Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 2

QUESTION 1

Student Name: Niharika Ahuja
Roll Number: 18111045
Date: March 13, 2019

My solution to problem 1

1 Part 1:

Posterior predictive distribution for a new input $\mathbf{x}_*$ is

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}) = N(f_*|\tilde{\mathbf{k}}_*\tilde{\mathbf{K}}^{-1}\mathbf{t}, k(x_*, x_*) - \tilde{k}_*^T\tilde{\mathbf{K}}^{-1}\tilde{k}_*)$$

With $(\mathbf{Z}, \mathbf{t})$ as pseudo training input, expression for the posterior predictive distribution will be:

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}, \mathbf{Z}) = \int p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}, \mathbf{Z}, \mathbf{t})p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z})d\mathbf{t}$$

Calculating $p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z})$,

$$\begin{aligned} p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z}) &\propto p(\mathbf{t}|\mathbf{Z})p(\mathbf{f}|\mathbf{X}, \mathbf{Z}, \mathbf{t}) \\ &\propto N(\mathbf{t}|0, \tilde{\mathbf{K}})N(\mathbf{f}|\mathbf{A}\mathbf{t}, \Sigma) \end{aligned}$$

Here, $\mathbf{A} = \mathbf{Q}^T\tilde{\mathbf{K}}^{-1}$ of size N $\times$ M, where $\mathbf{Q}$ is a matrix with its columns as, $\tilde{k}_1$, $\tilde{k}_2$,...,$\tilde{k}_n$, of size M $\times$ N. Σ is a diagonal matrix of size N $\times$ N, with each diagonal entry as, $k(x_i, x_i) - \tilde{k}_i^T\tilde{\mathbf{K}}^{-1}\tilde{k}_i$.
Using linear gaussian model,

$$p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z}) = N(t|\mu_t, \Sigma_t)$$

where,

$$\begin{aligned} \Sigma_t &= (\tilde{\mathbf{K}}^{-1} + \mathbf{A}\Sigma^{-1}\mathbf{A})^{-1} \\ \mu_t &= \Sigma_t\mathbf{A}^T\Sigma^{-1}\mathbf{f} \end{aligned}$$

Let, $\mathbf{s} = \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}_*$

$$\begin{aligned} p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}, \mathbf{Z}) &= \int p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f}, \mathbf{Z}, \mathbf{t})p(\mathbf{t}|\mathbf{X}, \mathbf{f}, \mathbf{Z})d\mathbf{t} \\ &= \int N(f_*|\tilde{\mathbf{k}}_*\tilde{\mathbf{K}}^{-1}\mathbf{t}, k(x_*, x_*) - \tilde{k}_*^T\tilde{\mathbf{K}}^{-1}\tilde{k}_*)N(t|\mu_t, \Sigma_t)d\mathbf{t} \\ &= N(y_*|\mathbf{s}^T\mu_t, \mathbf{s}^T\Sigma_t\mathbf{s} + k(x_*, x_*) - \tilde{k}_*^T\tilde{\mathbf{K}}^{-1}\tilde{k}_*) \end{aligned}$$

Comparing this posterior predictive for y_* with the usual GPs posterior predictive for y_* in terms of computational cost. In this case, computational cost will be $O(N + M^3)$ as it requires inversion of the diagonal matrix Σ and $\tilde{\mathbf{K}}$. In the usual case, cost is $O(N^3)$ for inverting $\mathbf{K}$.

2 Part 2

Using the same expression for $\mathbf{A}$ and Σ, Expression for the marginal likelihood will be:

$$\begin{aligned}
p(\mathbf{f}|\mathbf{X},\mathbf{Z}) &= \int p(\mathbf{f},\mathbf{t}|\mathbf{X},\mathbf{Z})d\mathbf{t} \\
&= \int p(\mathbf{f}|\mathbf{X},\mathbf{Z},\mathbf{t})p(\mathbf{t}|0,\tilde{\mathbf{K}})d\mathbf{t} \\
&= \int N(\mathbf{t}|0,\tilde{\mathbf{K}})N(\mathbf{f}|\mathbf{A}\mathbf{t},\Sigma)d\mathbf{t} \\
&= N(\mathbf{f}|0,\mathbf{A}\tilde{\mathbf{K}}\mathbf{A}^T+\Sigma)
\end{aligned}$$

Thus, MLE objective will be: $\arg\min_{\mathbf{Z}} -\log p(\mathbf{f}|\mathbf{X},\mathbf{Z})$

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 2

QUESTION 2

Student Name: Niharika Ahuja
Roll Number: 18111045
Date: March 13, 2019

My solution to problem 2
Local latent variables are $\{c_n, \mathbf{z}_n\}_{n=1}^N$ and global parameters, $\theta = \{\pi_m, \mu_m, \mathbf{W}_m, \sigma_m^2\}_{m=1}^M$.

1. **EM1**

   (a) Conditional posterior of latent variables

$$\begin{aligned}
p(c_n = m|\mathbf{x}_n, \theta) &= \int p(c_n = m, \mathbf{z}_n|\mathbf{x}_n, \theta) d\mathbf{z}_n \\
&= \int \frac{p(\mathbf{x}_n|c_n = m, \mathbf{z}_n, \theta)p(c_n = m, \mathbf{z}_n)}{p(\mathbf{x}_n|\theta)} d\mathbf{z}_n \\
&= \int \frac{p(\mathbf{x}_n|c_n = m, \mathbf{z}_n, \theta)p(\mathbf{z}_n|c_n = m)p(c_n = m)}{p(\mathbf{x}_n|\theta)} d\mathbf{z}_n \\
&\propto \pi_m \int N(\mathbf{x}_n|\mu_m + \mathbf{W}_m\mathbf{z}_n, \sigma_m^2\mathbf{I}_D)N(\mathbf{z}_n|0, \mathbf{I}_K) d\mathbf{z}_n
\end{aligned}$$

   Using linear gaussian model, we get

$$p(c_n = m|\mathbf{x}_n, \theta) \propto \pi_m N(\mathbf{x}_n|\mu_m, \mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D)$$

   Thus, conditional posterior is

$$p(c_n = m|\mathbf{x}_n, \theta) = \frac{\pi_m N(\mathbf{x}_n|\mu_m, \mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D)}{\sum_{i=1}^M \pi_i N(\mathbf{x}_n|\mu_i, \mathbf{W}_i\mathbf{W}_i^T + \sigma_i^2\mathbf{I}_D)}$$

