ISYE 6740, Fall 2020, Homework 3

100 points + 15 bonus points

Farshad Rafiei

1. Density estimation: Psychological experiments. (45 points)

We will use this data to study whether or not the two brain regions are likely to be independent of each other and considering different types of political view **For this question; you can use the proper package for histogram and KDE; no need to write your own.** The data set n90pol.csv contains information on 90 university students who participated in a psychological experiment designed to look for relationships between the size of different regions of the brain and political views. The variables amygdala and acc indicate the volume of two particular brain regions known to be involved in emotions and decision-making, the amygdala and the anterior cingulate cortex; more exactly, these are residuals from the predicted volume, after adjusting for height, sex, and similar body-type variables. The variable orientation gives the students' locations on a five-point scale from 1 (very conservative) to 5 (very liberal). Note that in the dataset, we only have observations for orientation from 2 to 5.

(a) (10 points) Form the 1-dimensional histogram and KDE to estimate the distributions of amygdala and acc, respectively. For this question, you can ignore the variable orientation.

Answer: Figure (1) shows the histogram for amygdala and acc using 20 bins. Figure (2) shows the KDE for those distributions. Bandwidth is set to 0.0075 for KDE plots.

(b) (10 points) Form 2-dimensional histogram for the pairs of variables (amygdala, acc). Decide on a suitable number of bins so you can see the shape of the distribution clearly. Also use kernel-density-estimation (KDE) to estimate the 2-dimensional density function of (amygdala, acc). Use a simple multi-dimensional Gaussian kernel, for

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2,$$

where x_1 and x_2 are the two dimensions respectively

$$K(x) = \frac{1}{2\pi} e^{-\frac{(x_1)^2+(x_2)^2}{2}}.$$

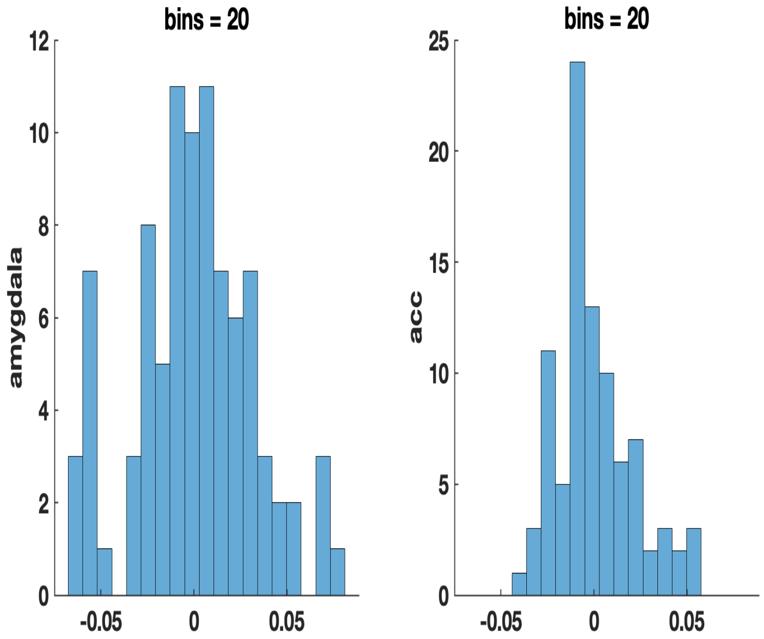

Figure 1: Histograms associated with amygdala and acc distributions.

Recall in this case, the kernel density estimator (KDE) for a density is given by

$$p(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{h} K\left(\frac{x^i - x}{h}\right),$$

where x^i are two-dimensional vectors, $h > 0$ is the kernel bandwidth. Set an appropriate h so you can see the shape of the distribution clearly. Plot the contour plot (like the ones in slides) for your estimated density. For this question, you can ignore the variable orientation.

Answer: Figure (3) shows the 2-dimensional histogram for pairs of variables (amygdala, acc). Number of bins is set to 20 for both variables. Figure (4) shows the 2-dimensional KDE for pairs of variables (amygdala, acc) as well as the contour plot. Bandwidth is set to 0.0075 for both variables.

(c) (10 points) Using (a) and (b), using KDE estimators, verify whether or not the variables amygdala and acc are independent? You can tell this by checking do we approximately have $p(\text{amygdala}, \text{acc}) = p(\text{amygdala})p(\text{acc})$? To verify this, please show three plots: the map for $p(\text{amygdala}, \text{acc})$, the map for $p(\text{amygdala})p(\text{acc})$ and the error map $|p(\text{amygdala}, \text{acc}) - p(\text{amygdala})p(\text{acc})|$. Comment on your results and whether this helps us to find out whether the two parts of brains (for emotions and decision-making) functions independently or they are related.

