CSED524-Homework 2

Inhyuk Na

2018/10/24

1 Factor Analysis

In the lecture, we've learned about factor analysis model. In factor analysis model, we model real-valued D-dimensional (observed) data $\boldsymbol{x}$ by using an real-valued d-dimensional (hidden) factor $\boldsymbol{s}$. Formally, we can describe as

$$\boldsymbol{x_n} = \boldsymbol{A}\boldsymbol{s_n} + \boldsymbol{\epsilon_n},$$

where $\boldsymbol{A}$ is known as the factor loading matrix. We assume that $\boldsymbol{s_n}$ is $\mathcal{N}(0, \boldsymbol{I})$ distributed, and ϵ_n is $\mathcal{N}(0, \boldsymbol{\Sigma})$ distributed, where $\boldsymbol{\Sigma}$ is a diagonal matrix.

In this report, we introduce the two typical factor analysis models, which are probabilistic PCA(PPCA) and mixture of PPCA. We also explain the detailed derivations of them. Then we show experiments about our implementations of mixtures of PPCA.

2 Probabilistic PCA

Probabilistic PCA(PPCA) is a factor analysis model when $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$. In PPCA, the subspace defined by the columns of $\boldsymbol{A}$ corresponds to principal subspace. To find $\boldsymbol{A}$ and $\boldsymbol{\sigma}$, we can apply EM optimization as follows (the following terms are slightly different with the lecture slides, but it is basically indicates same meaning):

E step:

$$\begin{aligned}\mathbb{E}[\boldsymbol{s}_n|\boldsymbol{x}_n] &= \boldsymbol{\Phi}\boldsymbol{x}_n,\\ \mathbb{E}[\boldsymbol{s}_n\boldsymbol{s}_n^T|\boldsymbol{x}_n] &= \boldsymbol{I} - \boldsymbol{\Phi}\boldsymbol{A} + \boldsymbol{\Phi}\boldsymbol{x_n}\boldsymbol{x_n}^T\boldsymbol{\Phi}^T, \text{ where } \boldsymbol{\Phi} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T + \sigma^2\boldsymbol{I})^{-1}.\end{aligned}$$

M step:

$$\boldsymbol{A} \leftarrow (\sum_{n=1}^{N} \boldsymbol{x_n}(\mathbb{E}[\boldsymbol{s_n}|\boldsymbol{x_n}])^T)(\sum_{n=1}^{N} \mathbb{E}[\boldsymbol{s_n}\boldsymbol{s_n}^T|\boldsymbol{x_n}])^{-1},$$

$$\sigma^2 \leftarrow \frac{1}{ND}\sum_{n=1}^{N}\left(\boldsymbol{x_n}^T\boldsymbol{x_n} - 2\boldsymbol{x_n}^T\boldsymbol{A}\mathbb{E}(\boldsymbol{s_n}|\boldsymbol{x_n}) + tr(\boldsymbol{A}^T\boldsymbol{A}\mathbb{E}(\boldsymbol{s_n}\boldsymbol{s_n}^T|\boldsymbol{x_n}))\right).$$

Proof. For E step, we have to know $p(\boldsymbol{s}|\boldsymbol{x}, \theta)$ and $\mathbb{E}\mathcal{L}_C$. Before we calculate this, note some properties of the Normal Distribution [1]:

For the jointly normal vector

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{s} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}, \begin{bmatrix} \boldsymbol{B} & \boldsymbol{D} \\ \boldsymbol{D}^T & \boldsymbol{C} \end{bmatrix} \right),$$

$$p(\boldsymbol{x}|\boldsymbol{s}) = \mathcal{N}(\boldsymbol{a} + \boldsymbol{D}\boldsymbol{C}^{-1}(\boldsymbol{s} - \boldsymbol{b}), \boldsymbol{B} - \boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{D}^T) \text{ and} \tag{1}$$
$$p(\boldsymbol{s}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{b} + \boldsymbol{D}^T\boldsymbol{B}^{-1}(\boldsymbol{x} - \boldsymbol{a}), \boldsymbol{C} - \boldsymbol{D}^T\boldsymbol{B}^{-1}\boldsymbol{D}). \tag{2}$$

In our case,

$$\begin{aligned}
\boldsymbol{a} &= \boldsymbol{b} = 0, \\
\\
\boldsymbol{B} &= Cov(\boldsymbol{x}) \\
&= \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T) - \mathbb{E}(\boldsymbol{x})\mathbb{E}(\boldsymbol{x})^T \\
&= \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T) \\
&= \mathbb{E}((\boldsymbol{A}\boldsymbol{s} + \boldsymbol{\epsilon})(\boldsymbol{A}\boldsymbol{s} + \boldsymbol{\epsilon})^T) \\
&= \mathbb{E}(\boldsymbol{A}\boldsymbol{s}\boldsymbol{s}^T\boldsymbol{A}^T + \boldsymbol{A}\boldsymbol{s}\boldsymbol{\epsilon}^T + \boldsymbol{\epsilon}\boldsymbol{s}^T\boldsymbol{A}^T + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) \\
&= \boldsymbol{A}Cov(\boldsymbol{s})\boldsymbol{A}^T + \boldsymbol{A}\mathbb{E}(\boldsymbol{s})\mathbb{E}(\boldsymbol{\epsilon}^T) + \mathbb{E}(\boldsymbol{\epsilon})\mathbb{E}(\boldsymbol{s}^T)\boldsymbol{A}^T + Cov(\boldsymbol{\epsilon}) \text{ (since } \boldsymbol{\epsilon} \text{ and } \boldsymbol{s} \text{ are independent)} \\
&= \boldsymbol{A}\boldsymbol{A}^T + \sigma^2\boldsymbol{I}, \\
\\
\boldsymbol{C} &= Cov(\boldsymbol{s}) = \boldsymbol{I}, \text{ and} \\
\\
\boldsymbol{D} &= \mathbb{E}(\boldsymbol{x}\boldsymbol{s}^T) \\
&= \mathbb{E}((\boldsymbol{A}\boldsymbol{s} + \boldsymbol{\epsilon})\boldsymbol{s}^T) \\
&= \boldsymbol{A}Cov(\boldsymbol{s}) + \mathbb{E}(\boldsymbol{\epsilon})\mathbb{E}(\boldsymbol{s}^T) \text{ (since } \boldsymbol{\epsilon} \text{ and } \boldsymbol{s} \text{ are independent)} \\
&= \boldsymbol{A}.
\end{aligned}$$