   (b) **CLL:**

$$\log p(\mathbf{X}, \mathbf{C}|\theta) = \sum_{n=1}^N \sum_{m=1}^M c_{nm} \log p(\mathbf{x}_n, c_n = m|\theta)$$

   Using the answer from first part, we get

$$\begin{aligned}
p(\mathbf{x}_n, c_n = m|\theta) &= p(c_n = m|\mathbf{x}_n, \theta)p(\mathbf{x}_n|\theta) \\
&= \pi_m N(\mathbf{x}_n|\mu_m, \mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D)
\end{aligned}$$

   Thus, CLL is

$$\sum_{n=1}^N \sum_{m=1}^M c_{nm}(\log \pi_m + \log N(\mathbf{x}_n|\mu_m, \mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D))$$

   Let $\Sigma_m = \mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D$,

$$\sum_{n=1}^N \sum_{m=1}^M c_{nm}(\log \pi_m - \frac{1}{2}\log|\Sigma_m| - \frac{1}{2}(\mathbf{x}_n - \mu_m)^T\Sigma_m^{-1}(\mathbf{x}_n - \mu_m))$$

ECLL:

$$\sum_{n=1}^{N}\sum_{m=1}^{M} E[c_{nm}](\log \pi_m + \log N(\mathbf{x}_n|\mu_m, \mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D))$$

(c) Here,

$$E[c_{nm}] = \gamma_{nm} = 0 \times p(c_{nm} = 0|\mathbf{x}_n, \theta^{old}) + 1 \times p(c_{nm} = 1|\mathbf{x}_n, \theta^{old})$$

$$\gamma_{nm} = \frac{\pi_m N(\mathbf{x}_n|\mu_m, \mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D)}{\sum_{i=1}^{M}\pi_i N(\mathbf{x}_n|\mu_i, \mathbf{W}_i\mathbf{W}_i^T + \sigma_i^2\mathbf{I}_D)}$$

(d) After maximizing ECLL, we get M step update equations for θ,

$$E[c_{nm}] = \gamma_{nm}$$

$$N_m = \sum_{n=1}^{N}\gamma_{nm}$$

$$\pi_m = \frac{N_m}{N}$$

$$\mu_m = \frac{1}{N_m}\sum_{n=1}^{N}\gamma_{nm}\mathbf{x}_n$$

Let $\mathbf{S}_m = \frac{1}{N_m}\sum_n \gamma_{nm}(\mathbf{x}_n - \mu_m)(\mathbf{x}_n - \mu_m)^T$. $\mathbf{W}$ and σ^2 are calculated by doing eigen decomposition of matrix $\mathbf{S}_m$.

$$\mathbf{W}_m = \mathbf{U}_{mK}(\mathbf{L}_{mK} - \sigma_m^2 I)^{1/2}\mathbf{R}_m \tag{2.3.2}$$

$$\sigma_m^2 = \frac{1}{D-K}\sum_{i=K+1}^{D}\lambda_i$$

Where, $\mathbf{U}_{mK}$ is $D \times K$ matrix of top K eigen vectors of converged $\mathbf{S}_m$, $\mathbf{L}_{mK}$ is $K \times K$ diagonal matrix of top K eigen values. , R_m is a $K \times K$ arbitrary rotation matrix.

(e) Overall sketch of the EM algorithm

i. Initialize $\theta = \{\pi_m, \mu_m, \mathbf{W}_m, \sigma_m^2\}_{m=1}^{M}$ as $\theta^{(0)}$, set t=1

ii. E step: Computing the expectation of each $\mathbf{c}_n$ as given in part (c).

$$E[c_{nm}^{(t)}] = \gamma_{nm}^{(t)}$$

iii. M step: Update equations are:

$$\pi_m = \frac{N_m}{N}$$

$$\mu_m = \frac{1}{N_m}\sum_{n=1}^{N}\gamma_{nm}^{(t)}\mathbf{x}_n$$

Now, $\mathbf{S}_m = \frac{1}{N_m}\sum_n \gamma_{nm}^{(t)}(\mathbf{x}_n - \mu_m)(\mathbf{x}_n - \mu_m)^T$ and eigen decomposition is done to obtain $\mathbf{W}$ and σ^2.

iv. Set $t = t + 1$ and go to step 2 if not yet converged

(f) Corresponding stepwise EM algorithm:

i. Initialize $\theta = \{\pi_m, \mu_m, \mathbf{W}_m, \sigma_m^2\}_{m=1}^M$ as $\theta^{(0)}$, set t=1

ii. Sample a mini batch from training examples.

iii. E step:

$$Q(\theta, \theta^{old}) = E[\log p(\mathbf{X}, \mathbf{C}|\theta)] = \sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm} \log p(\mathbf{x}_n, c_n = m|\theta)$$

$$Q_t = (1 - \gamma_t)Q_{t-1} + \gamma_t \sum_{n=1}^{N_t}\sum_{m=1}^{M} c_{nm} \log p(\mathbf{x}_n, c_n = m|\theta)$$

iv. M step(using equations as given in part (d)):

$$\theta_t = (1 - \gamma_t)\theta_{t-1} + \gamma_t \arg\max_\theta Q(\theta, \theta^{(t-1)})$$

v. Set $t = t + 1$ and go to step 2 if not yet converged

2. **EM2**

(a) Conditional posterior of latent variables

$$p(c_n = m, \mathbf{z}_n|\mathbf{x}_n, \theta) = p(\mathbf{z}_n|c_n = m, \mathbf{x}_n, \theta)p(c_n = m|\mathbf{x}_n, \theta)$$

Now, calculating

$$\begin{aligned} p(\mathbf{z}_n|c_n = m, \mathbf{x}_n, \theta) &\propto p(\mathbf{x}_n|c_n = m, \mathbf{z}_n, \theta)p(\mathbf{z}_n|c_n = m) \\ &\propto N(\mathbf{x}_n|\mu_m + \mathbf{W}_m\mathbf{z}_n, \sigma_m^2\mathbf{I}_D)N(\mathbf{z}_n|0, \mathbf{I}_K) \end{aligned}$$

