Homework 3

Conceptual questions

1. Please compare the pros and cons of KDE over histogram and give at least one advantage and disadvantage to each.

One advantage of Histogram over KDE is that it is less memory intensive ($1/\Delta^n$) compared to KDE which must store each point to form kernel density (*mn*). One disadvantage of histogram is that its representation can easily be biased by the bin size choice, where as KDE is smoothed and thus is not sensitive to choice of bin size. One advantage of KDE over histogram is that it has a smaller integrated risk over the same number if samples giving it a better statistical guarantee. One disadvantage of KDE compared to histogram is that it is more computationally expensive to derive *p(x)* than histogram especially when using cross-validation.

2. Why you cannot use maximum likelihood estimation to directly estimate GMM? Then how to estimate the model of GMM?

We don't know the mixture components (z_i, latent "hidden" variable) and thus cannot compute the log-likelihood function directly to make a direct estimate for GMM. Since x can come from any of the latent components, we have to reverse engineer the probability of x throughout all the possible given z_i states (*p(x|z)p(z)*).To do this, given the information we have we make an estimated guess on the expectation of the latent factors when calculating the log-likelihood using Expectation Maximization (EM). We are making the "best guess" to represent z_i given the data we have.

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

In EM, when computing the posterior distribution, we use the Bayes rule in this way:

Joint distribution: p(x,z)
Prior/Marginal dist: p(z)
Likelihood "x given z": p(x|z)
Posterior "z given z": p(z|x)
Normalization constant: p(x)

Bayes Rule Definition:
p(z|x) = Likelihood * prior / Normalization constant = p(x|z) * p(z) / p(x)

Since p(x|z) * p(z) is equal to the joint distribution by definition of conditional probability, equivalent to:
P(x,z) / Summation (p(x,z'))

Plug-in GMM terms for posterior p(z|x):
$\prod z\ N(X|\mu z, \sum z) / \sum z, \prod z\ '\ N(X|\ \mu z', \sum z')$

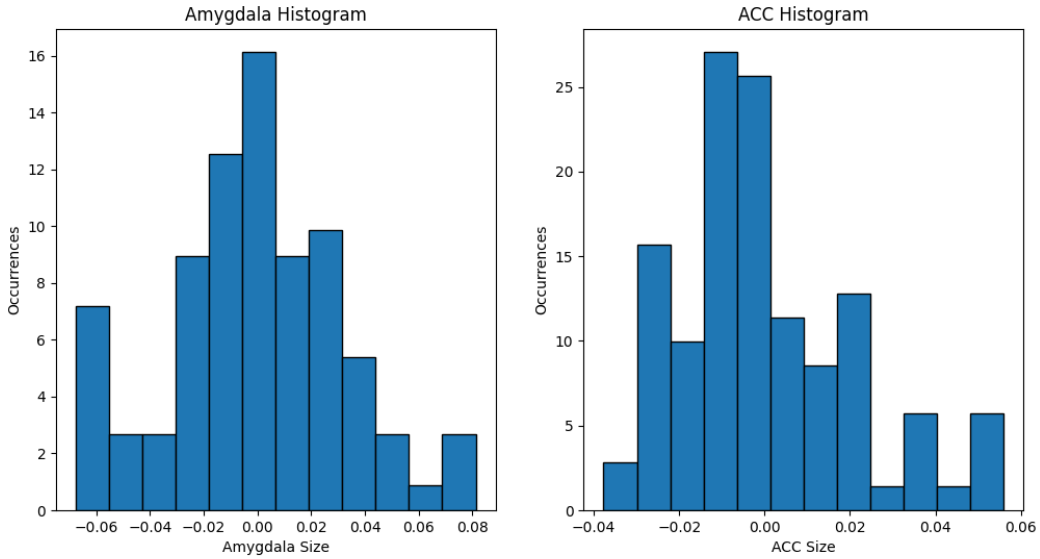
Next, E-step:
Where $q(z_1 \ldots z_m)$ represent the posterior of the latent variables at each iteration:
$q(z_1 \ldots z_m) = \prod p(z_i|x^i, 0^t)$

