CSED524-Homework 3

Inhyuk Na

2018/11/19

1 Variational Mixture of Gaussians

In the lecture, we've learned about variational Mixture of Gaussians. This model is basically similar with Mixture of Gaussians, because we assume latent variable $\boldsymbol{z}$ and several Gaussian clusters. Thus, the evidence of this model is as follows:

$$p(\boldsymbol{x_n}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, {\boldsymbol{\Lambda_k}}^{-1}).$$

However, in this model we assume that $\boldsymbol{\pi}$(mixing parameters), $\boldsymbol{\mu_k}$, and $\boldsymbol{\Lambda_k}$(component densities parameter) are also random variables. For $\boldsymbol{\pi}$, we assume Dirichlet prior such that

$$p(\boldsymbol{\pi}) = Dir(\alpha_0, ..., \alpha_0).$$

For $\boldsymbol{\mu_k}$ and $\boldsymbol{\Lambda_k}$, we assume Gaussian-Wishart prior such that

$$\begin{aligned} p(\boldsymbol{\mu_k}, \boldsymbol{\Lambda_k}) &= p(\boldsymbol{\mu_k}|\boldsymbol{\Lambda_k})p(\boldsymbol{\Lambda_k}) \\ &= \mathcal{N}(\boldsymbol{\mu_k}|m_0, (\beta\boldsymbol{\Lambda_k})^{-1})\mathcal{W}(\boldsymbol{\Lambda_k}|\boldsymbol{W_0}, v_0). \end{aligned}$$

Thus, we assume that variational distribution factorizes as

$$q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{Z})q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu_k}|\boldsymbol{\Lambda_k})q(\boldsymbol{\Lambda_k}).$$

2 Variational EM

We apply Variationial EM optimization for this model. The variational lower-bound is

$$\begin{aligned} \log p(\boldsymbol{X}) &\leq \sum_{\boldsymbol{Z}} \int\int\int q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \log\Big(\frac{p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}{q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}\Big) d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\Lambda} \\ &= \mathbb{E}\log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) + \mathbb{E}\log p(\boldsymbol{Z}|\boldsymbol{\pi}) + \mathbb{E}\log p(\boldsymbol{\pi}) + \mathbb{E}\log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) - \mathbb{E}\log q(\boldsymbol{Z}) - \mathbb{E}\log q(\boldsymbol{\pi}) - \mathbb{E}\log q(\boldsymbol{\mu}, \boldsymbol{\Lambda}). \end{aligned} \tag{1}$$

We can find the $q^*(\boldsymbol{Z})$ and $q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ for variational EM optimization as follows:

Variational E step:

$$q^*(\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{k,n}^{z_{k,n}},$$

where

$$r_{k,n} = \frac{\rho_{k,n}}{\sum_{j=1}^{K}\rho_{j,n}},$$

$$\log\rho_{k,n} = \mathbb{E}_{\boldsymbol{\pi}}[\log\pi_k] + \frac{1}{2}\mathbb{E}_{\Lambda}[\log|\boldsymbol{\Lambda_k}|] - \frac{D}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}[(\boldsymbol{x_n}-\boldsymbol{\mu_k})^T\boldsymbol{\Lambda_k}(\boldsymbol{x_n}-\boldsymbol{\mu_k})],$$

$$\mathbb{E}_{\boldsymbol{\pi}}\{\log\pi_k\} = \psi(\alpha_k) - \psi(\alpha_1+\cdots+\alpha_K),$$

$$\mathbb{E}_{\boldsymbol{\Lambda}}\{\log|\boldsymbol{\Lambda_k}|\} = \sum_{i=1}^{D}\psi\big(\frac{\nu_k+1-i}{2}\big) + D\log 2 + \log|\boldsymbol{W_k}|, \text{ and}$$

$$\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\{(\boldsymbol{x_n}-\boldsymbol{\mu_k})^T\boldsymbol{\Lambda_k}(\boldsymbol{x_n}-\boldsymbol{\mu_k})\} = D\beta_k^{-1} + \nu_k(\boldsymbol{x_n}-\boldsymbol{m_k})^T\boldsymbol{W_k}(\boldsymbol{x_n}-\boldsymbol{m_k}).$$

