CSED524-Homework 1

Inhyuk Na

2018/10/03

1 Mixture of Gaussians

In the lecture, we've learned about Mixture of Gaussians, which is a typical clustering model. It is also a kind of Finite Mixture Models, and assumes that each component densities of Finite Mixture Model are Gaussians. It can be represented as

$$p(\boldsymbol{x_n}|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}).$$

To achieve maximum log-likelihood, we can repeat the following steps iteratively, starting from appropriately initialized parameters.

Step 1: Compute responsibilities

$$r_{k,n} = \frac{\pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})}.$$

Step 2: Update parameters

$$\begin{aligned}
\boldsymbol{\mu_k} &= \frac{1}{N_k} \sum_{n=1}^{N} r_{k,n} \boldsymbol{x_n}, \\
\boldsymbol{\Sigma_k} &= \frac{1}{N_k} \sum_{n=1}^{N} r_{k,n} (\boldsymbol{x_n} - \boldsymbol{\mu_k})(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T, \text{ and} \\
\pi_k &= \frac{N_k}{N}
\end{aligned}$$

, where $N_k = \sum_{n=1}^{K} r_{k,n}$.

In this report, we first introduce the detailed derivations of the above method via two different ways: the first is direct maximization of likelihood and the second is to use EM optimization. Then we show experimental results of our own implementation of the method.

2 Derivation via direct maximization of likelihood

The log-likelihood of Mixture of Gaussians is given by

$$\mathcal{L} = \log p(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\boldsymbol{x_n}|\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{n=1}^{N} \log(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})).$$

To estimate the maximum value, we find each parameters which have an extreme value of $\mathcal{L}$. In other words, we find $\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}$ and $\boldsymbol{\pi}$ where

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu_k}} = 0,\\
&\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma_k}^{-1}} = 0, \text{ and}\\
&\frac{\partial(\mathcal{L} + \lambda(1 - \sum\limits_{k=1}^{K} \pi_k))}{\partial \boldsymbol{\pi}} = 0 \text{ (via Lagrange multiplier method).}
\end{aligned}$$

For $\boldsymbol{\mu_k}$,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu_k}} &= \frac{\partial}{\partial \boldsymbol{\mu_k}} \sum_{n=1}^{N} \log(\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j}))\\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \boldsymbol{\mu_k}} \log(\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j}))\\
&= \sum_{n=1}^{N} [\frac{1}{\sum\limits_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})} \frac{\partial}{\partial \boldsymbol{\mu_k}} \{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})\}]\\
&= \sum_{n=1}^{N} [\frac{1}{\sum\limits_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})} \frac{\partial}{\partial \boldsymbol{\mu_k}} \pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})]\\
&= \sum_{n=1}^{N} [\frac{\pi_k}{\sum\limits_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})} \frac{\partial}{\partial \boldsymbol{\mu_k}} \{\frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma_k}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\boldsymbol{x_n}-\boldsymbol{\mu_k})^T \boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu_k}))\}]\\
&= \sum_{n=1}^{N} [\frac{\pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})}{\sum\limits_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})} \frac{\partial}{\partial \boldsymbol{\mu_k}} \{-\frac{1}{2}(\boldsymbol{x_n}-\boldsymbol{\mu_k})^T \boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu_k})\}]\\
&= \sum_{n=1}^{N} [r_{k,n} \frac{\partial}{\partial \boldsymbol{\mu_k}} \{-\frac{1}{2}(\boldsymbol{x_n}-\boldsymbol{\mu_k})^T \boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu_k})\}]\\
&= \sum_{n=1}^{N} [r_{k,n}(-\frac{1}{2})\{(\boldsymbol{x_n}-\boldsymbol{\mu_k})^T \boldsymbol{\Sigma_k}^{-1} \frac{\partial(\boldsymbol{x_n}-\boldsymbol{\mu_k})}{\partial \boldsymbol{\mu_k}} + (\boldsymbol{x_n}-\boldsymbol{\mu_k})^T \boldsymbol{\Sigma_k}^{-1T} \frac{\partial(\boldsymbol{x_n}-\boldsymbol{\mu_k})}{\partial \boldsymbol{\mu_k}}\}]\\
&= \sum_{n=1}^{N} [r_{k,n}(-\frac{1}{2})\{-(\boldsymbol{x_n}-\boldsymbol{\mu_k})^T \boldsymbol{\Sigma_k}^{-1} - (\boldsymbol{x_n}-\boldsymbol{\mu_k})^T \boldsymbol{\Sigma_k}^{-1}\}]\\
&= \sum_{n=1}^{N} \{r_{k,n} \boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu_k})\}.
\end{aligned}$$

