gamma_{n'k'}^{z_{n'k'}} \\ &= \gamma_{nk}\end{aligned}$$

Therefore,

$$\log q^*(\pi) = \sum_k (\alpha_k - 1)\log \pi_k \tag{33}$$

where

$$\alpha_k = \alpha_{0k} + \sum_n \gamma_{nk}$$

- about $q^*(\mu, \mathbf{\Lambda})$

$$\log q^*(\mu, \mathbf{\Lambda}) = \log p(\mu, \mathbf{\Lambda}) + < \log p(\mathbf{X}|\mathbf{Z}, \mu, \mathbf{\Lambda}) >_{q(\mathbf{Z})} + const$$

Here, we look at each term.

$$\begin{aligned}\log p(\mu, \mathbf{\Lambda}) &= \sum_k \{\log N(\mu_{\mathbf{k}}|\mathbf{m_0}, (\beta_{\mathbf{0}}\mathbf{\Lambda_k})^{-\mathbf{1}}) + \log W(\mathbf{\Lambda_k}|\mathbf{W_0}, \nu_{\mathbf{0}})\} \\ < \log p(\mathbf{X}|\mathbf{Z}, \mu, \mathbf{\Lambda}) >_{q(\mathbf{Z})} &= \sum_k \sum_n < z_{nk} >_{q(\mathbf{Z})} \log N(\mathbf{x_n}|\mu_{\mathbf{n}}, \mathbf{\Lambda}^{-\mathbf{1}}) \\ &= \sum_k \sum_n \gamma_{nk} \log N(\mathbf{x_n}|\mu_{\mathbf{n}}, \mathbf{\Lambda_k^{-1}})\end{aligned}$$

Then, we look at about component k.

$$\begin{aligned}\log q^*(\mu_{\mathbf{k}}, \mathbf{\Lambda_k}) &= \log N(\mu_{\mathbf{k}}|\mathbf{m_0}, (\beta_{\mathbf{0}}\mathbf{\Lambda_k})^{-\mathbf{1}}) + \log W(\mathbf{\Lambda_k}|\mathbf{W_0}, \nu_{\mathbf{0}}) \\ &\quad + \sum_n \gamma_{nk} \log N(\mathbf{x_n}|\mu_{\mathbf{n}}, \mathbf{\Lambda_k^{-1}}) + const \\ &= -\frac{\beta_0}{2}(\mu_{\mathbf{k}} - \mathbf{m_0})^{\mathbf{T}}\mathbf{\Lambda_k}(\mu_{\mathbf{k}} - \mathbf{m_0}) + \frac{1}{2}\log|\mathbf{\Lambda_k}| \\ &\quad -\frac{1}{2}Tr(\mathbf{W_0^{-1}}\mathbf{\Lambda_k}) + \frac{\nu_0 - D - 1}{2}\log|\mathbf{\Lambda_k}| \\ &\quad -\frac{1}{2}\sum_n \gamma_{nk}(\mathbf{x_n} - \mu_{\mathbf{k}})^{\mathbf{T}}\mathbf{\Lambda_k}(\mathbf{x_n} - \mu_{\mathbf{k}}) + \frac{1}{2}\sum_n \gamma_{nk}\log|\mathbf{\Lambda_k}| + const.\end{aligned}$$

Using the product rule of probability, we express $q^*(\mu_\mathbf{k}, \mathbf{\Lambda_k})$ such as $q^*(\mu_\mathbf{k}, \mathbf{\Lambda_k}) = q^*(\mu_\mathbf{k}|\mathbf{\Lambda_k})q^*(\mathbf{\Lambda_k})$. At first, we only consider terms on the right side which depend on μ_k.

$$\begin{aligned}
\log q^*(\mu_\mathbf{k}|\mathbf{\Lambda_k}) &= -\frac{\beta_0}{2}(\mu_\mathbf{k}^\mathbf{T}\mathbf{\Lambda_k}\mu_\mathbf{k} - \mathbf{2}\mu_\mathbf{k}^\mathbf{T}\mathbf{\Lambda_k}\mathbf{m_0}) - \frac{1}{2}\sum_n \gamma_{nk}\{\mu_\mathbf{k}^\mathbf{T}\mathbf{\Lambda_k}\mu_\mathbf{k} - \mu_\mathbf{k}^\mathbf{T}\mathbf{\Lambda_k}\mathbf{x_n}\} \\
&\quad +const. \\
&= -\frac{1}{2}\mu_\mathbf{k}^\mathbf{T}(\beta_\mathbf{0} + \sum_\mathbf{n}\gamma_\mathbf{nk})\mathbf{\Lambda_k}\mu_\mathbf{k} + \mu_\mathbf{k}^\mathbf{T}\mathbf{\Lambda_k}(\beta_\mathbf{0}\mathbf{m_0} + \sum_\mathbf{n}\gamma_\mathbf{nk}\mathbf{x_n}) + const.
\end{aligned}$$

