Assignment 1

JOINT DISTRIBUTION

$$p(\mu, c, x) = \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{N} p(c_i)\, p(x_i | c_i, \mu) \tag{1}$$

The log of joint distribution is

$$\begin{aligned} log\ p(\mu, x, c) &= log\ p(\mu)p(c)p(x \mid c, \mu) && (2)\\ &= \sum_k log\ p(\mu_k) + \sum_i \big[log\ p(c_i) + log\ p(x_i \mid c_i, \mu)\big] && (3) \end{aligned}$$

VARIATIONAL DISTRIBUTION

$$\begin{aligned} log\ q(c, \mu) &= log\ q(c) + log\ q(\mu) && (4)\\ &= \sum_i log\ p(c_i; \phi_i) + \sum_k log\ p(\mu_k; m_k, s_k^2) && (5) \end{aligned}$$

ELBO
The ELBO can be exressed as

$$\begin{aligned} \mathcal{L}\left(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi}\right) &= E_q[\log p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{c})] - E_q[\log q(\boldsymbol{\mu}, \boldsymbol{c})] && (6)\\ &= \sum_{k=1}^{K} E_q\left[\log p\left(\boldsymbol{\mu}_k\right)\right] + \sum_{i=1}^{N} E_q\left[\log p\left(c_i\right)\right] + \sum_{i=1}^{N} E_q\left[\log p\left(\boldsymbol{x}_i | c_i, \boldsymbol{\mu}\right)\right] && (7)\\ &- \sum_{k=1}^{K} E_q\left[\log q\left(\boldsymbol{\mu}_k\right)\right] - \sum_{i=1}^{N} E_q\left[\log q\left(c_i\right)\right] && (8) \end{aligned}$$

PART:1 Prior on μ_k Mixture Location

$$\begin{aligned} log\ p(\mu_k) &= log\ \left\{\frac{1}{(2\pi\sigma^2)^{\frac{P}{2}}} e^{-\big[\frac{(\mu_k - \alpha)^T(\mu_k - \alpha)}{2\sigma^2}\big]}\right\} && (9)\\ &= log\ \frac{1}{(2\pi\sigma^2)^{\frac{P}{2}}} - \big[\frac{\mu_k^T\mu_k}{2\sigma^2} - \frac{\alpha\mu_k}{\sigma^2} + \frac{\alpha^2}{2\sigma^2}\big] && (10)\\ &= -\frac{P}{2} log(2\pi\sigma^2) - \frac{\mu_k^T\mu_k}{2\sigma^2} + \frac{\alpha\mu_k}{\sigma^2} - \frac{\alpha^2}{2\sigma^2} && (11) \end{aligned}$$

The expected log prior over mixture locations is:

$$\begin{aligned} \sum_{k=1}^{K} E_q(log\ p(\mu_k)) &= \sum_{k=1}^{K} \big[E_q\big[-\frac{P}{2}log(2\pi\sigma^2) - \frac{\mu_k^T\mu_k}{2\sigma^2} + \frac{\alpha\mu_k}{\sigma^2} - \frac{\alpha^2}{2\sigma^2}\big]\big] && (12)\\ &= \sum_{k=1}^{K} \big[-\frac{P}{2}log(2\pi\sigma^2) - E_q\big[\frac{\mu_k^T\mu_k}{2\sigma^2}\big] + E_q\big[\frac{\alpha\mu_k}{\sigma^2}\big] - \frac{\alpha^2}{2\sigma^2}\big] && (13)\\ &= \sum_{k=1}^{K} \big[-\frac{P}{2}log(2\pi\sigma^2) - \frac{E_q\big[\mu_k^T\mu_k\big]}{2\sigma^2} + \frac{\alpha E_q\big[\mu_k\big]}{\sigma^2} - \frac{\alpha^2}{2\sigma^2}\big] && (14) \end{aligned}$$