For each xi, use bayes t_i compute the posterior $p(z_i = k|x^i)$ for each k,
$T_k^i = p(z_i = k|\ x^i, 0^t)$, plug in relevant GMM terms:
$\prod k\ N(x^i|\mu k, \sum k) / \sum k'=1 \ldots k, \prod k', N(x^i|\ \mu k', \sum k')$

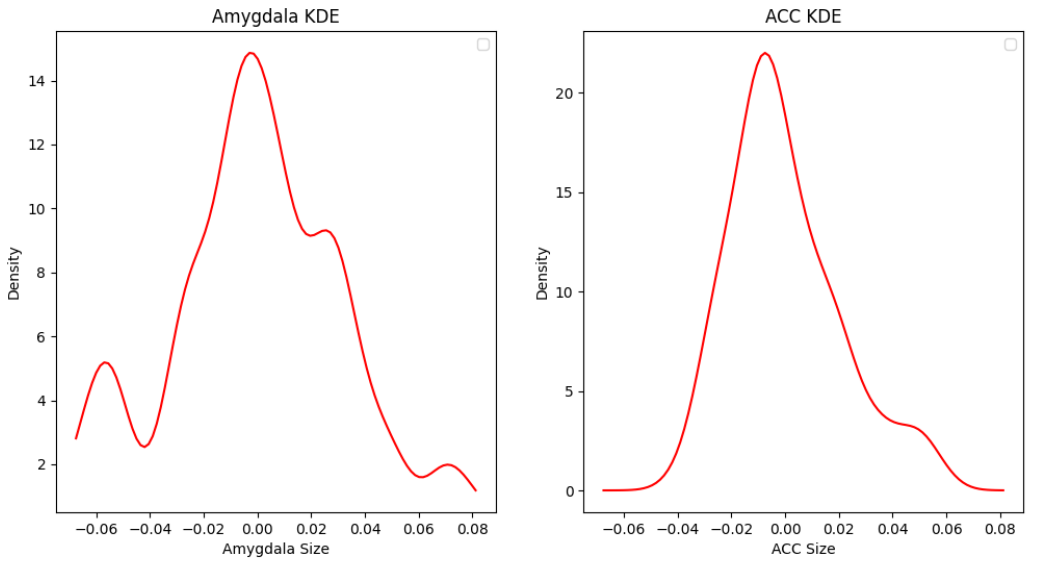
Density estimation: Psychological experiments.

Part A: 1-dimensional histogram and KDE for the distributions of Amygdala and ACC

Below are 1-dimensional histograms for Amygdala and ACC sizes respectively. Each has a bin size of 12, which allows the shape of the data to be seen:

