CS 698: Assignment 4

Ronghao Yang
ID: 20511820
Session: 4:00pm - 5:20pm

November 16, 2017

1 Question 1

1.1 Question 1a

When S_k's are constrained to be diagonal, we can write each S_k in the vector form.

$$S_k = [s_{k1}, s_{k2}, s_{k3}, ..., s_{kd}], \text{ where } s_{ki} \text{ is } i\text{th diagonal entry in } S_k$$

Therefore, when computing the responsibilities, we can write it in a simpler form, since when S_k is diagonal, $S_k^{-1} = \frac{1}{S_k}$, by doing so, the new method is less expensive than computing the inverse. Moreover, the determinant of a diagonal matrix is just the product of its diagonal entries, $\mid S_k \mid = prod(S_k)$. Therefore, computing the responsibility can be simplified to the following:

$$r_{ik} = \pi_k prod(S_k)^{-1/2} exp(-\frac{1}{2}(x_i - \mu_k)^T \frac{1}{S_k}(x_i - \mu_k))$$

In the first step, when updating the covariance matrix. The steps were derived by taking derivative of the cost function and take the derivative of S_k to 0. The cost function is the following:

$$min_\theta \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik}[-log\pi_k + \frac{1}{2}log \mid S_k \mid + \frac{1}{2}(x_i - \mu_k)^T S_k^{-1}(x_i - \mu_k)]$$

Since all S_k's are constrained to be diagonal, as has been stated before, they can be written in the vector form, where s_{ki} is ith diagonal entry in S_k, instead of taking the derivative of S_k, here we take the derivatives of s_{ki}'s, by doing so, we maintain the constraint of the covariance matrices being diagonal.

$$\frac{\partial}{\partial s_{kj}} = \sum_{i=1}^{n} r_{ik}[\frac{1}{2}\frac{\frac{|S_k|}{s_{kj}}}{|S_k|} + \frac{1}{2}(x_{ij} - \mu_{kj})^2 \frac{-1}{(s_{kj})^2}] = 0$$

$$\frac{\partial}{\partial s_{kj}} = \sum_{i=1}^{n} r_{ik}[\frac{1}{2}\frac{1}{s_{kj}} - \frac{1}{2}\frac{(x_{ij} - \mu_{kj})^2}{(s_{kj})^2}] = 0$$

$$\sum_{i=1}^{n} r_{ik}\frac{1}{2}\frac{1}{s_{kj}} = \sum_{i=1}^{n} r_{ik}\frac{1}{2}\frac{(x_{ij}-\mu_{kj})^2}{(s_{kj})^2}$$

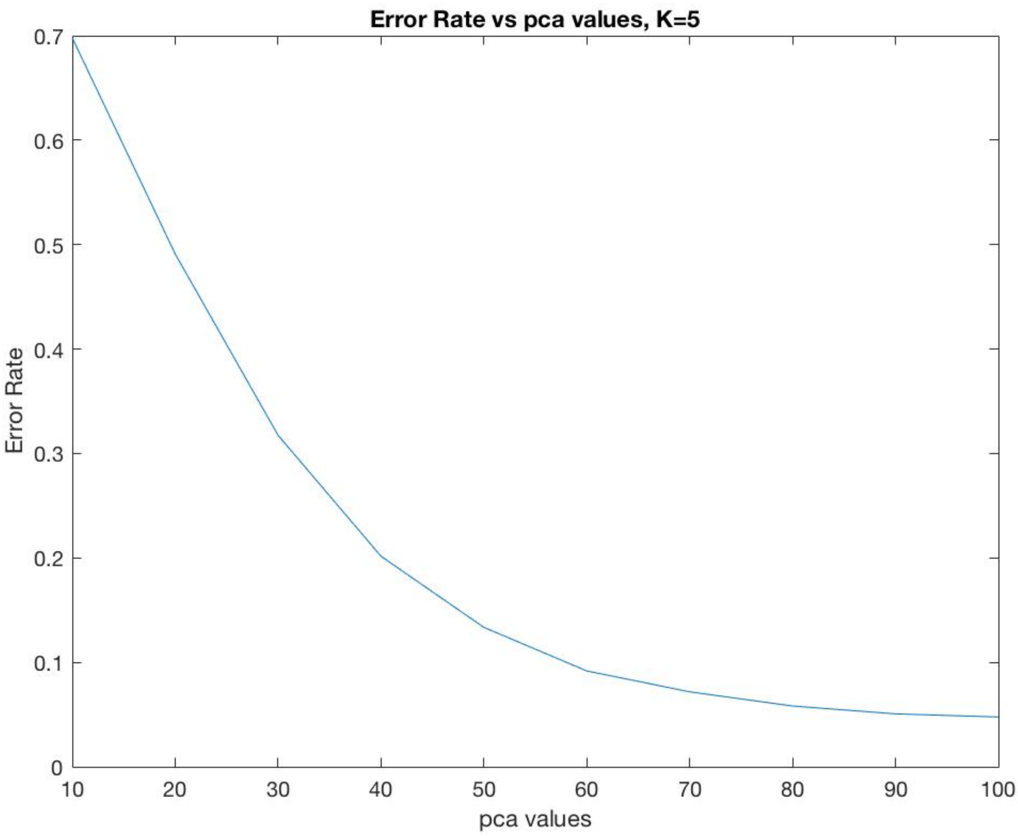
$$\sum_{i=1}^{n} r_{ik}s_{kj} = \sum_{i=1}^{n} r_{ik}(x_{ij}-\mu_{kj})^2$$

