Homework 3

1. Conceptual Questions

1. Based on the outline given in the lecture, show mathemtically that the maximum likelihood estimate (MLE) for Gaussian mean and variance parameters are given by
$\hat{\mu} = \frac{1}{m}\Sigma_{i=1}^{m} x^i, \quad \hat{\sigma}^2 = \frac{1}{m}\Sigma_{i=1}^{m}(x^i - \hat{\mu})^2$

Note: For this derivation, you will also need to show that these estimates for μ and σ are maximum.

First, start with the Gausian distribution:
$p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$

Then, take the log to get the log likelihood:
$l(\mu, \sigma, D) = \frac{-m}{2} log(2\pi) - \frac{m}{2} log(\sigma^2) - \Sigma_{i=1}^{m} \frac{(x^i-\mu)^2}{2\sigma^2}$

Maximize $l(\mu, \sigma, D)$ with respect to μ and σ^2 and set to 0 to get the estimates:
$\frac{\partial l}{\partial \mu} = \frac{\partial}{\partial \mu} \frac{-1}{2\sigma^2} \Sigma_{i=1}^{m}(x^i - \mu)^2 = \frac{-1}{2\sigma^2} \Sigma_{i=1}^{m} 2(x^i - \mu)(-1)$
$0 = \frac{1}{\sigma^2}(\Sigma_{i=1}^{m} x^i - \Sigma_{i=1}^{m} \mu)$
$\Sigma_{i=1}^{m} \mu = \Sigma_{i=1}^{m} x^i$
$\hat{\mu} = \frac{1}{m}\Sigma_{i=1}^{m} x^i$

$\frac{\partial l}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}[\frac{-m}{2} log(\sigma^2) - \frac{\Sigma_{i=1}^{m}(x^i-\mu)^2}{2\sigma^2}] = 0$
Skipping a few algebraic steps, the estimate for $\hat{\sigma}^2 = \frac{1}{m}\Sigma_{i=1}^{m}(x^i - \mu)^2$

To determine if these are maximum values, we have to take the second derivatives and show that it is negative in order for the estimates to be maximums.
$\frac{\partial^2 l}{\partial \mu^2} = \frac{1}{\sigma^2}(\Sigma_{i=1}^{m} x^i - \Sigma_{i=1}^{m} \mu) = \frac{-m\mu}{\sigma^2} \longrightarrow$ this is always negative so $\hat{\mu}$ is a maximum

$\frac{\partial^2 l}{\partial (\sigma^2)^2} = \frac{-m}{2\sigma^2} \longrightarrow$ this is also always negative so $\hat{\sigma}^2$ is a maximum.

2. Please compare the pros and cons of KDE as opposed to histograms, and give at least one advantage and disadvantage to each.

Histograms

- Advantage: Easy to interpret and understand, especially when presenting to non-technical stakeholders
- Disadvantage: Output depends on where you put the bins so the estimates can become noisy and histogram could look sparse and difficult to interpret

KDE

- Advantage: Smaller errors when comparing estimated density function vs true density function
- Disadvantage: Computationally heavy as you need to evaluate m functions

3. For the EM algorithm for GMM, please show how to use Bayes rule to drive τ_k^i in closed-form expression.

Bayes Rule

$P(z|k) = \frac{P(x|z)P(z)}{P(x)}$

Following Bayes Rule:

$\tau_k^i := p(z^i = k|x^i, \theta^i) = \frac{p(x^i|z^i=k)p(z^i=k)}{\Sigma_{k'=1...K}p(z^i=k',x^i)}$

$p(x|z) = p(x^i|z^i = k) = N(x|\mu_k, \Sigma_k)$

$p(z) = p(z^i = k) = \pi_k$

$p(x) = \Sigma_{k'=1...K}p(z^i = k', x^i) = \Sigma_{k'=1...K}\pi_{k'} N(x|\mu_k', \Sigma_k')$

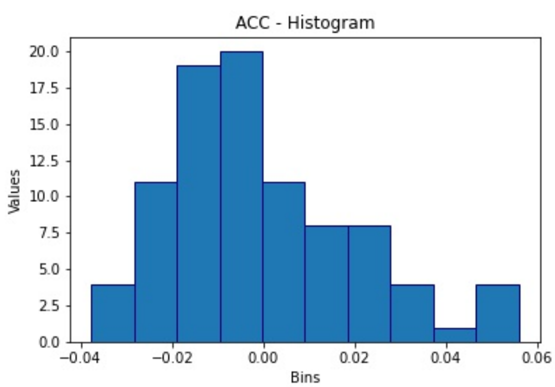
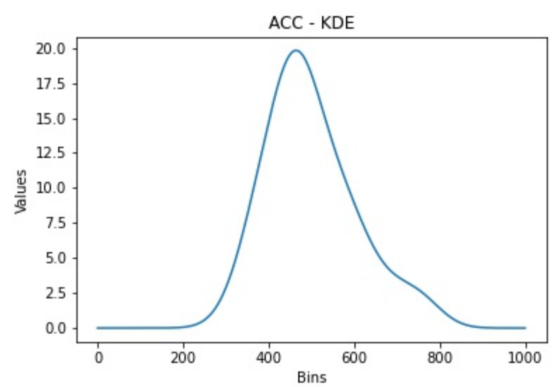
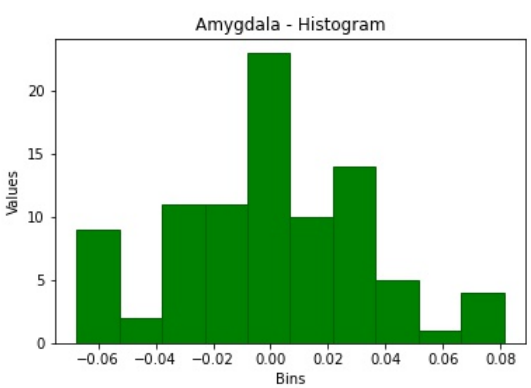
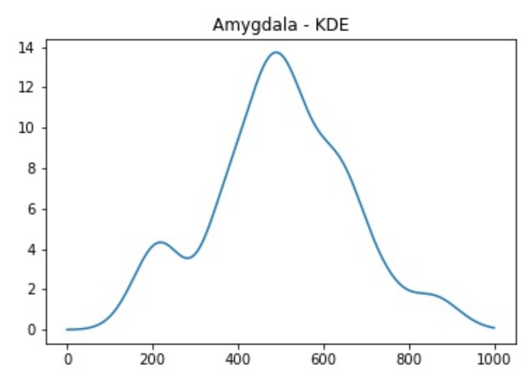
Thus, $\tau_k^i = \frac{\pi_k N(x|\mu_k,\Sigma_k)}{\Sigma_{k'=1...K}\pi_{k'} N(x|\mu_k',\Sigma_k')}$

2. Density Estimation: Psychological Experiments

Part A

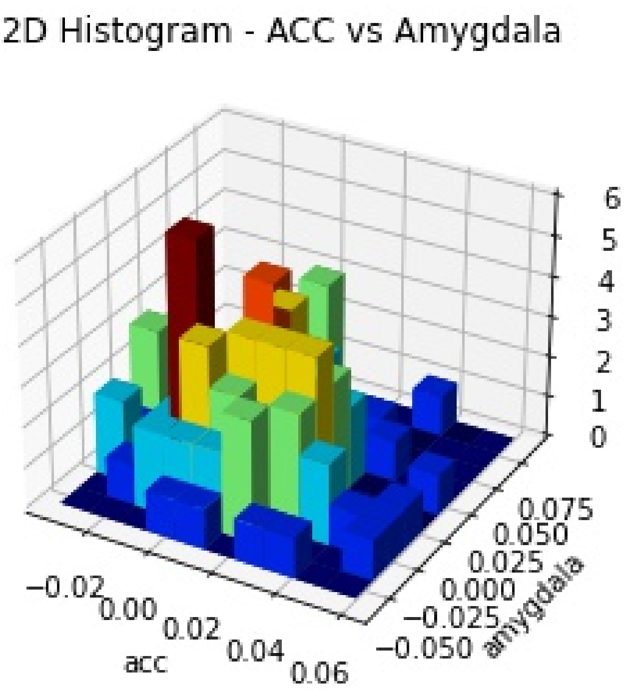
Form the 1-dimensional histogram and KDE to estimate the distributions of amygdala and acc, respectively. For this question, you can ignore the variable orientation. Decide on a suitable number of bins so you can see the shape of the distribution clearly. Set an appropriate kernel bandwidth h > 0.

For the histograms, I chose 10 bins.

	Histogram	KDE
ACC		
Amygdala		

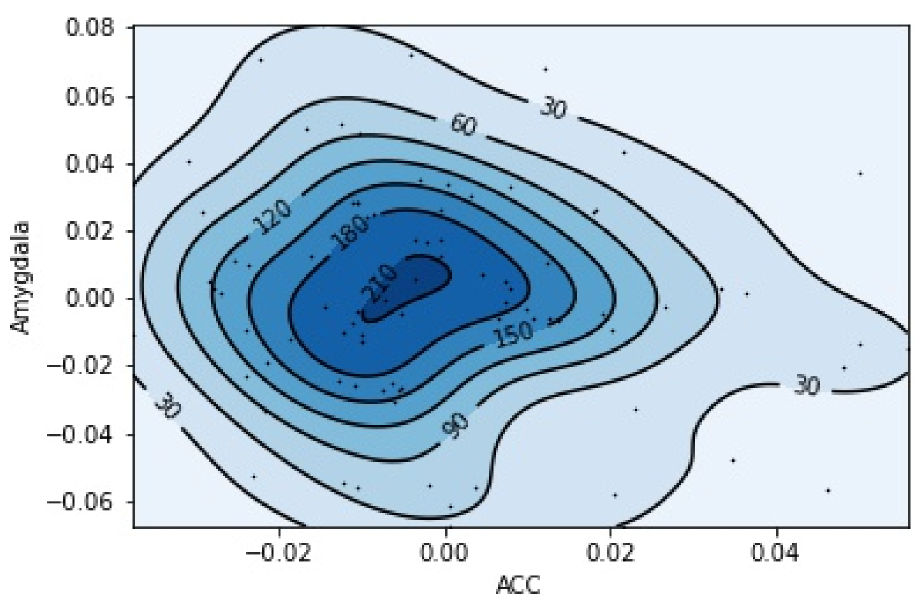
Part B

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.



Part C

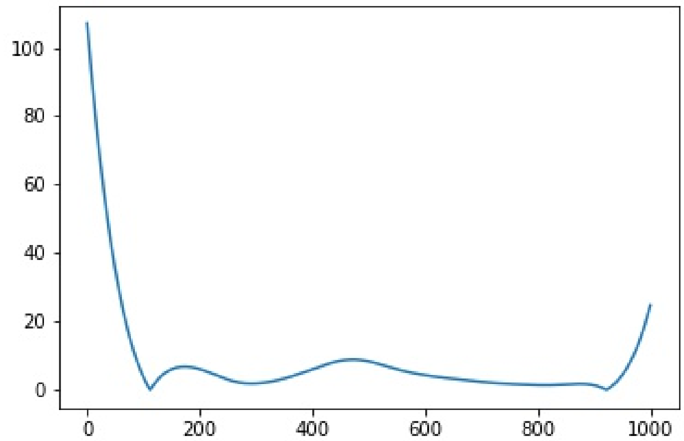
Use kernel-density-estimation (KDE) to estimate the 2-dimensional density function of (amygdala, acc) (this means for this question, you can ignore the variable orientation). Set an appropriate kernel bandwidth h > 0. Please show the two-dimensional KDE (e.g., two-dimensional heat-map, two-dimensional contour plot, etc.)



Please explain what you have observed: is the distribution unimodal or bi-modal? Are there any outliers?
The data appears to be unimodal and does have some outliers.

Please explain based on the results, can you infer that the two variables (amygdala, acc) are likely to be independent or not?
To check if the two distributions are independent, I must check if $p(acc, amygdala) = p(acc) * p(amygdala)$.

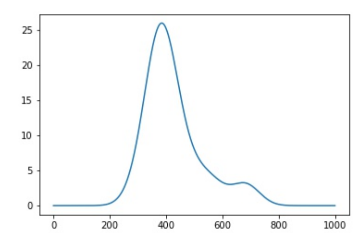
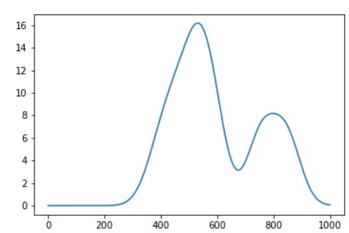
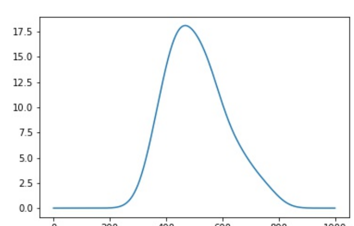
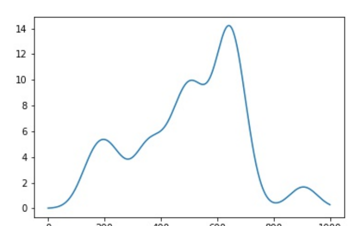
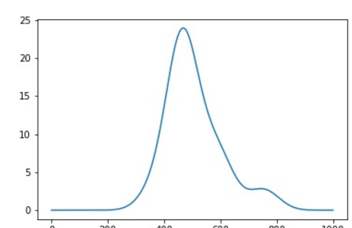
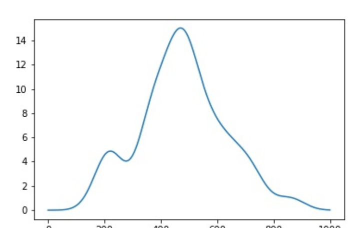
Below is the plot of the absolute error between the joint distribution and the product of the marginal distributions. If amygdala and acc are independent, I would expect this difference to be 0:

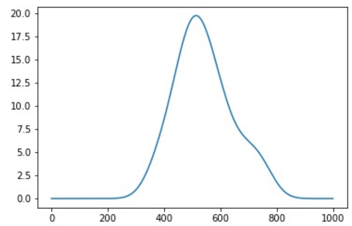
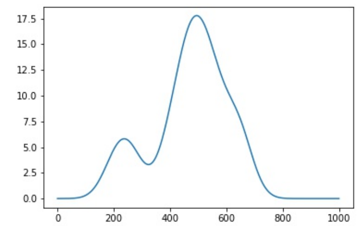


Since the difference between the two distributions is not 0, I can conclude that amygdala and acc are not independent.

Part D

We will consider the variable orientation and consider conditional distributions. Please plot the estimated conditional distribution of amygdala conditioning on political orientation: p(amygdala|orientation = c), c = 2, . . . , 5, using KDE. Set an appropriate kernel bandwidth h > 0. Do the same for the volume of the acc: plot p(acc|orientation = c), c = 2, . . . , 5 using KDE.

Orientation	ACC	Amygdala
2		
3		
4		

Orientation	ACC	Amygdala
5		

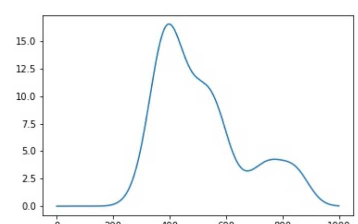
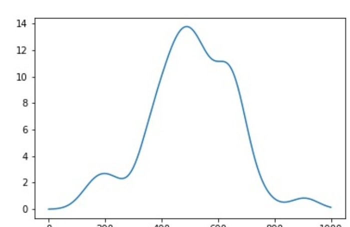
Now please explain based on the results, can you infer that the conditional distribution of amygdala and acc, respectively, are different from c = 2,...,5? This is a type of scientific question one could infer from the data: Whether or not there is a difference between brain structure and political view.
Based on these graphs, I would infer that there is indeed a relationship between the size of brain structures and political views. The distribution of amygdala size differs dramatically given different political orientations. There is a difference in ACC distribution, but not as drastic compare to that of amygdala.

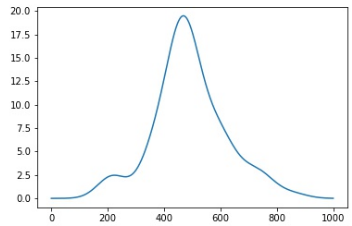
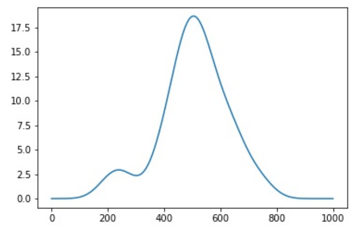
Conditional Sample Mean

	c = 2	c = 3	c = 4	c = 5
Amygdala	0.02	0	0	0.01
ACC	-0.01	0	0	-0.01

Part E

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(amygdala, acc|orientation = c), c = 2, . . . , 5. You will use two-dimensional KDE to achieve the goal; et an appropriate kernel band- width h > 0. Please show the two-dimensional KDE (e.g., two-dimensional heat-map, two-dimensional contour plot, etc.).

Orientation	Joint Distribution
2	
3	

Orientation	Joint Distribution
4	
5	

Please explain based on the results, can you infer that the conditional distribution of two variables (amygdala, acc) are different from c = 2, . . . , 5? This is a type of scientific question one could infer from the data: Whether or not there is a difference between brain structure and political view.
Based on the shape of the joint distributions, I can infer that the conditional distribution of amygdala and ACC are indeed different based on political view. It looks like the distributions for c=4,5 are similar in shape - perhaps these political orientations are more similar to each other than to c=2 or c=3.

3. Implementing EM for MNIST Dataset

Part A

Write down detailed expression of the E-step and M-step in the EM algorithm.
Used Slide 13 in Module 7 notes as a reference.

Expectation Step:

$$\tau_k^i = p(z^i = 1 | D, \mu, \Sigma) = \frac{\pi_k N(x^i|\mu_k,\Sigma_k)}{\Sigma_{k'=1}^K \pi_k' N(x^i|\mu_k',\Sigma_k')} = \frac{\frac{\pi_k}{\sqrt{|\Sigma_k|}} e^{\frac{-1}{2}(x^i-\mu_k)^T \Sigma_k^{-1}(x^i-\mu_k)}}{\Sigma_{k'=1}^K \frac{\pi_{k'}}{\sqrt{|\Sigma_{k'}|}} e^{\frac{-1}{2}(x^i-\mu_{k'})^T \Sigma_{k'}^{-1}(x^i-\mu_{k'})}}$$

Maximization Step:
Use the definition of τ_k^i outlined in E step.

$$\pi_k = \frac{1}{m} \Sigma_i \tau_k^i$$

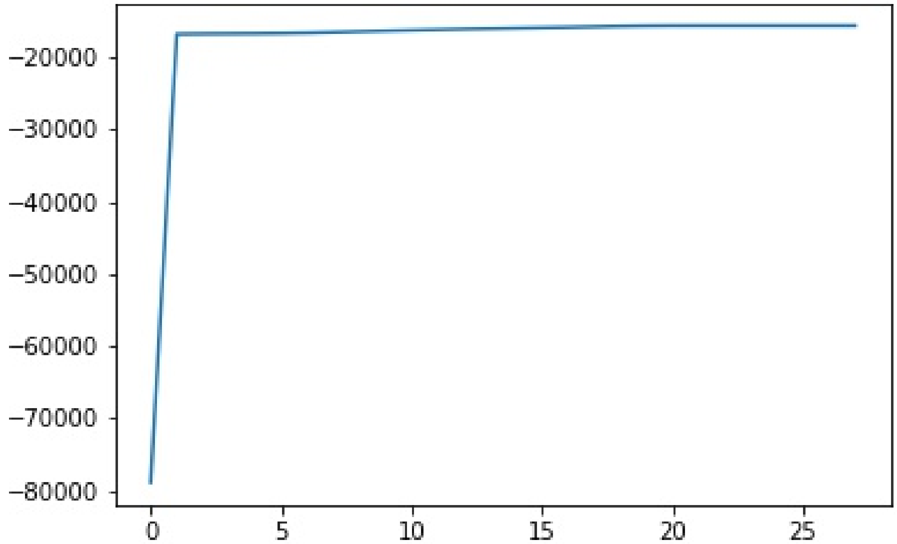
$$\mu_k = \frac{\Sigma_i \tau_k^i x^i}{\Sigma_i \tau_k^i}$$

$$\Sigma_k = \frac{\Sigma_i \tau_k^i (x^i-\mu_k)(x^i-\mu_k)}{\Sigma_i \tau_k^i}$$

Part B

Implement EM algorithm yourself. Plot the log-likelihood function vs the number of iterations to show your algorithm is converging.

EM algorithm converged in 27 iterations.

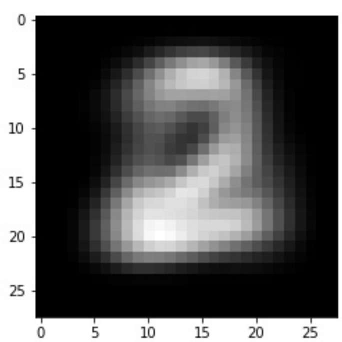


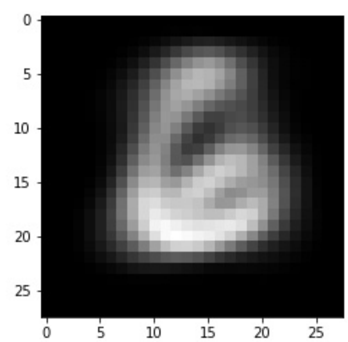
Part C

Report the fitted GMM model when EM terminates. For the mean of each component, map these back to the original space and reformat the vectors to make them into 28-by-28 matrices and show images. Ideally, you should be able to see these means correspond to "average" images. You can report the two 4-by-4 covariance matrices by visualizing their intensities (e.g., using a gray scaled image or heat map).

$\pi_1 = 0.48677765, \pi_2 = 0.51322235$

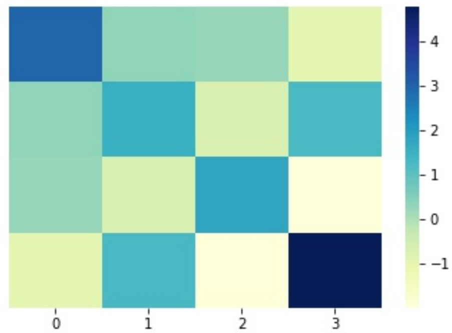
Mean Images

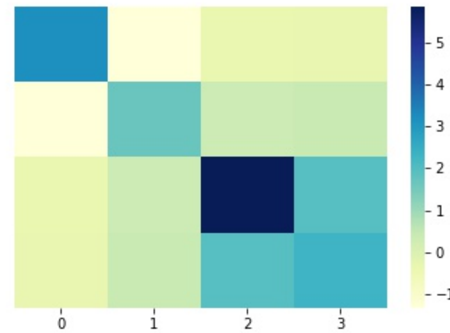




Covariance Heat Maps

Covariance Heat Maps





Part D

Use τki 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 code from previous assignments). Find the mis-classification rate for digits “2” and “6” respectively, and compare with GMM. Which model achieves better performance overall?

GMM Misclassification Rate = 0.03769
K Means Misclassification Rate = 0.06231

The EM algorithm achieves the best performance overall.