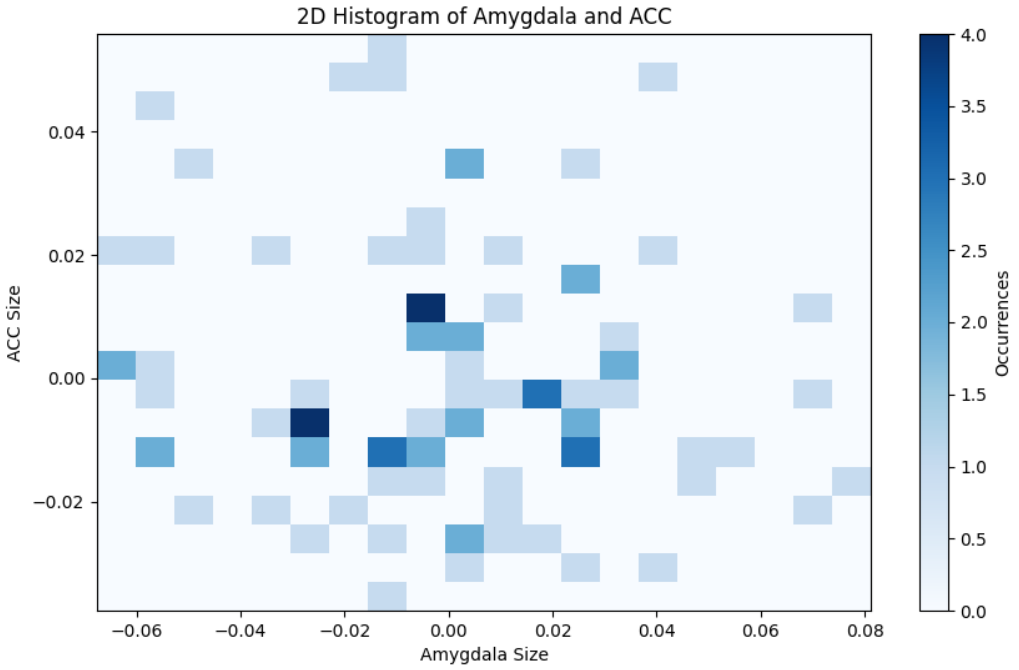


Below are 1-dimensional KDE plots for Amygdala and ACC respectively. Each has a bandwidth (h) value of .0075, chosen with the intent to represent the shape of the data but not be over-sensitive to outliers:



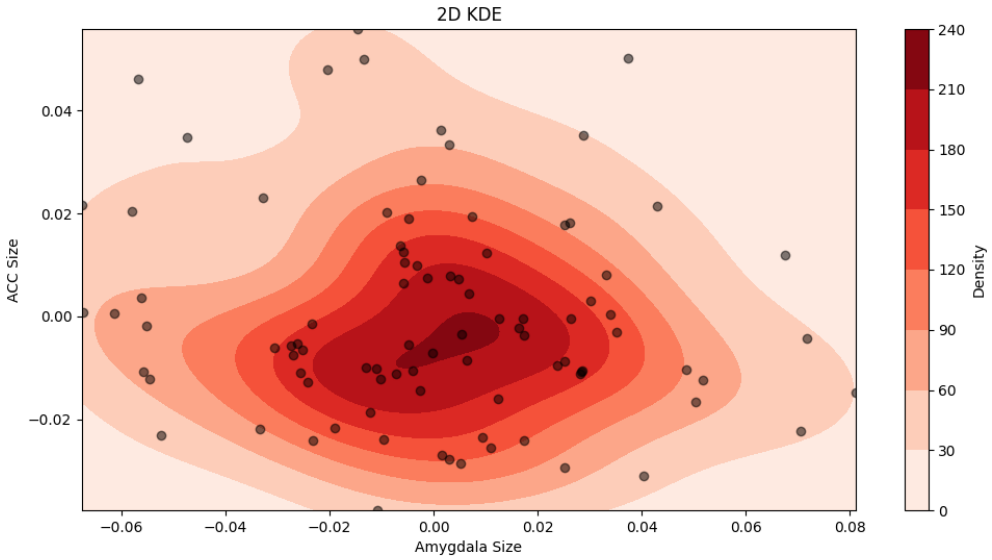
Part B: 2-dimensional histograms

Below is a 2-D histogram of Amygdala and ACC with bin sizes of 20 for both variables. The intensity of the color in each bin represents the value of occurrences.



Part C: Estimate of the 2-dimensional density function of Amygdala and ACC

Below is the 2-dimensional KDE representation shown by the contour plot.



Is the distribution unimodal or bi-modal?

Based on the plot above, the data appears to be unimodal centered roughly around (0,0).

Are there any outliers?

