Homework Assignment 03

Due: Oct 28, 2023

**1.** Estimation of maximum likelihood using parametric models.

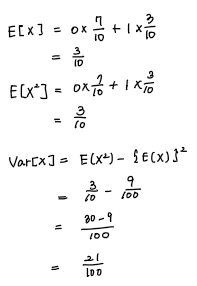
Suppose you are standing at an assembly line and examining the products whether they meet the quality requirements. You have observed 20 products come along the line, and 30% of which are identified as scrap according to the next sequence:

Scrap, good, good, good, good, scrap, good, scrap, good, good,

scrap, scrap, good, good, good, good, scrap, good, good, good

Assuming that you are estimating a random variable 𝑋 from these observations, answer the following questions.

(a) (1 pt.) Let us define 𝑿=𝟏 if a product is scrap; 𝑿=𝟎 if a product is good. What is 𝑬[𝑿] and 𝑽𝒂𝒓[𝑿]?

****

Using Python and external libraries, such as scipy.stats and seaborn, one can easily model and utilize the parametric density estimators that we studied in the class. See the provided code snippets, each of which basically has the below structure: # Instantiating a density model representing a random variable. # Sampling (𝑁=100000) by random sample generation according to the density model (the .rvs() function). # Plotting a histogram of the generated samples (using the seaborn library).

You are provided with the code for Bernoulli, binomial, Poisson, normal, and exponential distributions. Examine and understand the code and outcomes.

(b) (3 pt.) What does each of the histograms represents considering the choice of the density model? (There is no fixed answer.)

Sample answer: The resultant histograms represent their best fit in terms of MLE to the observations given the parametric assumption of underlying data distribution. E.g., Poisson: in the unit period (implicitly assuming length = 20 observations), visualizing the chance of how many times you would see scraps.

Sol)  
The Bernoulli distribution models a binary outcome, “scrap(1)” and “good(0)”. Bernoulli distribution represents the probability of scap and good. If you add two probabilities, you get 1.

The binomial distribution is used when you have a fixed number of independent Bernoulli and you want to model the number of scrap of those trials. The histogram shows the distribution of the number of 'scrap' products out of the 20 observations. You get a bell-shaped graph around the average with the mean and standard deviation that comes out.

The Poisson distribution is often used to model the number of events in a fixed interval of time. In this context, it's used to model how many times you would see 'scrap' products in a fixed length of 20 observations. The histogram shows the likelihood of observing different numbers of 'scrap' products within those 20 observations

The exponential distribution models the time between events in a Poisson process. In this context, it could be used to model the time between observing 'scrap' products. The histogram shows the likelihood of different time intervals between observing 'scrap' products.

The code fits a normal distribution to the observed data. It estimates the mean and variance of the data to define the normal distribution. The histogram for the normal distribution represents the estimated probability distribution of the data.

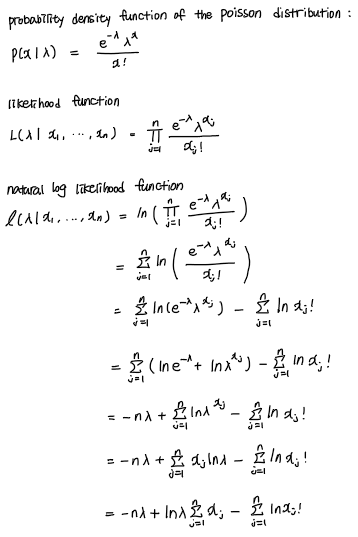
(c) (2 pt.) Of the five distributions, which one(s) can you relate to the central limit theorem. Explain your reason in your own words.  
Sol)

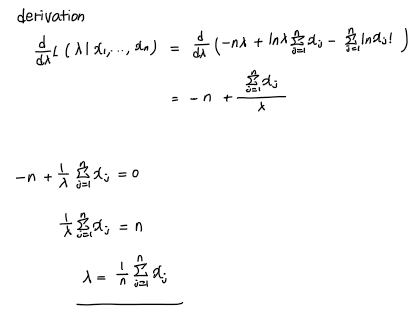
The central limit theorem states that when you take the average of a large number of random variables with the same probability distribution, the distribution of these averages will tend to resemble a normal distribution, provided that the sample size is sufficiently large. Among the five distributions you've mentioned, the one that most closely resembles a normal distribution is the binomial distribution. This is because it exhibits a bell-shaped curve around its mean, which is located at the center of the distribution.