Therefore, the posterior distribution is

$$p(\boldsymbol{s}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{s}|\boldsymbol{\Phi}\boldsymbol{x}, \boldsymbol{I} - \boldsymbol{\Phi}\boldsymbol{A}),$$

where $\boldsymbol{\Phi} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T + \sigma^2\boldsymbol{I})^{-1}$. For further calculation, we need the following two statistics:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{s}_n|\boldsymbol{x}_n] &= \boldsymbol{\Phi}\boldsymbol{x}_n, \\
\mathbb{E}[\boldsymbol{s}_n\boldsymbol{s}_n^T|\boldsymbol{x}_n] &= Cov(\boldsymbol{s_n}|\boldsymbol{x_n}) + \mathbb{E}[\boldsymbol{s}_n|\boldsymbol{x}_n](\mathbb{E}[\boldsymbol{s}_n|\boldsymbol{x}_n])^T \\
&= \boldsymbol{I} - \boldsymbol{\Phi}\boldsymbol{A} + \boldsymbol{\Phi}\boldsymbol{x_n}\boldsymbol{x_n}^T\boldsymbol{\Phi}^T.
\end{aligned} \tag{3}$$

Now we need to calculate the expectation of complete-data log likelihood,

$$\begin{aligned}
\mathbb{E}\mathcal{L}_C &= \mathbb{E}(\log(p(\boldsymbol{X},\boldsymbol{S}|\boldsymbol{A},\sigma^2))) \\
&= \mathbb{E}(\sum_{n=1}^{N}\log p(\boldsymbol{x_n}|\boldsymbol{s_n}) + \sum_{n=1}^{N}\log p(\boldsymbol{s_n})) \\
&= \mathbb{E}\left[\sum_{n=1}^{N}\log(\frac{1}{((2\pi)^D|\sigma^2\boldsymbol{I}|)^{1/2}}\exp\{-\frac{1}{2}(\boldsymbol{x_n}-\boldsymbol{A}\boldsymbol{s_n})^T(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{x_n}-\boldsymbol{A}\boldsymbol{s_n})\})\right] + C \\
&= -\frac{N}{2}\log((2\pi)^D|\sigma^2\boldsymbol{I}|) - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}\left[\frac{1}{\sigma^2}(\boldsymbol{x_n}-\boldsymbol{A}\boldsymbol{s_n})^T(\boldsymbol{x_n}-\boldsymbol{A}\boldsymbol{s_n})\right] + C \\
&= -\frac{ND}{2}\log(\sigma^2) - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}\left[\frac{1}{\sigma^2}(\boldsymbol{x_n}^T\boldsymbol{x_n} - 2\boldsymbol{x_n}^T\boldsymbol{A}\boldsymbol{s_n} + \boldsymbol{s_n}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{s_n})\right] + C \\
&= -\frac{ND}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(\boldsymbol{x_n}^T\boldsymbol{x_n} - 2\boldsymbol{x_n}^T\boldsymbol{A}\mathbb{E}(\boldsymbol{s_n}|\boldsymbol{x_n}) + tr(\boldsymbol{A}^T\boldsymbol{A}\mathbb{E}(\boldsymbol{s_n}\boldsymbol{s_n}^T|\boldsymbol{x_n}))\right) + C\ [2, pg.6], \quad (4)
\end{aligned}$$

where C represents the constant terms with respect to the parameters of the model.

For M step, we update $\boldsymbol{A}$ and σ^2 to maximize (4). To achieve this, we solve $\frac{\partial\mathbb{E}\mathcal{L}_C}{\partial\boldsymbol{A}} = 0$ and $\frac{\partial\mathbb{E}\mathcal{L}_C}{\partial\sigma^2} = 0$. From

$$\begin{aligned}
\frac{\partial\mathbb{E}\mathcal{L}_C}{\partial\boldsymbol{A}} &= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left[-2\boldsymbol{x_n}\mathbb{E}(\boldsymbol{s_n}|\boldsymbol{x_n})^T + (\boldsymbol{A}\mathbb{E}(\boldsymbol{s_n}\boldsymbol{s_n}^T|\boldsymbol{x_n}) + \boldsymbol{A}\mathbb{E}(\boldsymbol{s_n}\boldsymbol{s_n}^T|\boldsymbol{x_n})^T)\right]\ [2, pg.12] \\
&= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left[-2\boldsymbol{x_n}\mathbb{E}(\boldsymbol{s_n}|\boldsymbol{x_n})^T + 2\boldsymbol{A}\mathbb{E}(\boldsymbol{s_n}\boldsymbol{s_n}^T|\boldsymbol{x_n})\right] \\
&= 0,
\end{aligned}$$

we can obtain

$$\boldsymbol{A}^{new} = \left(\sum_{n=1}^{N}\boldsymbol{x_n}\mathbb{E}(\boldsymbol{s_n}|\boldsymbol{x_n})^T\right)\left(\sum_{n=1}^{N}\mathbb{E}(\boldsymbol{s_n}\boldsymbol{s_n}^T|\boldsymbol{x_n})\right)^{-1}. \quad (5)$$

Similarly, from

$$\begin{aligned}
\frac{\partial\mathbb{E}\mathcal{L}_C}{\partial\sigma^2} &= -\frac{ND}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=1}^{N}\left(\boldsymbol{x_n}^T\boldsymbol{x_n} - 2\boldsymbol{x_n}^T\boldsymbol{A}\mathbb{E}(\boldsymbol{s_n}|\boldsymbol{x_n}) + tr(\boldsymbol{A}^T\boldsymbol{A}\mathbb{E}(\boldsymbol{s_n}\boldsymbol{s_n}^T|\boldsymbol{x_n}))\right) \\
&= 0 \text{ and } (5),
\end{aligned}$$

we can obtain

$$\sigma^{2^{new}} = \frac{1}{ND}\sum_{n=1}^{N}\left(\boldsymbol{x_n}^T\boldsymbol{x_n} - 2\boldsymbol{x_n}^T\boldsymbol{A}\mathbb{E}(\boldsymbol{s_n}|\boldsymbol{x_n}) + tr(\boldsymbol{A}^T\boldsymbol{A}\mathbb{E}(\boldsymbol{s_n}\boldsymbol{s_n}^T|\boldsymbol{x_n}))\right).$$