Variational M step:

$$q^*(\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{k=1}^{K}\mathcal{N}\big(\boldsymbol{\mu_k}|\boldsymbol{m_k},(\beta_k\boldsymbol{\Lambda_k})^{-1}\big)\mathcal{W}(\boldsymbol{\Lambda_k}|\nu_k),$$

where

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

$$\overline{\boldsymbol{x_k}} = \frac{1}{N_k}\sum_{n=1}^{N} r_{k,n}\boldsymbol{x_n},$$

$$\boldsymbol{Z_k} = \frac{1}{N_k}\sum_{n=1}^{N} r_{k,n}(\boldsymbol{x_n}-\overline{\boldsymbol{x_k}})(\boldsymbol{x_n}-\overline{\boldsymbol{x_k}})^T,$$

$$\beta_k = \beta_0 + N_k,$$

$$\boldsymbol{m_k} = \frac{1}{\beta_k}(\beta_0\boldsymbol{m_0} + N_k\overline{\boldsymbol{x_k}}),$$

$$\nu_k = \nu_0 + N_k, \text{ and}$$

$$\boldsymbol{W_k}^{-1} = \boldsymbol{W_0}^{-1} + N_k\boldsymbol{Z_k} + \frac{\beta_0 N_k}{\beta_0+N_k}(\overline{\boldsymbol{x_k}}-\boldsymbol{m_0})(\overline{\boldsymbol{x_k}}-\boldsymbol{m_0})^T.$$

$$q^*(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K}\pi_k^{\alpha_k-1},$$

where

$$\boldsymbol{\alpha} = [\alpha_1,...,\alpha_K]^T \text{ and } \alpha_k = \alpha_0 + N_k.$$

Proof. Before we derive, recall that

$$\begin{aligned}
\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \log(p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})) \\
&= \sum_{k=1}^{K}\big[\sum_{n=1}^{N}\log p(\boldsymbol{x_n}|z_{k,n}, \boldsymbol{\mu_k}, \boldsymbol{\Lambda_k}) + \sum_{n=1}^{N}\log p(z_{k,n}|\pi_k) + \log p(\pi_k) + \log p(\boldsymbol{\mu_k}|\boldsymbol{\Lambda_k}) + \log p(\boldsymbol{\Lambda_k})\big] \\
&= \sum_{k=1}^{K}\big[\sum_{n=1}^{N}\log \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, {\boldsymbol{\Lambda_k}}^{-1})^{z_{k,n}} + \sum_{n=1}^{N}\log(\pi_k^{z_{k,n}}) \\
&\quad + \log \mathcal{D}(\pi_k|\alpha_0) + \log \mathcal{N}(\boldsymbol{\mu_k}|\boldsymbol{m_0}, (\beta_0\boldsymbol{\Lambda_k})^{-1}) + \log \mathcal{W}(\boldsymbol{\Lambda_k}|\boldsymbol{W_0}, \nu_0)\big]. \qquad (2)
\end{aligned}$$

For variational E step, we have to compute $\log q^*(\boldsymbol{Z}) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}(\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})) + const.$ We use above equation and only consider the terms which depend on $z_{k,n}$.