Therefore, $q^*(\mu_\mathbf{k}|\mathbf{\Lambda_k})$ is a Gaussian distribution. So, $\log q^*(\mu_\mathbf{k}|\mathbf{\Lambda_k})$ is given such as

$$\begin{aligned}
\log q^*(\mu_\mathbf{k}|\mathbf{\Lambda_k}) &= -\frac{\beta_k}{2}(\mu_\mathbf{k} - \mathbf{m_k})^\mathbf{T}\mathbf{\Lambda_k}(\mu_\mathbf{k} - \mathbf{m_k}) \\
&\quad +\frac{1}{2}\log|\mathbf{\Lambda_k}| + const. \qquad (34)
\end{aligned}$$

where

$$\begin{aligned}
\beta_k &= \beta_0 + \sum_n \gamma_{nk} \\
\mathbf{m_k} &= \frac{1}{\beta_k}(\beta_0\mathbf{m_0} + \sum_n \gamma_{nk}\mathbf{x_n})
\end{aligned}$$

Next, we consider about $q^*(\mathbf{\Lambda_k})$.

$$\begin{aligned}
\log q^*(\mathbf{\Lambda_k}) &= \log q^*(\mu_\mathbf{k}\mathbf{\Lambda_k}) - \log q^*(\mu_\mathbf{k}|\mathbf{\Lambda_k}) \\
&= -\frac{\beta_0}{2}(\mu_\mathbf{k} - \mathbf{m_0})^\mathbf{T}\mathbf{\Lambda_k}(\mu_\mathbf{k} - \mathbf{m_0}) + \frac{1}{2}\log|\mathbf{\Lambda_k}| \\
&\quad -\frac{1}{2}Tr(\mathbf{W_0^{-1}}\mathbf{\Lambda_k}) + \frac{\nu_0 - D - 1}{2}\log|\mathbf{\Lambda_k}| \\
&\quad -\frac{1}{2}\sum_n \gamma_{nk}(\mathbf{x_n} - \mu_\mathbf{k})^\mathbf{T}\mathbf{\Lambda_k}(\mathbf{x_n} - \mu_\mathbf{k}) + \frac{1}{2}\sum_n \gamma_{nk}\log|\mathbf{\Lambda_k}| \\
&\quad +\frac{\beta_k}{2}(\mu_\mathbf{k} - \mathbf{m_k})^\mathbf{T}\mathbf{\Lambda_k}(\mu_\mathbf{k} - \mathbf{m_k}) \\
&\quad -\frac{1}{2}\log|\mathbf{\Lambda_k}| + const. \\
&= -\frac{\beta_0}{2}Tr\{(\mu_\mathbf{k} - \mathbf{m_0})(\mu_\mathbf{k} - \mathbf{m_0})^\mathbf{T}\mathbf{\Lambda_k}\} - \frac{1}{2}Tr(\mathbf{W_0^{-1}}\mathbf{\Lambda_k}) \\
&\quad -\frac{1}{2}\sum_n \gamma_{nk}Tr\{(\mathbf{x_n} - \mu_\mathbf{k})(\mathbf{x_n} - \mu_\mathbf{k})^\mathbf{T}\mathbf{\Lambda_k}\} + \frac{\beta_k}{2}Tr\{(\mu_\mathbf{k} - \mathbf{m_k})(\mu_\mathbf{k} - \mathbf{m_k})^\mathbf{T}\mathbf{\Lambda_k}\} \\
&\quad +\frac{\nu_0 + \sum_n \gamma_{nk} - D - 1}{2}\log|\mathbf{\Lambda_k}| + const. \\
&= -\frac{1}{2}Tr\{(\beta_0(\mu_\mathbf{k} - \mathbf{m_0})(\mu_\mathbf{k} - \mathbf{m_0})^\mathbf{T} + \mathbf{W_0^{-1}} \\
&\quad +\sum_n \gamma_{nk}(\mathbf{x_n} - \mu_\mathbf{k})(\mathbf{x_n} - \mu_\mathbf{k})^\mathbf{T} - \beta_\mathbf{k}(\mu_\mathbf{k} - \mathbf{m_k})(\mu_\mathbf{k} - \mathbf{m_k})^\mathbf{T})\mathbf{\Lambda_k}\} \\
&\quad +\frac{\nu_0 + \sum_n \gamma_{nk} - D - 1}{2}\log|\mathbf{\Lambda_k}| + const.
\end{aligned}$$

Therefore, $\log q^*(\mathbf{\Lambda_k})$ is a Wishart distribution, given by

$$\log q^*(\mathbf{\Lambda_k}) = -\frac{1}{2}Tr(\mathbf{W_k^{-1}}\mathbf{\Lambda_k}) + \frac{\nu_k - D - 1}{2}\log|\mathbf{\Lambda_k}| + const. \quad (35)$$