□

3 Mixtures of PPCA

Mixtures of PPCA model assumes both Gaussian and categorical latent variables, so to perform clustering and dimensionality reduction simultaneously. This model assumes

$$p(\boldsymbol{s_n}) = \mathcal{N}(\boldsymbol{s_n}|0, \boldsymbol{I}) \text{ and}$$

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

for the latent variables. This model also assumes the distribution over observed variables conditioned on latent variables as

$$p(\boldsymbol{x_n}|\boldsymbol{s_n}, \boldsymbol{z_n}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{A_k}\boldsymbol{s_n} + \boldsymbol{\mu_k}, \sigma_k^2\boldsymbol{I})^{z_{k,n}}.$$

We can apply EM optimization as follows:

E step:

$$\begin{aligned}
r_{k,n} = p(z_{k,n} = 1|\boldsymbol{x_n}) &= \frac{\pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{A_k}\boldsymbol{A_k}^T + \sigma_k^2\boldsymbol{I})}{\sum\limits_j \pi_j \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j}, \boldsymbol{A_j}\boldsymbol{A_j}^T + \sigma_j^2\boldsymbol{I})}, \\
\mathbb{E}[\boldsymbol{s_n}|z_{k,n} = 1, \boldsymbol{x_n}] &= \boldsymbol{\Phi_k}(\boldsymbol{x_n} - \boldsymbol{\mu_k}), \\
\mathbb{E}[\boldsymbol{s_n}\boldsymbol{s_n}^T|z_{k,n} = 1, \boldsymbol{x_n}] &= \boldsymbol{I} - \boldsymbol{\Phi_k}\boldsymbol{A_k} + \boldsymbol{\Phi_k}(\boldsymbol{x_n} - \boldsymbol{\mu_k})(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T\boldsymbol{\Phi_k}^T,
\end{aligned}$$

where $\boldsymbol{\Phi_k} = \boldsymbol{A_k}^T(\boldsymbol{A_k}\boldsymbol{A_k}^T + \sigma_k^2\boldsymbol{I})^{-1}$.

M step:

$$\begin{aligned}
\boldsymbol{A_k} &\leftarrow \left(\sum_{n=1}^{N} r_{k,n}(\boldsymbol{x_n} - \boldsymbol{\mu_k})\mathbb{E}[\boldsymbol{s_n}|z_{k,n} = 1, \boldsymbol{x_n}]^T\right)\left(\sum_{n=1}^{N} r_{k,n}\mathbb{E}[\boldsymbol{s_n}\boldsymbol{s_n}^T|z_{k,n} = 1, \boldsymbol{x_n}]\right)^{-1}, \\
\boldsymbol{\mu_k} &\leftarrow \frac{\sum\limits_{n=1}^{N} r_{k,n}(\boldsymbol{x_n} - \boldsymbol{A_k}\mathbb{E}[\boldsymbol{s_n}|z_{k,n} = 1, \boldsymbol{x_n}])}{\sum\limits_{n=1}^{N} r_{k,n}}, \\
\sigma_k^2 &\leftarrow \frac{1}{D\sum\limits_{n=1}^{N} r_{k,n}}\Big[\sum_{n=1}^{N} r_{k,n}(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T(\boldsymbol{x_n} - \boldsymbol{\mu_k}) \\
&\quad - \sum_{n=1}^{N} 2r_{k,n}\mathbb{E}[\boldsymbol{s_n}|z_{k,n} = 1, \boldsymbol{x_n}]^T\boldsymbol{A_k}^T(\boldsymbol{x_n} - \boldsymbol{\mu_k}) + \sum_{n=1}^{N} r_{k,n}tr(\mathbb{E}[\boldsymbol{s_n}\boldsymbol{s_n}^T|z_{k,n} = 1, \boldsymbol{x_n}]\boldsymbol{A_k}^T\boldsymbol{A_k})\Big], \\
\pi_k &\leftarrow \frac{\sum\limits_{n=1}^{N} r_{k,n}}{N}.
\end{aligned}$$