$$\begin{aligned}
\log q^*(\boldsymbol{Z}) &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}(\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})) + const \\
&= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}(\sum_{k=1}^{K}\big[\sum_{n=1}^{N}\log \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, {\boldsymbol{\Lambda_k}}^{-1})^{z_{k,n}} + \sum_{n=1}^{N}\log(\pi_k^{z_{k,n}})\big]) + const \\
&= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}(\sum_{k=1}^{K}\big[\sum_{n=1}^{N} z_{k,n}\{\log(\frac{1}{(2\pi)^{D/2}|{\boldsymbol{\Lambda_k}}^{-1}|^{1/2}}) - \frac{1}{2}(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T\boldsymbol{\Lambda_k}(\boldsymbol{x_n} - \boldsymbol{\mu_k})\} + \sum_{n=1}^{N} z_{k,n}\log\pi_k\big]) + const \\
&= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}(\sum_{k=1}^{K}\big[\sum_{n=1}^{N} z_{k,n}\{\frac{1}{2}\log|\boldsymbol{\Lambda_k}| - \frac{D}{2}\log 2\pi - \frac{1}{2}(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T\boldsymbol{\Lambda_k}(\boldsymbol{x_n} - \boldsymbol{\mu_k})\} + \sum_{n=1}^{N} z_{k,n}\log\pi_k\big]) + const \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N} z_{k,n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}\big[\frac{1}{2}\log|\boldsymbol{\Lambda_k}| - \frac{D}{2}\log 2\pi - \frac{1}{2}(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T\boldsymbol{\Lambda_k}(\boldsymbol{x_n} - \boldsymbol{\mu_k}) + \log\pi_k\big] + const \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N} z_{k,n}\big[\mathbb{E}_{\boldsymbol{\pi}}[\log\pi_k] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}}[\log|\boldsymbol{\Lambda_k}|] - \frac{D}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}[(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T\boldsymbol{\Lambda_k}(\boldsymbol{x_n} - \boldsymbol{\mu_k})]\big] + const.
\end{aligned}$$

We define $\rho_{k,n}$ as

$$\log\rho_{k,n} = \mathbb{E}_{\boldsymbol{\pi}}[\log\pi_k] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}}[\log|\boldsymbol{\Lambda_k}|] - \frac{D}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}[(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T\boldsymbol{\Lambda_k}(\boldsymbol{x_n} - \boldsymbol{\mu_k})].$$

After the substitution to the above equation, we can get

$$q^*(Z) \propto \prod_{n=1}^{N}\prod_{k=1}^{K}\rho_{k,n}^{z_{k,n}}.$$

We normalize $\rho_{k,n}$ as

$$r_{k,n} = \frac{\rho_{k,n}}{\sum_{j=1}^{K}\rho_{j,n}},$$

then we can finally get

$$q^*(\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K}r_{k,n}^{z_{k,n}}.$$

In addition, by properties of Dirichlet and Gaussian-Wishart distribution, we can compute the $\rho_{k,n}$ with

$$\mathbb{E}_{\boldsymbol{\pi}}[\log \pi_k] = \psi(\alpha_k) - \psi(\alpha_1 + \cdots + \alpha_K),$$

$$\mathbb{E}_{\boldsymbol{\Lambda}}[\log |\boldsymbol{\Lambda_k}|] = \sum_{i=1}^{D} \psi\big(\frac{\nu_k + 1 - i}{2}\big) + D\log 2 + \log |\boldsymbol{W_k}|, \text{ and}$$

$$\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}[(\boldsymbol{x_n} - \boldsymbol{\mu_k})^T \boldsymbol{\Lambda_k} (\boldsymbol{x_n} - \boldsymbol{\mu_k})] = D\beta_k^{-1} + \nu_k (\boldsymbol{x_n} - \boldsymbol{m_k})^T \boldsymbol{W_k} (\boldsymbol{x_n} - \boldsymbol{m_k}).$$

For variational M step, we have to compute $\log q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathbb{E}_{\boldsymbol{Z}}(\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})) + const.$ We can see that the second and the third terms only depend on $\boldsymbol{pi}$, and do not depend on $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. Therefore, $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$.