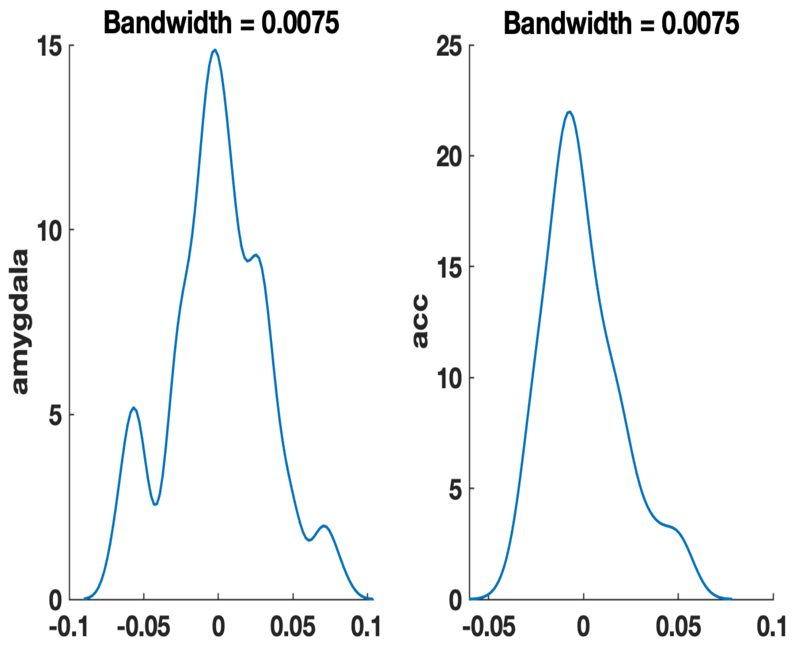

Figure 2: KDE associated with amygdala and acc distributions.

Answer: $p(\mathsf{amygdala}, \mathsf{acc})$ was showed on Figure (4). Figure (5) and Figure (6) show the distribution of $p(\mathsf{amygdala})p(\mathsf{acc})$ and $|p(\mathsf{amygdala}, \mathsf{acc}) - p(\mathsf{amygdala})p(\mathsf{acc})|$, respectively. As can be seen in Figure (6), the error map does not equal to zero in many locations. This indicates that amygdala and anterior cingulate cortex (acc) does not perform independently. In fact, this is a well-known fact that none of the brain regions work completely independent of other regions. It is assumed to be a huge network with a lot of direct and indirect connections between different nodes.

(d) (5 points) Now we will consider the variable orientation. We will estimate the conditional distribution of the volume of the amygdala, conditioning on political orientation: $p(\mathsf{amygdala}|\mathsf{orientation} = c)$, $c = 2, \ldots, 5$. Do the same for the volume of the acc: Plot $p(\mathsf{acc}|\mathsf{orientation} = c)$, $c = 2, \ldots, 5$. You will use KDE to achieve the goal. (Note that the conditional distribution can be understood as fitting a distribution for the data with the same (fixed) orientation. Thus there should be 4 one-dimensional distribution functions to show for this question.)

Answer: Figure (7) shows the amygdala distribution conditioned on political orientation of the participants. Also, Figure (8) shows the distribution of acc conditioned on political orientation. Bandwidth is set to 0.008 for both density estimates.

(e) (5 points) Again we will consider the variable orientation. We will estimate the conditional *joint* distribution of the volume of the amygdala and acc, conditioning on a function of political orientation: $p(\mathsf{amygdala}, \mathsf{acc}|\mathsf{orientation} = c)$, $c = 2, \ldots, 5$. You will

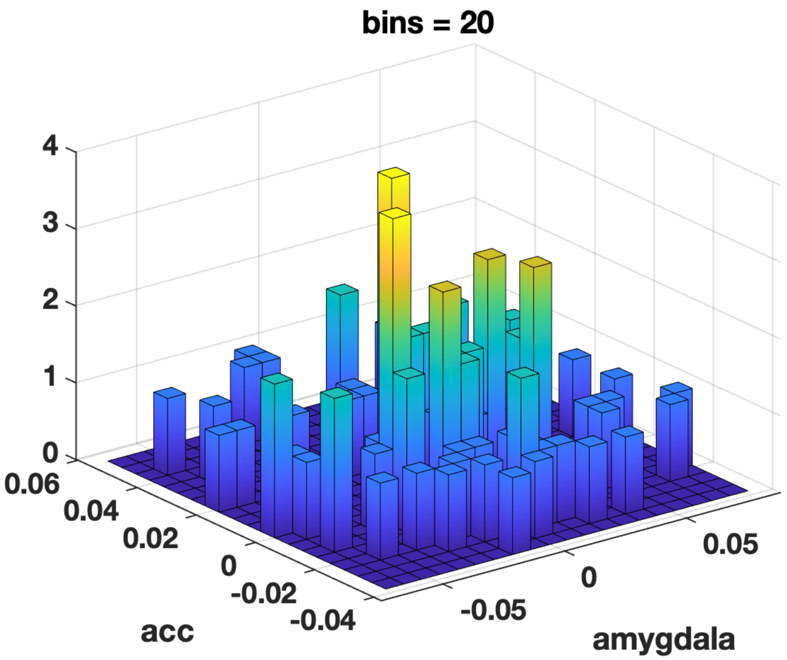

Figure 3: Histograms associated with joint distribution of (amygdala, acc).

use two-dimensional KDE to achieve the goal.

Answer: Figure (9) shows the conditional joint distribution of $p(\mathsf{amygdala}, \mathsf{acc}|\mathsf{orientation} = c)$. The joint distribution is separated for different political orientations and shown separately.