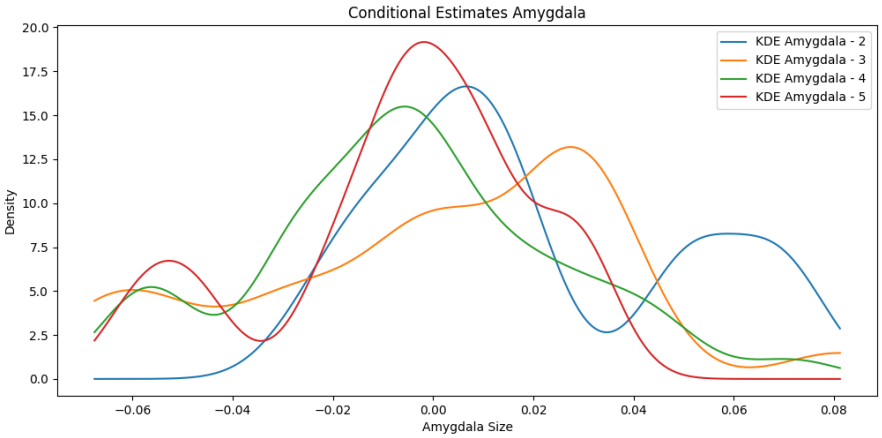
There do appear to be a few outliers at the higher end of ACC size and upper end of Amygdala size, as well as a few on the higher end of ACC size and high end of Amygdala size. In total, 4 points appear to be outliers.

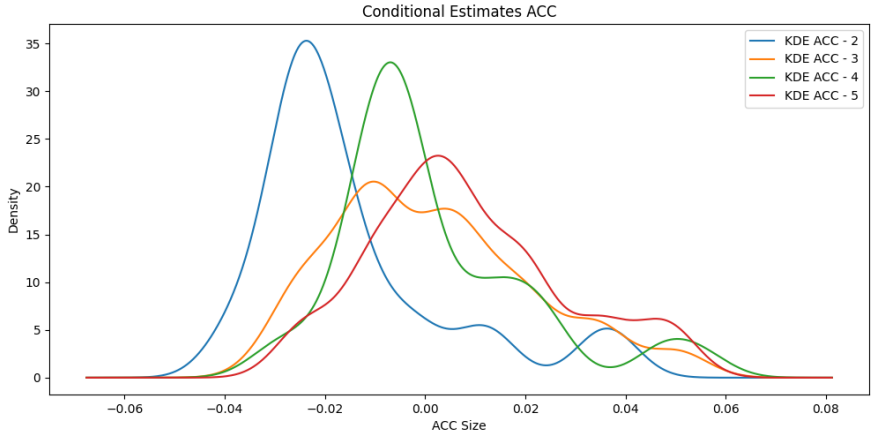
Are the two variables likely to be independent or not?

The two variables are not independent because the joint probability is not equal to the product of the marginal probabilities (p(amygdala|acc) != p(amygdala) * p(acc)). This was computed by deriving the joint probability of the two variables, and then multiplying out the marginal values for each occurrence.

Part D: Estimated conditional distributions based on orientation

Below are the conditional distributions for each variable separated by orientation with a bandwidth of .3:





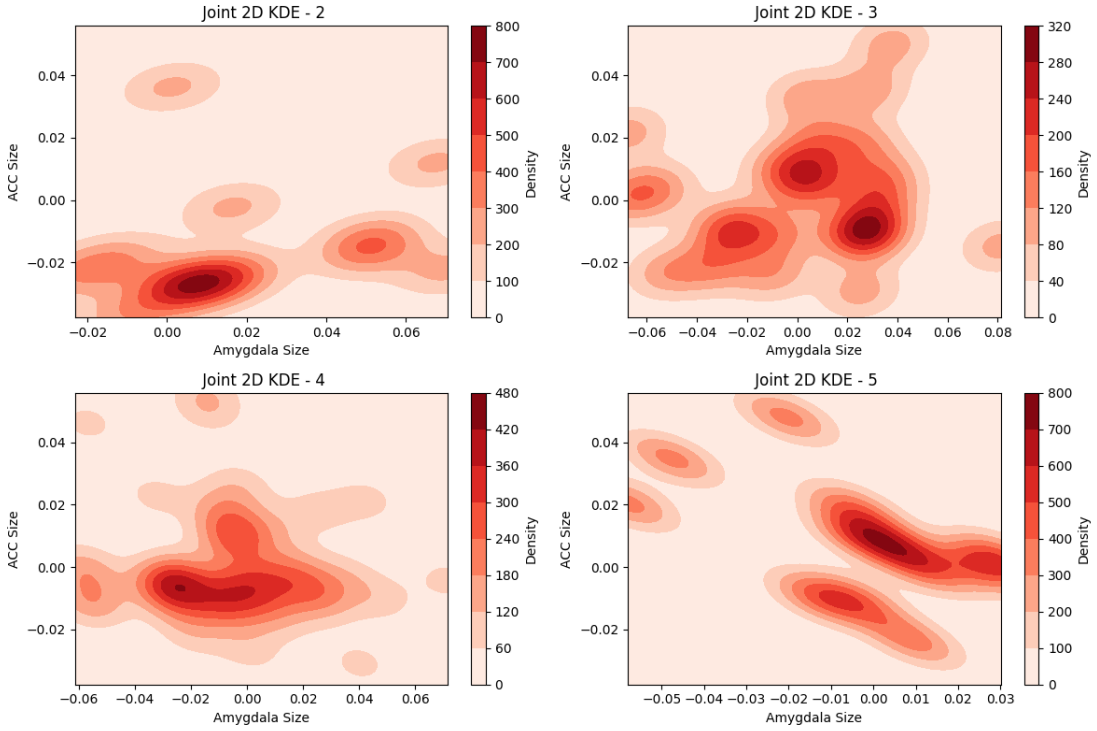
	c=2	c=3	c=4	c=5
Amygdala	.0191	.001	-.005	-.006
ACC	-.0148	.002	.001	.008

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 the above distributions, it appears that there is a relationship between ACC size and political orientation, while it appears less likely that there is a relationship between Amygdala size and political orientation. I do not believe there is a relationship between Amygdala and orientation because there is a significant amount of overlap amongst the pdfs of the various conditional distributions. Although they are not all identical in shape, there appears to be too much similarity between the groups to justify a meaningful difference between orientations. In contrast, the pdfs of the various ACC conditional distributions have more distinguished shapes and means. The derivations seem to indicate distinct structural differences between political groups. As indicated in the question, modelling the conditional distributions provided us with much more context than just the mean point-estimates across groups.

Part E: Conditional Joint Distribution

Below are the conditional joint distributions of both variables for each orientation:



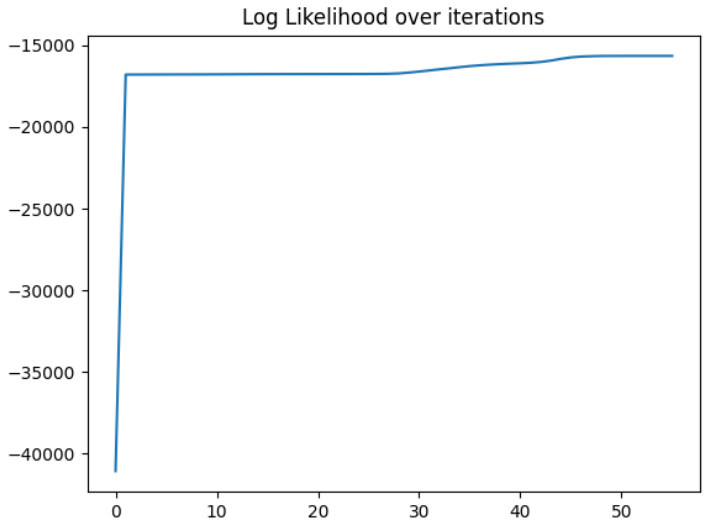
Please explain based on the results, can you infer that the conditional distribution of two variables (amygdala, acc) are different from c = 2, . . . , 5? This is a type of scientific question one could infer from the data: Whether or not there is a difference between brain structure and political view.