We first calculate $q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. By removing unrelated terms, we can see that

$$\begin{aligned}
\log q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \mathbb{E}_{\boldsymbol{Z}}\Big[\sum_{k=1}^{K} \big[\sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x_n} | \boldsymbol{\mu_k}, \boldsymbol{\Lambda_k}^{-1})^{z_{k,n}} + \log \mathcal{N}(\boldsymbol{\mu_k} | \boldsymbol{m_0}, (\beta_0 \boldsymbol{\Lambda_k})^{-1}) + \log \mathcal{W}(\boldsymbol{\Lambda_k} | \boldsymbol{W_0}, \nu_0) \big] \Big] + const \\
&= \sum_{k=1}^{K} \big[\sum_{n=1}^{N} r_{k,n} \log \mathcal{N}(\boldsymbol{x_n} | \boldsymbol{\mu_k}, \boldsymbol{\Lambda_k}^{-1}) + \log \mathcal{N}(\boldsymbol{\mu_k} | \boldsymbol{m_0}, (\beta_0 \boldsymbol{\Lambda_k})^{-1}) + \log \mathcal{W}(\boldsymbol{\Lambda_k} | \boldsymbol{W_0}, \nu_0) \big] + const
\end{aligned}$$

This can be regarded as Gaussian evidences updated(the first term) over Gaussian-Wishart prior(the second and the third terms). Therefore, we can use the formula in [1, pg.3]. After substitution, we can get

$$q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}\big(\boldsymbol{\mu_k} | \boldsymbol{m_k}, (\beta_k \boldsymbol{\Lambda_k})^{-1}\big) \mathcal{W}(\boldsymbol{\Lambda_k} | \nu_k),$$

where

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

$$\overline{\boldsymbol{x_k}} = \frac{1}{N_k} \sum_{n=1}^{N} r_{k,n} \boldsymbol{x_n},$$

$$\boldsymbol{Z_k} = \frac{1}{N_k} \sum_{n=1}^{N} r_{k,n} (\boldsymbol{x_n} - \overline{\boldsymbol{x_k}})(\boldsymbol{x_n} - \overline{\boldsymbol{x_k}})^T,$$

$$\beta_k = \beta_0 + N_k,$$

$$\boldsymbol{m_k} = \frac{1}{\beta_k} (\beta_0 \boldsymbol{m_0} + N_k \overline{\boldsymbol{x_k}}),$$

$$\nu_k = \nu_0 + N_k, \text{ and}$$

$$\boldsymbol{W_k}^{-1} = \boldsymbol{W_0}^{-1} + N_k \boldsymbol{Z_k} + \frac{\beta_0 N_k}{\beta_0 + N_k} (\overline{\boldsymbol{x_k}} - \boldsymbol{m_0})(\overline{\boldsymbol{x_k}} - \boldsymbol{m_0})^T.$$

For $q(\boldsymbol{\pi})$, we can see that

$$\begin{aligned}\log q^*(\boldsymbol{\pi}) &= \mathbb{E}_{\boldsymbol{Z}}\Big[\sum_{k=1}^{K}\big[\sum_{n=1}^{N}\log(\pi_k^{z_{k,n}}) + \log\mathcal{D}(\pi_k|\alpha_0)\big]\Big] + const \\ &= \sum_{k=1}^{K}\big[\sum_{n=1}^{N} r_{k,n}\log(\pi_k) + \log\mathcal{D}(\pi_k|\alpha_0)\big] + const \\ &= \sum_{k=1}^{K}\big[(\alpha_0 + N_k - 1)\log(\pi_k)\big] + const. \end{aligned} \tag{3}$$

We can finally get

$$q^*(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}),$$

where

$$\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_K]^T \text{ and } \alpha_k = \alpha_0 + N_k.$$

□

3 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 used the old faithful data in `https://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat`. Before the experiment, we normalized this data to be in (-1, 1). We set the $\alpha_0 = 0.1$, $\beta_0 = 1$, $\boldsymbol{W_0} = \begin{bmatrix} 0.05 & 0 \\ 0 & 0.05 \end{bmatrix}$, $\nu_0 = 50$, and $\boldsymbol{m_0}$ as the randomly picked data. The number of clusters is 20. The result is shown in Figure 1. Each ellipse corresponds to the mean value of Λ_k and μ_k. The transparency corresponds with the mean value of π_k. We can see that the most of ellipses are reduced, and only two ellipse remains.

We also experiments with $\alpha_0 = 10$. The result is shown in Figure 2. We can see that some ellipses remain even the computation converges.

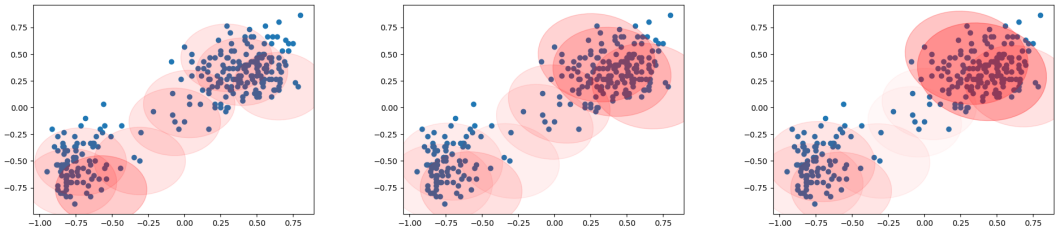
4 Considerations

4.1 Model Selection

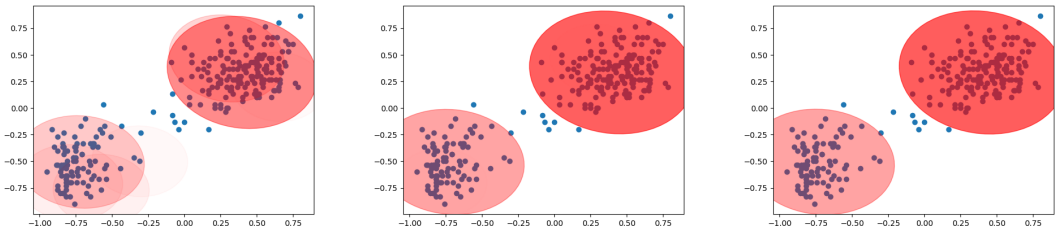
The old faithful data are composed of two clusters, so it is appropriate to apply MoG model to this data. This time, we used variational MoG.

4.2 Variational MoG vs MoG

MoG with EM algorithm also finds the clusters of data, but it has some limitations. First, it just estimate the point of parameters. Therefore, it does not consider the uncertainty of parameters. One of the effect of this is that this method is sensitive to the initialization. However, the variational method estimate the distribution of parameters, so it is known to be less sensitive than normal MoG, and variational MoG

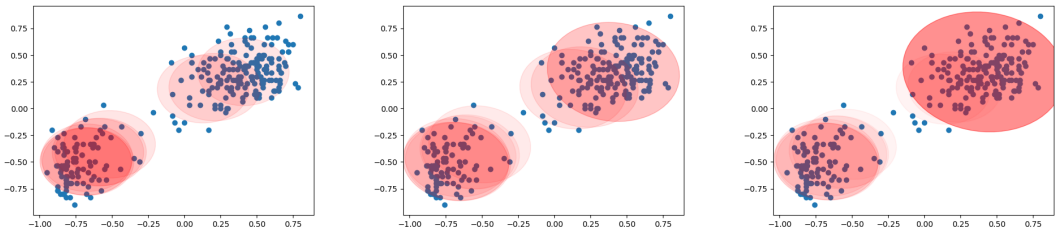


(a) 0th iteration (b) 5th iteration (c) 10th iteration



(d) 15th iteration (e) 20th iteration (f) 25th iteration

Figure 1: Result of the experiment with $\alpha_0 = 0.1$



(a) 0th iteration (b) 20th iteration (c) 40th iteration

Figure 2: Result of the experiment with $\alpha_0 = 10$

is more robust than normal MoG(even in small data set). Thus, we don't have to care the number of clusters when we use variational MoG, because this prunes the useless clusters itself. However, variational MoG is more complicated, so we have to appropriately choose the model depend on the problem which we target.

5 Conclusion

We could review and derived the optimization method of variational MoG. We also implemented the method and verified with simple experiment. We could see successful results for applying our method. Thus we considered about the model selection and comparison with normal MoG.

References

[1] Tom SF Haines. Gaussian conjugate prior cheat sheet.