(f) (5 points) Using (d) and (e), evaluate whether or not the two variables are likely to be conditionally independent. To verify this, please show three plots: the map for

$$p(\mathsf{amygdala}, \mathsf{acc}|\mathsf{orientation} = c),$$

the map for

$$p(\mathsf{amygdala}|\mathsf{orientation} = c)p(\mathsf{acc}|\mathsf{orientation} = c)$$

and the error map

$$|p(\mathsf{amygdala}, \mathsf{acc}|\mathsf{orientation} = c) - p(\mathsf{amygdala}|\mathsf{orientation} = c)p(\mathsf{acc}|\mathsf{orientation} = c)|,$$

$c = 2, \dots, 5$. Comment on your results and whether this helps us to find out whether the two parts of brains (for emotions and decision-making) functions independently or they are related, conditionally on the political orientation (i.e., considering different types of personality).

Answer: Figure (10) shows the distribution associated with the multiplication of conditional distributions, $p(\mathsf{amygdala}|\mathsf{orientation} = c)p(\mathsf{acc}|\mathsf{orientation} = c)$, $c = 2, \dots, 5$.

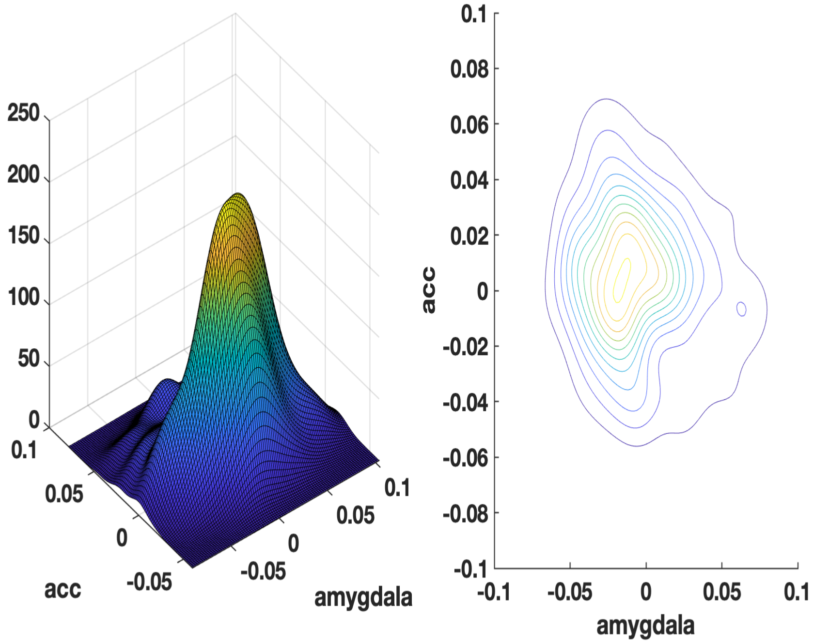

Figure 4: KDE associated with joint distribution of (amygdala, acc).

Finally, Figure (11) shows the error map. None of the errors maps show a constant zero probability map. This indicates that amygdala and anterior cingulate cortex (acc) do not work independently for any population with different political orientation.

2. Implementing EM for MNIST dataset, with PCA for dimensionality reduction. (55 points)

Implement the EM algorithm for fitting a Gaussian mixture model for the MNIST dataset. We reduce the dataset to be only two cases, of digits "2" and "6" only. Thus, you will fit GMM with $C = 2$. Use the data file data.mat or data.dat. True label of the data are also provided in label.mat and label.dat

The matrix images is of size 784-by-1990, i.e., there are totally 1990 images, and each column of the matrix corresponds to one image of size 28-by-28 pixels (the image is vectorized; the original image can be recovered by map the vector into a matrix).

First use PCA to reduce the dimensionality of the data before applying to EM. We will put all "6" and "2" digits together, to project the original data into 5-dimensional vectors. Now implement EM algorithm for the projected data (with 5-dimensions).

(a) (5 points) Select from data one raw image of "2" and "6" and visualize them, respectively.

Answer: Figure (12) shows two random instance from each group of numbers.

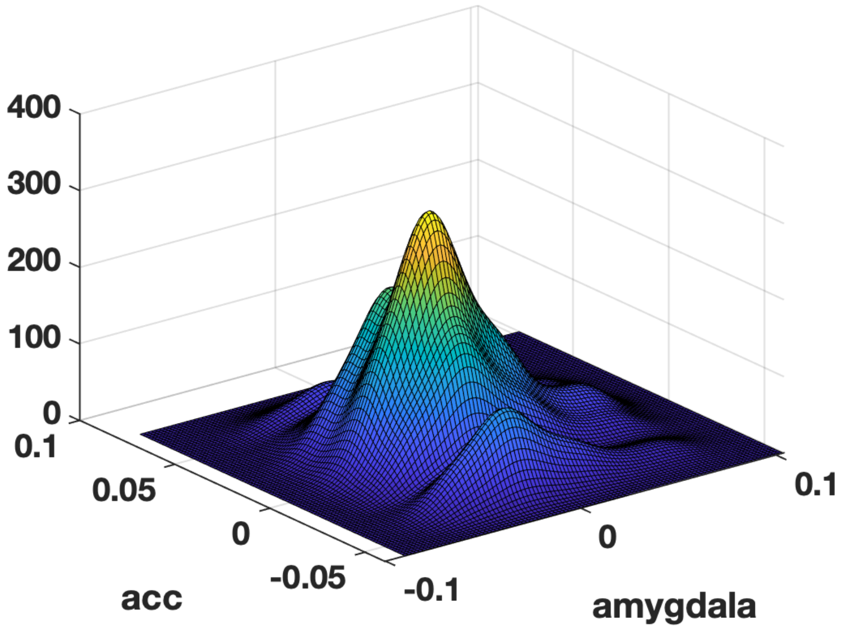

Figure 5: KDE associated with $p(\mathsf{amygdala})p(\mathsf{acc})$ distribution.