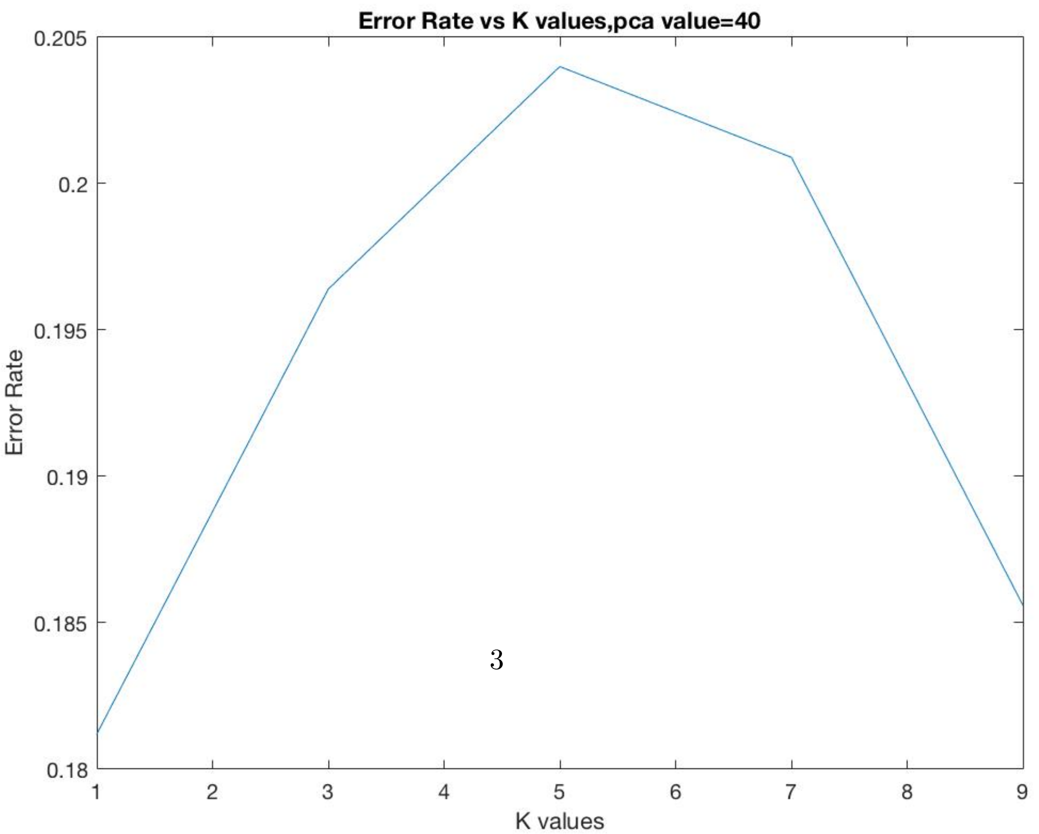
$$\text{So, } s_{kj} = \frac{\sum_{i=1}^{n} r_{ik}(x_{ij}-\mu_{kj})^2}{\sum_{i=1}^{n} r_{ik}}$$

$$s_j = \frac{\sum_{i=1}^{n} r_{ik}(x_{ij}-\mu_{kj})^2}{\sum_{i=1}^{n} r_{ik}} = \frac{\sum_{i=1}^{n} r_{ik}x_{ij}^2}{\sum_{i=1}^{n} r_{ik}} - \mu_{kj}^2$$

For space complexity of this algorithm, we need to store matrix r and matrix S at each iteration, it takes $O(nk + dk)$ space. For time complexity of this algorithm, computing the covariance matrix costs $O(ndK)$, overall, the time complexity is $O(ndK)$.