where

$$\begin{aligned}
\mathbf{W_k^{-1}} &= \beta_0(\mu_\mathbf{k} - \mathbf{m_0})(\mu_\mathbf{k} - \mathbf{m_0})^\mathbf{T} + \mathbf{W_0^{-1}} \\
&\quad + \sum_n \gamma_{nk}(\mathbf{x_n} - \mu_\mathbf{k})(\mathbf{x_n} - \mu_\mathbf{k})^\mathbf{T} - \beta_\mathbf{k}(\mu_\mathbf{k} - \mathbf{m_k})(\mu_\mathbf{k} - \mathbf{m_k})^\mathbf{T} \\
&= \mathbf{W_0^{-1}} + \beta_0(\mu_\mathbf{k}\mu_\mathbf{k}^\mathbf{T} - \mu_\mathbf{k}\mathbf{m_0^T} - \mathbf{m_0}\mu_\mathbf{k}^\mathbf{T} + \mathbf{m_0}\mathbf{m_0^T}) \\
&\quad + \sum_n \gamma_{nk}\mathbf{x_n}\mathbf{x_n^T} - \sum_n \gamma_{nk}(\mathbf{x_n}\mu_\mathbf{k}^\mathbf{T} + \mu_\mathbf{k}\mathbf{x_n^T}) + \sum_n \gamma_{nk}\mu_\mathbf{k}\mu_\mathbf{k}^\mathbf{T} \\
&\quad -(\beta_0 + \sum_n \gamma_{nk})\mu_\mathbf{k}\mu_\mathbf{k}^\mathbf{T} + \mu_\mathbf{k}(\beta_0\mathbf{m_0} + \sum_n \gamma_{nk}\mathbf{x_n})^T \\
&\quad +(\beta_0\mathbf{m_0} + \sum_n \gamma_{nk}\mathbf{x_n})\mu_k^T - \beta_k\mathbf{m_k}\mathbf{m_k^T} \\
&= \mathbf{W_0^{-1}} + \beta_0\mathbf{m_0}\mathbf{m_0^T} + \sum_n \gamma_{nk}\mathbf{x_n}\mathbf{x_n^T} - \beta_k\mathbf{m_k}\mathbf{m_k^T} \\
\nu_k &= \nu_0 + \sum_n \gamma_{nk}
\end{aligned}$$

From the above, we express $q^*(\mu, \mathbf{\Lambda})$, as follows.

$$q^*(\mu, \mathbf{\Lambda}) = \prod_k q^*(\mu_\mathbf{k}, \mathbf{\Lambda_k}) \quad (36)$$

$$= \prod_k N(\mu_\mathbf{k}|\mathbf{m_0}, (\beta_\mathbf{0}\mathbf{\Lambda_k})^{-\mathbf{1}})\, W(\mathbf{\Lambda_k}|\mathbf{W_k}, \nu_\mathbf{k}) \quad (37)$$

4 Implement of VB algorithm

I assumed 6 classes of Gaussian Mixtures, that is $K = 6$. I gave the initial parameters of prior distributions, such as the parameter α_0 of dirichlet distribution is 0.001, the parameters $\mathbf{m_0}$, β_0 of the prior distribution with the means μ_k given precision $\mathbf{\Lambda}_k$ are $\mathbf{0}$, 1, the parameter $\mathbf{W}_0$, ν_0 of Wishart distribution are the identify matrix, the number of data.

I iterated VB-E step and VB-Mstep until the variational lower bound converges at $O(10^{-5})$. I showed the variational lower bound at each iterations.

```
iter    lower bound
0       -278626163.927253
100     -148474.342581
200     -148251.254414
300     -148046.524745
400     -147810.823523
500     -147471.198794
507     -147459.101514
```

Then, last estimated parameters was recored in "params.dat", and the probability of being each class for each data was recored in "z.csv". I used python for this implement. The code of this implement was written in "VB_gmm.py".

I showed the data classified by coloring each data points. The figure shows that only 4 of all 6 classes are effective.

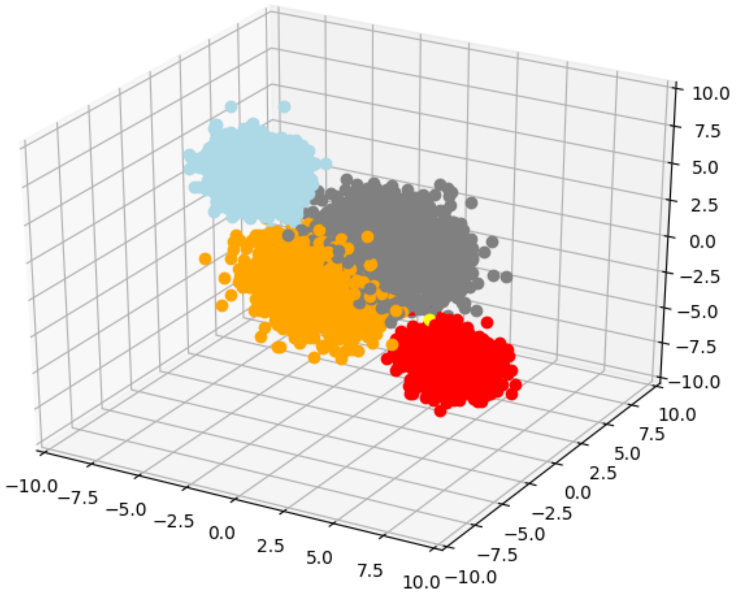

Figure 2: VB algorithm for x.csv