Using Linear gaussian model,

$$p(\mathbf{z}_n|c_n = m, \mathbf{x}_n, \theta) = N(\mathbf{z}_n|\mu_p, \Sigma_p)$$

where,

$$\Sigma_p = \check{(}I_K + \frac{1}{\sigma_m^2}\mathbf{W}_m^T\mathbf{W}_m)^{-1}$$

$$\mu_p = \Sigma_p\frac{1}{\sigma_m^2}\mathbf{W}_m^T(\mathbf{x}_n - \mu_m)$$

The expected values will be,

$$E[\mathbf{z}_n] = \mu_p$$

$$E[\mathbf{z}_n\mathbf{z}_n^T] = E[\mathbf{z}_n]E[\mathbf{z}_n]^T + cov(\mathbf{z}_n) = \mu_p\mu_p^T + \Sigma_p$$

$p(c_n = m|\mathbf{x}_n, \theta)$ has the same expression as in EM1. So,$E[c_{nm}] = \gamma_{nm}$

(b) **CLL:**

$$\begin{aligned}
\log p(\mathbf{X}, \mathbf{Z}, \mathbf{C}|\theta) &= \sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm} \log p(\mathbf{x}_n, \mathbf{z}_n, c_n = m|\theta) \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm}(\log N(\mathbf{x}_n|\mu_m + \mathbf{W}_m\mathbf{z}_n, \sigma_m^2\mathbf{I}_D) + \log N(\mathbf{z}_n|0, \mathbf{I}_K) + \log \pi_m) \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm}\Big(-(\frac{D}{2}\log\sigma_m^2 + \frac{||\mathbf{x}_n - \mu_m||^2}{2\sigma_m^2} \\
&+ \frac{1}{2\sigma_m^2} tr(\mathbf{z}_n\mathbf{z}_n^T\mathbf{W}_m^T\mathbf{W}_m) - \frac{1}{\sigma_m^2}\mathbf{z}_n^T\mathbf{W}_m^T(\mathbf{x}_n - \mu_m) + \frac{1}{2}tr(\mathbf{z}_n\mathbf{z}_n^T)) + \log\pi_m\Big)
\end{aligned}$$

ECLL:

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{m=1}^{M} E[c_{nm}]\Big(-(\frac{D}{2}\log\sigma_m^2 + \frac{||\mathbf{x}_n - \mu_m||^2}{2\sigma_m^2} + \frac{1}{2\sigma_m^2} tr(E[\mathbf{z}_n\mathbf{z}_n^T]\mathbf{W}_m^T\mathbf{W}_m) \\
&\qquad - \frac{1}{\sigma_m^2}E[\mathbf{z}_n^T]\mathbf{W}_m^T(\mathbf{x}_n - \mu_m) \ + \frac{1}{2}tr(E[\mathbf{z}_n\mathbf{z}_n^T])) + \log\pi_m\Big)
\end{aligned}$$

(c) Here,

$$E[c_{nm}] = \gamma_{nm} = \frac{\pi_m N(\mathbf{x}_n|\mu_m, \mathbf{W}_m\mathbf{W}_m^T + \sigma_m^2\mathbf{I}_D)}{\sum_{i=1}^{M}\pi_i N(\mathbf{x}_n|\mu_i, \mathbf{W}_i\mathbf{W}_i^T + \sigma_i^2\mathbf{I}_D)}$$

and,

$$E[\mathbf{z}_n] = \mu_p$$
$$E[\mathbf{z}_n\mathbf{z}_n^T] = E[\mathbf{z}_n]E[\mathbf{z}_n]^T + cov(\mathbf{z}_n) = \mu_p\mu_p^T + \Sigma_p$$

(d) M step, the update equations are as follows, where $N_m = \sum_{n=1}^{N}\gamma_{nm}$:

$$\pi_m = \frac{N_m}{N}$$

$$\mathbf{W}_m = (\sum_{n=1}^{N}\gamma_{nm}(\mathbf{x}_n - \mu_m)E[\mathbf{z}_n]^T)(\sum_{n=1}^{N}\gamma_{nm}E[\mathbf{z}_n\mathbf{z}_n^T])^{-1}$$

$$\sigma_m^2 = \frac{1}{N_m D}\sum_{n=1}^{N}\gamma_{nm}(||\mathbf{x}_n - \mu_m||^2 - 2E[\mathbf{z}_n]^T\mathbf{W}_m^T(\mathbf{x}_n - \mu_m) + tr(E[\mathbf{z}_n\mathbf{z}_n^T]\mathbf{W}_m^T\mathbf{W}_m))$$

$$\mu_m = \frac{1}{N_m}\sum_{n=1}^{N}\gamma_{nm}(\mathbf{x}_n - \mathbf{W}_m E[\mathbf{z}_n])$$

(e) Overall sketch of the EM algorithm:

i. Initialize $\theta = \{\pi_m, \mu_m, \mathbf{W}_m, \sigma_m^2\}_{m=1}^{M}$ as $\theta^{(0)}$, set t=1

ii. E step: Computing the expectation of each $\mathbf{c}_n$ and $\mathbf{z}_n$ as given in part (c).

$$E[c_{nm}^{(t)}] = \gamma_{nm}^{(t)}$$

$$E[\mathbf{z}_n^{(t)]=\mu_p^{(t)}}$$

$$E[\mathbf{z}_n^{(t)}\mathbf{z}_n^{(t)T}] = \mu_p^{(t)}\mu_p^{(t)T} + \Sigma_p^{(t)}$$

iii. M step: Update equations are same as done in part (d).

iv. Set $t = t + 1$ and go to step 2 if not yet converged

(f) Corresponding stepwise EM algorithm:

i. Initialize $\theta = \{\pi_m, \mu_m, \mathbf{W}_m, \sigma_m^2\}_{m=1}^{M}$ as $\theta^{(0)}$, set t=1

ii. Sample a mini batch from training examples.

iii. E step:

$$Q(\theta, \theta^{old}) = E[\log p(\mathbf{X}, \mathbf{C}|\theta)] = \sum_{n=1}^{N} \sum_{m=1}^{M} c_{nm} \log p(\mathbf{x}_n, c_n = m|\theta)$$

$$Q_t = (1 - \gamma_t) Q_{t-1} + \gamma_t \sum_{n=1}^{N_t} \sum_{m=1}^{M} c_{nm} \log p(\mathbf{x}_n, c_n = m|\theta)$$

iv. M step(using equations as given in part (d)):

$$\theta_t = (1 - \gamma_t)\theta_{t-1} + \gamma_t \arg\max_{\theta} Q(\theta, \theta^{(t-1)})$$

v. Set $t = t + 1$ and go to step 2 if not yet converged

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Niharika Ahuja
Roll Number: 18111045
Date: March 13, 2019