1.2 Question 1b





When I was tuning my parameters (K and pca_value), when $K = 5$, $pca\_value = 100$, the error rate has reached 4.67%.

2 Question 2

2.1 Question 2a

To predict $Z_{t_{n+1}} \mid Z_{t_1}, Z_{t_2}, ..., Z_{t_{n-1}}, Z_{t_n}$, given the fact that

For a joint Gaussian Distribution $\mathcal{N}(\mu, K)$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$$

Where μ_1 is the mean vector of $Z_{t_1}, Z_{t_2}....Z_{t_n}$, μ_2 is the mean vector of $Z_{t_{n+1}}$. And K_{11} is the covariance matrix of $Z_{t_1}...Z_{t_n}$, K_{12} is the covariance matrix of $Z_{t_1}...Z_{t_n}, Z_{t_{n+1}}$, K_{22} is the variance matrix of $Z_{t_{n+1}}$.

Then the conditional distribution of $Z_{t_{n+1}}$ is

$$\mathcal{N}(\mu', K')$$

Where

$$\mu' = \mu_2 + K_{21}K_{11}^{-1}(Z - \mu_1), K' = K_{22} - K_{21}K_{11}^{-1}K_{12},\ Z = (Z_{t_1}...Z_{t_n})$$

2.2 Question 2b

Question 2.2. $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \right)$

where $X_1 = (Z_{t_1}, Z_{t_2} \cdots Z_{t_n})'$ $X_2 = (Z_{t_{n+1}}, Z_{t_{n+2}}, \cdots Z_{t_{n+p}})'$

$\therefore X_2 \mid X_1 \sim N(U, K)$

$U = U_2 + K_{21} K_{11}^{-1} (X_1 - U_1)$

$K = K_{22} - K_{21} K_{11}^{-1} K_{12}$

Let $Z = (Z_{t_{n+2}}, \cdots Z_{t_{n+p}})$

Then $\begin{pmatrix} Z_{t_{n+1}} \\ Z \end{pmatrix} \sim N(U, K)$, where U, K are calculated above

Then $Z_{t_{n+1}} \sim (U', \Sigma')$

where $U' = [U_2 + K_{21} K_{11}^{-1} (X_1 - U_1)]_{11}$

$\Sigma' = [K_{22} - K_{21} K_{11}^{-1} K_{12}]_{11}$

For u', $U_{2\,11}$ is the mean of $Z_{t_{n+1}}$

$[K_{21} K_{11}^{-1} (X_1 - U_1)]_{11}$, we take the first row of K_{21},

which equals $[K_{21}[1,:] K_{11}^{-1} (X_1 - U_1)]_{11} = K_{21}' K_{11}^{-1} (Z' - U_1')$

Therefore $u' = U_2' + K_{21}' K_{11}^{-1} (Z' - U_1')$, where U_2 is mean of $Z_{t_{n+1}}$

K_{21}', is covariance matrix of $Z_{t_{n+1}}, Z_{t_1}, Z_{t_2} \cdots Z_{t_n}$,

K_1' is covariance matrix of $Z_{t_1} \cdots Z_{t_n}$.

$Z' = (Z_{t_1} \cdots Z_{t_n})$, U_1' is mean of $(Z_{t_1} \cdots Z_{t_n})$

By similar analysis, $\Sigma' = K_{22}' - K_{21}' K_{11}'^{-1} K_{12}'$

where K_{22}' is covariance matrix of $Z_{t_{n+1}}$

K_{21}' is covariance matrix of $Z_{t_{n+1}}, Z_{t_1} \cdots Z_{t_n}$

K_{11} is covariance matrix of $Z_{t_1} \cdots Z_{t_n}$

K_{12} is covariance matrix of $Z_{t_1} \cdots Z_{t_n}, Z_{t_{n+1}}$

Therefore, both 2a and 2b give the same results

3 Question 3

3.1 Question 3a

Question 3.1.