Since $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu_k}}$ has to be 0, we can derive

$$\sum_{n=1}^{N}\{r_{k,n}\boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu_k})\}=0,$$

$$\sum_{n=1}^{N}r_{k,n}\boldsymbol{\Sigma_k}^{-1}\boldsymbol{x_n}=\sum_{n=1}^{N}r_{k,n}\boldsymbol{\Sigma_k}^{-1}\boldsymbol{\mu_k},$$

$$\boldsymbol{\mu_k}=\frac{\sum\limits_{n=1}^{N}r_{k,n}\boldsymbol{x_n}}{\sum\limits_{n=1}^{N}r_{k,n}}.$$

For $\boldsymbol{\Sigma_k}$,

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Sigma_k}^{-1}} &= \frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}\sum_{n=1}^{N}\log(\sum_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j}))\\
&= \sum_{n=1}^{N}\frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}\log(\sum_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j}))\\
&= \sum_{n=1}^{N}[\frac{1}{\sum\limits_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})}\frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}\{\sum_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})\}]\\
&= \sum_{n=1}^{N}[\frac{1}{\sum\limits_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})}\frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}\pi_k\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})]\\
&= \sum_{n=1}^{N}[\frac{\pi_k}{\sum\limits_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})}\frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}\{\frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma_k}|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\boldsymbol{x_n}-\boldsymbol{\mu_k})^T\boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu_k}))\}]\\
&= \sum_{n=1}^{N}[\frac{\pi_k}{\sum\limits_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})}\{\frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}(\frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma_k}|^{\frac{1}{2}}})\exp(-\frac{1}{2}(\boldsymbol{x_n}-\boldsymbol{\mu_k})^T\boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu_k}))\\
&+\frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma_k}|^{\frac{1}{2}}}\frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}(\exp(-\frac{1}{2}(\boldsymbol{x_n}-\boldsymbol{\mu_k})^T\boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu_k})))\}]. \qquad (1)
\end{aligned}$$

Since

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}(\frac{1}{|\boldsymbol{\Sigma_k}|^{\frac{1}{2}}}) &= \frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}(|\boldsymbol{\Sigma_k}^{-1}|^{\frac{1}{2}})\\
&= \frac{1}{2}|\boldsymbol{\Sigma_k}^{-1}|^{-\frac{1}{2}}\frac{\partial|\boldsymbol{\Sigma_k}^{-1}|}{\partial\boldsymbol{\Sigma_k}^{-1}}\\
&= \frac{1}{2}|\boldsymbol{\Sigma_k}^{-1}|^{-\frac{1}{2}}|\boldsymbol{\Sigma_k}^{-1}|\boldsymbol{\Sigma_k}^{T}\\
&= \frac{1}{2}|\boldsymbol{\Sigma_k}|^{-\frac{1}{2}}\boldsymbol{\Sigma_k}, \qquad (2)
\end{aligned}$$