Proof. For E step, we derive complete-data log likelihood as

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_C] &= \mathbb{E}[\sum_{n=1}^{N} \log p(\boldsymbol{x_n}, \boldsymbol{s_n}, \boldsymbol{z_n})] \\
&= \mathbb{E}\left[\sum_{n=1}^{N} \{\log p(\boldsymbol{x_n}|\boldsymbol{s_n}, \boldsymbol{z_n}) + \log p(\boldsymbol{s_n}) + \log p(\boldsymbol{z_n})\}\right] \\
&= \sum_{n=1}^{N} \mathbb{E}\left[\sum_{k=1}^{K} z_{k,n} \log \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{A_k}\boldsymbol{s_n} + \boldsymbol{\mu}, \sigma_k \boldsymbol{I}) + \log \mathcal{N}(\boldsymbol{s_n}|0, \boldsymbol{I}) + \sum_{k=1}^{K} z_{k,n} \log \pi_k\right] \\
&= \sum_{n=1}^{N} \mathbb{E}\Big[\sum_{k=1}^{K} z_{k,n} \log\left(\frac{1}{((2\pi)^D|\sigma_k^2 \boldsymbol{I}|)^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x_n} - \boldsymbol{A_k}\boldsymbol{s_n} - \boldsymbol{\mu_k})^T(\sigma_k^2 \boldsymbol{I})^{-1}(\boldsymbol{x_n} - \boldsymbol{A_k}\boldsymbol{s_n} - \boldsymbol{\mu_k})\right)\right) \\
&\quad + \sum_{k=1}^{K} z_{k,n} \log \pi_k\Big] + C \\
&= \sum_{n=1}^{N} \mathbb{E}\Big[\sum_{k=1}^{K} z_{k,n} \log\left(\frac{1}{(2\pi\sigma_k^2)^{D/2}} \exp\left(-\frac{1}{2\sigma_k^2}(\boldsymbol{x_n} - \boldsymbol{A_k}\boldsymbol{s_n} - \boldsymbol{\mu_k})^T(\boldsymbol{x_n} - \boldsymbol{A_k}\boldsymbol{s_n} - \boldsymbol{\mu_k})\right)\right) \\
&\quad + \sum_{k=1}^{K} z_{k,n} \log \pi_k\Big] + C \\
&= \sum_{n=1}^{N} \Big[\sum_{k=1}^{K} \mathbb{E}(z_{k,n}) \log\left(\frac{1}{(2\pi\sigma_k^2)^{D/2}}\right) + \sum_{k=1}^{K} \mathbb{E}\left[z_{k,n}\left(-\frac{1}{2\sigma_k^2}(\boldsymbol{x_n} - \boldsymbol{A_k}\boldsymbol{s_n} - \boldsymbol{\mu_k})^T(\boldsymbol{x_n} - \boldsymbol{A_k}\boldsymbol{s_n} - \boldsymbol{\mu_k})\right)\right] \\
&\quad + \sum_{k=1}^{K} \mathbb{E}(z_{k,n}) \log \pi_k\Big] + C \\
&= \sum_{n=1}^{N} \Big[\sum_{k=1}^{K} \mathbb{E}(z_{k,n}) \log\left(\frac{1}{(2\pi\sigma_k^2)^{D/2}}\right) + \sum_{k=1}^{K} \mathbb{E}\Big[z_{k,n}\Big(-\frac{1}{2\sigma_k^2}(\boldsymbol{x_n}^T\boldsymbol{x_n} - \boldsymbol{x_n}^T\boldsymbol{A_k}\boldsymbol{s_n} - \boldsymbol{x_n}^T\boldsymbol{\mu_k} \\
&\quad - \boldsymbol{s_n}^T\boldsymbol{A_k}^T\boldsymbol{x_n} + \boldsymbol{s_n}^T\boldsymbol{A_k}^T\boldsymbol{A_k}\boldsymbol{s_n} + \boldsymbol{s_n}^T\boldsymbol{A_k}^T\boldsymbol{\mu_k} - \boldsymbol{\mu_k}^T\boldsymbol{x_n} + \boldsymbol{\mu_k}^T\boldsymbol{A_k}\boldsymbol{s_n} + \boldsymbol{\mu_k}^T\boldsymbol{\mu_k})\Big)\Big] + \sum_{k=1}^{K} \mathbb{E}(z_{k,n}) \log \pi_k\Big] + C \\
&= \sum_{n=1}^{N} \Big[\sum_{k=1}^{K} \mathbb{E}(z_{k,n}) \log\left(\frac{1}{(2\pi\sigma_k^2)^{D/2}}\right) + \sum_{k=1}^{K} \mathbb{E}\Big[z_{k,n}\Big(-\frac{1}{2\sigma_k^2}(\boldsymbol{x_n}^T\boldsymbol{x_n} - \boldsymbol{x_n}^T\boldsymbol{\mu_k} \\
&\quad - 2\boldsymbol{s_n}^T\boldsymbol{A_k}^T\boldsymbol{x_n} + tr(\boldsymbol{s_n}\boldsymbol{s_n}^T\boldsymbol{A_k}^T\boldsymbol{A_k}) + \boldsymbol{s_n}^T\boldsymbol{A_k}^T\boldsymbol{\mu_k} - \boldsymbol{\mu_k}^T\boldsymbol{x_n} + \boldsymbol{\mu_k}^T\boldsymbol{A_k}\boldsymbol{s_n} + \boldsymbol{\mu_k}^T\boldsymbol{\mu_k})\Big)\Big] + \sum_{k=1}^{K} \mathbb{E}(z_{k,n}) \log \pi_k\Big] + C \\
&= \sum_{n=1}^{N} \Big[\sum_{k=1}^{K} \mathbb{E}(z_{k,n}) \log\left(\frac{1}{(2\pi\sigma_k^2)^{D/2}}\right) + \sum_{k=1}^{K} \Big[-\frac{1}{2\sigma_k^2}(\mathbb{E}(z_{k,n})\boldsymbol{x_n}^T\boldsymbol{x_n} - \mathbb{E}(z_{k,n})\boldsymbol{x_n}^T\boldsymbol{\mu_k} \\
&\quad - 2\mathbb{E}(z_{k,n}\boldsymbol{s_n})^T\boldsymbol{A_k}^T\boldsymbol{x_n} + tr(\mathbb{E}(z_{k,n}\boldsymbol{s_n}\boldsymbol{s_n}^T)\boldsymbol{A_k}^T\boldsymbol{A_k}) + \mathbb{E}(z_{k,n}\boldsymbol{s_n})^T\boldsymbol{A_k}^T\boldsymbol{\mu_k} - \mathbb{E}(z_{k,n})\boldsymbol{\mu_k}^T\boldsymbol{x_n} \\
&\quad + \boldsymbol{\mu_k}^T\boldsymbol{A_k}\mathbb{E}(z_{k,n}\boldsymbol{s_n}) + \mathbb{E}(z_{k,n})\boldsymbol{\mu_k}^T\boldsymbol{\mu_k})\Big] + \sum_{k=1}^{K} \mathbb{E}(z_{k,n}) \log \pi_k\Big] + C
\end{aligned}$$

, where C represents the constant terms with respect to the parameters of the model. We use the term $r_{k,n}$ for $\mathbb{E}[z_{k,n}]$, and this is

$$\begin{aligned}
r_{k,n} &= \mathbb{E}[z_{k,n}] \\
&= \sum_{z_{k,n}} z_{k,n} p(z_{k,n}|\boldsymbol{x_n}) \\
&= p(z_{k,n} = 1|\boldsymbol{x_n}). \qquad (6)
\end{aligned}$$

When $z_{k,n} = 1$, $\begin{bmatrix}\boldsymbol{x_n}\\ \boldsymbol{s_n}\end{bmatrix}$ follows $\mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu_k}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{A_k}\boldsymbol{A_k}^T + \sigma_k^2\boldsymbol{I} & \boldsymbol{A_k}\\ \boldsymbol{A_k}^T & \boldsymbol{I}\end{bmatrix}\right)$, as explained in section 2. Therefore $p(\boldsymbol{x_n}|z_{k,n} = 1) = \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{A_k}\boldsymbol{A_k}^T + \sigma_k^2\boldsymbol{I})$[2, pg.40]. Using this we can finish the caclculation of (7),

$$
\begin{aligned}
r_{k,n} = p(z_{k,n} = 1|\boldsymbol{x_n}) &= \frac{p(\boldsymbol{x_n}|z_{k,n} = 1)p(z_{k,n} = 1)}{\sum\limits_j p(\boldsymbol{x_n}|z_{j,n} = 1)p(z_{j,n} = 1)}\\
&= \frac{\pi_k\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{A_k}\boldsymbol{A_k}^T + \sigma_k^2\boldsymbol{I})}{\sum\limits_j \pi_j\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_j}, \boldsymbol{A_j}\boldsymbol{A_j}^T + \sigma_j^2\boldsymbol{I})}.
\end{aligned}
$$

For $\mathbb{E}[z_{k,n}\boldsymbol{s_n}]$, we can derive

$$
\begin{aligned}
\mathbb{E}[z_{k,n}\boldsymbol{s_n}] &= \int_{\boldsymbol{s_n}} \sum_{z_{k,n}} (z_{k,n}\boldsymbol{s_n}p(z_{k,n}, \boldsymbol{s_n}|\boldsymbol{x_n}))\\
&= \int_{\boldsymbol{s_n}} (\boldsymbol{s_n}p(z_{k,n} = 1, \boldsymbol{s_n}|\boldsymbol{x_n}))\\
&= \int_{\boldsymbol{s_n}} (\boldsymbol{s_n}p(z_{k,n} = 1|\boldsymbol{x_n})p(\boldsymbol{s_n}|z_{k,n} = 1, \boldsymbol{x_n}))\\
&= r_{k,n}\mathbb{E}[\boldsymbol{s_n}|z_{k,n} = 1, \boldsymbol{x_n}]. \qquad (7)
\end{aligned}
$$

With joint distribution $\begin{bmatrix}\boldsymbol{x_n}\\ \boldsymbol{s_n}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu_k}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{A_k}\boldsymbol{A_k}^T + \sigma_k^2\boldsymbol{I} & \boldsymbol{A_k}\\ \boldsymbol{A_k}^T & \boldsymbol{I}\end{bmatrix}\right)$ and [2, pg.41],

$$
\mathbb{E}[\boldsymbol{s_n}|z_{k,n} = 1, \boldsymbol{x_n}] = \boldsymbol{\Phi_k}(\boldsymbol{x_n} - \boldsymbol{\mu_k})
$$

is hold, here $\boldsymbol{\Phi_k} = \boldsymbol{A_k}^T(\boldsymbol{A_k}\boldsymbol{A_k}^T + \sigma_k^2\boldsymbol{I})^{-1}$. Similarly,

$$
\mathbb{E}[z_{k,n}\boldsymbol{s_n}\boldsymbol{s_n}^T] = r_{k,n}\mathbb{E}[\boldsymbol{s_n}\boldsymbol{s_n}^T|z_{k,n} = 1, \boldsymbol{x_n}]
$$

is hold, and similar with (3),

$$
\mathbb{E}[\boldsymbol{s_n}\boldsymbol{s_n}^T|z_{k,n} = 1, \boldsymbol{x_n}] = \boldsymbol{I} - \boldsymbol{\Phi_k}\boldsymbol{A_k} + \boldsymbol{\Phi_k}(\boldsymbol{x_n} - \boldsymbol{\mu_k})(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T\boldsymbol{\Phi_k}^T
$$

is also hold.

For M step, we solve the equations $\frac{\partial\mathbb{E}\mathcal{L}_\mathcal{C}}{\partial\boldsymbol{A_k}} = 0$, $\frac{\partial\mathbb{E}\mathcal{L}_\mathcal{C}}{\partial\boldsymbol{\mu_k}} = 0$, $\frac{\partial\mathbb{E}\mathcal{L}_\mathcal{C}}{\partial\sigma_k^2} = 0$, and $\frac{\partial\mathbb{E}\mathcal{L}_\mathcal{C}}{\partial\pi_k} = 0$. With $\frac{\partial\mathbb{E}\mathcal{L}_\mathcal{C}}{\partial\boldsymbol{A_k}} = 0$ and (6),

$$
\begin{aligned}
\frac{\partial\mathbb{E}\mathcal{L}_\mathcal{C}}{\partial\boldsymbol{A_k}} &= \sum_{n=1}^N\left[-\frac{1}{2\sigma_k^2}(-2\boldsymbol{x_n}\mathbb{E}(z_{k,n}\boldsymbol{s_n})^T + \boldsymbol{A_k}\mathbb{E}(z_{k,n}\boldsymbol{s_n}\boldsymbol{s_n}^T)^T + \boldsymbol{A_k}\mathbb{E}(z_{k,n}\boldsymbol{s_n}\boldsymbol{s_n}^T) + \boldsymbol{\mu_k}\mathbb{E}(z_{k,n}\boldsymbol{s_n})^T + \boldsymbol{\mu_k}\mathbb{E}(z_{k,n}\boldsymbol{s_n})^T)\right]\\
&= \sum_{n=1}^N\left[-\frac{1}{2\sigma_k^2}(-2\boldsymbol{x_n}\mathbb{E}(z_{k,n}\boldsymbol{s_n})^T + 2\boldsymbol{A_k}\mathbb{E}(z_{k,n}\boldsymbol{s_n}\boldsymbol{s_n}^T) + 2\boldsymbol{\mu_k}\mathbb{E}(z_{k,n}\boldsymbol{s_n})^T)\right]\\
&= \boldsymbol{0} \text{ [2, pg.10, 13].}
\end{aligned}
$$

Then we can obtain

$$
\begin{aligned}
\boldsymbol{A_k} &= \left(\sum_{n=1}^N(\boldsymbol{x_n} - \boldsymbol{\mu_k})\mathbb{E}(z_{k,n}\boldsymbol{s_n})^T\right)\left(\sum_{n=1}^N\mathbb{E}(z_{k,n}\boldsymbol{s_n}\boldsymbol{s_n})^T\right)^{-1}\\
&= \left(\sum_{n=1}^N r_{k,n}(\boldsymbol{x_n} - \boldsymbol{\mu_k})\mathbb{E}[\boldsymbol{s_n}|z_{k,n} = 1, \boldsymbol{x_n}]^T\right)\left(\sum_{n=1}^N r_{k,n}\mathbb{E}[\boldsymbol{s_n}\boldsymbol{s_n}^T|z_{k,n} = 1, \boldsymbol{x_n}]\right)^{-1}.
\end{aligned}
$$

With $\frac{\partial \mathbb{E}\mathcal{L}_{\mathcal{C}}}{\partial \boldsymbol{\mu_k}} = 0$ and (6),

$$\begin{aligned}
\frac{\partial \mathbb{E}\mathcal{L}_{\mathcal{C}}}{\partial \boldsymbol{\mu_k}} &= \sum_{n=1}^{N} \left[-\frac{1}{2\sigma_k^2}(-\mathbb{E}(z_{k,n})\boldsymbol{x_n} + \boldsymbol{A_k}\mathbb{E}(z_{k,n}\boldsymbol{s_n}) - \mathbb{E}(z_{k,n})\boldsymbol{x_n} + \boldsymbol{A_k}\mathbb{E}(z_{k,n}\boldsymbol{s_n}) + 2\mathbb{E}(z_{k,n})\boldsymbol{\mu_k})\right] \\
&= \sum_{n=1}^{N} \left[-\frac{1}{2\sigma_k^2}(-2\mathbb{E}(z_{k,n})\boldsymbol{x_n} + 2\boldsymbol{A_k}\mathbb{E}(z_{k,n}\boldsymbol{s_n}) + 2\mathbb{E}(z_{k,n})\boldsymbol{\mu_k})\right] \\
&= \mathbf{0} \text{ [2, pg.10, 11].}
\end{aligned}$$

Then we can obtain

$$\begin{aligned}
\boldsymbol{\mu_k} &= \frac{\sum\limits_{n=1}^{N} \mathbb{E}(z_{k,n})\boldsymbol{x_n} - \boldsymbol{A_k}\mathbb{E}(z_{k,n}\boldsymbol{s_n})}{\sum\limits_{n=1}^{N} \mathbb{E}(z_{k,n})} \\
&= \frac{\sum\limits_{n=1}^{N} r_{k,n}(\boldsymbol{x_n} - \boldsymbol{A_k}\mathbb{E}[\boldsymbol{s_n}|z_{k,n} = 1, \boldsymbol{x_n}])}{\sum\limits_{n=1}^{N} r_{k,n}}.
\end{aligned}$$

With $\frac{\partial \mathbb{E}\mathcal{L}_{\mathcal{C}}}{\partial \sigma_k^2} = 0$ and (6) (in this derivation, we start from the middle line of (6)),

$$\begin{aligned}
\frac{\partial \mathbb{E}\mathcal{L}_{\mathcal{C}}}{\partial \sigma_k^2} &= \frac{\partial \sum\limits_{n=1}^{N} \left[\mathbb{E}(z_{k,n}) \log\left(\frac{1}{(2\pi\sigma_k^2)^{D/2}}\right) + \mathbb{E}\left[z_{k,n}\left(-\frac{1}{2\sigma_k^2}(\boldsymbol{x_n} - \boldsymbol{A_k s_n} - \boldsymbol{\mu_k})^T(\boldsymbol{x_n} - \boldsymbol{A_k s_n} - \boldsymbol{\mu_k})\right)\right]\right]}{\partial \sigma_k^2} \\
&= \sum_{n=1}^{N} -\frac{D}{2}\mathbb{E}(z_{k,n})\frac{1}{\sigma_k^2} + \sum_{n=1}^{N} \frac{1}{2\sigma_k^4}\mathbb{E}\left[z_{k,n}\left((\boldsymbol{x_n} - \boldsymbol{A_k s_n} - \boldsymbol{\mu_k})^T(\boldsymbol{x_n} - \boldsymbol{A_k s_n} - \boldsymbol{\mu_k})\right)\right] \\
&= 0.
\end{aligned}$$

Then we can obtain

$$\begin{aligned}
\sigma_k^2 &= \frac{1}{D\sum\limits_{n=1}^{N} \mathbb{E}(z_{k,n})} \sum_{n=1}^{N} \mathbb{E}\left[z_{k,n}\left((\boldsymbol{x_n} - \boldsymbol{A_k s_n} - \boldsymbol{\mu_k})^T(\boldsymbol{x_n} - \boldsymbol{A_k s_n} - \boldsymbol{\mu_k})\right)\right] \\
&= \frac{1}{D\sum\limits_{n=1}^{N} r_{k,n}} \sum_{n=1}^{N} \mathbb{E}\left[z_{k,n}\left((\boldsymbol{x_n} - \boldsymbol{\mu_k})^T(\boldsymbol{x_n} - \boldsymbol{\mu_k}) - 2\boldsymbol{s_n}^T\boldsymbol{A_k}^T(\boldsymbol{x_n} - \boldsymbol{\mu_k}) + \boldsymbol{s_n}^T\boldsymbol{A_k}^T\boldsymbol{A_k s_n}\right)\right] \\
&= \frac{1}{D\sum\limits_{n=1}^{N} r_{k,n}} \sum_{n=1}^{N} \{\mathbb{E}(z_{k,n})(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T(\boldsymbol{x_n} - \boldsymbol{\mu_k}) - 2\mathbb{E}(z_{k,n}\boldsymbol{s_n})^T\boldsymbol{A_k}^T(\boldsymbol{x_n} - \boldsymbol{\mu_k}) + tr(\mathbb{E}(z_{k,n}\boldsymbol{s_n}\boldsymbol{s_n}^T)\boldsymbol{A_k}^T\boldsymbol{A_k})\} \\
&= \frac{1}{D\sum\limits_{n=1}^{N} r_{k,n}} \Big[\sum_{n=1}^{N} r_{k,n}(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T(\boldsymbol{x_n} - \boldsymbol{\mu_k}) \\
&\quad - \sum_{n=1}^{N} 2r_{k,n}\mathbb{E}[\boldsymbol{s_n}|z_{k,n} = 1, \boldsymbol{x_n}]^T\boldsymbol{A_k}^T(\boldsymbol{x_n} - \boldsymbol{\mu_k}) + \sum_{n=1}^{N} r_{k,n} tr(\mathbb{E}[\boldsymbol{s_n}\boldsymbol{s_n}^T|z_{k,n} = 1, \boldsymbol{x_n}]\boldsymbol{A_k}^T\boldsymbol{A_k})\Big].
\end{aligned} \tag{8}$$

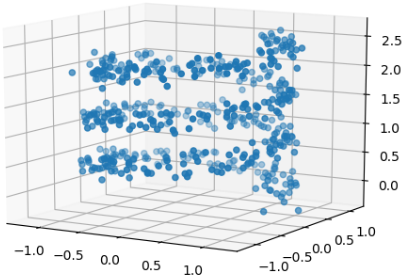


Figure 1: Dataset of the first experiment

With $\frac{\partial \mathbb{E}\mathcal{L}_{\mathcal{C}}}{\partial \pi_k} = 0$ and (6) with Lagrange multiplier $\lambda(\sum\limits_{j=1}^{K} \pi_j - 1)$,

$$\begin{aligned} \frac{\partial \mathbb{E}\mathcal{L}_{\mathcal{C}} + \lambda(\sum\limits_{j=1}^{K} \pi_j - 1)}{\partial \pi_k} &= \sum_{n=1}^{N} \mathbb{E}(z_{k,n}) \frac{1}{\pi_k} + \lambda \\ &= 0. \end{aligned} \tag{9}$$

Then we can obtain

$$\pi_k = \frac{\sum\limits_{n=1}^{N} r_{k,n}}{-\lambda}.$$

□

Since $\sum\limits_{j=1}^{K} \sum\limits_{n=1}^{N} r_{k,n} = N$, we can decide π_k as

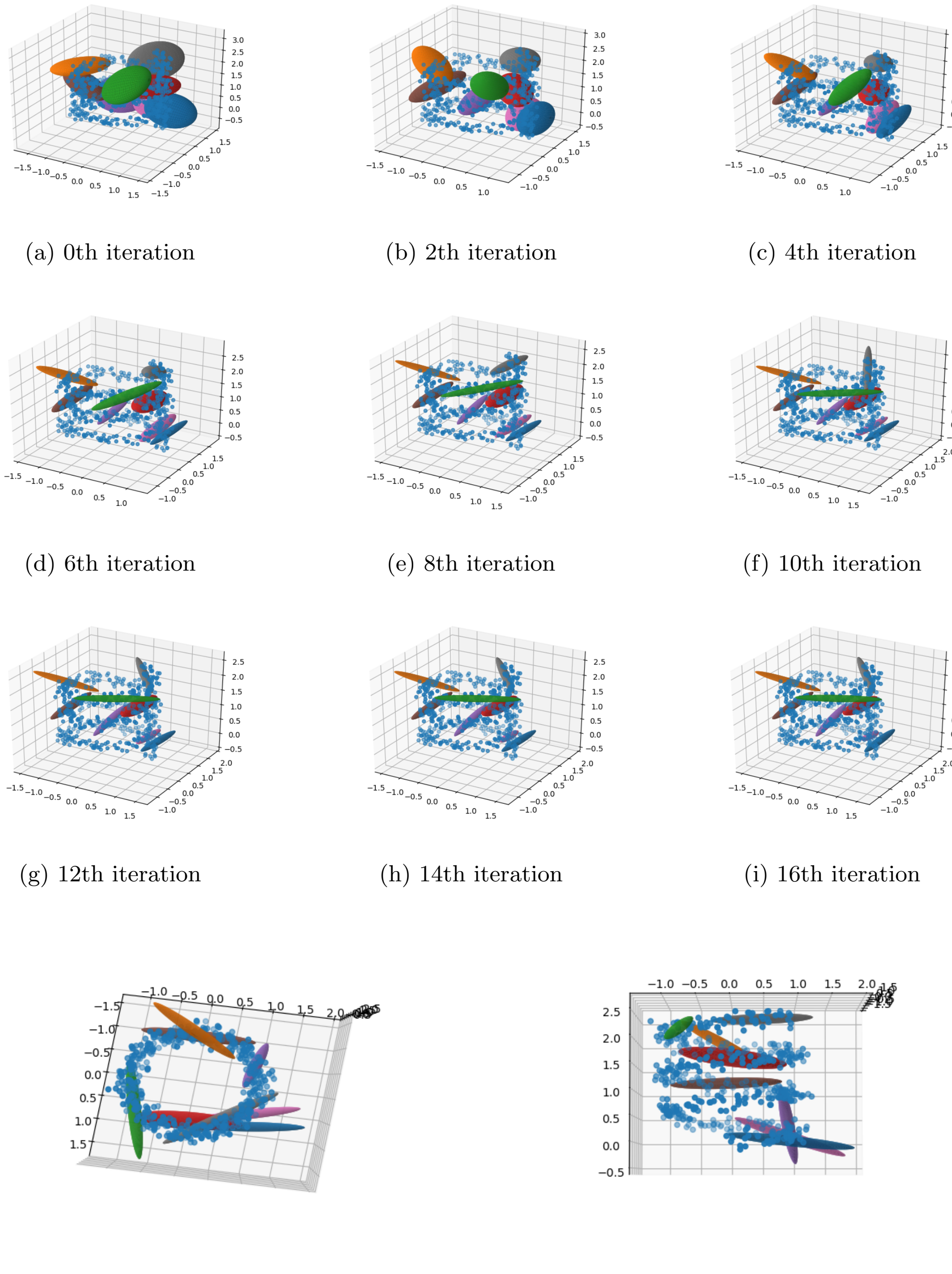
$$\pi_k = \frac{\sum\limits_{n=1}^{N} r_{k,n}}{N}.$$

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 plotting codes.

We experimented in two different ways. For the first experiment, we first reproduced the result of Tipping and Bishop's experiments[3]. In this experiment, we generated 500 three dimensional spiral data with Gaussian noise. The shape of dataset is shown in Figure 1. We also followed the other settings. We set the number of clusters as 8, the hidden data's dimension as 1. The result is shown in Figure 2. The ellipsoid represents $\boldsymbol{A_k}$, $\boldsymbol{\mu_k}$, and σ_k. To get each ellipsoid, we first decide the shape of ellipsoid by σ_k, rotate the ellipsoid by $\boldsymbol{A_k}$, and then moved the ellipsoid by $\boldsymbol{\mu_k}$.

For the second experiment, we generated the another dataset, which is shown in Figure 3. We followed the another settings of the first experiment. The result is shown in Figure 4.



(a) 0th iteration

(b) 2th iteration

(c) 4th iteration

(d) 6th iteration

(e) 8th iteration

(f) 10th iteration

(g) 12th iteration

(h) 14th iteration

(i) 16th iteration

(j) 16th iteration in another viewpoint

(k) 16th iteration in another viewpoint

Figure 2: Result of the first experiment

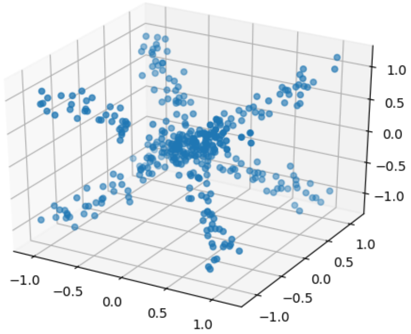
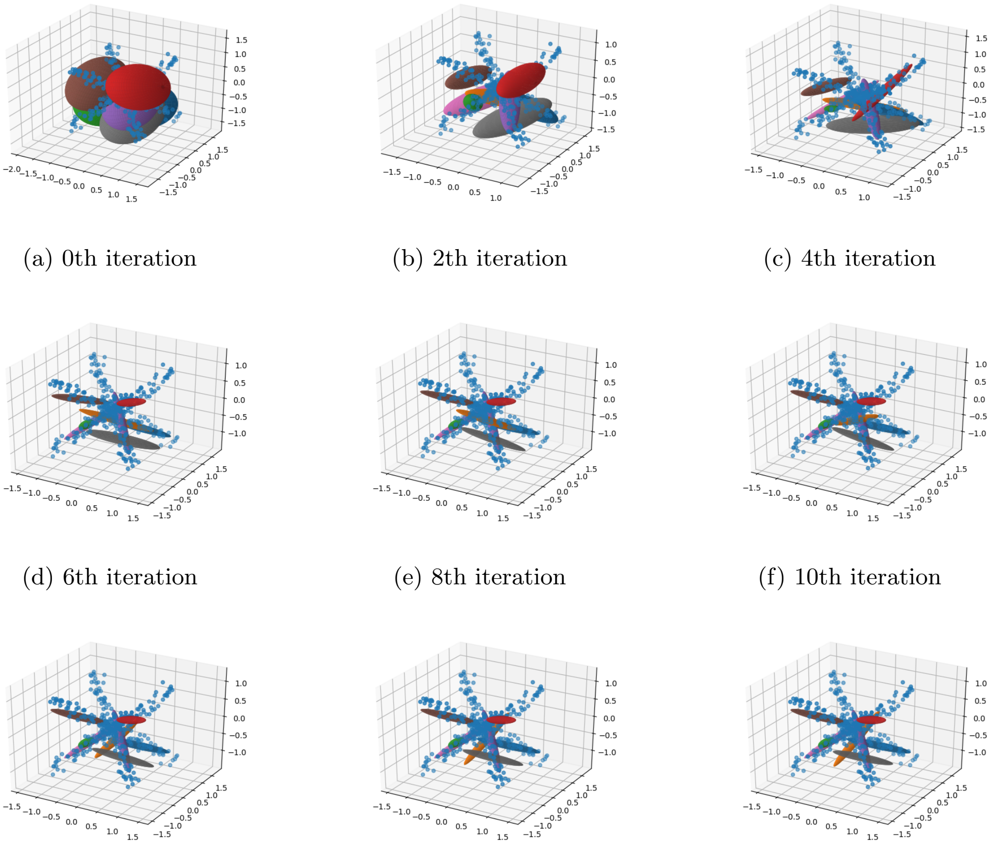


Figure 3: Dataset of the second experiment


(a) 0th iteration

(b) 2th iteration

(c) 4th iteration

(d) 6th iteration

(e) 8th iteration

(f) 10th iteration

(g) 12th iteration

(h) 14th iteration

(i) 16th iteration

Figure 4: Result of the second experiment

4.1 Considerations

In the result, the direction and location of ellipsoids follow the shape of datasets. Therefore we can regard that the mixture of PPCA could find the density of dataset. However, during the experiments, we could see the quality of results highly depend on the initial parameters. In this experiment we just initialized $\boldsymbol{A_k}$, σ_k, and π_k as the same small value and $\boldsymbol{\mu_k}$ as the randomly picked data of the datasets. The result quality mainly depend on the initial value of $\boldsymbol{\mu_k}$. Therefore, imporoved techniques for initializing parameters are needed. We suggest initialize $\boldsymbol{\mu_k}$ by another clustering methods.(e.g. K-means clustering) Thus, we didn't change the number of clusters in this experiments, but choosing the number of clusters can also be an interesting problem.

5 Conclusion

We could review and derived the optimization method of PPCA and mixture of PPCA. We also implemented the method and verified with two simple experiment. We could see successful results for representing the dataset by mixture of PPCA. Thus we considered about the result and further problems.

References

[1] Chuong B Do. More on multivariate gaussians. 2008.

[2] Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook (version: November 15, 2012), 2012.

[3] Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.