$$\tilde{x}_i = f(x_i, \epsilon_i) = x_i + \epsilon_i$$

Then $\min\limits_{w\in\mathbb{R}^d} \sum_{i=1}^{n} E[(y_i - w^T\tilde{x}_i)^2]$

$$(y_i - w^T\tilde{x}_i)^2 = [(y_i - w^T x_i) - w^T\epsilon_i]^2 \quad = (y_i - w^T x_i)^2 + (w^T\epsilon_i)^2 - 2(y_i - w^T x_i)(w^T\epsilon_i)$$

For $w^T\epsilon_i = \sum_{j}^{d} w_j \epsilon_{ij}$ Then $(w^T\epsilon_i)^2 = (\sum_{j=1}^{d} w_j \epsilon_{ij})^2$

$$= \sum_{j=1}^{d} w_j^2 \epsilon_{ij}^2 + \sum_{m=1}^{d}\sum_{n=1}^{d} w_m w_n \epsilon_{im}\epsilon_{in}$$

$$\therefore E[(w^T\epsilon_i)^2] = \sum_{j=1}^{d} w_j^2 E(\epsilon_{ij})^2 + \sum_{m=1}^{d}\sum_{n=1}^{d} w_m w_n E(\epsilon_{im}\epsilon_{in})$$

Since $Var(x) = E[x^2] - E[x]^2$

$\therefore E[x^2] = Var(x) + E[x]^2$

$\therefore E(\epsilon_{ij})^2 = Var(\epsilon_{ij}) + E(\epsilon_{ij})^2 = \lambda$

$$\therefore E[(w^T\epsilon_i)^2] = \sum_{j=1}^{d} w_j^2 \cdot \lambda$$

$$\therefore \min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} E[(y_i - w^T\tilde{x}_i)^2] = \min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda\|w\|^2$$

3.2 Question 3b

Question 3.2 $\hat{x}_i = f(x_i, \epsilon_i) = x_i \odot \epsilon_i$

$$(y_i - w^T\hat{x}_i)^2 = [y_i - \sum_j w_j x_{ij} \epsilon_{ij}]^2$$

$$= y_i^2 - 2y_i \sum_j w_j x_{ij} \epsilon_{ij} + (\sum_j w_j x_{ij} \epsilon_{ij})^2$$

$$(\sum_j w_j x_{ij} \epsilon_{ij})^2 = \sum_j w_j x_{ij} \epsilon_{ij} \sum_k w_k x_{ik} \epsilon_{ik}$$

$$E(y_i - w^T\hat{x}_i)^2 = E[y_i^2 - 2y_i \sum_j w_j x_{ij} \epsilon_{ij} + (\sum_j w_j x_{ij} \epsilon_{ij})^2]$$

$$\cancel{= E(y_i)^2 - 2y}$$

$$= y_i^2 - 2y_i \sum_j w_j x_{ij} E(\epsilon_{ij}) + 2\sum_j \sum_k w_j w_k x_{ij} x_{ik} E(\epsilon_{ij} \epsilon_{ik}) + \sum_j w_j^2 x_{ij}^2 E(\epsilon_{ij})^2$$

$\because$ $p\epsilon_{ij} \sim$ Bernoulli(p)

$\therefore$ $E(\epsilon_{ij}) = 1$ $\quad$ $E(\epsilon_{ij} \epsilon_{ik}) = E(\epsilon_{ij}) E(\epsilon_{ik}) = 1$

$\therefore$ $E(\epsilon_{ij})^2 = p \cdot (\frac{1}{p})^2 = \frac{1}{p}$

$$\therefore E(y_i - w^T\hat{x}_i)^2 = y_i^2 - 2y_i \sum_j w_j x_{ij} + 2\sum_j \sum_k w_j w_k x_{ij} x_{ik} E(\epsilon_{ij} \epsilon_{ik}) + \frac{1}{p} \sum_j w_j^2 x_{ij}^2$$

$$= (y_i - w^T x_i)^2 + \frac{1-p}{p} (w^T x_i)^2$$

$$\therefore \min_{w \in \mathbb{R}^d} \sum_{i=1}^n E[(y_i - w^T\hat{x}_i)^2] = \min_{w \in \mathbb{R}^d} \sum_{i=1}^n (y_i - w^T x_i)^2 + \frac{1-p}{p} (w^T x_i)^2$$