We can substitute (2) into (1). After substitution, (1) becomes to

$$
\begin{aligned}
(1) = & \sum_{n=1}^{N}[\frac{\pi_k}{\sum\limits_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})}\{(\frac{1}{(2\pi)^{\frac{D}{2}}\left|\boldsymbol{\Sigma_k}\right|^{\frac{1}{2}}})\exp(-\frac{1}{2}\left(\boldsymbol{x_n}-\boldsymbol{\mu_k}\right)^{T}\boldsymbol{\Sigma_k}^{-1}\left(\boldsymbol{x_n}-\boldsymbol{\mu_k}\right))\frac{1}{2}\boldsymbol{\Sigma_k} \\
& +\frac{1}{(2\pi)^{\frac{D}{2}}\left|\boldsymbol{\Sigma_k}\right|^{\frac{1}{2}}}\frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}(\exp(-\frac{1}{2}\left(\boldsymbol{x_n}-\boldsymbol{\mu_k}\right)^{T}\boldsymbol{\Sigma_k}^{-1}\left(\boldsymbol{x_n}-\boldsymbol{\mu_k}\right)))\}] \\
= & \sum_{n=1}^{N}[\frac{\pi_k}{\sum\limits_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})}\{\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})\frac{1}{2}\boldsymbol{\Sigma_k} \\
& +\frac{1}{(2\pi)^{\frac{D}{2}}\left|\boldsymbol{\Sigma_k}\right|^{\frac{1}{2}}}\frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}(\exp(-\frac{1}{2}\left(\boldsymbol{x_n}-\boldsymbol{\mu_k}\right)^{T}\boldsymbol{\Sigma_k}^{-1}\left(\boldsymbol{x_n}-\boldsymbol{\mu_k}\right)))\}]. \\
= & \sum_{n=1}^{N}[\frac{\pi_k}{\sum\limits_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})}\{\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})\frac{1}{2}\boldsymbol{\Sigma_k}+\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})\frac{\partial}{\partial\boldsymbol{\Sigma_k}^{-1}}(-\frac{1}{2}\left(\boldsymbol{x_n}-\boldsymbol{\mu_k}\right)^{T}\boldsymbol{\Sigma_k}^{-1}\left(\boldsymbol{x_n}-\boldsymbol{\mu_k}\right))\}]. \\
= & \sum_{n=1}^{N}\{r_{k,n}\frac{1}{2}(\boldsymbol{\Sigma_k}-(\boldsymbol{x_n}-\boldsymbol{\mu_k})(\boldsymbol{x_n}-\boldsymbol{\mu_k})^{T})\}.
\end{aligned}
\tag{3}
$$

(3) has to be 0, so we can obtain

$$
\boldsymbol{\Sigma_k}=\frac{1}{N_k}\sum_{n=1}^{N}r_{k,n}(\boldsymbol{x_n}-\boldsymbol{\mu_k})(\boldsymbol{x_n}-\boldsymbol{\mu_k})^{T}).
$$

For $\boldsymbol{\pi}$,

$$
\begin{aligned}
\frac{\partial(\mathcal{L}+\lambda(1-\sum\limits_{k=1}^{K}\pi_k))}{\partial\pi_k} & =\frac{\partial}{\partial\pi_k}\sum_{n=1}^{N}\log(\sum_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j}))-\lambda \\
& =\sum_{n=1}^{N}\frac{\partial}{\partial\pi_k}\log(\sum_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j}))-\lambda \\
& =\sum_{n=1}^{N}[\frac{1}{\sum\limits_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})}\frac{\partial}{\partial\pi_k}\{\sum_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})\}]-\lambda \\
& =\sum_{n=1}^{N}\frac{\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})}{\sum\limits_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})}-\lambda.
\end{aligned}
\tag{4}
$$

(4) has to be 0, so we can obtain

$$
\sum_{n=1}^{N}\frac{\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})}{\sum\limits_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})}=\lambda.
$$

After multiply both side by π_k,

$$
N_k=\lambda\pi,
$$

$$
\pi=\frac{N_k}{\lambda}.
$$

Since $\sum k = 1^K \pi_k = 1$, $\lambda = \sum\limits_{k=1}^{K} N_k = N$. Therefore,

$$\pi = \frac{N_k}{N}.$$

Like above, we can get approximation of an extreme value of Mixture of Gaussians with alternatively updating $r_{k,n}$ and $\{\boldsymbol{\mu}, \boldsymbol{\Sigma_k}, \boldsymbol{\pi}\}$, iteratively.

3 Derivation via EM optimization

For EM optimization, we have to represent the complete data log-likelihood. The complete data log-likelihood is given by

$$\mathcal{L}_c = \sum_{n=1}^{N} \log p(\boldsymbol{x_n}, \boldsymbol{z_n} | \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{n=1}^{N} \{\log p(\boldsymbol{x_n} | \boldsymbol{\pi}, \boldsymbol{\theta}) p(\boldsymbol{z_n} | \boldsymbol{\pi})\}.$$

Let $\boldsymbol{z_n}$ to be one-hot encoded vector so that $\boldsymbol{z_n} \in \{0, 1\}^K$, and $\pi_k = p(z_{k,n} = 1)$. We can represent probabilities as follows:

$$p(\boldsymbol{x_n} | \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x_n} | \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})^{z_{k,n}},$$

$$p(\boldsymbol{z_n} | \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{k,n}}.$$