(b) (10 points) Write down detailed expression of the E-step and M-step in the EM algorithm (hint: when computing τ_k^i, you can drop the $(2\pi)^{n/2}$ factor from the numerator and denominator expression, since it will be canceled out; this can help avoid some numerical issues in computation).

Answer: Let's assume that we have m data points x^i, $i = 1, \ldots, m$. For this question, we only have two Gaussians to fit, one for 2 and one for 6. Therefore, we set $k = 2$. To perform the EM algorithm, we first need to initialize the priors (π_k, $k = 1, 2$), means (μ_k, $k = 1, 2$) and covariance matrices (Σ_k, $k = 1, 2$) associated with each Gaussian. The, we will iterate expectation and maximization steps until the algorithm converges.

Expectation step

The goal in this step is to check how likely is the data, based on the parameters we already have (π_k, μ_k, Σ_k; $k = 1, 2$). Basically, we compute the probability of a data point Under a certain distribution, which is created based on the given parameters. This can be done as following:

$$\tau_k^i = \frac{\pi_k \mathcal{N}(x^i|\mu_k, \Sigma_k)}{\sum_{j=1}^{2} \mathcal{N}(x^i|\mu_j, \Sigma_j)} \tag{1}$$

In this equation, $\mathcal{N}(x^i|\mu_k, \Sigma_k)$ shows the multivariate normal probability density function, defined as following:

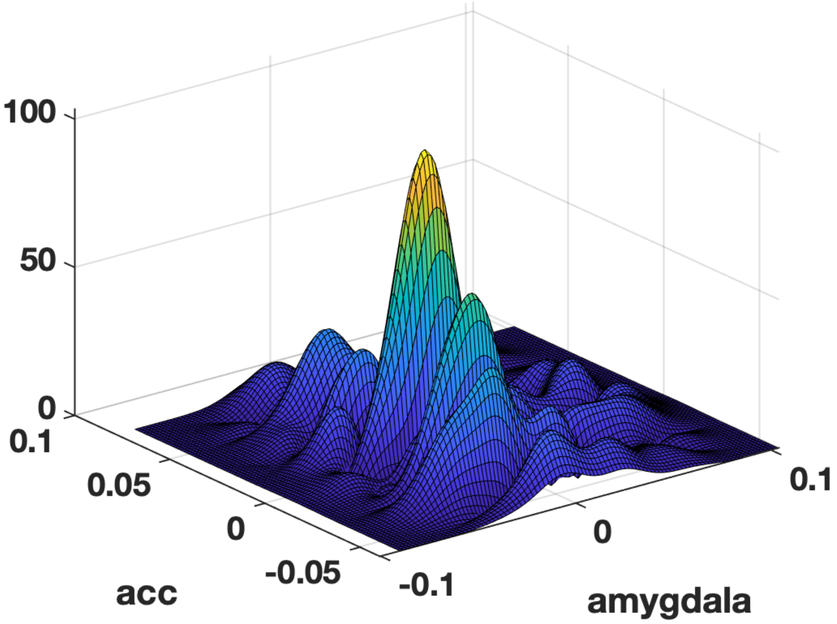

Figure 6: KDE associated with error map distribution of $|p(\mathsf{amygdala}, \mathsf{acc}) - p(\mathsf{amygdala})p(\mathsf{acc})|$.

$$\mathcal{N}(x^i|\mu_k, \Sigma_k) = \frac{1}{\sqrt{\Sigma}} exp(-\frac{1}{2}(x^i - \mu_k)\Sigma_k^{-1}(x^i - \mu_k)^T) \tag{2}$$

Equation (1) shows how τ_k becomes updated at this step, given the parameters (π_k, μ_k, Σ_k; $k = 1, 2$).

Maximization step

Now that we have new posteriors for each data point, we can update the parameters (π_k, μ_k, Σ_k; $k = 1, 2$) based on these posteriors. This can be done as following:

$$\pi_k = \frac{\sum_i \tau_k^i}{m} \tag{3}$$

$$\mu_k = \frac{\sum_i \tau_k^i x^i}{\sum_i \tau_k^i} \tag{4}$$

$$\mu_k = \frac{\sum_i \tau_k^i (x^i - \mu_k)(x^i - \mu_k)^T}{\sum_i \tau_k^i} \tag{5}$$

We updated the mean and covariance of the Gaussians using maximum likelihood estimates.

