Pattern Recognition

Daiki Tanaka 6930-30-3041

July 19, 2018

1 ML estimation

1.1 Derive the update formulas of the parameters by letting the partial derivative of the lower bound w.r.t. each parameter equal to zero.

In ML estimation, we aim to estimate

$$\theta^\star = arg\max_\theta p(X;\theta) \tag{1}$$

$p(X;\theta)$ is given as

$$\begin{aligned}
\log p(X;\theta) &= \log \int p(X,Z;\theta)dZ \\
&= \log \int q(Z)\frac{p(X,Z;\theta)}{q(Z)}dZ \\
&\geq \int q(Z)\log\frac{p(X,Z;\theta)}{q(Z)}dZ \\
&= \int q(Z)\log p(X,Z;\theta)dZ - \int q(Z)\log q(Z)dZ \\
&= \int q(Z|X;\theta^{old})\log p(X,Z;\theta)dZ + const.(\because q(Z) = q(Z|X;\theta^{old})) \\
&= Q(\theta,\theta^{old}) + const.
\end{aligned}$$

In M-Step, we maximize lower bound w.r.t θ. Therefore, optimal $\theta^\star$ is given as

$$\theta^\star = arg\max_\theta Q(\theta,\theta^{old}) \tag{2}$$

This is calculated in M-step.

1.1.1 preparation

In this section, I prepare for EM algorithm of GMM.

$$p(z_k = 1) = \pi_k \tag{3}$$

1. p(z) The prior distribution of z is given as

$$p(z) = \prod_{k=1}^{K} p(z_k) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{4}$$

   where K is the number of gaussian distributions.

2. $p(x|z)$ $p(x|z_k = 1)$ is given as

$$p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k) \tag{5}$$

   Then, $p(x|z)$ is given as

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} \tag{6}$$

3. p(x) Further, p(x) is given as

$$p(x) = \sum_{k=1}^{K} p(x|z_k)p(z_k) = \sum_{k=1}^{K} \pi_k p(x|\mu_k, \Sigma_k) \tag{7}$$

4. $p(z_k = 1|x)$ Finally, $\gamma(z_k) \equiv p(z_k = 1|x)$ is given by Bayes' theorem as follows.

$$\begin{aligned}
\gamma(z_k) \equiv p(z_k = 1|x) &= \frac{p(z_k = 1, x)}{p(x)} \\
&= \frac{p(z_k = 1)p(x|z_k = 1)}{p(x)} \\
&= \frac{\pi_k p(x|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k p(x|\mu_k, \Sigma_k)}
\end{aligned}$$

Here, from $p(x)$, $p(X|\mu, \Sigma)$ is given as

$$\log p(X|\mu, \Sigma) = \sum_{n=1}^{N} \log(\sum_{k=1}^{K} \pi_k p(x_n|\mu_k, \Sigma_k)) \tag{8}$$

1.1.2 caluculate $\mu^\star$

By letting the partial derivative of log likelihood $\log p(X|\mu, \Sigma)$ to zero, we can get optimal μ.

$$\begin{aligned}\frac{\partial}{\partial \mu_k} \log p(X|\mu, \Sigma) &= \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \log(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)) \\ &= \sum_{n=1}^{N} \frac{\pi_k \frac{\partial}{\partial \mu_k} \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}\end{aligned} \tag{9}$$

Here,

$$\begin{aligned}\frac{\partial}{\partial \mu_k} \mathcal{N}(x|\mu_k, \Sigma_k) &= \mathcal{N}(x|\mu_k, \Sigma_k) \frac{\partial}{\partial \mu_k} \log \mathcal{N}(x|\mu_k, \Sigma_k) \\ &= \mathcal{N} \frac{\partial}{\partial \mu_k} (-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)) \\ &= \mathcal{N}(-1)(-\frac{1}{2}) 2\Sigma_k^{-1}(x - \mu_k) \\ &= \mathcal{N}\Sigma_k^{-1}(x - \mu_k)\end{aligned} \tag{10}$$