PART:2 Prior on c_i Mixture assignments

$$log\ p(c_i) = -log(K) \tag{15}$$

The expected log prior over mixture assignments:

$$\sum_{i=1}^{N} E_q\big[log\ p(c_i)\big] = \sum_{i=1}^{N} -log(K) \tag{16}$$

PART:3 Log Likelihood

$$\begin{aligned}
log\ p(x_i \mid c_i, \mu) &= log\ \prod_k p(x_i \mid \mu_k)^{c_{i,k}} &(17)\\
&= \sum_k c_{i,k}\big[log\ p(x_i \mid \mu_k)\big] &(18)\\
&= \sum_k c_{i,k}\big[log\ \Big\{\frac{1}{(2\pi\lambda^2)^{\frac{P}{2}}} e^{-\big[\frac{(x_i-\mu_k)^T(x_i-\mu_k)}{2\lambda^2}\big]}\Big\}\big] &(19)\\
&= \sum_k c_{i,k}\big[-\frac{P}{2}log(2\pi\lambda^2) - \big[\frac{(x_i-\mu_k)^T(x_i-\mu_k)}{2\lambda^2}\big]\big] &(20)\\
&= \sum_k c_{i,k}\big[-\frac{P}{2}log(2\pi\lambda^2) - \big[\frac{(x_i-\mu_k)^T(x_i-\mu_k)}{2\lambda^2}\big]\big] &(21)\\
&= \sum_k c_{i,k}\big[-\frac{P}{2}log(2\pi\lambda^2) - \frac{x_i^T x_i}{2\lambda^2} + \frac{x_i^T\mu_k}{\lambda^2} - \frac{\mu_k^T\mu_k}{2\lambda^2}\big] &(22)
\end{aligned}$$

The expected log likelihood will be

$$\begin{aligned}
\sum_{i=1}^{N} E_q log\ p(x_i \mid c_i, \mu) &= \sum_{i=1}^{N}\big[E_q\left[\sum_{k=1}^{K} c_{i,k}\log p\left(\boldsymbol{x}_i|\boldsymbol{\mu}_k\right)\right]\big] &(23)\\
&= \sum_{i=1}^{N}\big[E_q \sum_k c_{i,k}\big[-\frac{P}{2}log(2\pi\lambda^2) - \frac{x_i^T x_i}{2\lambda^2} + \frac{x_i^T\mu_k}{\lambda^2} - \frac{\mu_k^T\mu_k}{2\lambda^2}\big]\big] &(24)
\end{aligned}$$

Components that do not depend on μ_k are constants.

$$= \sum_{i=1}^{N}\big[\phi_{i,k}\left(\frac{x_i^T\mathbb{E}\left[\mu_k\right]}{\lambda^2} - \frac{\mathbb{E}\left[\mu_k^T\mu_k\right]}{2\lambda^2}\right)\big] + \text{ const.} \tag{25}$$

PART:4 Variational posterior on c_i Mixture assignments

$$\begin{aligned}
log\ p(c_i;\phi_i) &= log\ \prod_k p(c_i = k;\phi_i) &(26)\\
&= log\ \prod_k (\phi_{i,k})^{c_{i,k}} &(27)\\
&= \sum_k c_{i,k} log(\phi_{i,k}) &(28)
\end{aligned}$$

The entropy of each variational assignment posterior is:

$$\begin{aligned}
\sum_{i=1}^{N}\big[E_q \log\ p(c_i;\phi_i)\big] &= \sum_{i=1}^{N}\big[E_q[log\ \prod_k p(c_i = k;\phi_i)] &(29)\\
&= \sum_{i=1}^{N} E_q\big[\sum_k c_{i,k} log(\phi_{i,k})\big] &(30)\\
&= \sum_{i=1}^{N}\sum_k E_q\big[c_{i,k}\big] log(\phi_{i,k}) &(31)\\
&= \sum_{i=1}^{N}\sum_k \phi_{i,k} log(\phi_{i,k})\big] &(32)
\end{aligned}$$

