Ryan_Plain_HW3_Report

February 20, 2022

1 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

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x^i, \quad \hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m}(x^i - \hat{\mu})^2$$

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

Gaussian distribution has two sets of parameters (μ, σ)

The likeihood of one data point is:

$$p(x^i|\mu, \sigma) \propto exp(-\frac{1}{2\sigma^2}(x^i - \mu)^2)$$

The paramter estimates are:

$$\hat{\mu}_{MLE} = \frac{1}{m}\sum_{i=1}^{m} x^i, \quad \hat{\sigma}^2_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(x^i - \hat{\mu})^2$$

The objective function is log likelihood:

$$l(\mu, \sigma; D) = log\prod_{i=1}^{m}\frac{1}{(2\pi)^{\frac{1}{2}}\sigma}exp(-\frac{1}{2\sigma^2}(x^i - \mu)^2)$$

$$= -\frac{m}{2}log(2\pi) - \frac{m}{2}log(\sigma^2) - \sum_{i=1}^{m}\frac{(x^i - \mu)^2}{2\sigma^2}$$

Maximize $l(\mu, \sigma; D)$ with respect to μ, σ

Take derivates w.r.t. μ, σ^2

$$\frac{\partial l}{\partial \mu} = 0$$

$$\frac{\partial l}{\partial \sigma^2} = 0$$

$$= -\frac{m}{2}log(2\pi) - \frac{m}{2}log(\sigma^2) - \sum_{i=1}^{m}\frac{(x^i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}[(x_1 + ... + x_n) - n\mu]$$

$$0 = \frac{1}{\sigma^2}[(x_1 + ... + x_n) - n\mu]$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}[(x_1-\mu)^2) + ... + (x_n-\mu)^2]$$

$$0 = -\frac{n}{\sigma} + \frac{1}{\sigma^3}[(x_1-\mu)^2) + ... + (x_n-\mu)^2]$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$$

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

- KDE
  - Pros
    - smooth density where boxes are located in histogram
    - reduces noise of arbitrary bin size (advantage over histogram)
    - Better in high dimensional data (advantage to histogram)
  - Cons
    - Have to hold all the data in memory
    - Parameters increase with m, more expensive computation
      - Histogram would have the advantage depending on the size of bins, and memory requirement

3.) For the EM algorithm for GMM, please show how to use Bayes rule to derive τ_k^i in closed-form expression. let $\theta = (\pi_k, \mu_k, \Sigma_k), k = 1, ..., K)$

Maximize: $\theta^{t+1} = argmax_\theta f(\theta)$

`- Prior`

$$p(z)$$

`- Likelihood`

$$p(x|z) = N(x|\mu_z, \Sigma_z)$$

- Posterior

$$p(z|x) = \frac{\pi_z N(X|\mu_z, \Sigma_z)}{\Sigma_z, \pi_z, N(X|\mu_{z'}, \Sigma_{z'}}$$

$$\tau_k^i = p(z^i = k|x^i, \theta^t) = \frac{p(x^i|z^i = k)p(z^i = k)}{\Sigma_{k'=1..K} p(z^i = k', x^i)}$$

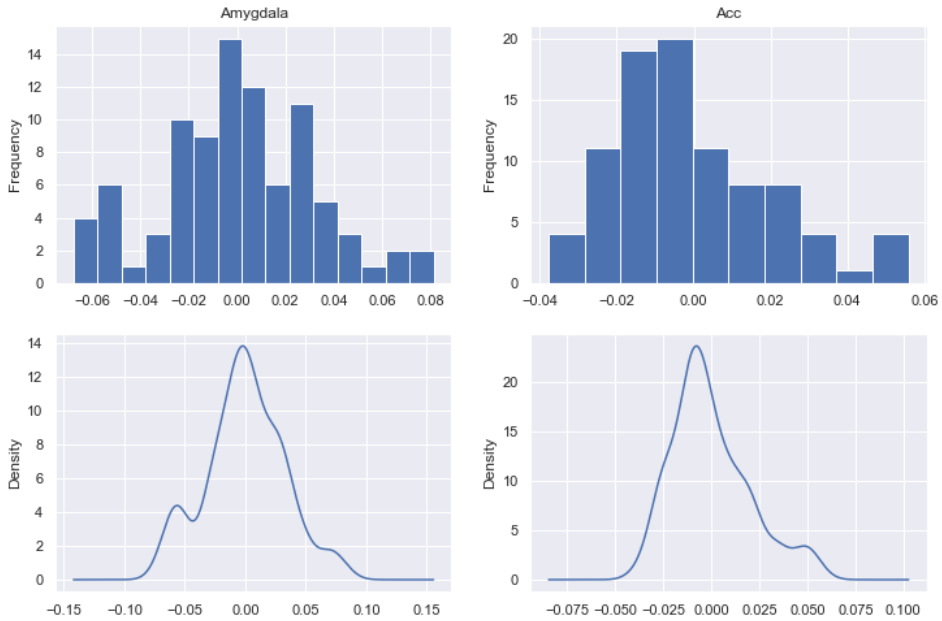
$$= \frac{\pi_z N(X|\mu_z, \Sigma_z)}{\Sigma_z, \pi_z, N(X|\mu_{z'}, \Sigma_{z'})}$$

---

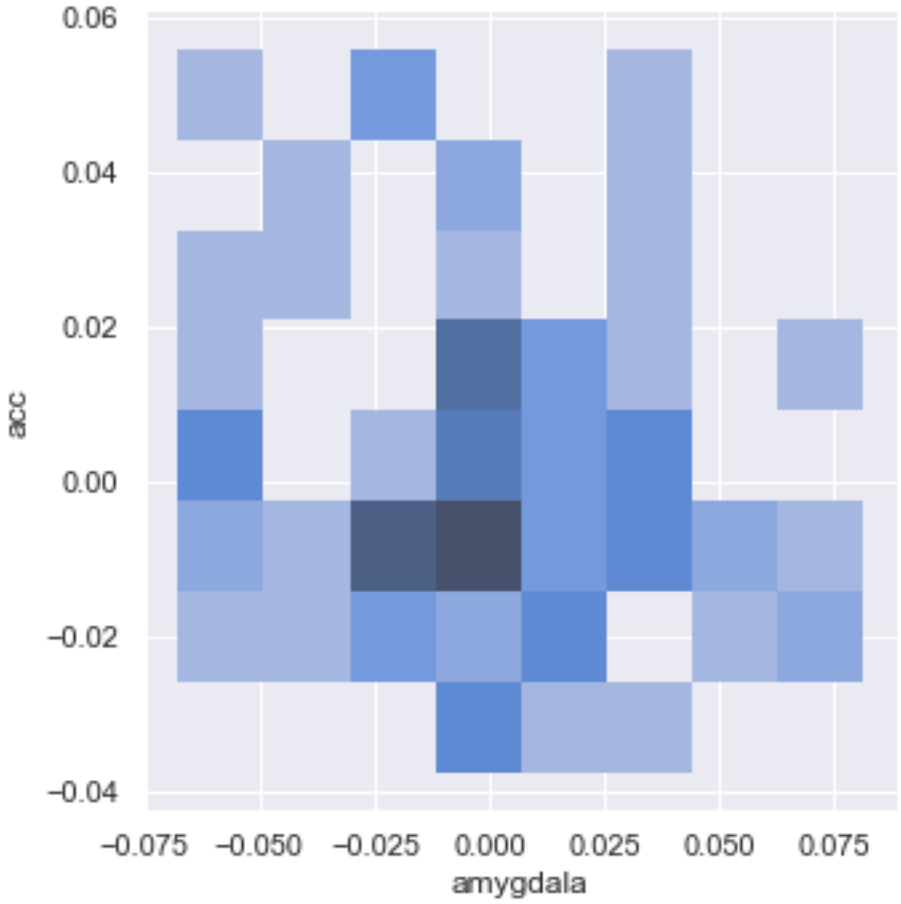
2 2. Density estimation: Psychological experiments

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 orienta- tion. Decide on a suitable number of bins so you can see the shape of the distribution clearly. Set an appropriate kernel bandwidth h > 0 To get the Histogram and KDE, manually checking different parameters for bin size and bandwidth were used. I wasn't able to use the rule of thumb Silverman method since this did not match a normal distribution.

The bandwith selected of 0.3, gives enough balance to capture modes without over interpolating the data an blurring out these points.

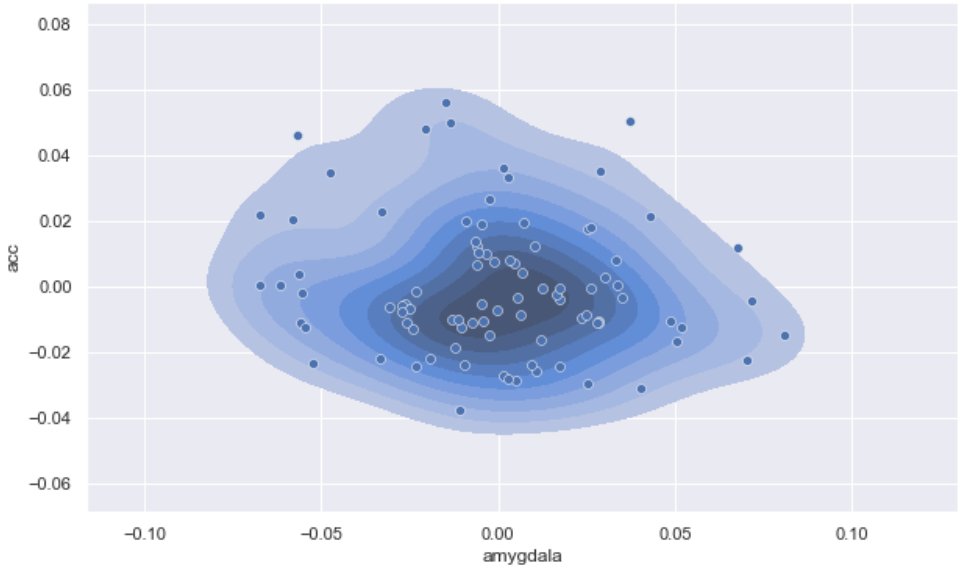


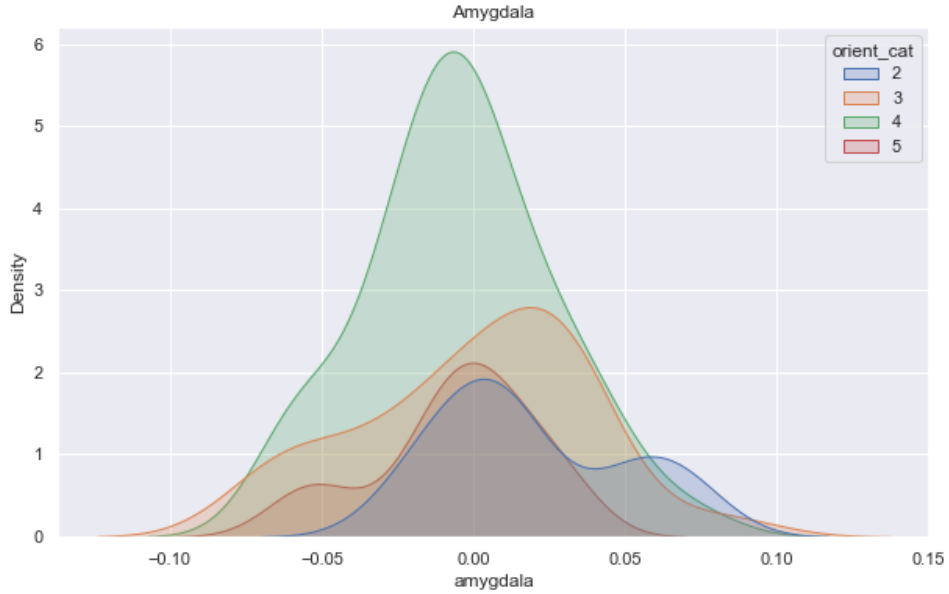
b.) Form 2-dimensional histogram for the pairs of variables (amygdala, acc). Decide on a suitable number of bins so you can see the shape of the distribution clearly. The number of bins selected was **8**. This was able to best identify where the the modal points were jointly distributed. It also demonstrates how much data is needed to even fill out a bin size as small as that, limiting the effectiveness of multi-dimensional histograms. Lowering the bins further diminishes seeing how spread out the data is.



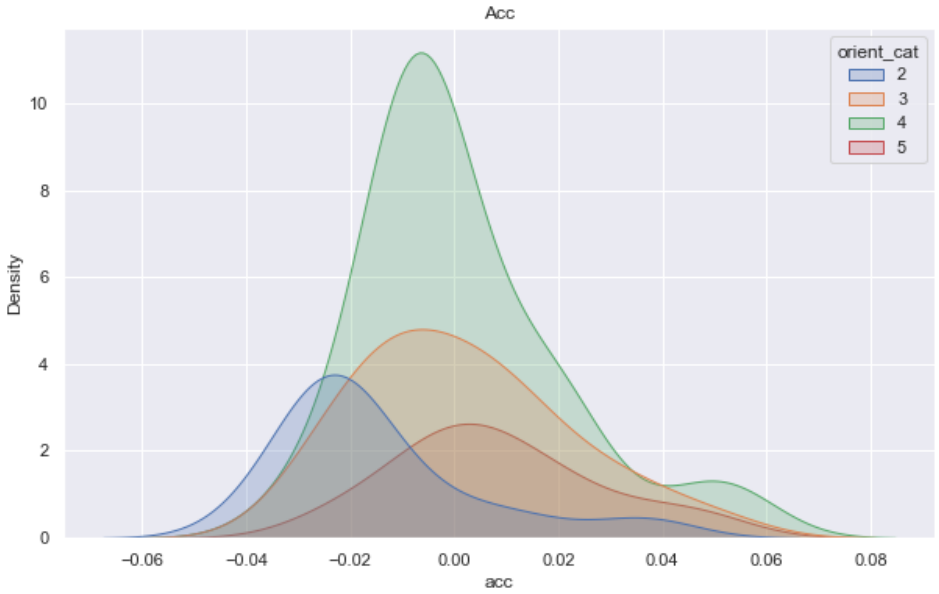
c.) Use kernel-density-estimation (KDE) to estimate the 2-dimensional density function of (amygdala, acc) (this means for this question, you can ignore the variable orientation). Set an appropriate kernel bandwidth h > 0. At all the levels of h checked manually, the joint distribution appears to be unimodal. Additionally, you can see that there are several outliers, visuzalized as points that are outside the scope of the KDE plot.

Looking at this visually, I would determine that there is a strong relationship between the variables and they are not independent. Next steps would be to test statistically how strong their dependence is.


0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
acc
−0.10
−0.05
0.00
0.05
0.10
amygdala


Amygdala
orient_cat
2
3
4
5
6
5
4
3
2
1
0
Density
−0.10
−0.05
0.00
0.05
0.10
0.15
amygdala

d.)



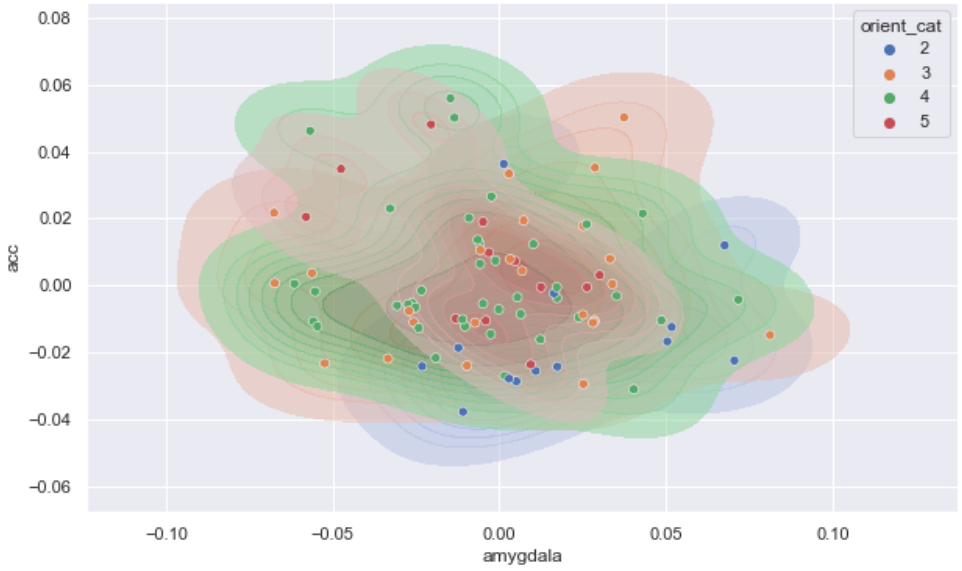
orientation	amygdala	acc
2	0.0190615	-0.0147692
3	0.0005875	0.00167083
4	-0.00471951	0.00130976
5	-0.00569167	0.00814167

Conditional Sample Means We can conclude that the distributions of the individual variables is conditional on the orientation. Orientation 4 appears to be the most common, and with 3 make up most of the values. This logically makes sence due to the fact that the orientations are more mild political opinions, and would (hopefully) contain more samples. Although some skewness, they appear to be more Gaussian.

Orientation 2 and 5 are more extreme political views. This is seen clearly with `acc` and orientation 2, where there is a strong right skewed modal around -0.025.

The table shows empirically the difference in sample means conditioned on orientation. The scale of `amygdala` is positive and negative values split between 3 and 4. `acc` has the only negative mean with orietation 2, and the difference between 2 and 5 is large.

3.) Joint Conditional Probability .



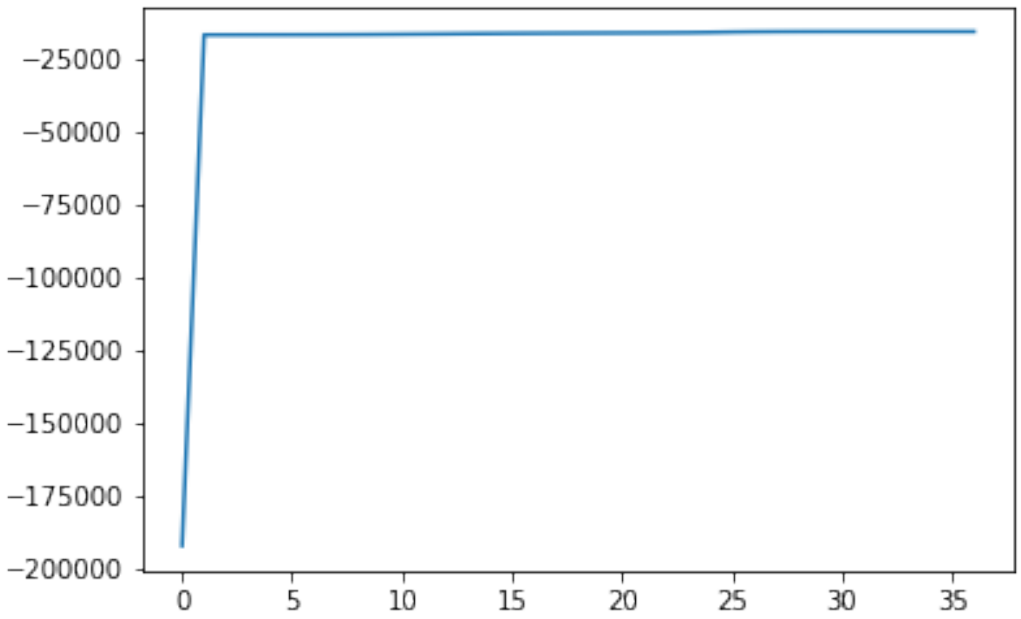
From the data presented, it can be infered that the distribution of political orientation is conditioned on the brain regions `amygdala` and `acc`. This though is just an inference based on the data provided. It is possible that bringing in more data points shows that the predictor variables are correlated strongly with another more powerful predictor. That information would be missing in this analysis, and attributing the signal to the wrong feature.

---

3 3. Implementing EM for MNIST dataset

a.)

b.) Plotting the log liklihood vs the number of iterations. The algorithm appears to have converged quickly, and it is good to see that it remained stable until meeting the threshold.


−25000
−50000
−75000
−100000
−125000
−150000
−175000
−200000
0
5
10
15
20
25
30
35