Finally,

$$\begin{aligned}\frac{\partial}{\partial \mu_k} \log p(X|\mu, \Sigma) &= \sum_{n=1}^{N} \frac{\pi_k \frac{\partial}{\partial \mu_k} \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} \\ &= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} \Sigma_k^{-1}(x_n - \mu_k) \\ &= \sum_{n=1}^{N} \gamma(z_{nk}) \Sigma_k^{-1}(x_n - \mu_k) \\ &= 0\end{aligned} \tag{11}$$

Then, by multiplying Σ_k, we can get $\mu_k^\star$

$$\mu_k^\star = \frac{1}{N_k} \sum_n \gamma(z_{nk} x_n) \tag{12}$$

1.1.3 caluculate $\Sigma^\star$

By letting the partial derivative of log likelihood $\log p(X|\mu, \Sigma)$ to zero, we can get optimal Σ.

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_k} \log p(X|\mu, \Sigma) &= \frac{\partial}{\partial \Sigma_k} \sum_{n=1}^{N} \log(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)) \\
&= \sum_{n=1}^{N} \frac{\pi_k \frac{\partial}{\partial \Sigma_k} \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} \\
&= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \frac{\partial}{\partial \Sigma_k} \log \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \{-\frac{1}{2}\Sigma_k^{-1} + \frac{1}{2}(\Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1})\} \\
&= 0
\end{aligned} \tag{13}$$

Then we can get

$$\Sigma^\star = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{14}$$

1.1.4 caluculate $\pi^\star$

π has a constraint as follows

$$\sum_{k=1}^{K} \pi_k = 1 \tag{15}$$

we can use Method of Lagrange multipliers.

$$
\begin{aligned}
\frac{\partial}{\partial \pi_k}\{\log p(X|\mu,\Sigma) + \lambda(\sum_{k=1}^{K}\pi_k - 1)\} &= \frac{\partial}{\partial \pi_k}\log p(X|\mu,\Sigma) + \frac{\partial}{\partial \pi_k}\lambda(\sum_{k=1}^{K}\pi_k - 1) \\
&= \frac{\partial}{\partial \pi_k}\sum_{n=1}^{N}\log(\sum_{k=1}^{K}\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)) + \lambda \\
&= \sum_{n=1}^{N}\frac{\mathcal{N}(x_n|\mu_k,\Sigma_k))}{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k))} + \lambda \\
&= \sum_{n=1}^{N}\frac{\gamma(z_{nk})}{\pi_k} + \lambda \\
&= 0
\end{aligned} \tag{16}
$$

Then, we get

$$
\sum_{n=1}^{N}\gamma(z_{nk}) = -\lambda\pi_k \tag{17}
$$

Here, the number of sample N is

$$
N = \sum_{k=1}^{K}\sum_{n=1}^{N}\gamma(z_{nk}) = -\lambda\sum_{k=1}^{K}\pi_k = -\lambda \tag{18}
$$

Finally, we get

$$
\pi_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})}{N} \tag{19}
$$

1.2 Implement the EM algorithm.

I implement EM algorithm to data in x.csv. I set the number of clusters $k = 4$ and default parameters of clusters $\pi_k = \frac{1}{4}$, μ_k as the data picked from dataset and Σ_k as identity matrix. The posterior distributions of Z when given X are in z.csv, and Figure 1 shows the result.

2 Bayesian estimation

2.1 Derive the variational posteriors of the parameters by using the formulas.

As similar to EM-algorithm, our final goal is to find latent variables Z that maximize log likelihood of observed data X, $\log p(X)$. Like EM algorithm,

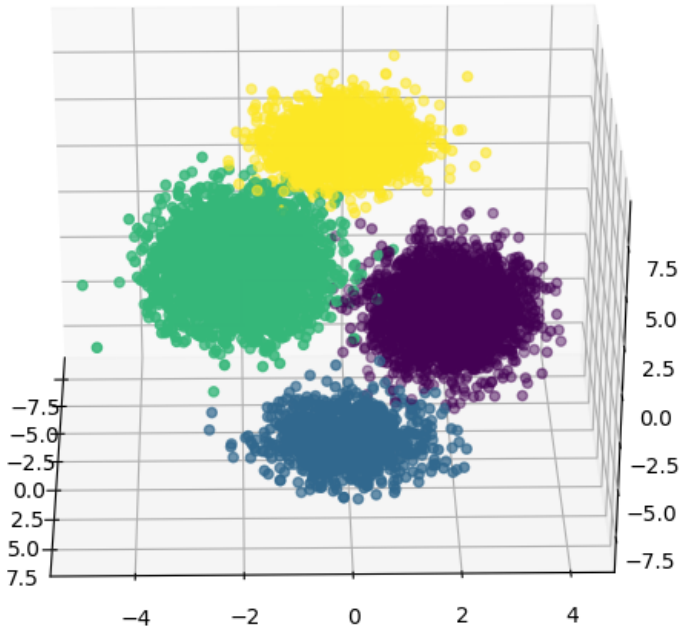

Figure 1: The result of EM-algorithm applying.

$\log p(X)$ can be divided into

$$\begin{aligned}\log p(X) &= \mathcal{L}(q) + KL(q||p) \\ &= \int q(Z) \log\{\frac{p(X,Z)}{q(Z)}\} dZ - \int q(Z) \log\{\frac{p(Z|X)}{q(Z)}\} dZ\end{aligned} \tag{20}$$

We want maximize $\mathcal{L}(q)$, thus minimize $KL(p||q)$, but we assume that it is difficult to get $p(Z|X)$ directly. Therefore, we introduce assumption that $p(Z)$ is divided into several groups.

$$p(Z) = \prod_{i=1}^{M} q_i(Z_i) \tag{21}$$

Then, to maximize $\mathcal{L}(q)$ w.r.t $q_i(Z_i)$, we rewrite $\mathcal{L}(q)$ as follows.

$$\begin{aligned}
\mathcal{L}(q) &= \int q(Z) \log\{\frac{p(X,Z)}{q(Z)}\} dZ \\
&= \int \prod_i q_i(Z_i) \{\log \frac{p(X,Z)}{\prod_i q_i(Z_i)}\} dZ \\
&= \int \prod_i q_i(Z_i) \{\log p(X,Z) - \sum_i \log q_i(Z_i)\} dZ \\
&= \int q_j(Z_j) \{\int \log p(X,Z) \prod_{i \neq j} q_i(Z_i) dZ_i\} dZ_j - \int q_j(Z_j) \log q_j(Z_j) dZ_j + const.
\end{aligned} \tag{22}$$

First item of the last row is obtained by marginalization w.r.t $q_j(Z_j)$.

Here, we define $\log \hat{p}(X, Z_j)$ as below. Note that we write $q_i(Z_i)$ as q_i.

$$\begin{aligned}
\log \hat{p}(X, Z_j) &= \int \log p(X,Z) \prod_{i \neq j} q_i dZ_i + const. \\
&= \mathbb{E}_{i \neq j}[\log p(X,Z)] + const.
\end{aligned} \tag{23}$$

Then we get

$$\mathcal{L}(q) = \int q_j \log \hat{p}(X, Z_j) dZ_j - \int q_j \log q_j dZ_j + const. \tag{24}$$

Actually, this is negative KL divergence between $q_j(Z_j)$ and $\hat{p}(X, Z_j)$. Therefore, to maximize $\mathcal{L}(q)$, we should minimize KL divergence, that is, we should set $q_j(Z_j) = \hat{p}(X, Z_j)$. Therefore, optimal $q_j^\star(Z_j)$ is given as

$$\log q_j^\star(Z_j) = \mathbb{E}_{i \neq j}[\log p(X,Z)] + const. \tag{25}$$

2.1.1 preparation

We assume the set of observed data as $X = \{x_1, ..., x_N\}$ and latent variables as $Z = \{z_1, ..., z_N\}$.

Here, the posterior distribution of Z when given π is

$$p(Z|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{26}$$

The conditional probability of X when given latent variables and parameters is

$$p(X|Z,\mu,\Sigma) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}(x_n|\mu_k,\Sigma_k)^{z_{nk}} \tag{27}$$

Next, we introduce the prior distributions of parameters μ, Σ, π. The prior distribution of π is Dirichlet distribution as follows.

$$p(\pi) = Dir(\pi|\alpha_0) = C(\alpha_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1} \tag{28}$$

The prior dustrubutions of μ, Σ is Gaussian-Wishart distribution as follows.

$$p(\mu,\Lambda) = p(\mu|\Lambda)p(\Lambda) = \prod_{k=1}^{K}\mathcal{N}(\mu_k|m_0,(\beta_0\Lambda)^{-1})\mathcal{W}(\Lambda_k|W_0,\nu_0) \tag{29}$$

Lastly, simultaneous distribution of parameters is

$$p(X,Z,\pi,\mu,\Lambda) = p(X|Z,\mu,\Lambda)p(Z|\pi)p(\pi)p(\mu|\Lambda)p(\Lambda) \tag{30}$$

Note that we observe only X.

Now, we assume that $q(Z,\pi,\mu,\Lambda)$ can be divided into several groups.

$$q(Z,\pi,\mu,\Lambda) = q(Z)q(\pi,\mu,\Sigma) \tag{31}$$

Here, I obtain following statistics.

2.1.2 Derive the variational posteriors of Z

From formula (25), we can get $\log q^\star(Z)$.

$$\begin{aligned}\log q^\star(Z) &= \mathbb{E}_{\pi,\mu,\Lambda}[\log p(X,Z,\pi,\mu,\Lambda)] + const \\ &= \mathbb{E}_{\pi,\mu,\Lambda}[\log\{p(X|Z,\mu,\Lambda)p(Z|\pi)p(\pi)p(\mu|\Lambda)p(\Lambda)\}] + const \\ &= \mathbb{E}_{\mu,\Lambda}[\log\{p(X|Z,\mu,\Lambda)] + \mathbb{E}_{\pi}[\log p(Z|\pi)] + const'\end{aligned} \tag{32}$$

Here,

$$\begin{aligned}\mathbb{E}_{\mu,\Lambda}[\log\{p(X|Z,\mu,\Lambda)] &= \mathbb{E}_{\mu,\Lambda}[\log\prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}(x_n|\mu_k,\Lambda_k)^{z_{nk}}] \\ &= \mathbb{E}_{\mu,\Lambda}[\sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\log\mathcal{N}(x_n|\mu_k,\Lambda_k)]\end{aligned} \tag{33}$$

Additionally,

$$\mathbb{E}_\pi[\log\{p(Z|\pi)] = \mathbb{E}_\pi[\log \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}}] = \mathbb{E}_\pi[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log\pi_k] \tag{34}$$

Finally, we get

$$\begin{aligned}\log q^\star(Z) &= \mathbb{E}_{\mu,\Lambda}[\log\{p(X|Z,\mu,\Lambda)] + \mathbb{E}_\pi[\log p(Z|\pi)] + const \\ &= \mathbb{E}_{\mu,\Lambda}[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log\mathcal{N}(x_n|\mu_k,\Lambda_k)] + \mathbb{E}_\pi[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log\pi_k] + const \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}s_{nk} + const\end{aligned} \tag{35}$$

where,

$$\begin{aligned}s_{nk} \equiv \mathbb{E}_\pi[\log\pi_k] - \frac{d}{2}\log 2\pi + \mathbb{E}_\Lambda[\frac{1}{2}\log|\Lambda_k|] \\ - \frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T)\Lambda_k(x_n-\mu_k)]\end{aligned} \tag{36}$$

d is the dimension of data. From (35), we get

$$q^\star(Z) \propto \prod_n\prod_k s_{nk}^{z_{nk}} \tag{37}$$

This distribution must be normalized and $z_{nk} \in \{0,1\}$, $\sum_k z_{nk} = 1$. Therefore,

$$q^\star(Z) = \prod_n\prod_k r_{nk}^{z_{nk}} \tag{38}$$

where,

$$r_{nk} = \frac{s_{nk}}{\sum_{j=1}^{K} s_{nj}} \tag{39}$$

2.1.3 Derive the variational posteriors of π, μ, Λ

Next, we consider $q(\pi,\mu,\Lambda)$. From (25),

$$\begin{aligned}\log q^\star(\pi,\mu,\Lambda) &= \mathbb{E}_Z[\log p(X,Z,\pi,\mu,\Lambda)] + const \\ &= \log p(\pi) + \log p(\mu,\Lambda) + \mathbb{E}_Z[\log p(Z|\pi)] \\ &+ \mathbb{E}_Z[\log\{p(X|Z,\mu,\Lambda)] + const\end{aligned} \tag{40}$$

Here,

$$q(\pi, \mu, \Lambda) = q(\pi)q(\mu, \Lambda) = q(\pi) \prod_{k=1}^{K} q(\mu_k, \Lambda_k) \tag{41}$$

From above (40), we extract items on π.

$$\log q^\star(\pi) = (\alpha_0 - 1) \sum_{k=1}^{K} \log \pi_k + \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \log \pi_k + const \tag{42}$$

Therefore,

$$q^\star(\pi) = Dir(\pi|\alpha) \tag{43}$$

Here, the k-th element of α is

$$\alpha_k = \alpha_0 + N_k \tag{44}$$

Finally, we consider $q^\star(\mu_k, \Lambda_k)$. In order to derive the optimal solution for $q(\mu_k, \Lambda_k)$ we start with the result (40) and keep only those term which depend on μ_k or Λ_k to give

$$\begin{aligned}
\log q^\star(\mu_k, \Lambda_k) &= \log q(\mu_k, \Lambda_k) + \sum_n \mathbb{E}[z_{nk}] \log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1}) + const \\
&= \log \mathcal{N}(\mu_k|m_0, (\beta_0 \Lambda_k)^{-1}) + \log \mathcal{W}(\Lambda_k|W_0, \nu_0) \\
&+ \sum_n \mathbb{E}[z_{nk}] \log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1}) + const. \\
&= \frac{\log |\Lambda_k|}{2} - \frac{\beta_0}{2}(\mu_k - m_0)^T \Lambda_k (\mu_k - m_0) - \mathrm{Tr}(\Lambda_k W_0^{-1}) \\
&+ \frac{\nu_0 - d - 1}{2} \log |\Lambda_k| - \frac{1}{2} \sum_n \mathbb{E}[z_{nk}](x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \\
&+ \frac{1}{2}(\sum_n \mathbb{E}[z_{nk}]) \log |\Lambda_k| + const.
\end{aligned} \tag{45}$$

By the product rule pf probability, $\log q^\star(\mu_k, \Lambda_k) = \log q^\star(\mu_k|\Lambda_k) + \log q^\star(\Lambda_k)$. First, I calculate the distribution for μ_k. To do this, it is need to consider

terms on (45) which depend on μ_k, giving

$$\begin{aligned}\log q^\star(\mu_k|\Lambda_k) &= -\frac{1}{2}\mu_k^T\left[\beta_0 + \sum_n \mathbb{E}[z_{nk}]\right]\Lambda_k\mu_k \\ &\quad + \mu_k^T\Lambda_k\left[\beta_0 m_0 + \sum_n \mathbb{E}[z_{nk}]x_n\right] + const. \\ &= -\frac{1}{2}\mu_k^T[\beta_0 + N_k]\Lambda_k\mu_k + \mu_k^T\Lambda_k[\beta_0 m_0 + N_k\bar{x}_k] + const.\end{aligned} \tag{46}$$

We see that $\log q^\star(\mu_k|\Lambda_k)$ depends quadratically on μ_k and therefore $\log q^\star(\mu_k|\Lambda_k)$ is a Gaussian distribution. Thus,

$$q^\star(\mu_k|\Lambda_k) = \mathcal{N}(\mu_k|m_k, \beta_k\Lambda_k) \tag{47}$$

where,

$$\begin{aligned}\beta_k &= \beta_0 + N_k \\ m_k &= \frac{1}{\beta_k}(\beta_0 m_0 + N_k\bar{x}_k)\end{aligned} \tag{48}$$

Next, I determine $q^\star(\Lambda_k)$ using following relation

$$\log q^\star(\Lambda_k) = \log q^\star(\mu_k, \Lambda_k) - \log q^\star(\mu_k|\Lambda_k) \tag{49}$$

Considering terms on (45) which depend on Λ_k and (46), giving

$$\begin{aligned}\log q^\star(\Lambda_k) &= \frac{\log|\Lambda_k|}{2} + \frac{\beta_0}{2}(\mu_k - m_0)^T\Lambda_k(\mu_k - m_0) - \mathrm{Tr}(\Lambda_k W_0^{-1}) \\ &\quad + \frac{\nu_0 - d - 1}{2}\log|\Lambda_k| + \frac{1}{2}(\sum_n \mathbb{E}[z_{nk}])\log|\Lambda_k| \\ &\quad - \frac{1}{2}\sum_n \mathbb{E}[z_{nk}](x_n - \mu_k)^T\Lambda_k(x_n - \mu_k) \\ &\quad + \frac{\beta_k}{2}(x_n - \mu_k)^T\Lambda_k(x_n - \mu_k) - \frac{\log|\Lambda_k|}{2} + const. \\ &= \frac{\nu_0 - d - 1}{2}\log|\Lambda_k| - \mathrm{Tr}(\Lambda_k W_0^{-1}) + const.\end{aligned} \tag{50}$$

Here I define

$$\begin{aligned}W_k^{-1} &= W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)^T \\ \nu_k &= \nu_0 + N_k\end{aligned} \tag{51}$$

Thus,

$$q^{\star}(\Lambda_k) = \mathcal{W}(\Lambda_k|W_k, \nu_k) \tag{52}$$

$$q^{\star}(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k|m_k, (\beta_k\Lambda_k)^{-1})\mathcal{W}(\Lambda_k|W_k, \nu_k) \tag{53}$$

From (36) and (39), we obtain

$$r_{nk} \propto \tilde{\pi}_k\tilde{\Lambda}_k^{\frac{1}{2}} \exp\{-\frac{d}{2\beta_k} - \frac{\nu_k}{2}(x_n - m_k)^T W_k(x_n - m_k)\} \tag{54}$$

2.2 Implement the VB and/or GS algorithm

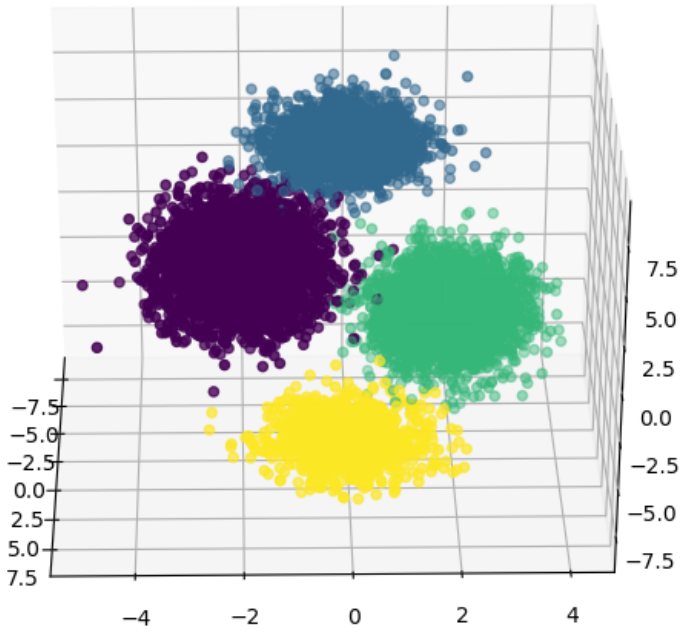

Figure 2: The result of Variational bayes applying.

Figure 2 shows the result of applying the variational bayes on x.csv. I set the number of clusters as 4 which is the same as EM-algorithm.

3 reference

https://qiita.com/ctgk/items/49d07215f700ecb03eeb

Christopher M. Bishop. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.