**2. Deriving maximum likelihood estimators.**

(a) (4 pt.) Find the MLE for the parameter 𝜆 of a Poisson distribution. Show your derivation steps.

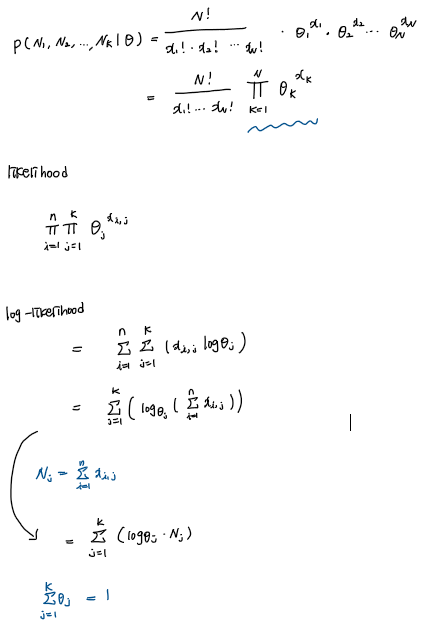
Sol)

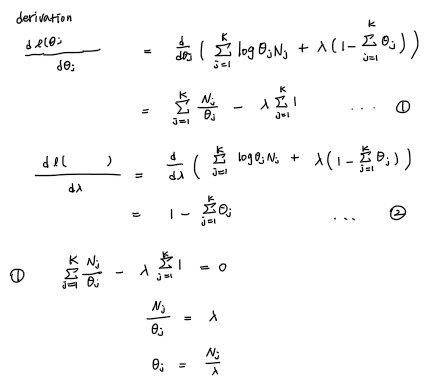


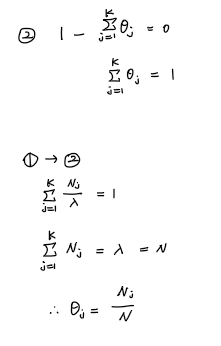


(b) (4 pt.) Find the MLE for the parameter 𝜃 ∶ 𝑘 ∈ {1,…,𝐾} of a multinomial distribution. Show your derivation steps.

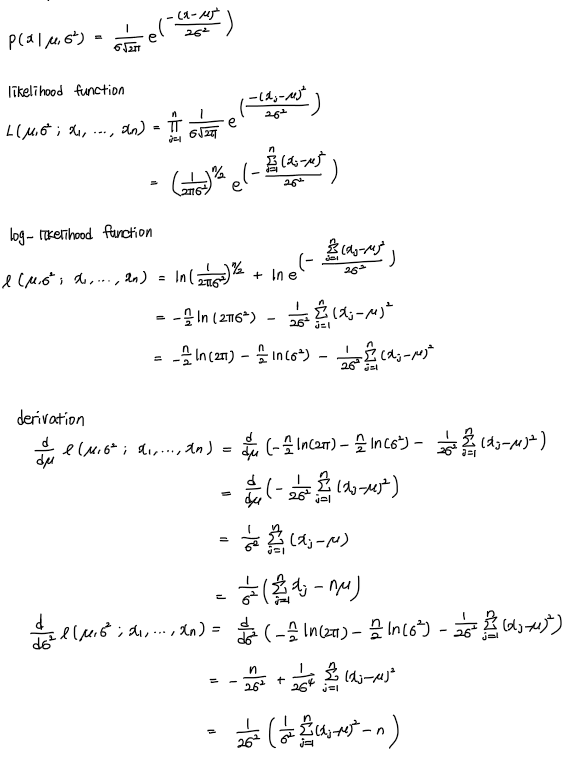
Sol)

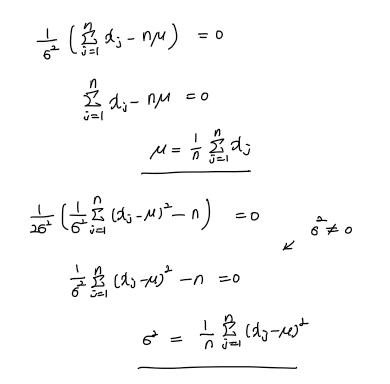




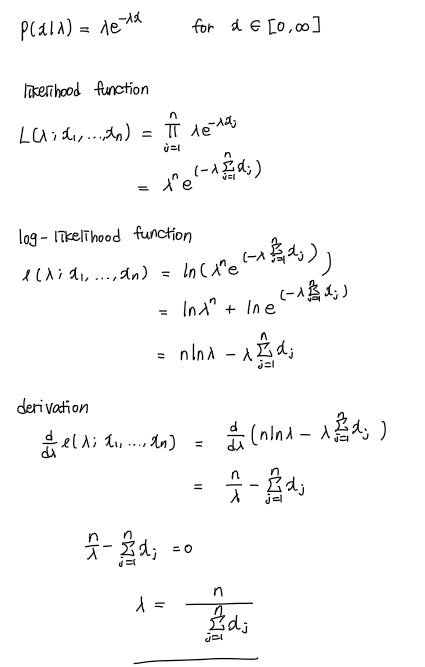


(c) (4 pt.) Find the MLE for the parameter 𝜇 of a normal distribution. Show your derivation steps.





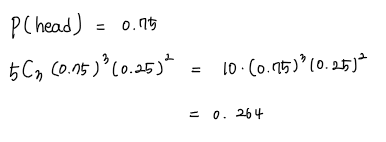
(d) (4 pt.) Find the MLE for the parameter 𝜆 of an exponential distribution. Show your derivation steps.

****

**3. Using parametric density models.**

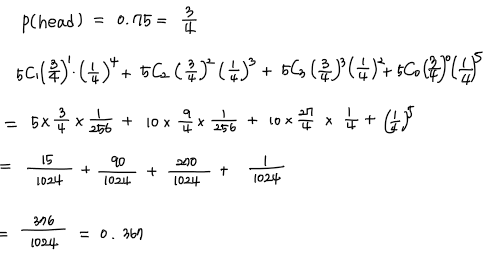
(a) (3 pt.) Consider an unfair coin that flips heads with probability 0.75. Assuming the events follow a binomial distribution, what is the probability of getting exactly 3 heads after 5 trials?

Sol)



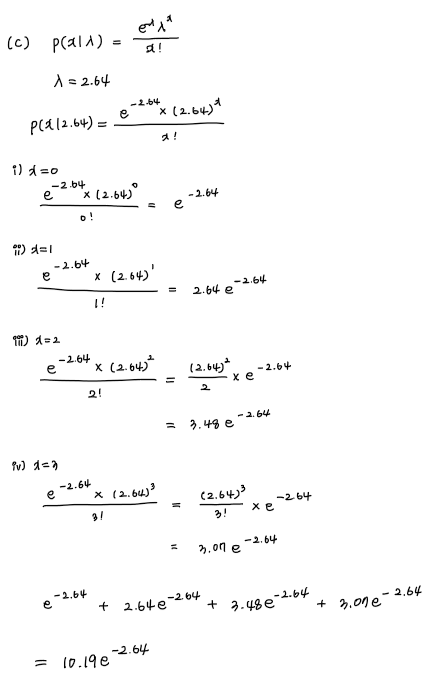
(b) (3 pt.) Considering the same unfair coin in Problem (a), what is the probability of getting less than 4 heads after 5 trials?

Sol)



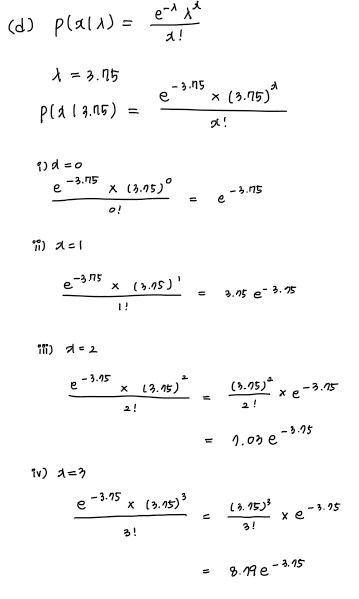
(c) (4 pt.) In the 2018 FIFA World Cup, there were 169 goals scored in 64 matches, for an average of 2.64 goals per match1. Assuming the number of goals per game follows a Poisson distribution, what is the probability that 𝑥 = 0, 1,2,3, and 4 goals are scored, respectively, in a game?

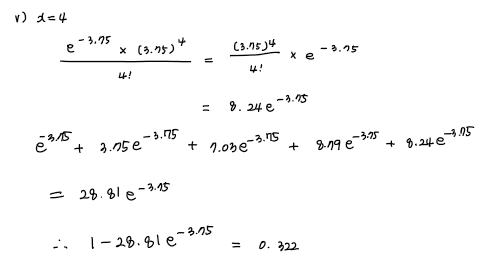
Sol)



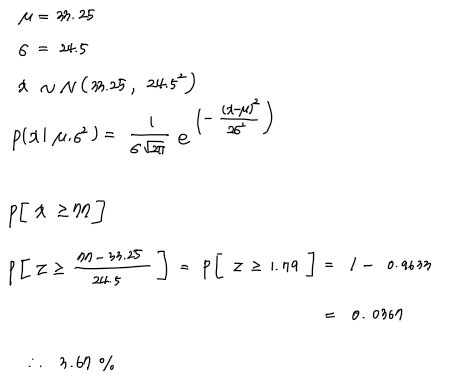
(d) (4 pt.) Our TA Jihyeon is seating near the front door of the Newton Hall and counting the number of people entering the building. She finds an average of 3.75 people enter the Newton Hall every minute. Assuming the number of people coming in follows a Poisson distribution, what is the probability that 5 or more people enter the building in a minute?

Sol)





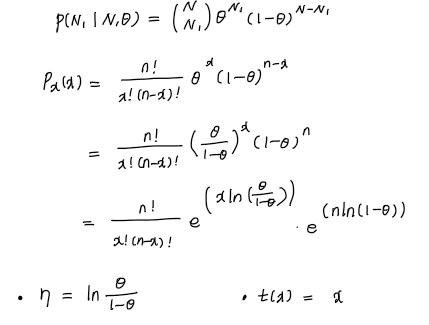
(e) (4 pt.) The entire CSEE students took a coding test, and their scores follow a normal distribution. If the distribution has mean 33.35 and standard deviation 24.5, what is the percentage of the students scored higher than 77?



**4. Exponential family.**

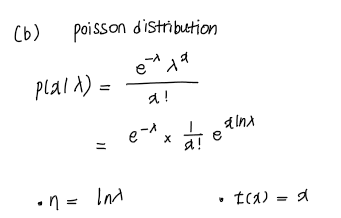
(a) (3 pt.) Show that the binomial distribution belongs to the exponential family of distributions. Give its natural (canonical) parameters and sufficient statistics.

Sol)



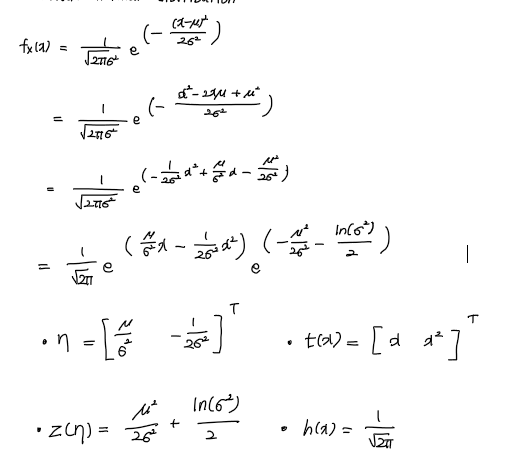
(b) (3 pt.) Show that the Poisson distribution belongs to the exponential family of distributions. Give its natural (canonical) parameters and sufficient statistics.

Sol)



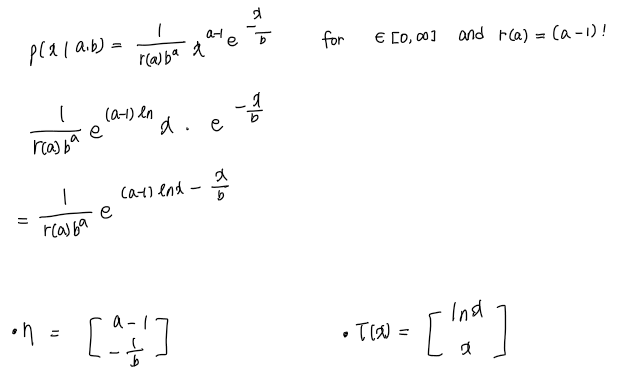
(c) (3 pt.) Show that the univariate normal distribution belongs to the exponential family of distributions. Give its natural (canonical) parameters and sufficient statistics.

Sol)



(d) (3 pt.) Show that the gamma distribution belongs to the exponential family of distributions. Give its natural (canonical) parameters and sufficient statistics.

Sol)



**5. Kernel density estimation.**

Find the dataset posted along with the assignment (hw3\_data.csv). This dataset contains the signals captured from 90 college students’ brains and their political orientation (views)2. In particular, the dataset includes the following four attributes:

Subject:

subject ID (not necessary for density estimation)

Amygdala: right amygdala size

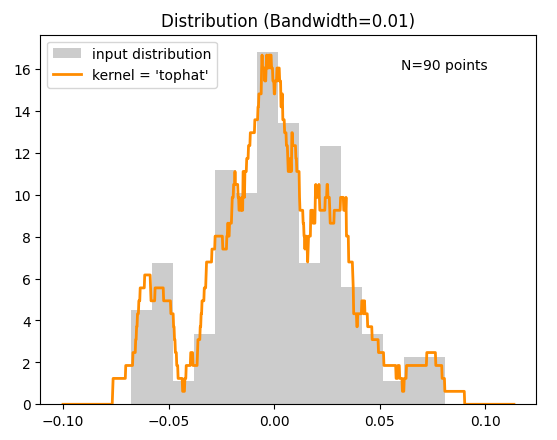
Acc: gray matter volume in the anterior cingulate cortex

Orientation: a self-reported five-point scale ranging from “very conservative” (1) to “very liberal” (5)

Note that amygdala and anterior cingulate cortex are two particular regions in human brain (one does not have to know the details of the brain structure for this analysis). The reported values are residuals (differences) from the predictive volume after adjusting height, sex, and other anatomical variables.

(a) (2 pt.) Write a Python script that reads the provided csv file and estimate 𝑃(𝑎𝑚𝑦𝑔𝑑𝑎𝑙𝑎). You may use the provided kernel density estimation function (kde1D()) with the Parzen window (kernel='tophat'). You may modify the function if you need. Try to find the best bandwidth. Report the resulting plot with the bandwidth value.

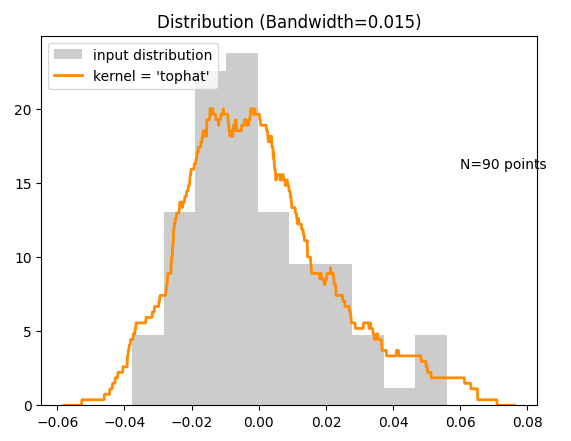
Sol)



Best bandwidth = 0.01

(b) (2 pt.) Repeat (a) to estimate 𝑃(𝑎𝑐𝑐). Report the resulting plot with the bandwidth value.

Sol)

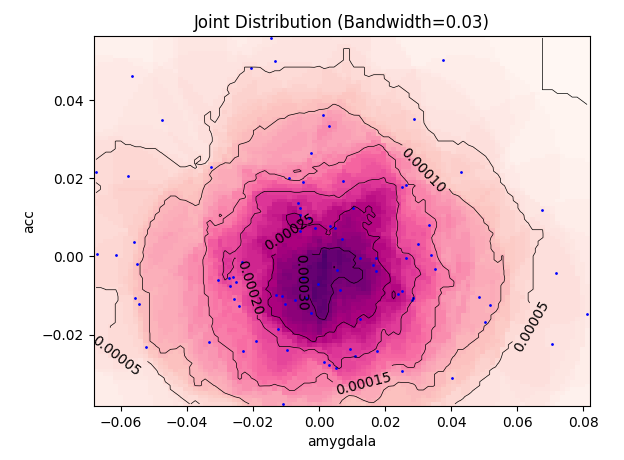


Best bandwidth = 0.015

(c) (2 pt.) Plot the joint distribution of 𝑎𝑚𝑦𝑔𝑑𝑎𝑙𝑎 and 𝑎𝑐𝑐. You may use the provided kernel density estimation function (kde2D()) with the Parzen window (kernel='tophat'). You may modify the function if you need. Report the resulting plot with the bandwidth.

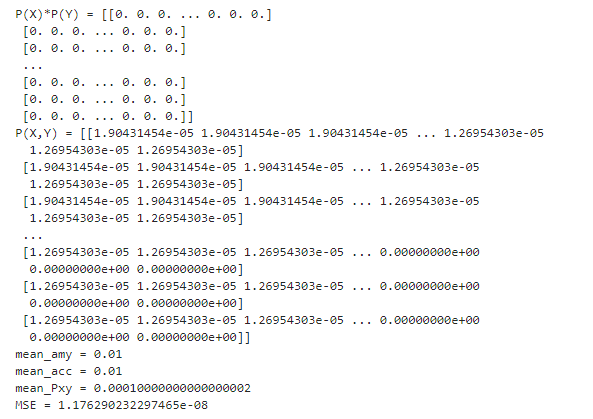
Sol)

Best bandwidth = 0.03



(d) (2 pt.) According to your estimation, are 𝑎𝑚𝑦𝑔𝑑𝑎𝑙𝑎 and 𝑎𝑐𝑐 independent? Note that statistical independence is defined as 𝑃(𝑋,𝑌) = 𝑃(𝑋)𝑃(𝑌).

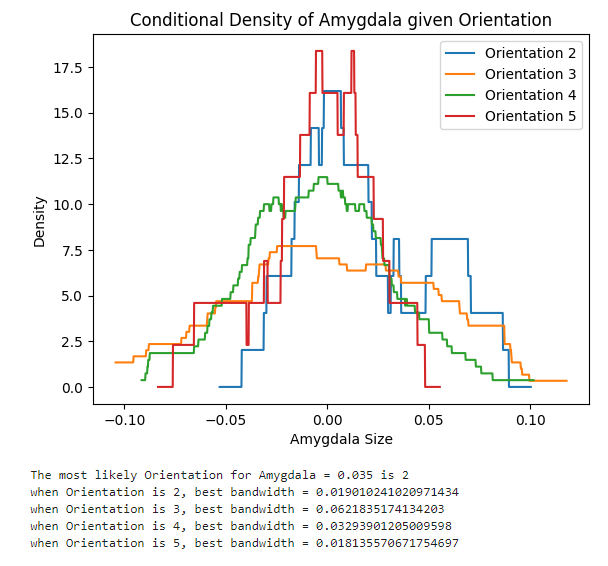
Sol)



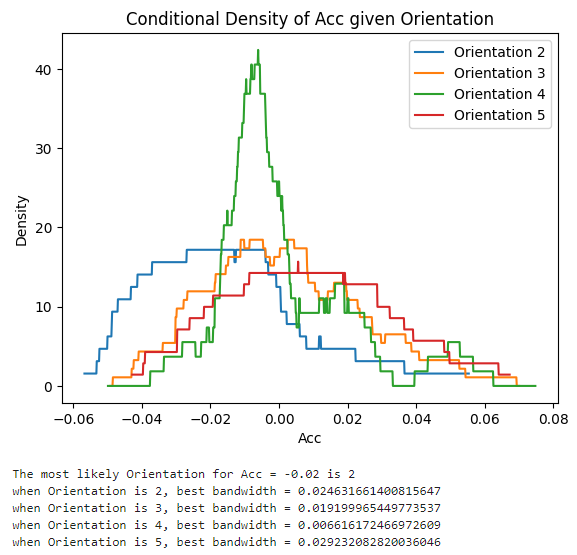
MSE is almost 0. So, 𝑎𝑚𝑦𝑔𝑑𝑎𝑙𝑎 and 𝑎𝑐𝑐 are statistical independent.

(e) (3 pt.) Find the conditional density of 𝑎𝑚𝑦𝑔𝑑𝑎𝑙𝑎 given 𝑜𝑟𝑖𝑒𝑛𝑡𝑎𝑡𝑖𝑜𝑛. You will need to estimate 𝑃(𝑎𝑚𝑦𝑔𝑑𝑎𝑙𝑎) for each value of 𝑜𝑟𝑖𝑒𝑛𝑡𝑎𝑡𝑖𝑜𝑛 (i.e., 𝑃(𝑎𝑚𝑦𝑔𝑑𝑎𝑙𝑎 | 𝑜𝑟𝑖𝑒𝑛𝑡𝑎𝑡𝑖𝑜𝑛)). Report the resulting plot. Say you have met with a student whose 𝑎𝑚𝑦𝑔𝑑𝑎𝑙𝑎 is 0.035. Which 𝑜𝑟𝑖𝑒𝑛𝑡𝑎𝑡𝑖𝑜𝑛 value between 1-5 is most likely for this student?

Sol)



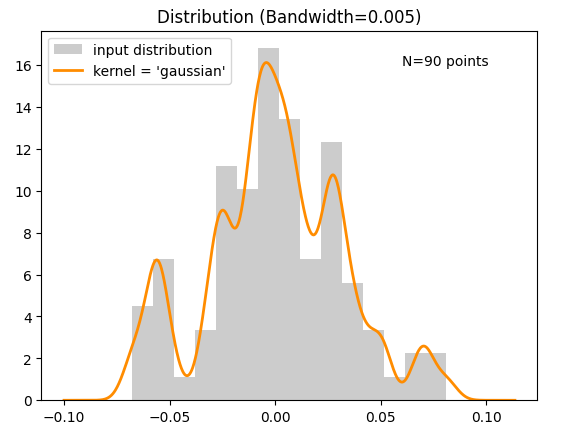
(f) (3 pt.) Repeat (e) to estimate 𝑃(𝑎𝑐𝑐 | 𝑜𝑟𝑖𝑒𝑛𝑡𝑎𝑡𝑖𝑜𝑛). Report the resulting plot. Say you have met with a student whose 𝑎𝑐𝑐 is -0.02. Which 𝑜𝑟𝑖𝑒𝑛𝑡𝑎𝑡𝑖𝑜𝑛 value between 1-5 is most likely for this student?

Sol)

(g) (10 pt.) Repeat (a)-(f) using the Gaussian kernel (kernel='gaussian'). Compare the results that you have obtained using the Parzen window and that of the Gaussian kernel.

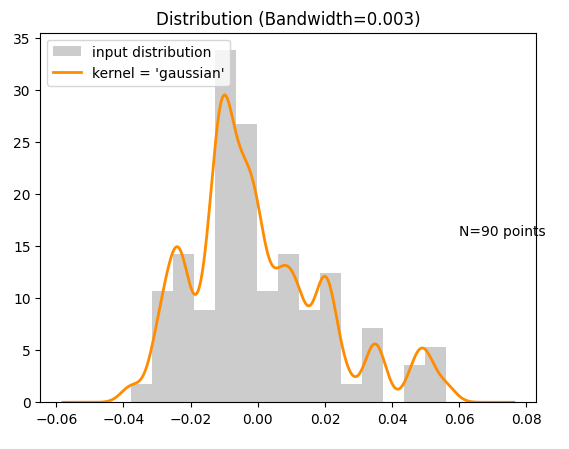
Sol)

(a)



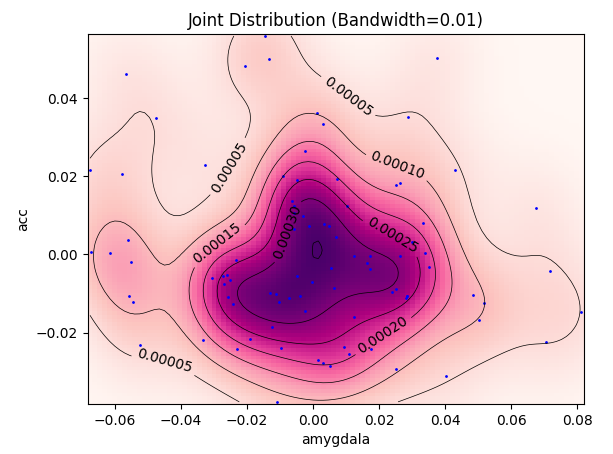
Best bandwidth = 0.005

(b)



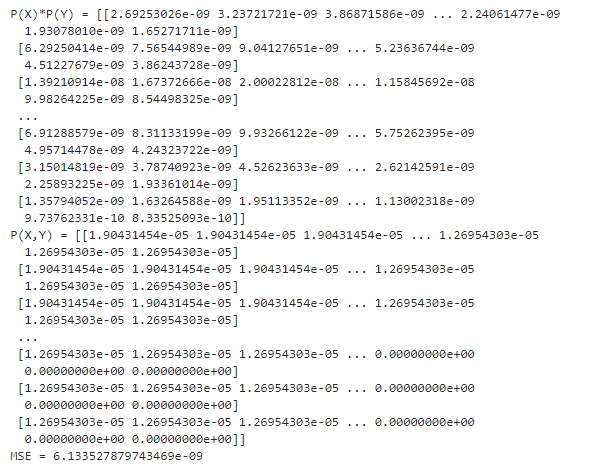
Best bandwidth = 0.003

(c)



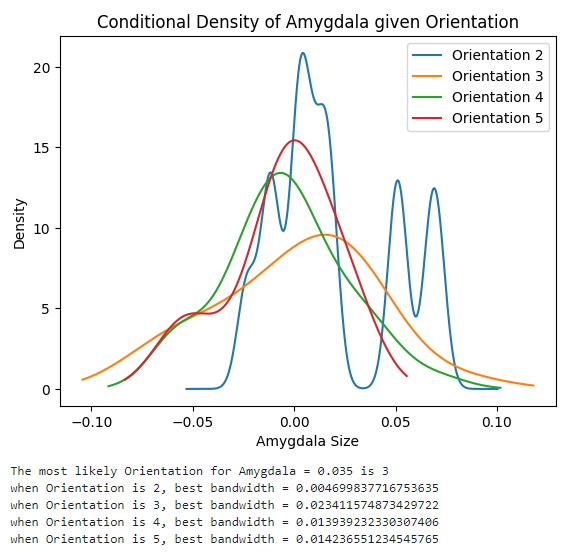
Best bandwidth = 0.01

(d)



MSE is almost 0. So, 𝑎𝑚𝑦𝑔𝑑𝑎𝑙𝑎 and 𝑎𝑐𝑐 are statistical independent.

(e)



(f)