QUESTION 3

My solution to problem 3

Deriving the mean-field VI algorithm for approximating the posterior distribution, $p(\mathbf{w}, \beta, \alpha_1, ..., \alpha_D|\mathbf{y}, \mathbf{X})$ by $q(\mathbf{z}|\phi)$.

$$q(\mathbf{z}|\phi) = q(\mathbf{w}|\phi_w)q(\beta|\phi_\beta)\prod_{d=1}^{D} q(\alpha_d|\phi_d)$$

Calculating the conditional posteriors,

1.

$$\begin{aligned} p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \beta, \alpha_1, ..., \alpha_D) &\propto p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha_1, ..., \alpha_D) \\ &\propto \prod_{n=1}^{N} N(\mathbf{y}|\mathbf{X}\mathbf{w}, \beta^{-1})N(\mathbf{w}|0, \boldsymbol{\alpha}^{-1}) \\ &= N(\mathbf{w}|\mu_N, \Sigma_N) \end{aligned}$$

Here,

$$\Sigma_N = (\beta\mathbf{X}^T\mathbf{X} + \alpha)^{-1}$$

and

$$\mu_N = \beta\Sigma_N\mathbf{X}^T\mathbf{y}$$

Since we need to take expectations only for natural parameters, so

$$\phi_w = E_{q\neq w}[\Sigma_N^{-1}, -\frac{1}{2}\Sigma_N^{-1}]^T$$

$$\phi_w = [\beta\mathbf{X}^T\mathbf{y}, -\frac{1}{2}E_{\phi_\beta}[\beta]\mathbf{X}^T\mathbf{X} + E_{\phi_\alpha}[\alpha]]^T$$

2.

$$\begin{aligned} p(\beta|\mathbf{y}, \mathbf{X}, \mathbf{w}, \alpha_1, ..., \alpha_D) &\propto p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta)p(\beta|a_o, b_o) \\ &= Gamma(\beta|a_o + \frac{N}{2}, b_o + \frac{\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2}{2}) \end{aligned}$$

Expectations for natural parameters in this case will be,

$$\phi_\beta = E_{q\neq\beta}[-b_o + \frac{\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2}{2}, a_o + \frac{N}{2} - 1]^T$$

$$\phi_\beta = [-b_o + \frac{\sum_{n=1}^{N} E_{\phi_\mathbf{w}}[(y_n - \mathbf{w}^T\mathbf{x}_n)^2}{2}], a_o + \frac{N}{2} - 1]^T$$

3.

$$p(\alpha_d|\mathbf{w}, \alpha_{-d}, \mathbf{y}, \mathbf{X}, \beta) \propto p(\mathbf{w}|\boldsymbol{\alpha})p(\alpha_d|e_o, f_o)$$
$$= Gamma(\alpha_d|e_o + \frac{1}{2}, f_o + \frac{w_d^2}{2})$$

Expectations for natural parameters in this case will be,

$$\phi_d = E_{q \neq \alpha_d}[-f_o + \frac{w_d^2}{2}, e_o + \frac{1}{2} - 1]^T$$

$$\phi_d = [-f_o + \frac{E_{\phi_w}[w_d^2]}{2}, e_o + \frac{1}{2} - 1]^T$$

So, the algorithm is as follows:

1. Initialize variational parameters
2. Update ϕ_w,ϕ_β,ϕ_d $\forall$ d, using the above equations.
3. Compute ELBO($E_q[\log p(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}, \mathbf{w}, \alpha)] - E_q[\log q(\mathbf{z})]$) and go to step 2 if not yet converged.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Niharika Ahuja
Roll Number: 18111045
Date: March 13, 2019

QUESTION 4

My solution to problem 4

VI for bayesian logistic regression

$$p(y_n|w, x_n) = \sigma(y_n w^T x_n)$$
$$p(w) = N(0, \lambda^{-1} I)$$
$$q(w|\phi) = N(w|\mu, \Sigma)$$

$$\begin{aligned}
ELBO &= E_q[\log p(Y, w) - \log q(w, \phi)] \\
&= E_q[\log p(Y|X, w) + \log p(w) - \log N(w|\mu, \Sigma)] \\
&= E_q[\sum_n \log \sigma(y_n w^T x_n) - \frac{\lambda}{2} w^T w - \frac{1}{2}\log\det\Sigma + \frac{D}{2}\log\lambda - \frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)] \\
&= E_q[f(X, Y, \mu, \Sigma, \lambda, w)]
\end{aligned}$$

1. Black-box VI based on score-function gradients

$$\nabla_\phi L(q) = E_q[\nabla_\phi \log q(w|\phi)(\log p(Y, w) - \log q(w|\phi))]$$

Gradients with respect to μ and L are as follows:

$$\begin{aligned}
\nabla_\mu L(q) &= E_q[-\Sigma^{-1}(w-\mu) f(X, Y, \mu, \Sigma, \lambda, w)] \\
&\approx \frac{1}{S}\sum_s -\Sigma^{-1}(w^{(s)}-\mu) f(X, Y, \mu, \Sigma, \lambda, w)
\end{aligned}$$

$$\begin{aligned}
\nabla_L L(q) &= E_q[-\frac{1}{2}\Big(\Sigma^{-1} - (-\Sigma^{-1}(w-\mu)(w-\mu)^T\Sigma^{-1})\Big) f(X, Y, \mu, \Sigma, \lambda, w) \times 2L] \\
&\approx \frac{1}{S}\Bigg(-\frac{1}{2}\Big(\Sigma^{-1} - (-\Sigma^{-1}(w-\mu)(w-\mu)^T\Sigma^{-1})\Big)\Bigg) f(X, Y, \mu, \Sigma, \lambda, w) \times 2L
\end{aligned}$$