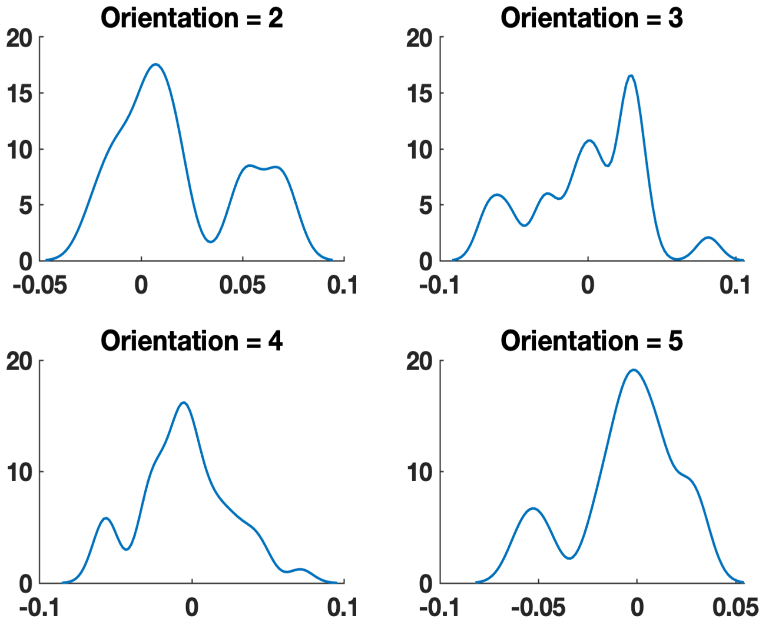

Figure 7: KDE associated with $p(\mathsf{amygdala}|\mathsf{orientation} = c)$, $c = 2, \ldots, 5$ distribution.

(c) (15 points) Implement EM algorithm yourself. Use the following initialization

- initialization for mean: random Gaussian vector with zero mean
- initialization for covariance: generate two Gaussian random matrix of size n-by-n: S_1 and S_2, and initialize the covariance matrix for the two components are $\Sigma_1 = S_1 S_1^T + I_n$, and $\Sigma_2 = S_2 S_2^T + I_n$, where I_n is an identity matrix of size n-by-n.

Plot the log-likelihood function versus the number of iterations to show your algorithm is converging.

Answer: Implementation of algorithm starts with initialization of parameters (π_k, μ_k, Σ_k; $k = 1, 2$) as the problem states. We then iterate through expectation and maximization steps, until the convergence achieved. The algorithm stops in 100 iterations or when the likelihood of the data, does not enhance during the last 5 iterations, whichever comes first. Multiple times running of algorithm indicates that it converges in way sooner than 100 iterations. Figure (13) shows the log likelihood function versus the number of iterations in one time running of the algorithm.

(d) (15 points points) Report, the fitted GMM model when EM has terminated in your algorithms as follows. Make sure to report the weights for each component, and the mean of each component (you can reformat the vector to make them into 28-by-28 matrices and show images). Ideally, you should be able to see these means corresponds

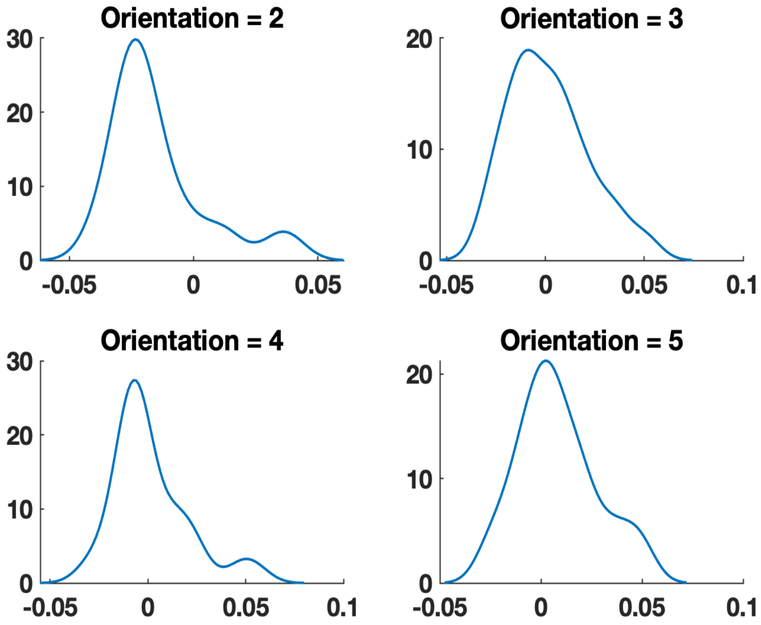

Figure 8: KDE associated with $p(\mathsf{acc}|\mathsf{orientation} = c)$, $c = 2, \ldots, 5$ distribution.

to "average" images. Report the two 784-by-784 covariance matrices by visualize their intensities.

Answer: The final weights achieved for each component are $\pi_1 = 0.4931$ (corresponds to "2"s) and $\pi_2 = 0.5069$ (corresponds to "6"s). The mean and covariance matrices are shown in Figure (14) and Figure (15), respectively.

(e) (10 points) Use the τ_k^i to infer the labels of the images, and compare with the true labels. Report the mis-classification rate for digits "2" and "6" respectively. Perform K-means clustering with $K = 2$ (you may call a package or use the code from your previous homework). Find out the mis-classification rate for digits "2" and "6" respectively, and compare with GMM. Which one achieves the better performance?

Answer: There are total number of 1032 images with "2" labels. EM algorithm was able to cluster 972 of these correctly. 60 of the instances was incorporated to the competing cluster with probability higher than 50%. Therefore, mis-classification rate for "2" images is equal to 0.0581 (or 5.81%). Also, 951 out of 958 images which were labeled as "6" were clustered correctly. Mis-classification rate for this component is 0.0073 (or 0.73%). Confusion matrix is shown in Figure (16).

For k-means algorithm, mis-classification rate for "2" and "6" components are 0.0543 (or 5.43%) and 0.0710 (or 7.1%), respectively. Confusion matrix is shown in Figure (17). k-means algorithm performs slightly better in clustering images with "2" labels. However, the overall performance of EM algorithm is better than k-means algorithm.