By using this, the complete data log-likelihood can be written as

$$\begin{aligned} \mathcal{L}_c &= \sum_{n=1}^{N} \log(\prod_{k=1}^{K} \pi_k^{z_{k,n}} \mathcal{N}(\boldsymbol{x_n} | \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})^{z_{k,n}}) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{k,n} \log(\pi_k \mathcal{N}(\boldsymbol{x_n} | \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})). \end{aligned}$$

For the E step, we have to calculate the expectation of the complete data log-likelihood. Therefore,

$$\begin{aligned} \mathop{\mathbb{E}}_{p(z|x)}[\mathcal{L}_c] &= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathop{\mathbb{E}}_{p(z|x)}[z_{k,n} \log(\pi_k \mathcal{N}(\boldsymbol{x_n} | \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}))] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{k,n} \log(\pi_k \mathcal{N}(\boldsymbol{x_n} | \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{k,n} \{\log(\pi_k) - \frac{2}{D}(2\pi) - \frac{1}{2} \log(|\Sigma_k|) - \frac{1}{2}(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T \boldsymbol{\Sigma_k}^{-1} (\boldsymbol{x_n} - \boldsymbol{\mu_k})\} \end{aligned} \quad (5)$$

, since

$$\begin{aligned}
\mathop{\mathbb{E}}_{p(z|x)}[z_{k,n}] &= p(z_{k,n}=1|\boldsymbol{x_n}) \\
&= \frac{p(z_{k,n}=1)p(\boldsymbol{x_n}|z_{k,n}=1)}{\sum\limits_{\boldsymbol{z_n}} p(\boldsymbol{x_n},\boldsymbol{z_n})} \\
&= \frac{p(z_{k,n}=1)p(\boldsymbol{x_n}|z_{k,n}=1)}{\sum\limits_{\boldsymbol{z_n}} p(\boldsymbol{x_n},\boldsymbol{z_n})} \\
&= \frac{p(z_{k,n}=1)p(\boldsymbol{x_n}|z_{k,n}=1)}{\sum\limits_{j=1}^{K} p(z_{j,n}=1)p(\boldsymbol{x_n}|z_{j,n}=1)} \\
&= \frac{\pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})}{\sum\limits_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})} \\
&= r_{k,n}.
\end{aligned}$$

For the M step, we have to update parameters to maximize the expectation of complete data log-likelihood. We obtain partial derivatives of (3) and update to new parameters such that each partial derivatives go to 0. Therefore, with $\frac{\partial(3)}{\partial \boldsymbol{\mu_k}}$,

$$\begin{aligned}
\frac{\partial\, \mathbb{E}_{p(z|x)}[\mathcal{L}_c]}{\partial \boldsymbol{\mu_k}} &= \sum_{n=1}^{N} r_{k,n} \boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x_n} - \boldsymbol{\mu_k}) \\
&= 0. \qquad (6)
\end{aligned}$$

We can obtain

$$\boldsymbol{\mu_k} = \frac{\sum\limits_{n=1}^{N} r_{k,n}\boldsymbol{x_n}}{\sum\limits_{n=1}^{N} r_{k,n}}.$$

Similarly, we can also obtain

$$\boldsymbol{\Sigma_k} = \frac{1}{N_k}\sum_{n=1}^{N} r_{k,n}(\boldsymbol{x_n} - \boldsymbol{\mu_k})(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T),\ \text{and}$$

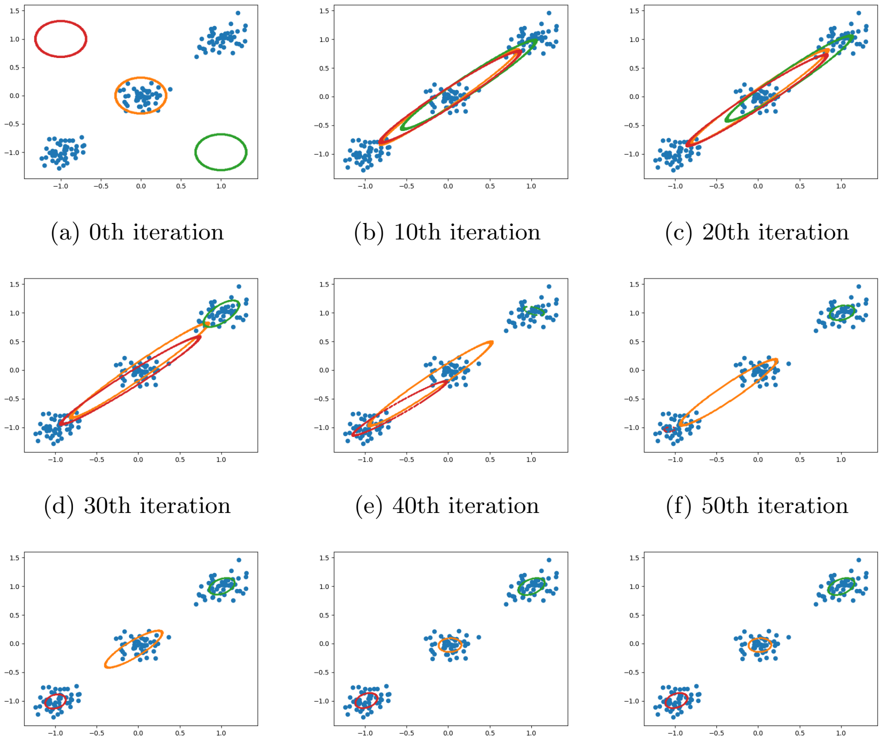
$$\pi = \frac{N_k}{N}$$

, which are same result with derivation via direct maximization of likelihood.

4 Experiments

We implemented the explained method via Python language with numpy library. We used matplotlib library to display the result. The source code is available in `https://github.com/inyukwo1/POSTECH_CSED524`. The implementation also includes experimental codes and 2D plotting codes.

For experiment, we made a toy dataset and applied the optimization method. The toy dataset is generated by three different 2D Gaussians, which have different means and covariance matrices. We generated 50 data each, so we could get total 150 data. We applied the optimization method to determine three clusters.



(a) 0th iteration

(b) 10th iteration

(c) 20th iteration

(d) 30th iteration

(e) 40th iteration

(f) 50th iteration

(g) 60th iteration

(h) 70th iteration

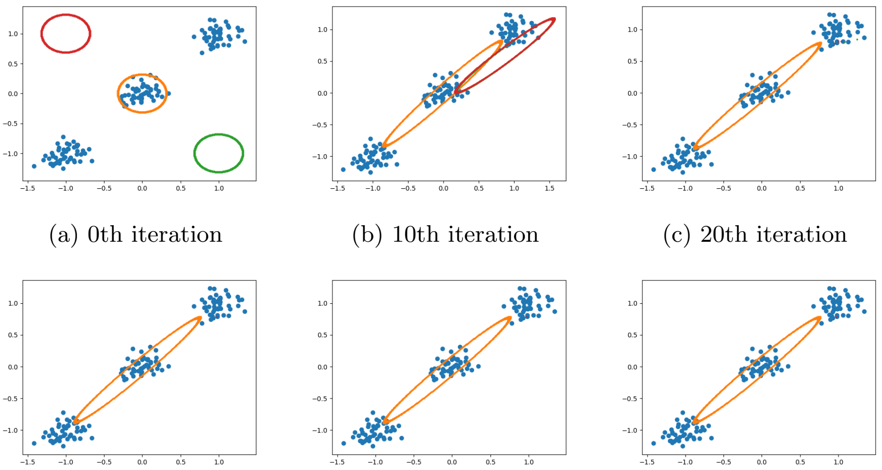
(i) 80th iteration

Figure 1: Successful result

Figure 1 shows the successful result to find three clusters. The blue dots represents data, and the other colored elipses represents the clusters which is found by optimization method. The eplises represent the set of points where $(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T \boldsymbol{\Sigma_k}^{-1} (\boldsymbol{x_n} - \boldsymbol{\mu_k})) = 1$. The optimization method often fails. Figure 2 shows the failed result. Failure depends on initialization of parameters and data. Thus, program sometimes occurs error when finding inverse matrices.

5 Conclusion

We could review and derived the optimization method of Mixture of Gaussians in two different ways. We also implemented the method and verified with a simple experiment. We could see both successful and unsuccessful results for finding clusters of a synthetic dataset.



(a) 0th iteration

(b) 10th iteration

(c) 20th iteration

(d) 30th iteration

(e) 40th iteration

(f) 50th iteration

Figure 2: Failed result