2. Reparametrization trick based on pathwise gradients Reparametrize $w = \mu + Lv$ where, $v \sim N(0, I)$,

$$ELBO = E_{q(v)}[\log p(Y, w) - \log q(w|\phi)]$$

After replacing $w = \mu + Lv$ in the ELBO EXPRESSION, Gradients with respect to μ and L are as follows:

$$\begin{aligned}
\nabla_\mu L(q) &= E_{q(v)} \sum_n (1 - \sigma(y_n(\mu + Lv)^T x_n)) y_n x_n - \lambda(\mu + Lv) \\
&\approx \frac{1}{S}\sum_s (1 - \sigma(y_n(\mu + Lv^{(s)})^T x_n)) y_n x_n - \lambda(\mu + Lv^{(s)})
\end{aligned}$$

$$
\begin{aligned}
\nabla_L L(q) &= E_{q(v)} \sum_n (1 - \sigma(y_n(\mu + Lv)^T x_n)) y_n x_n v^T - \lambda(\mu v^T + Lvv^T) - L^{-T} \\
&\approx \frac{1}{S} \sum_s (1 - \sigma(y_n(\mu + Lv^{(s)})^T x_n)) y_n x_n v^{(s)T} - \lambda(\mu v^{(s)T} + Lv^{(s)} v^{(s)T}) - L^{-T}
\end{aligned}
$$

Overall sketch of the VI algorithm:

(a) Initialize $\theta = \mu, L$

(b) Choose a mini batch of B examples. While sampling, choose S samples from q(w) in case of BBVI and from q(v) in case of re parametrization.

(c) Posing VI as a general gradient based optimization problem,

$$\theta_{new} = \theta_{old} + \eta(g_\theta)$$

where, g_θ are gradients calculated above.

(d) Go to step (b) until converged.

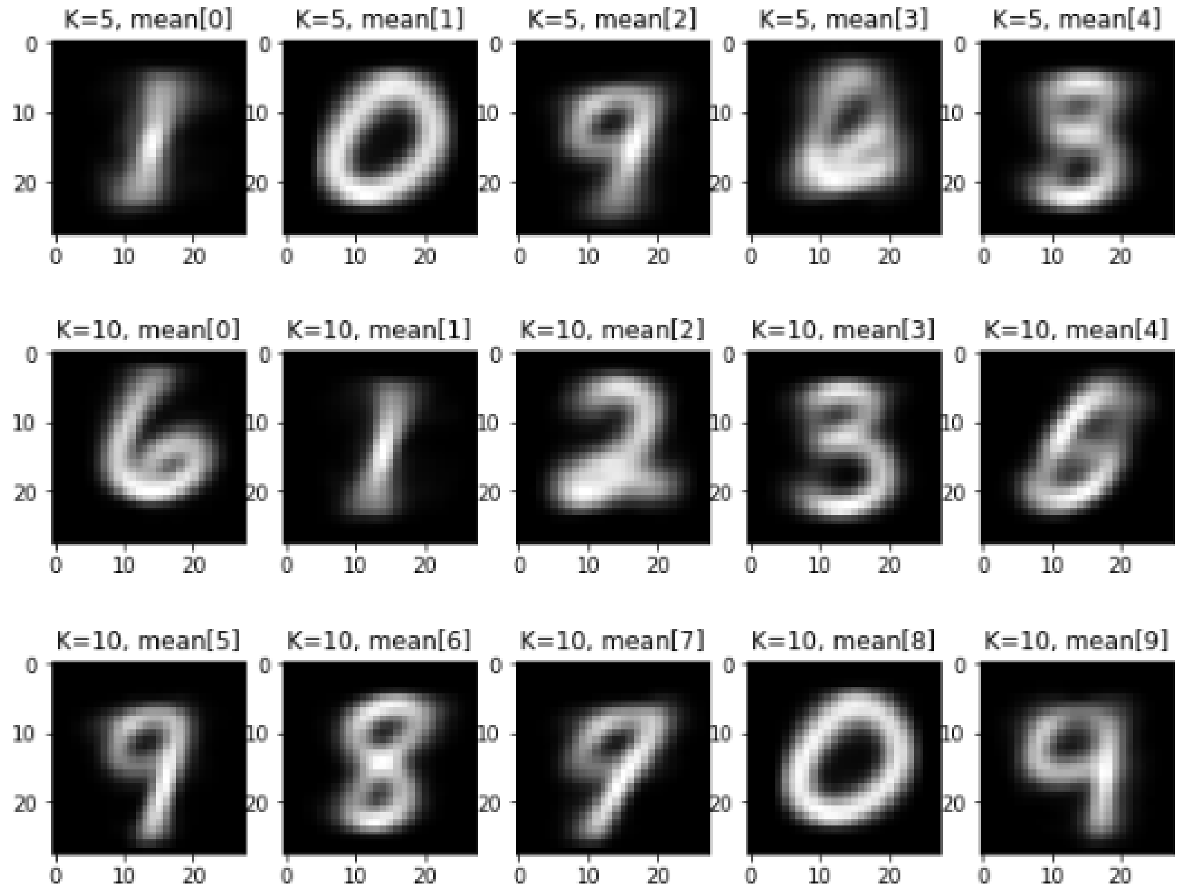
Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 2