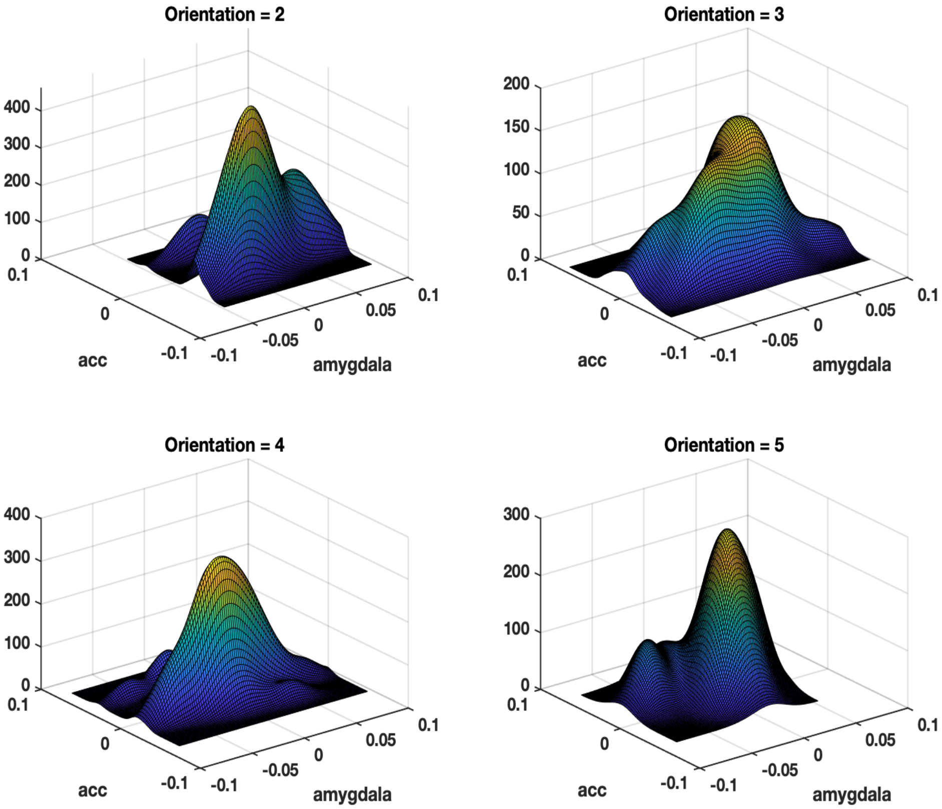

Figure 9: KDE associated with conditional joint distribution of $p(\mathsf{amygdala}, \mathsf{acc}|\mathsf{orientation} = c)$, $c = 2, \ldots, 5$.

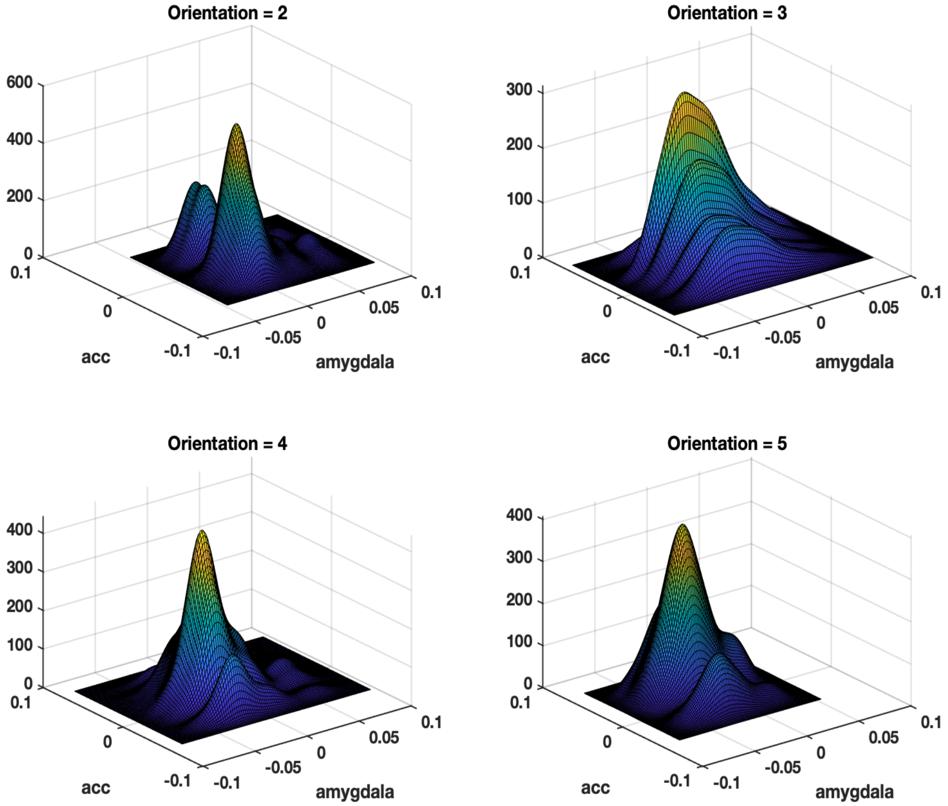

Figure 10: KDE associated with $p(\mathsf{amygdala}|\mathsf{orientation} = c)p(\mathsf{acc}|\mathsf{orientation} = c)$, $c = 2, \ldots, 5$ distribution.

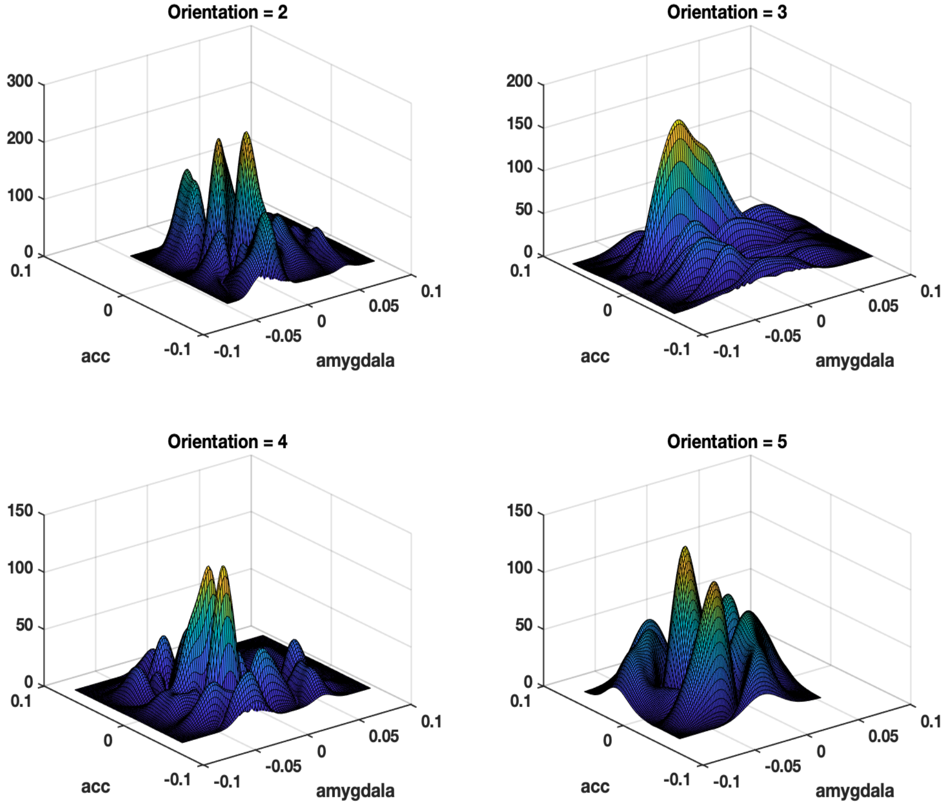

Figure 11: KDE associated with error map, $|p(\mathsf{amygdala}, \mathsf{acc}|\mathsf{orientation} = c) - p(\mathsf{amygdala}|\mathsf{orientation} = c)p(\mathsf{acc}|\mathsf{orientation} = c)|$, $c = 2, \ldots, 5$.

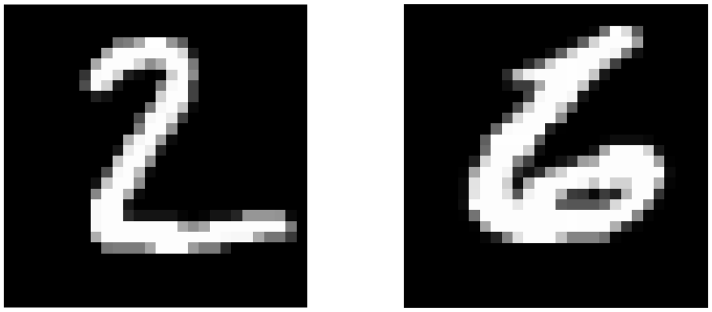

Figure 12: Representation of raw data in MNIST dataset.

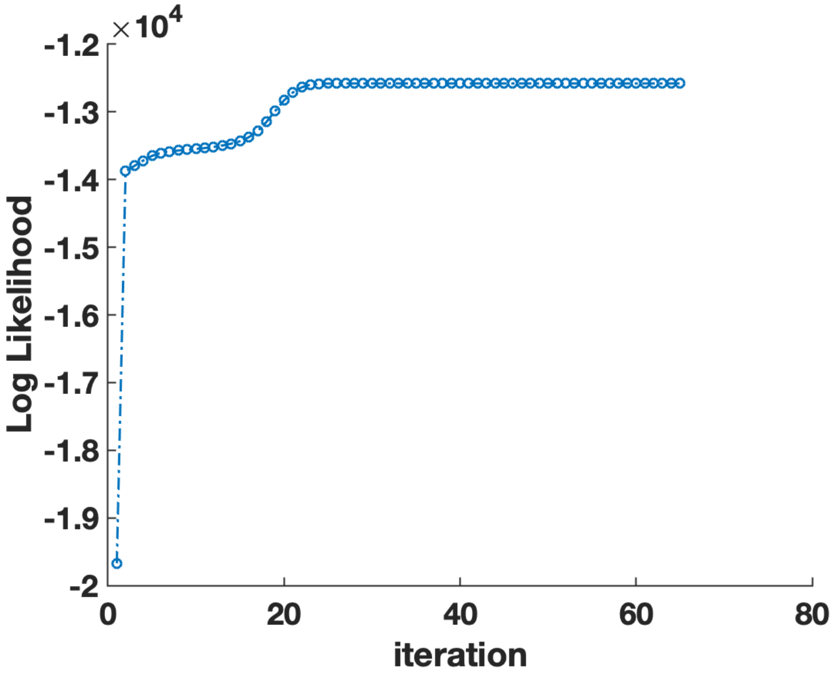

Figure 13: Log likelihood of the MNIST data using EM algorithm.

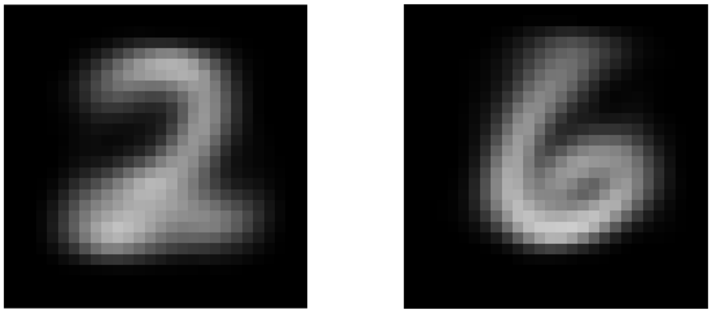

Figure 14: Visualization of $\mu_k; k = 1, 2$ resulted from EM algorithm associated with each component

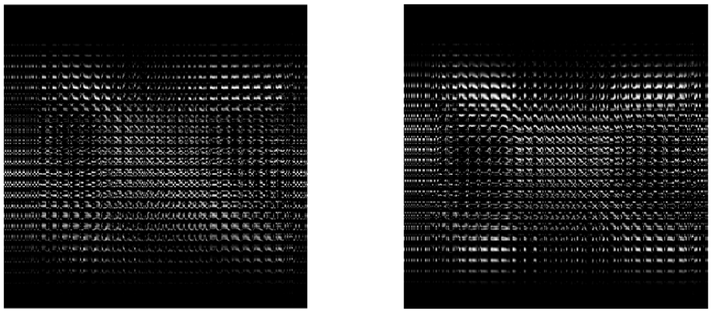

Figure 15: Visualization of $\Sigma_k; k = 1, 2$ resulted from EM algorithm associated with each component

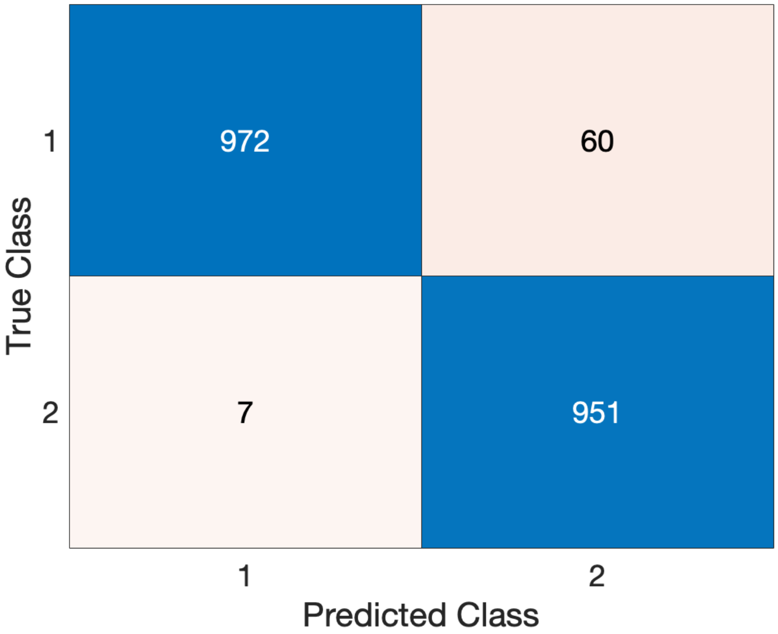

Figure 16: Confusion matrix for MNIST dataset when EM algorithm was used.

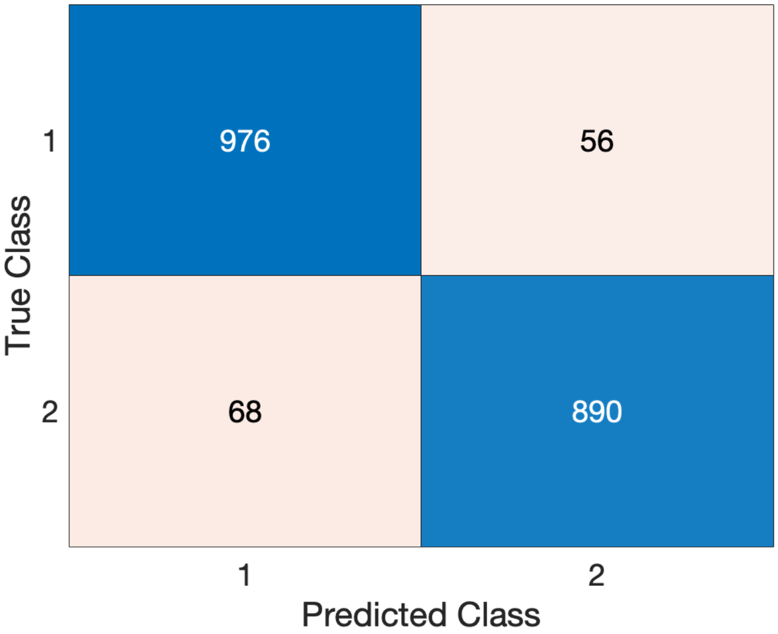

Figure 17: Confusion matrix for MNIST dataset when k-means algorithm was used.