Based on the joint conditional distributions, it appears difficult to infer the conditional distributions are different. The KDE representations across groups seem to demonstrate more bi-modal distributions than modelled before. There also appear to be more presence of outliers, and consequently more variably spread making it difficult to make conclusions about relative distinction. The shape of the data, potential presence of outliers, and small sample size of 90 all lead me to believe that we cannot reasonably conclude that the condition distribution of the two variables as it relates to political orientation are distinguishably different.

Implementing EM for MNIST dataset

Part A: EM implementation and log-likelihood

The plot below shows the log-likelihood over iterations of the EM algorithm. We can see that as we increase in iteration the likelihood function converges.



Part B: GMM

Below are the weights computed during the GMM computation:

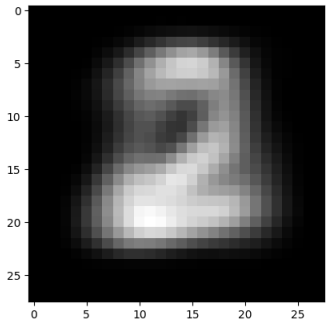
[0.51342529 0.48657471]

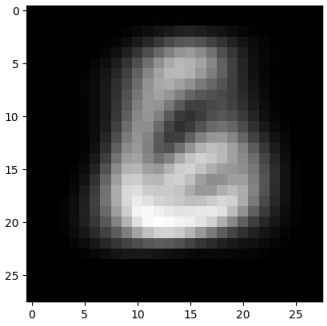
Below are the means of the 4 PCA components computed during the GMM computation for each group:

Means for “2’s”: [-6.86347489, 1.98139058, 0.07008971, -0.4764883]

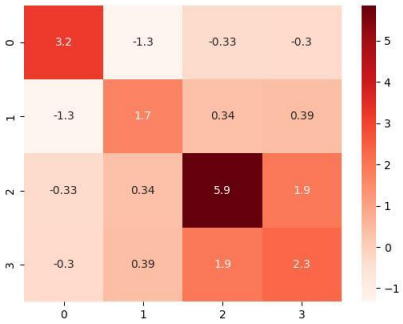
Means for “6’s”: [-7.00019555, -2.19889124, -0.08345029, 0.3135287]

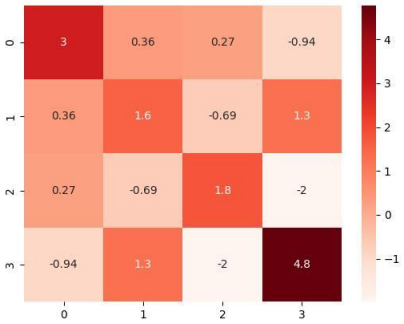
The images below are the 28x28 reshaped “average” images based on the GMM derivation. We can see that they form a generalized version of the 2 and 6 having been computed based upon the principal components.





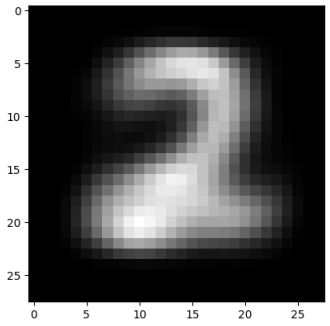
Below are the covariance matrices for the 2’s and 6’s respectively. We can see the differences in density between the principal components for each grouping by the color intensity.

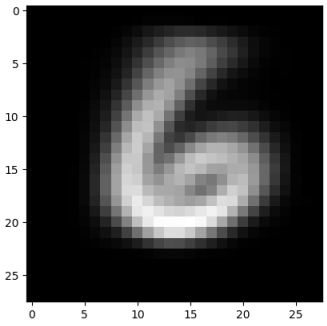




Part C: K-means

Below are the reconstructed 2 and 6 images based on the outputs of the k-means algorithm.





The table below shows the mismatch rate for the respective modelling techniques as computed during the construction processes:

Model Type	Mismatch Rate
GMM	.0376
K-Means	.0618

While both have relatively low mismatch rates, we can see that GMM performs better than K-means, with nearly half the mismatch rate. The lower mismatch rate indicates that GMM has better performance than K-means in the present classification.

BONUS:

Q1: