Pattern Recognition Adv Report(Yoshii)

6930-31-6255 Fujiwara Daigo

7/21/2019

1 ML estimation

1.1 Explanation of EM aglgorythm

The lower bound of log likelihood is :

$$\ln p(\boldsymbol{X};\boldsymbol{\theta}) \geq \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{1}$$

Now, we call the last estemated parameter $\boldsymbol{\theta}$ as $\boldsymbol{\theta}^{\text{old}}$
In EM algorythm, Expectstion step set $q(\boldsymbol{Z})$, which is estimated distribution of $\boldsymbol{Z}$, to $p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta}^{\text{old}})(= q^*(\boldsymbol{Z}))$. It means minimizing KL divergence between $q(\boldsymbol{Z})$ and $p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta}^{\text{old}})$, and also means maximamizing the lower bound for $q(\boldsymbol{Z})$, fixed $\boldsymbol{\theta}^{\text{old}}$. Then, the inequality of (1) become equal,thus:

$$\ln p(\boldsymbol{X};\boldsymbol{\theta}^{\text{old}}) = \int q^*(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}^{\text{old}})}{q^*(\boldsymbol{Z})} d\boldsymbol{Z} \tag{2}$$

And,in Maximization step, we maximize likelihood $\ln p(\boldsymbol{X};\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ fixed $q(\boldsymbol{Z})$, then $\boldsymbol{\theta}^{\text{old}}$ is updated to $\boldsymbol{\theta}^{\text{new}}$. Thus:

$$\boldsymbol{\theta}^{\text{new}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \left[\ln p\left(\boldsymbol{X};\boldsymbol{\theta}\right) = \int q^*(\boldsymbol{Z}) \ln \frac{p\left(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}\right)}{q^*(\boldsymbol{Z})} d\boldsymbol{Z} \right] \tag{3}$$

Finally, for this θ^{new}, this inequality holds:

$$\ln p(\boldsymbol{X};\boldsymbol{\theta}^{\text{new}}) \geq \int q^*(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}^{\text{new}})}{q^*(\boldsymbol{Z})} d\boldsymbol{Z} \tag{4}$$

Then in the next Expectation step, we maximamize the lower bound for $q(\boldsymbol{Z})$.

1.2 Derive the update formulas of the parameters

In GMM, $\boldsymbol{X}$ and $\boldsymbol{Z}$ Simultaneous distribution is given by prameter $\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}$, exprssed as bellow:

$$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Lambda},\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

$$\therefore \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{\ln \pi_k + \ln \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)\right\}$$

Then, the lower bound $L(q, \boldsymbol{\theta})$ is:

$$\begin{aligned}
L(q, \boldsymbol{\theta}) &= \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta})}{q(\boldsymbol{Z})} d\boldsymbol{Z} && (5)\\
&= \int q(\boldsymbol{Z}) \left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{\ln \pi_k + \ln \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)\right\} - \ln q(\boldsymbol{Z})\right] d\boldsymbol{Z} && (6)\\
&= \int q(\boldsymbol{Z}) \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{\ln \pi_k + \ln \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)\right\} d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) d\boldsymbol{Z} && (7)
\end{aligned}$$

$$\left(\ln \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) = -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k) - \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Lambda}_k|\right)$$

After Expectation step, $q(\boldsymbol{Z})$ is given as:

$$q^*(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} p\left(\mathbf{z}_n|\boldsymbol{x}_n; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\right) = \prod_{n=1}^{N}\prod_{k=1}^{K} \gamma_{nk}^{z_{nk}}$$

$$\gamma_{nk} = \frac{\pi_k N\left(x_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)}{\sum_{k'=1}^{K} \pi_{k'} N\left(\boldsymbol{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Lambda}_{k'}^{-1}\right)}$$

In Maximization step, $L(q, \boldsymbol{\theta})$ is maximized for parameters $\boldsymbol{\theta}$.
First, for the parameter π_k, using Lagrange multiplyer, minimizing $F = L - \lambda(\sum_{k=1}^{K} \pi_k - 1)$:

$$\begin{aligned}
\frac{\partial F}{\partial \pi_k} &= 0 && (8)\\
\therefore \int q(\boldsymbol{Z}) \sum_{n=1}^{N} z_{nk}\frac{1}{\pi_k} d\boldsymbol{Z} - \lambda &= 0 \quad (\because (7)) && (9)\\
\therefore \int q(\boldsymbol{Z}) \sum_{n=1}^{N} z_{nk} d\boldsymbol{Z} - \lambda\pi_k &= 0 && (10)\\
\therefore \sum_{k=1}^{K}\int q(\boldsymbol{Z}) \sum_{n=1}^{N} z_{nk} d\boldsymbol{Z} - \lambda &= 0 \quad (\because \sum_{k=1}^{K} \pi_k = 1) && (11)\\
\therefore \int q(\boldsymbol{Z}) \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} d\boldsymbol{Z} - \lambda &= 0 && (12)\\
\therefore \int q(\boldsymbol{Z}) N d\boldsymbol{Z} - \lambda &= 0 \quad (\because \sum_{k=1}^{K} z_{nk} = 1) && (13)\\
\therefore \lambda &= N && (14)
\end{aligned}$$

Therefore,

$$
\begin{aligned}
\pi_k^{\text{new}} &= \frac{1}{N}\int q(\boldsymbol{Z})\sum_{n=1}^{N} z_{nk}\,d\boldsymbol{Z} && (15)\\
&= \frac{1}{N}\sum_{n=1}^{N}\int q(\boldsymbol{Z}) z_{nk}\,d\boldsymbol{Z} && (16)\\
&= \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{\boldsymbol{z}_n} z_{nk}\left[\pi_k\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)\right]^{z_{nk}}}{\sum_{\boldsymbol{z}_n}\left[\pi_j\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Lambda}_j^{-1}\right)\right]^{z_{nj}}} && (17)\\
&= \frac{1}{N}\sum_{n=1}^{N}\frac{\pi_k\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)}{\sum_{j=1}^{K}\pi_j\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Lambda}_j^{-1}\right)} && (18)\\
&= \frac{1}{N}\sum_{n=1}^{N}\gamma_{nk} = \frac{N_k}{N} && (19)
\end{aligned}
$$

where,

$$
N_k = \sum_{n=1}^{N}\gamma_{nk}
$$

Second, for the parameter $\boldsymbol{\mu}_k$, minimizimg $L(q,\boldsymbol{\theta})$

$$
\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\mu}_k} &= 0 && (20)\\
\therefore \int q(\boldsymbol{Z})\sum_{n=1}^{N} z_{nk}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k-\boldsymbol{x}_n)\,d\boldsymbol{Z} &= 0 \quad (\because (7)) && (21)\\
\therefore \sum_{n=1}^{N}\int q(\boldsymbol{Z}) z_{nk}\,d\boldsymbol{Z}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k-\boldsymbol{x}_n) &= 0 && (22)\\
\therefore \sum_{n=1}^{N}\gamma_{nk}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k-\boldsymbol{x}_n) &= 0 && (23)\\
\therefore \sum_{n=1}^{N}\gamma_{nk}(\boldsymbol{\mu}_k-\boldsymbol{x}_n) &= 0 && (24)\\
\therefore \boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}\boldsymbol{x}_n && (25)
\end{aligned}
$$

Third, for the parameter $\boldsymbol{\Lambda}_k$, minimizimg $L(q, \boldsymbol{\theta})$

$$\frac{\partial L}{\partial \boldsymbol{\Lambda}_k} = 0 \tag{26}$$

$$\therefore \int q(\boldsymbol{Z}) \sum_{n=1}^{N} z_{nk} \left[-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T + \frac{1}{2}\boldsymbol{\Lambda}_k^{-1} \right] d\boldsymbol{Z} = 0 \quad (\because (7)) \tag{27}$$

$$\therefore \sum_{n=1}^{N} \int q(\boldsymbol{Z}) z_{nk} d\boldsymbol{Z} \left[-(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T + \boldsymbol{\Lambda}_k^{-1} \right] = 0 \tag{28}$$

$$\therefore \sum_{n=1}^{N} \gamma_{nk} \left[-(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T + \boldsymbol{\Lambda}_k^{-1} \right] = 0 \tag{29}$$

$$\therefore \boldsymbol{\Lambda}_k^{-1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \tag{30}$$

Then, in this update we use $\boldsymbol{\mu}_k^{\mathrm{new}}$ as $\boldsymbol{\mu}_k$, so update formula is:

$$\boldsymbol{\Lambda}_k^{-1\mathrm{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k^{\mathrm{new}})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{\mathrm{new}})^T \tag{31}$$

$$= \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n \boldsymbol{x}_n^T - \boldsymbol{\mu}_k^{\mathrm{new}} \boldsymbol{x}_n^T - \boldsymbol{x}_n \boldsymbol{\mu}_k^{\mathrm{newT}} + \boldsymbol{\mu}_k^{\mathrm{new}} \boldsymbol{\mu}_k^{\mathrm{newT}}) \tag{32}$$

$$= \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n \boldsymbol{x}_n^T - \boldsymbol{\mu}_k^{\mathrm{new}} \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n^T - \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n \boldsymbol{\mu}_k^{\mathrm{newT}} + \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{\mu}_k^{\mathrm{new}} \boldsymbol{\mu}_k^{\mathrm{newT}} \tag{33}$$

$$= \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n \boldsymbol{x}_n^T - \boldsymbol{\mu}_k^{\mathrm{new}} \boldsymbol{\mu}_k^{\mathrm{newT}} \quad (\because (25)) \tag{34}$$

update formulas are (19)(25)(34).

1.3 About implementation

I did the implementation of EM algorythm. This is scatter of EM algorythm, where each cluster is painted different color. The cluster is judged by the cluster which have most large probability. It seems to be good.

As a feature of the implementation, sometimes the update is get stuck in local optimal solution, and then the clustering is like that, two or more clusters are miss-judged as the one cluster, so that to some extent log likelihood is high, but it is not surfficiant. This phenomenon seems to be depend on initial value of parameters.

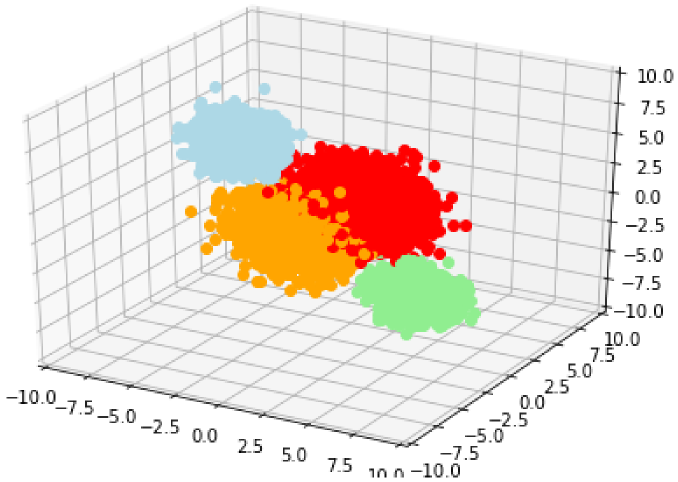

図 1　scatter and clustering of x by EM

2 Bayesian estimation

2.1 Explanation of VB aglgorythm

VB algorythm aims to maximize lower bound by minimizing KL divergence between $q(\boldsymbol{Z})$ and $p(\boldsymbol{Z}|\boldsymbol{X})$, as well as EM algorythm, but in VB algorythm, there is no certain expression of posteior $p(\boldsymbol{Z}|\boldsymbol{X})$.Thus, in this algorythm, indstead of using directly $p(\boldsymbol{Z}|\boldsymbol{X})$ as $q(\boldsymbol{Z})$, we use limited expression of $q(\boldsymbol{Z})$ and minimize KL divergence.
As such limited expression, we assume that $q(\boldsymbol{Z})$ can be factorized as bellow:

$$q(\mathbf{Z})=\prod_{i=1}^{M} q_{i}\left(\mathbf{Z}_{i}\right)$$

And lower bound is maximized by each $q_i\left(\mathbf{Z}_i\right)$ alternatingly.
The lower bound $L(q)$ is like that:

$$\begin{aligned} \ln p(\boldsymbol{X}) &= \ln \int p(\boldsymbol{X},\boldsymbol{Z})d\boldsymbol{Z} = \ln \int q(\boldsymbol{Z})\frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}d\boldsymbol{Z} \\ &\geq \int q(\boldsymbol{Z})\ln\frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}d\boldsymbol{Z} = L(q) \end{aligned} \tag{35}$$

$$L(q)=\int\left(\prod_{m=1}^{M} q\left(\boldsymbol{Z}_{m}\right)\right)\left(\ln p(\boldsymbol{X},\boldsymbol{Z})-\sum_{m=1}^{M}\ln q\left(\boldsymbol{Z}_{m}\right)\right) d\boldsymbol{Z}_{1} d\boldsymbol{Z}_{2}\cdots d\boldsymbol{Z}_{M} \tag{36}$$

$$=\int\left(\prod_{m=1, m\neq i}^{M} q\left(\boldsymbol{Z}_{m}\right)\right) q\left(\boldsymbol{Z}_{i}\right)\left(\ln p(\boldsymbol{X},\boldsymbol{Z})-\sum_{m=1}^{M}\ln q\left(\boldsymbol{Z}_{m}\right)\right) d\boldsymbol{Z}_{1} d\boldsymbol{Z}_{2}\cdots d\boldsymbol{Z}_{M} \tag{37}$$

$$=\int q\left(\boldsymbol{Z}_{i}\right)\left(\int\ln p(\boldsymbol{X},\boldsymbol{Z})\prod_{m=1, m\neq i}^{M} q\left(\boldsymbol{Z}_{m}\right) d\boldsymbol{Z}_{m}\right) d\boldsymbol{Z}_{i}-\int q(\boldsymbol{Z}_{i})\ln q\left(\boldsymbol{Z}_{i}\right) d\boldsymbol{Z}_{i}+\text{const} \tag{38}$$

$$=\int q\left(\boldsymbol{Z}_{i}\right)\ln\tilde{p}(\boldsymbol{X},\boldsymbol{Z}_{i}) d\boldsymbol{Z}_{i}-\int q(\boldsymbol{Z}_{i})\ln q\left(\boldsymbol{Z}_{i}\right) d\boldsymbol{Z}_{i}+\text{const} \tag{39}$$

$$=\int q\left(\boldsymbol{Z}_{i}\right)\ln\frac{\tilde{p}(\boldsymbol{X},\boldsymbol{Z}_{i})}{q(\boldsymbol{Z}_{i})} d\boldsymbol{Z}_{i}=KL(q(\boldsymbol{Z}_{i})||\tilde{p}(\boldsymbol{X},\boldsymbol{Z}_{i})) \tag{40}$$

where,

$$\ln\tilde{p}(\boldsymbol{X},\boldsymbol{Z}_{i})=\int\ln p(\boldsymbol{X},\boldsymbol{Z})\prod_{m=1, m\neq i}^{M} q\left(\boldsymbol{Z}_{m}\right) d\boldsymbol{Z}_{i}=\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]$$

The part of "const" is indipendent to $q(\boldsymbol{Z}_i)$. Thus, by maximizing $L(q)$, $q(\boldsymbol{Z}_i)$ is set as bellow:

$$\ln q^{*}(\boldsymbol{Z}_{i})=\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})] \tag{41}$$

Z_i includes paranmeters θ, which is treated random variables now, so maximizing for $q(\boldsymbol{Z})$ contains both of E step and M step.

2.2 Derive the variational posteriors of the parameters

Now, in GMM, latent variables is given as $\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}$. We assume that simultaneous distribution of them is factorized as bellow:

$$q(\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})=q(\mathbf{Z})q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$$

And of GMM is wriinten as this form:

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

$$p(\boldsymbol{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})=\prod_{n=1}^{N}\prod_{k=1}^{K} N\left(\boldsymbol{x}_{n}|\boldsymbol{\mu}_{k},\boldsymbol{\Lambda}_{k}^{-1}\right)^{z_{nk}}$$

$$p(\boldsymbol{Z}|\boldsymbol{\pi})=\prod_{n=1}^{N}\text{Categorical}\left(\mathbf{z}_{n}|\boldsymbol{\pi}\right)=\prod_{n=1}^{N}\prod_{k=1}^{K}\pi_{k}^{z_{nk}}$$

$$p(\boldsymbol{\pi})=\text{Dir}\left(\boldsymbol{\pi}|\boldsymbol{\alpha}_{0}\right)=\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_{0k}\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_{0k}\right)}\prod_{k=1}^{K}\pi_{k}^{\alpha_{0k}-1}$$

$$p(\boldsymbol{\mu},\Lambda)=\prod_{k=1}^{K} N\left(\boldsymbol{\mu}_{k}|\boldsymbol{m}_{0},\left(\beta_{0}\boldsymbol{\Lambda}_{k}\right)^{-1}\right) W\left(\boldsymbol{\Lambda}_{k}|\mathbf{W}_{0},v_{0}\right) \tag{42}$$

In E step, $q(\boldsymbol{Z})$ is updated by (41). The result is billow:

$$q^{\star}(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

where,

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}$$

$$\ln \rho_{nk} = \mathbb{E}\left[\ln \pi_k\right] + \frac{1}{2}\mathbb{E}\left[\ln |\boldsymbol{\Lambda}_k|\right] - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^{\mathrm{T}} \boldsymbol{\Lambda}_k \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\right]$$

In M step, distribution of parameters is updated by (41) (using (42) model).

$$\begin{aligned} \ln q^{\star}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \ln p(\boldsymbol{\pi}) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] \\ &\quad + \sum_{k=1}^{K} \ln p\left(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\right) + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}\left[z_{nk}\right] \ln \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) + \text{const.} \end{aligned} \tag{43}$$

(45) is separated into parts only depending on $\boldsymbol{\pi}$ and parts depending on each $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$. Since the left side is log part, it means that simultaneous distribution is expressed as bellow:

$$q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) \prod_{k=1}^{K} q\left(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\right)$$

Then first, deribve variational posterior of $\boldsymbol{\pi}$, by extracting $\boldsymbol{\pi}$ depending parts from (43) (using (42) model):

$$\ln q^{\star}(\boldsymbol{\pi}) = \ln p(\boldsymbol{\pi}) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \text{const} \tag{44}$$

$$= (\alpha_0 - 1)\sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk} \ln \pi_k + \text{const} \tag{45}$$

$$= \sum_{k=1}^{K}\left(\alpha_0 - 1 + \sum_{n=1}^{N} r_{nk}\right) \ln \pi_k + \text{const} \tag{46}$$

$$\therefore q^{\star}(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \tag{47}$$

where,

$$\alpha_k = \alpha_0 + N_k$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_{0k}$$

$$N_k = \sum_{n=1}^{N} r_{nk}$$

Second, deribve variational posterior of $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$. By extracting $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ depending parts from (43) (using (42) model):

$$\ln q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \ln p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + \sum_{n=1}^{N} \mathbb{E}[z_{nk}] \ln \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) + \text{const} \tag{48}$$

$$= \ln N\left(\boldsymbol{\mu}_k | \boldsymbol{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}\right) W(\boldsymbol{\Lambda}_k | \mathbf{W}_0, v_0) + \sum_{n=1}^{N} \mathbb{E}[z_{nk}] \ln \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) + \text{const} \tag{49}$$

$$\begin{aligned} &= -\frac{1}{2}(\boldsymbol{\mu}_k - \mathbf{m}_0)^T \beta_0 \Lambda_k (\boldsymbol{\mu}_k - \mathbf{m}_0) + \frac{1}{2}\ln|\beta_0 \Lambda_k| \\ &+ \frac{\nu_0 - D - 1}{2} \ln|\Lambda_k| - \frac{1}{2}\text{Tr}\left(W_0^{-1}\Lambda_k\right) \\ &- \sum_n \frac{\gamma_{nk}}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \Lambda_k (\mathbf{x}_n - \boldsymbol{\mu}_k) + \sum_n \frac{\gamma_{nk}}{2} \ln|\Lambda_k| + \text{const.} \end{aligned} \tag{50}$$

$$(※ \ln \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) = -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k) - \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Lambda}_k|)$$

Thus,

$$\ln q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}\left(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k | \mathbf{W}_k, \nu_k) \tag{51}$$

where,

$$\begin{aligned} \beta_k &= \beta_0 + N_k \\ \mathbf{m}_k &= \frac{1}{\beta_k}(\beta_0 \mathbf{m}_0 + N_k \overline{\mathbf{x}}_k) \\ \mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + N_k \mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\overline{\mathbf{x}}_k - \mathbf{m}_0)(\overline{\mathbf{x}}_k - \mathbf{m}_0)^{\mathrm{T}} \\ \nu_k &= \nu_0 + N_k \end{aligned}$$

$$\begin{aligned} \overline{\mathbf{x}}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n \\ \mathbf{S}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \overline{\mathbf{x}}_k)(\mathbf{x}_n - \overline{\mathbf{x}}_k)^{\mathrm{T}} \end{aligned}$$

The variational posterior formulas are (47) and (51).

2.3 About implementation

I did the implementation of VB algorythm. This is scatter of vb algorythm, where each cluster is painted different color. The cluster is judged by the cluster which have most large probability. It seems to be good.

As a feature of the implementation, compared with EM algorythm, the convergence is very slow for each iteration, but it seems to be less likely to be get stuck in local optimal solution. Staibly it comes to the similar clusterring like this scatter graph, and less depend on the initial distribution of prameters.

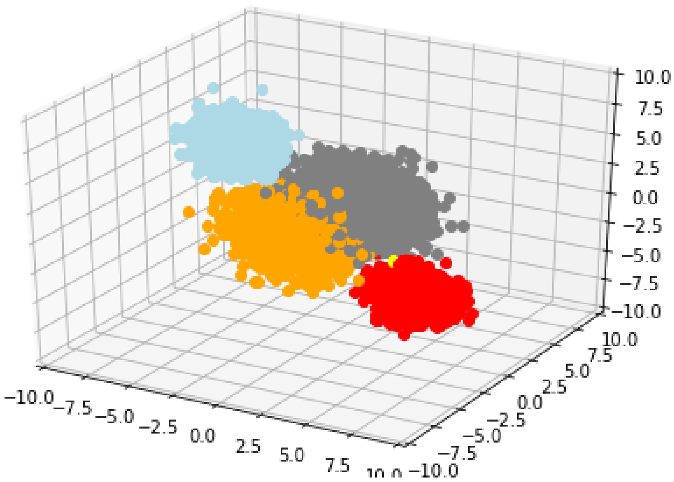

図 2 scatter and clustering of x by VB