PART:5 Variational posterior on μ_k Mixture Location

$$log\ q(\mu_k; m_k, s_k^2) = log\left\{\frac{1}{(2\pi s_k^2)^{\frac{P}{2}}} e^{[-\frac{(\mu_k - m_k)^T(\mu_k - m_k)}{2s_k^2}]}\right\} \tag{33}$$

$$= -\frac{P}{2} log\ (2\pi s_k^2) - \frac{(\mu_k - m_k)^T(\mu_k - m_k)}{2s_k^2} \tag{34}$$

$$= -\frac{P}{2} log(2\pi s_k^2) - [\frac{\mu_k^T\mu_k - 2m_k^T\mu_k + m_k^T m_k}{2s_k^2}] \tag{35}$$

$$= -\frac{P}{2} log(2\pi s_k^2) - \frac{\mu_k^T\mu_k}{2s_k^2} + \frac{m_k^T\mu_k}{s_k^2} - \frac{m_k^T m_k}{2s_k^2} \tag{36}$$

The entropy of each variational location posterior is

$$\sum_k E_q \log q\left(\boldsymbol{\mu}_k\right) = \sum_k E_q\big[-\frac{P}{2} log(2\pi s_k^2) - \frac{\mu_k^T\mu_k}{2s_k^2} + \frac{m_k^T\mu_k}{s_k^2} - \frac{m_k^T m_k}{2s_k^2}\big] \tag{37}$$

$$= \sum_k \big[-\frac{P}{2} log(2\pi s_k^2) - \frac{E_q[\mu_k^T\mu_k]}{2s_k^2} + \frac{m_k^T E_q[\mu_k]}{s_k^2} - \frac{m_k^T m_k}{2s_k^2}\big] \tag{38}$$

$$\text{substituting E}[\mu_k] = m_k \text{ and E}[\mu_k^T\mu_k] = m_k^T m_k + s_k^2 \tag{39}$$

$$= \sum_k \big[-\frac{P}{2} log(2\pi s_k^2) - \frac{1}{2}\big] \tag{40}$$

Substituting each term from the above derivations , setting $\alpha = 0$ and $\lambda = I$ and **ignoring the constants**.

$$ELBO \propto \sum_k -\frac{\mathbb{E}[\mu_k^T\mu_k]}{2\sigma^2} + \sum_i \phi_{i,k}\left(\frac{x_i^T\mathbb{E}\left[\mu_k\right]}{I^2} - \frac{\mathbb{E}\left[\mu_k^T\mu_k\right]}{2I^2}\right) \tag{41}$$

$$-\sum_i\sum_k\Big[\phi_{i,k}\log\ \phi_{ik}\Big] + \sum_k \frac{P}{2} log\ (2\pi s_k^2) \tag{42}$$

$$\text{where E}[\mu_k] = m_k \text{ and E}[\mu_k^T\mu_k] = m_k^T m_k + s_k^2 \tag{43}$$

Assignment 2

The optimal $q_j^*\left(z_j\right)$ given that all other factors are fixed is

$$q_j^*\left(z_j\right) \propto \exp\left\{E_{-j|}\left[\log p\left(z_j, z_{-j}, \boldsymbol{x}\right)\right]\right\} \tag{44}$$

Log joint distribution is

$$\begin{aligned}\log p\left(\boldsymbol{x}, c_i, \boldsymbol{c}_{-i}, \boldsymbol{\mu}\right) = \log p(\boldsymbol{\mu}) + \sum_{j\neq i}\left(\log p\left(c_j\right) + \log p\left(\boldsymbol{x}_j|c_j, \boldsymbol{\mu}\right)\right) \\ + \log p\left(c_i\right) + \log p\left(\boldsymbol{x}_i|c_i, \boldsymbol{\mu}\right)\end{aligned} \tag{45}$$

Terms that are not functions of c_i are constants, therefore

$$q^*\left(c_i\right) \propto \exp\left\{\log p\left(c_i\right) + E\left[\log p\left(\boldsymbol{x}_i|c_i, \boldsymbol{\mu}\right)\right]\right\} \tag{46}$$

$\log p(c_i) = -logK$ which is a constant. The second term is the expectation of log of the c_i'th Gaussian density. Recalling that c_i is an indicator vector, from **17** we have,

$$p\left(x_i|c_i, \mu\right) = \prod_{k=1}^{K} p\left(x_i|\mu_k\right)^{c_{ik}} \tag{47}$$

The expectation of the Gaussian in the variational update for cluster assignments can be written as

$$
\begin{aligned}
\mathbb{E}\left[\log p\left(x_i|c_i,\mu\right)\right] &= \sum_k c_{ik}\mathbb{E}(\left[\log p\left(x_i|\mu_k\right);m_k,s_k^2\right] \\
&= \sum_k c_{ik}\mathbb{E}\log\left\{\frac{1}{(2\pi\lambda^2)^{\frac{P}{2}}}e^{-\left[\frac{(x_i-\mu_k)^T(x_i-\mu_k)}{2\lambda^2}\right]}\right\} \\
&= \sum_k c_{ik}\mathbb{E}\Big\{\frac{-P}{2}\log(2\pi\lambda^2)-\big[\frac{(x_i^Tx_i-2\mu_kx_i^T+\mu_k^T\mu_k)}{2\lambda^2}\big]\Big\} \\
&= \sum_k c_{ik}\mathbb{E}\Big\{\frac{(x_i^Tx_i+2\mu_kx_i^T-\mu_k^T\mu_k)}{2\lambda^2}\Big\}+\text{ const.} \\
&= \sum_k c_{ik}\left(\frac{\mathbb{E}\left[\mu_k\right]x_i^T}{\lambda^2}-\frac{\mathbb{E}\left[\mu_k^T\mu_k\right]}{2\lambda^2}\right)+\text{ const.}
\end{aligned} \tag{48}
$$

Terms that are constant with respect to c_i can be ignored.Substituting 48 in 46 gives

$$
\phi_{i,k}\propto\exp\left\{\frac{\boldsymbol{x}_i^TE\left[\boldsymbol{\mu}_k\right]}{\lambda^2}-\frac{E\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right]}{2\lambda^2}\right\} \tag{49}
$$

Assignment 3

Similarly to the derivation of $q\left(c_i\right)$

$$
q^*\left(\boldsymbol{\mu}_k\right)\propto\exp\left\{\log p\left(\boldsymbol{\mu}_k\right)+\sum_{i=1}^N E_{q\left(\boldsymbol{\mu}_{-k},c_i\right)}\left[\log p\left(\boldsymbol{x}_i|c_i,\boldsymbol{\mu}\right)\right]\right\} \tag{50}
$$

$q^*\left(\boldsymbol{\mu}_k\right)$ is proportional to the product of Gaussians, therefore the resulting distribution is also a Gaussian.The expressions in the exponent are:

$$
\log p\left(\boldsymbol{\mu}_k\right)+\sum_{i=1}^N E_{q\left(\boldsymbol{\mu}_{-k},c_i\right)}\left[\log p\left(\boldsymbol{x}_i|c_i,\boldsymbol{\mu}\right)\right] \tag{51}
$$

$$
=\log p\left(\boldsymbol{\mu}_k\right)+\sum_{i=1}^N E_{q(\boldsymbol{\mu}-k,c_i)}\left[\sum_{j=1}^K c_{i,j}\log p\left(\boldsymbol{x}_i|\boldsymbol{\mu}_j\right)\right] \tag{52}
$$

$=\{$ components that do not depend on μ_k are constants .Linear term in μ_k is absent because $\alpha=0.\}$

$$
=\log p\left(\boldsymbol{\mu}_k\right)+\sum_{i=1}^N E_{q(\boldsymbol{\mu}-k,c_i)}\left[c_{i,k}\log p\left(\boldsymbol{x}_i|\boldsymbol{\mu}_k\right)\right]+\text{const} \tag{53}
$$

$$
=-\frac{\mu_k^T\boldsymbol{\mu}_k}{2\sigma^2}+\sum_{i=1}^N E_{q(c_i)}\left[c_{i,k}\right]\log p\left(\boldsymbol{x}_i|\boldsymbol{\mu}_k\right)+\text{ const} \tag{54}
$$

$$
=-\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2\sigma^2}+\sum_{i=1}^N\phi_{i,k}\log p\left(\boldsymbol{x}_i|\boldsymbol{\mu}_k\right)+\text{ const} \tag{55}
$$

$$
=-\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2\sigma^2}+\sum_{i=1}^N\phi_{i,k}\left(-\frac{1}{2\lambda^2}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)^T\left(\boldsymbol{x}_i-\boldsymbol{\mu}_k\right)\right)+\text{ const} \tag{56}
$$

$$
=-\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2\sigma^2}-\sum_{i=1}^N\phi_{i,k}\frac{\left(\boldsymbol{x}_i^T\boldsymbol{x}_i-2\boldsymbol{x}_i\boldsymbol{\mu}_k^T+\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right)}{2\lambda^2}+\text{const} \tag{57}
$$

Since the resulting distribution is also a gaussian with mean m_k and variance s_k^2

$$
p(\mu_k;m_k,s_k^2)=\frac{1}{(2\pi2s_k^2)^{\frac{P}{2}}}e^{\left[-\frac{(\mu_k-m_k)^T(\mu_k-m_k)}{2s_k^2}\right]} \tag{58}
$$

The exponents of the above equation :

$$- \frac{(\mu_k - m_k)^T(\mu_k - m_k)}{s_k^2} \tag{59}$$

$$= - \frac{[\mu_k^T \mu_k - 2m_k\mu_k^T + m_k^T m_k]}{2s_k^2} \tag{60}$$

Comparing Quadratic terms in μ_k 57 and 60

$$- \frac{\mu_k^T \mu_k}{2\sigma^2} - \sum_{i=1}^{N} \phi_{i,k} \frac{\mu_k^T \mu_k}{2\lambda^2} = - \frac{\mu_k^T \mu_k}{2s_k^2} \tag{61}$$

$$- \frac{1}{2}\mu_k^T \mu_k \Big[\frac{1}{\sigma^2} + \frac{\sum_{i=1}^{N} \phi_{i,k}}{\lambda^2} \Big] = - \frac{\mu_k^T \mu_k}{2s_k^2} \tag{62}$$

By inspection ,

$$s_k^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{\sum_{i=1}^{N} \phi_{i,k}}{\lambda^2}} \tag{63}$$

$$\tag{64}$$

In the exercise $\lambda = 1$, the result is further simplified to

$$s_k^2 = \frac{1}{\frac{1}{\sigma^2} + \sum_{i=1}^{N} \phi_{i,k}} \tag{65}$$

$$\tag{66}$$

Comparing Linear terms in μ_k 57 and 60

$$\frac{\sum_{i=1}^{N} \phi_{i,k} x_i \mu_k^T}{\lambda^2} = \frac{m_k \mu_k^T}{s_k^2} \tag{67}$$

By inspection μ_k can be written as

$$m_k = \frac{\sum_{i=1}^{N} \phi_{i,k} x_i}{\lambda^2} * s_k^2 \tag{68}$$

Subsituting λ=1 and s_k^2 from 67, the final update is

$$m_k = \frac{\sum_{i=1}^{N} \phi_{i,k} x_i}{\frac{1}{\sigma^2} + \sum_{i=1}^{N} \phi_{i,k}} \tag{69}$$

Assignment 4

The results from Assignment 1,2,3 were used to complete Assignment 4. The plots of the true data, the posterior and the ELBO can be seen for three randomly generated datasets. In all the datasets the ELBO can be seen to increase monotonically.

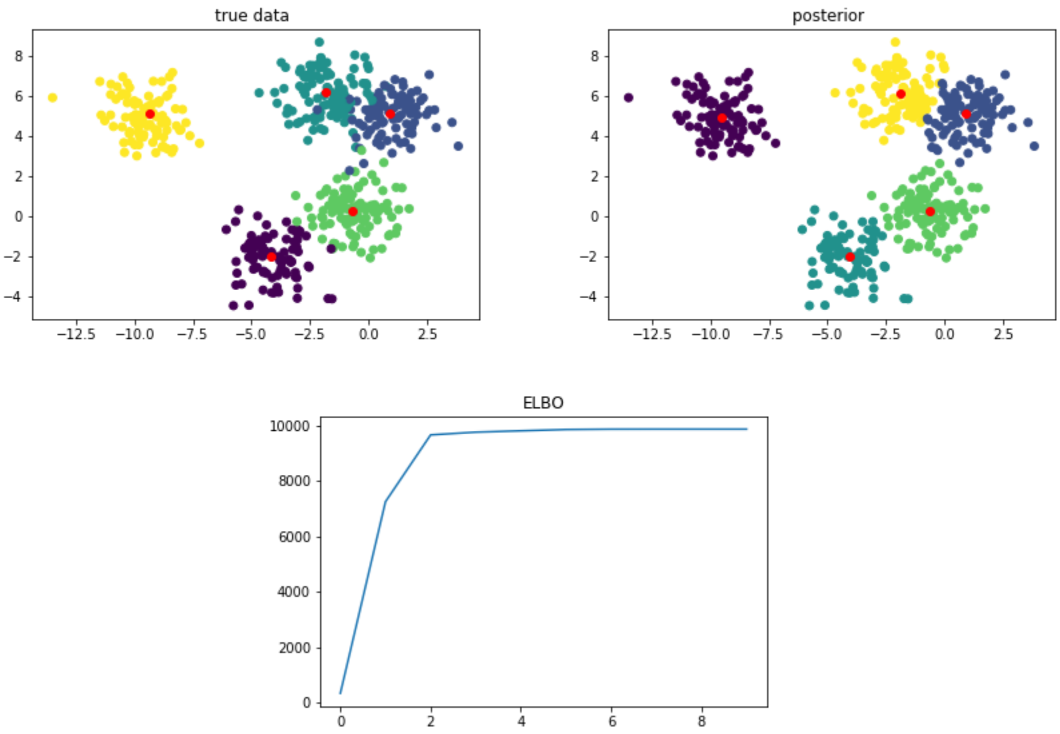

Figure 1: Dataset 1 with 5 clusters

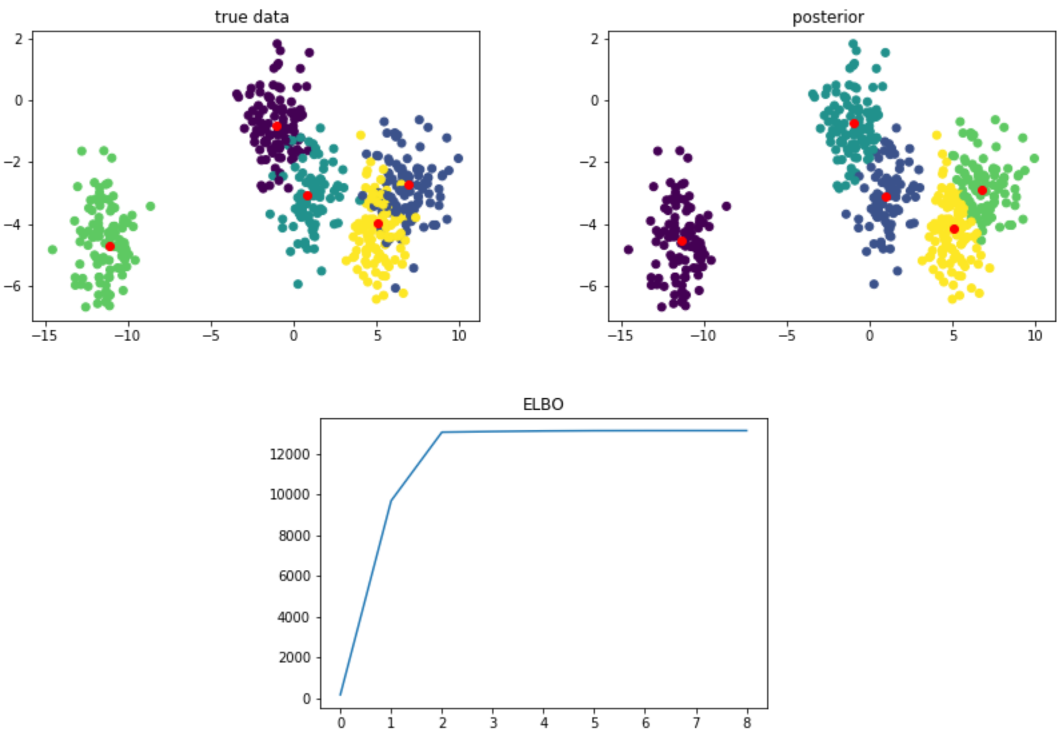

Figure 2: Dataset 2 with 5 clusters

For the last dataset , the number of clusters was increased to 7.

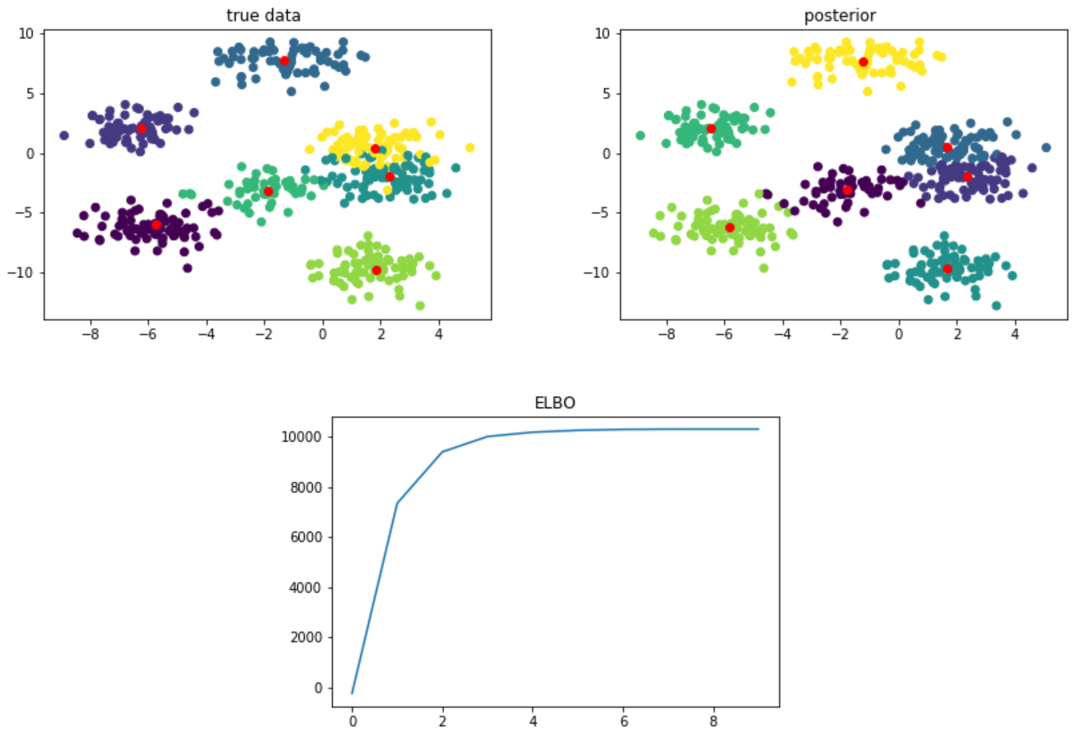

Figure 3: Dataset 3 with 7 clusters

References

[1] Blei, David M., Alp Kucukelbir, and Jon D. McAulie. ”Variational in- ference: A review for statisticians.” Journal of the American Statistical Association just-accepted (2017).

[2] Petersen, K. B. Pedersen, M. S. (2008). The Matrix Cookbook Technical University of Denmark (Version 20081110)

[3] https://stats.stackexchange.com/questions/369718/variational-inference-computation-of-elbo-and-cavi-algorithm?rq=1