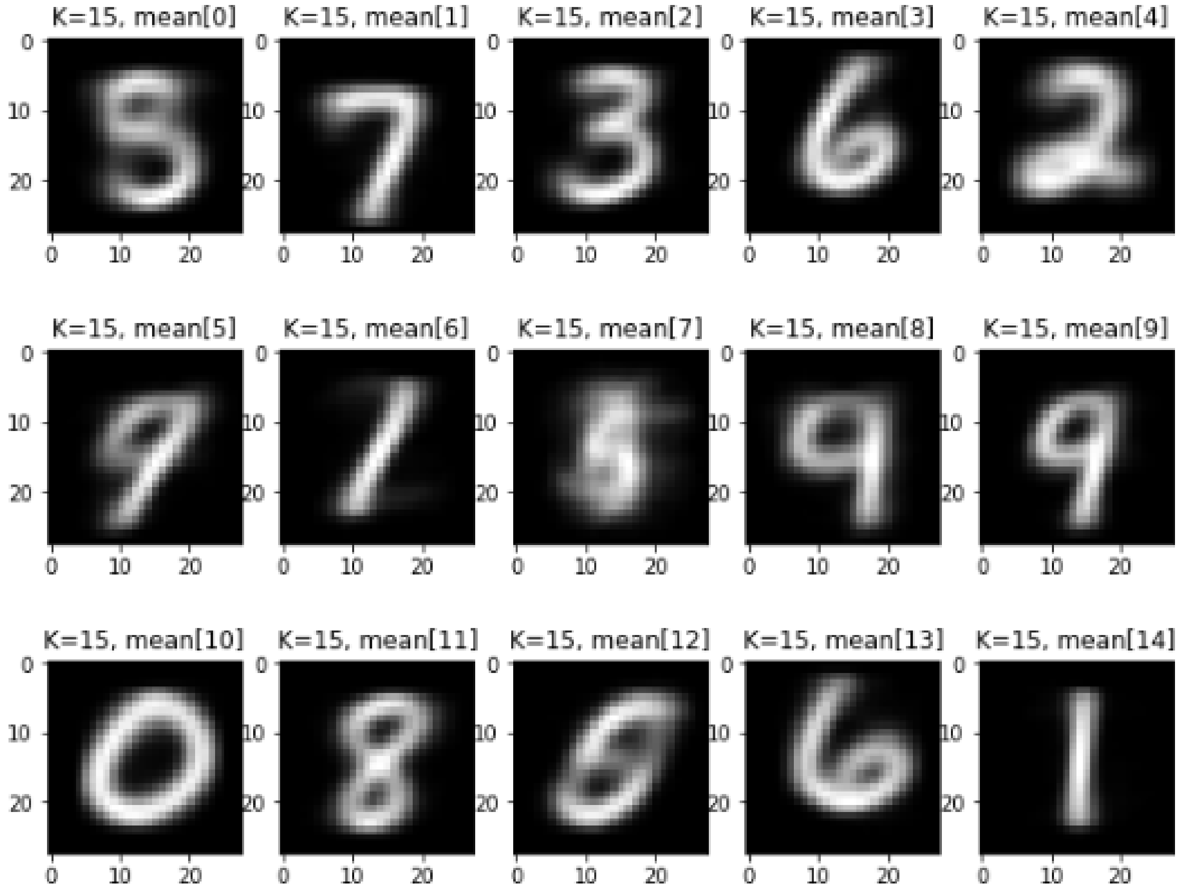
QUESTION 5

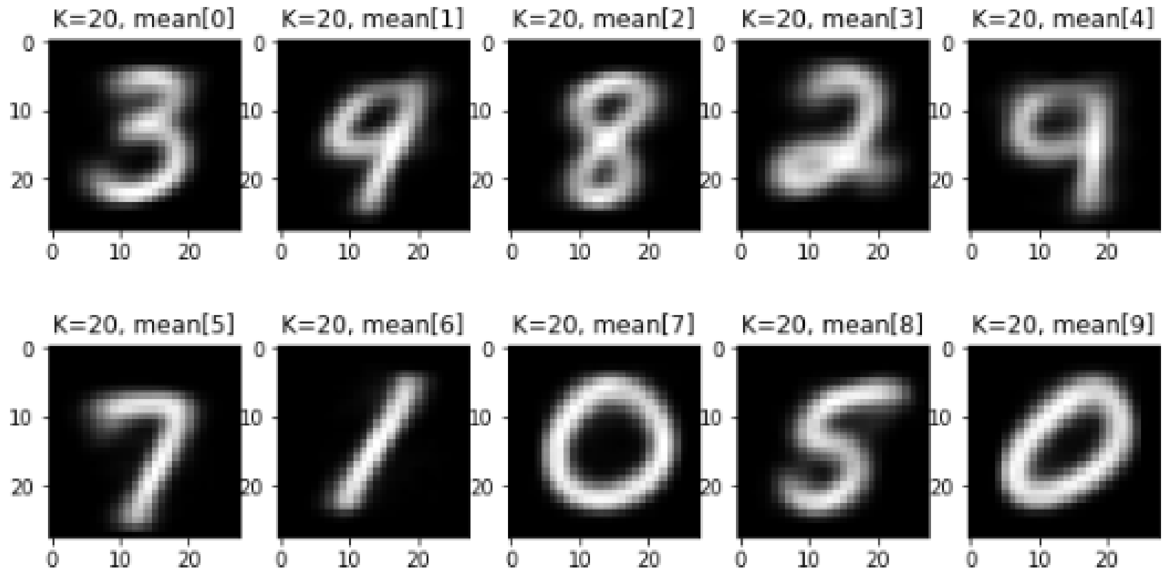
Student Name: Niharika Ahuja
Roll Number: 18111045
Date: March 13, 2019

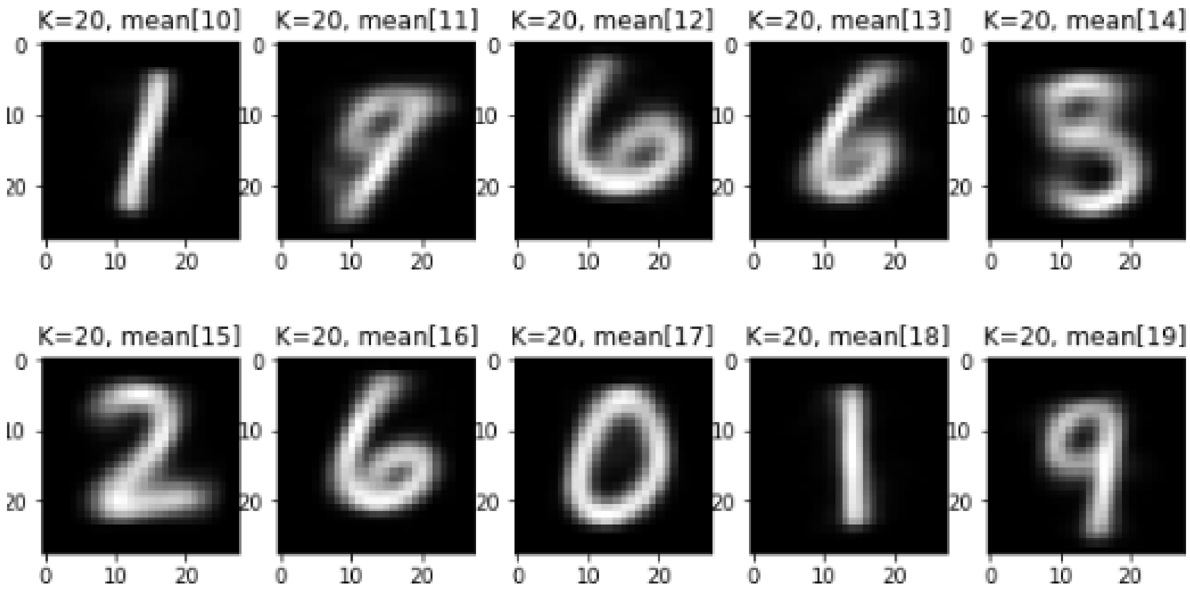
My solution to problem 5

Implementing EM algorithm for GMM using the usual update equations and the ones provided in the question. Plots for cluster means are as follows:

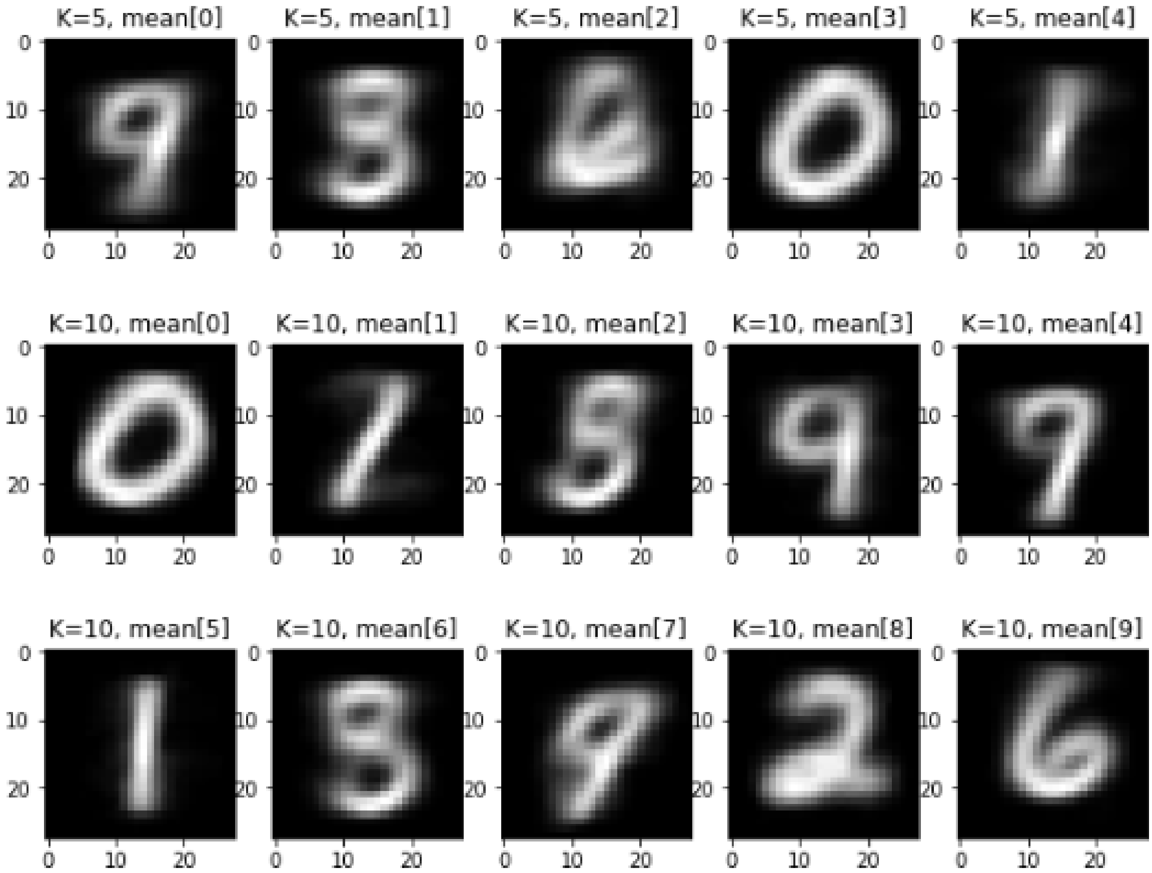


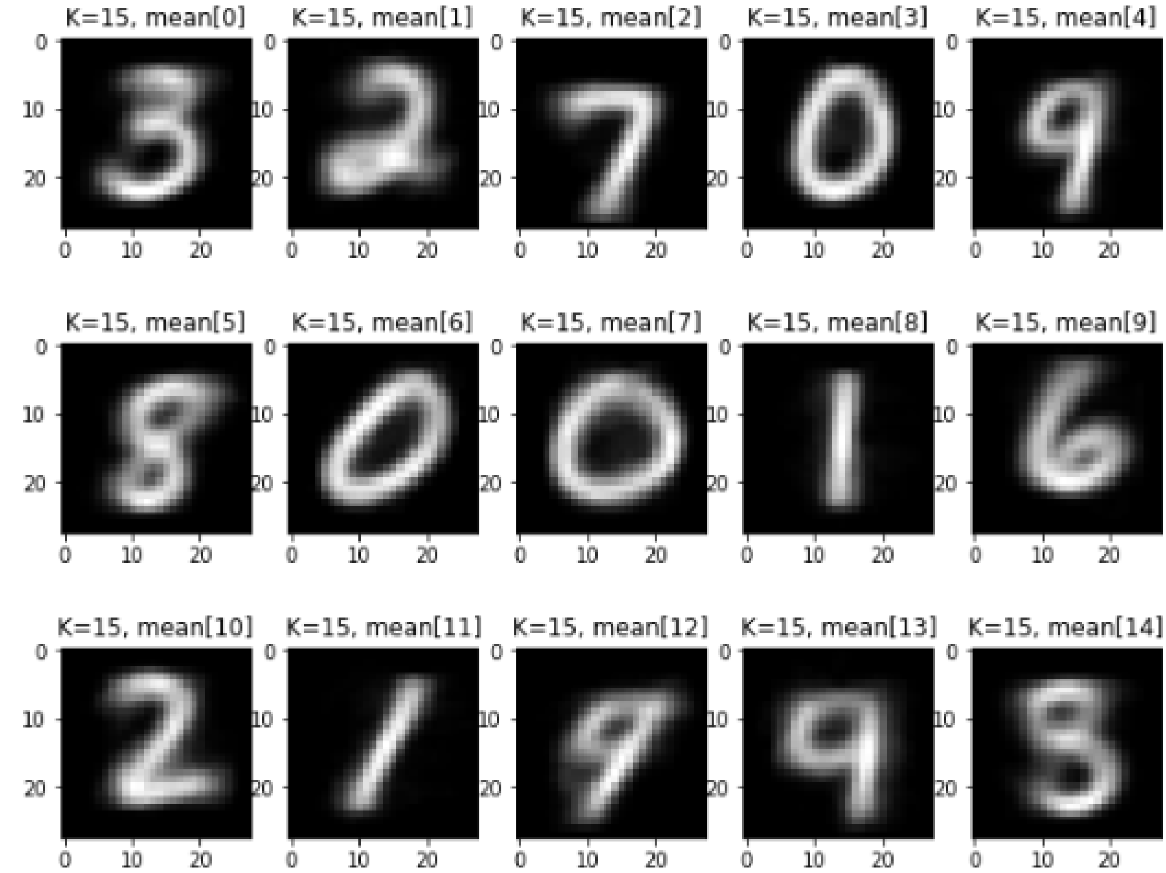

K=15, mean[0]
K=15, mean[1]
K=15, mean[2]
K=15, mean[3]
K=15, mean[4]
K=15, mean[5]
K=15, mean[6]
K=15, mean[7]
K=15, mean[8]
K=15, mean[9]
K=15, mean[10]
K=15, mean[11]
K=15, mean[12]
K=15, mean[13]
K=15, mean[14]

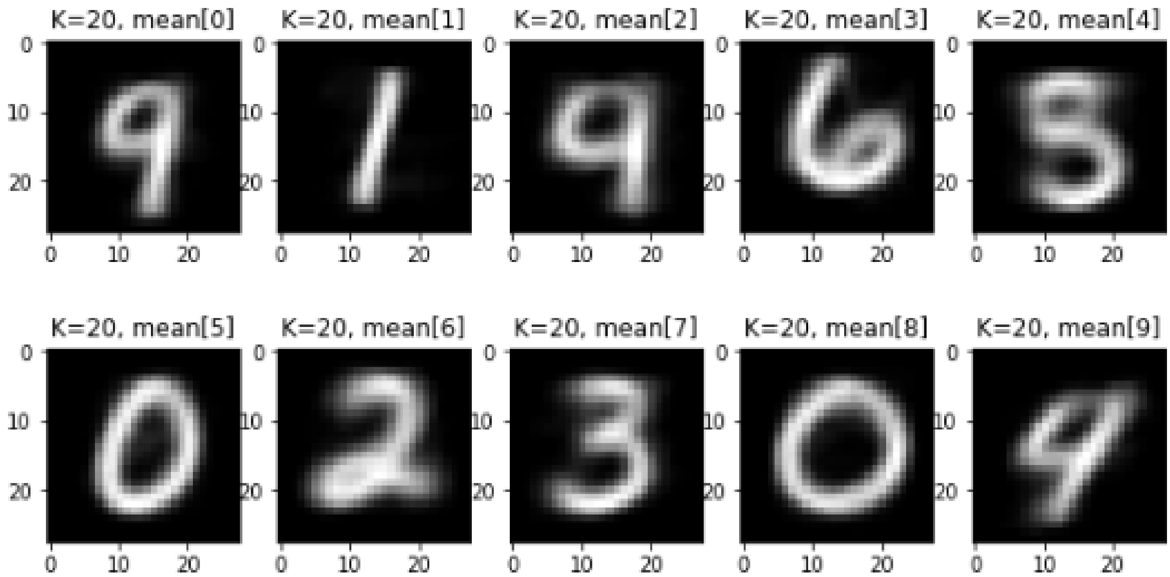

K=20, mean[0]
K=20, mean[1]
K=20, mean[2]
K=20, mean[3]
K=20, mean[4]
K=20, mean[5]
K=20, mean[6]
K=20, mean[7]
K=20, mean[8]
K=20, mean[9]

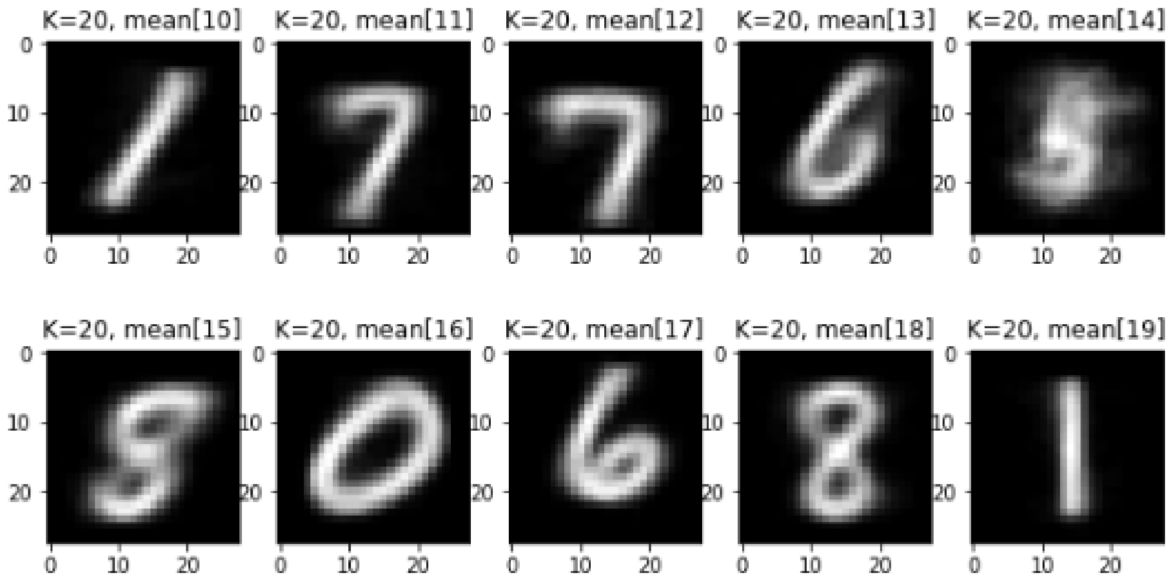


For stepwise or online EM, mini batch size is 100. The plots for cluster means are as follows:




K=15, mean[0]
K=15, mean[1]
K=15, mean[2]
K=15, mean[3]
K=15, mean[4]
K=15, mean[5]
K=15, mean[6]
K=15, mean[7]
K=15, mean[8]
K=15, mean[9]
K=15, mean[10]
K=15, mean[11]
K=15, mean[12]
K=15, mean[13]
K=15, mean[14]


K=20, mean[0]
K=20, mean[1]
K=20, mean[2]
K=20, mean[3]
K=20, mean[4]
K=20, mean[5]
K=20, mean[6]
K=20, mean[7]
K=20, mean[8]
K=20, mean[9]


K=20, mean[10]
K=20, mean[11]
K=20, mean[12]
K=20, mean[13]
K=20, mean[14]
K=20, mean[15]
K=20, mean[16]
K=20, mean[17]
K=20, mean[18]
K=20, mean[19]