ISYE6740-HW3

pkubsad

February 2021

1. Order of faces using ISOMAP [50 points]

(a) (10 points) Visualize the similarity graph (you can either show the adjacency matrix, or similar to the lecture slides, visualize the graph using graph visualization packages such as Gephi (https://gephi.org) and illustrate a few images corresponds to nodes at different parts of the graph, e.g., mark them by hand or use software packages).

Answer:
I calculated the threshold epsilon such that every node has atleast 50 nieghbours. But this resulted in some nodes having close to 700 to 800 nodes. As discussed in piazza, https://piazza.com/class/khkx47q2d3p5ln?cid=356$_f$60.*Iswitchedbyimplementationtousekneighbor*

(a) Adjacency/Similarity matrix

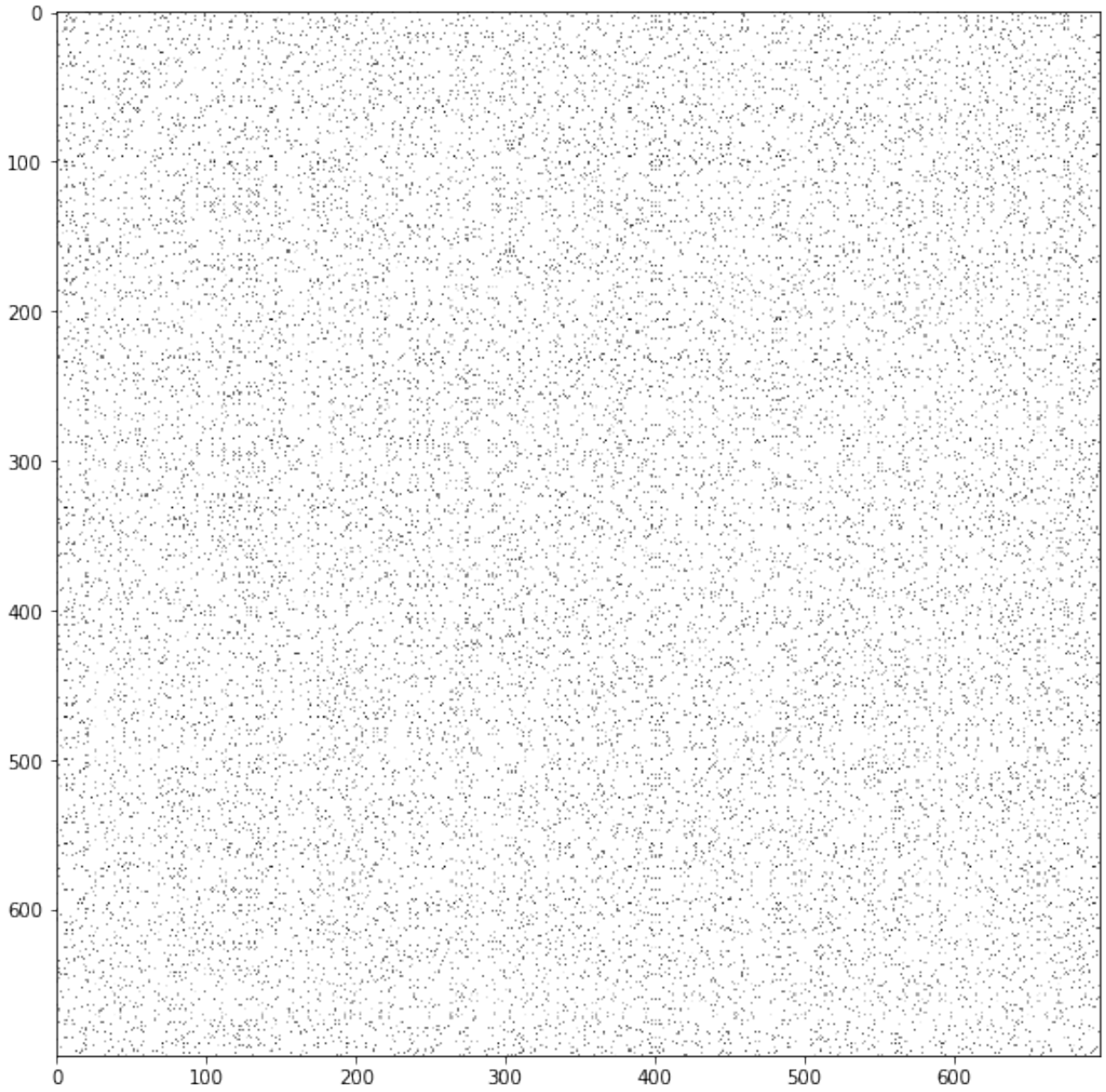

Figure 1: Adjacency/Similarity matrix

(b) Adjacency as network of graph

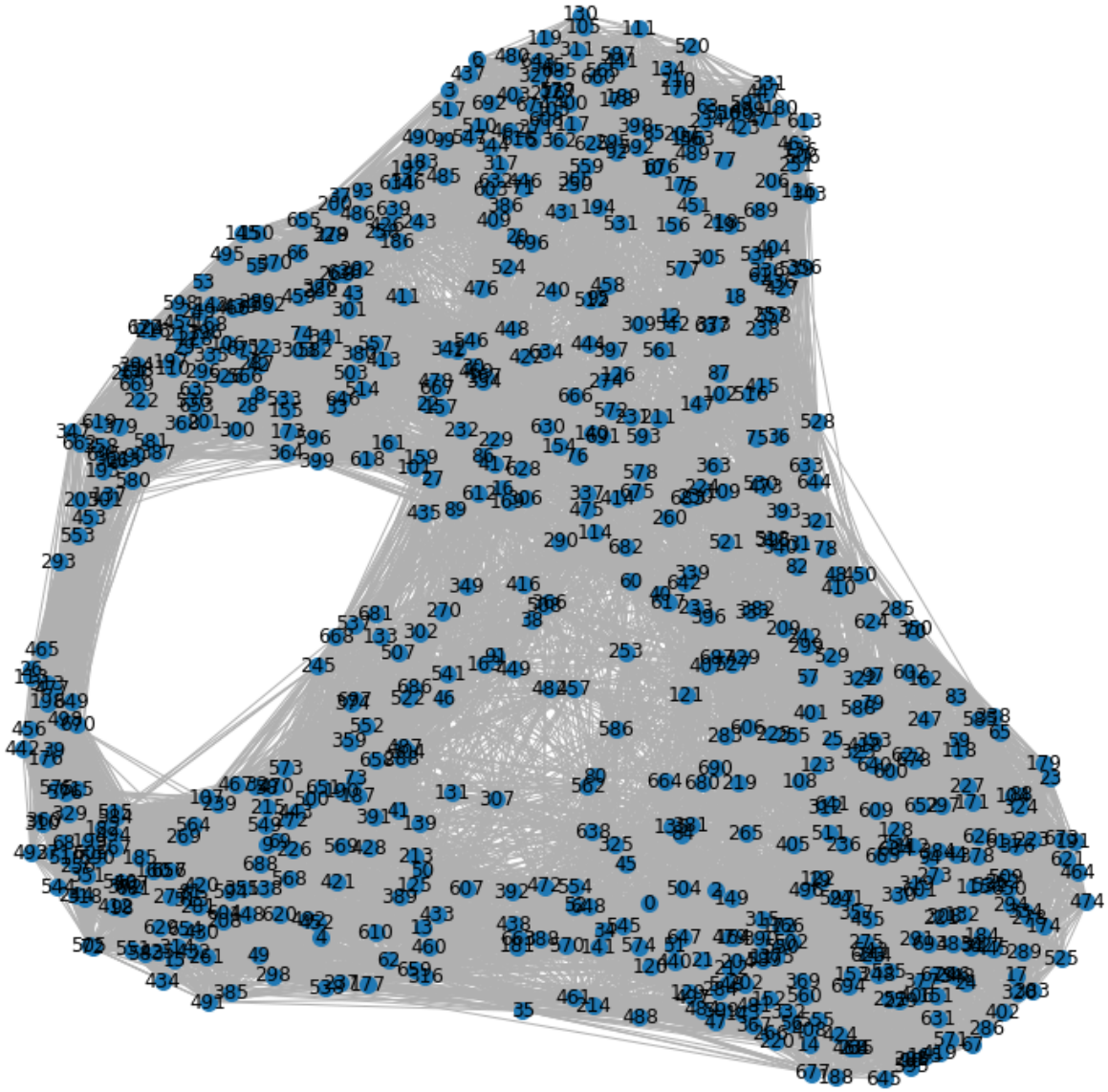

Figure 2: Adjacency matrix as network of graph

(c) Sample Images in adjacency graph,(I could not add the image in the graph, so am printing them separately

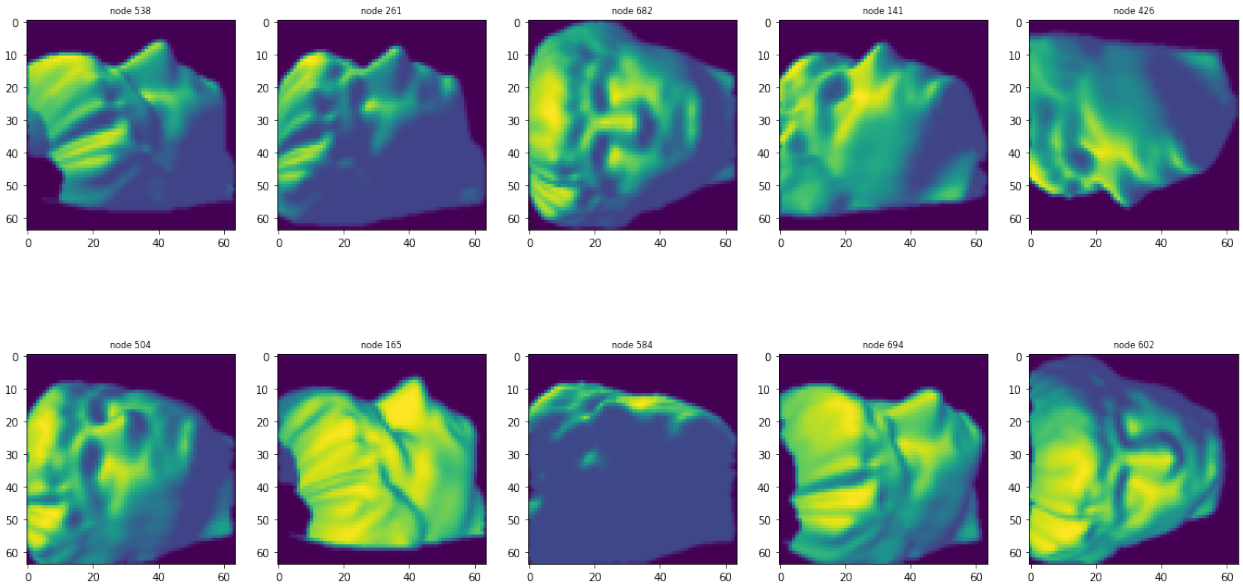

Figure 3: Images in adjacency graph

(b) (20 points) Implement the ISOMAP algorithm yourself to obtain a two-dimensional low-dimensional embedding. Plot the embeddings using a scatter plot, similar to the plots in lecture slides. Find a few images in the embedding space and show what these images look like. Comment on do you see any visual similarity among them and their arrangement, similar to what you seen in the paper?

Answer:
Observations looking at the output graph and running the program again and again:

(a) The faces pointing in a particular direction are grouped together.

(b) Faces looking in left direction are at the bottom of the graph.

(c) Faces looking in the right direction are at the top of the graph.

(d) Faces looking up are to the left, looking down are to the right of the graph.

(e) Everytime we run the program, the positions of the faces migth change, but they are grouped together.

The network graph I obtained for isomap with euclidean distance:

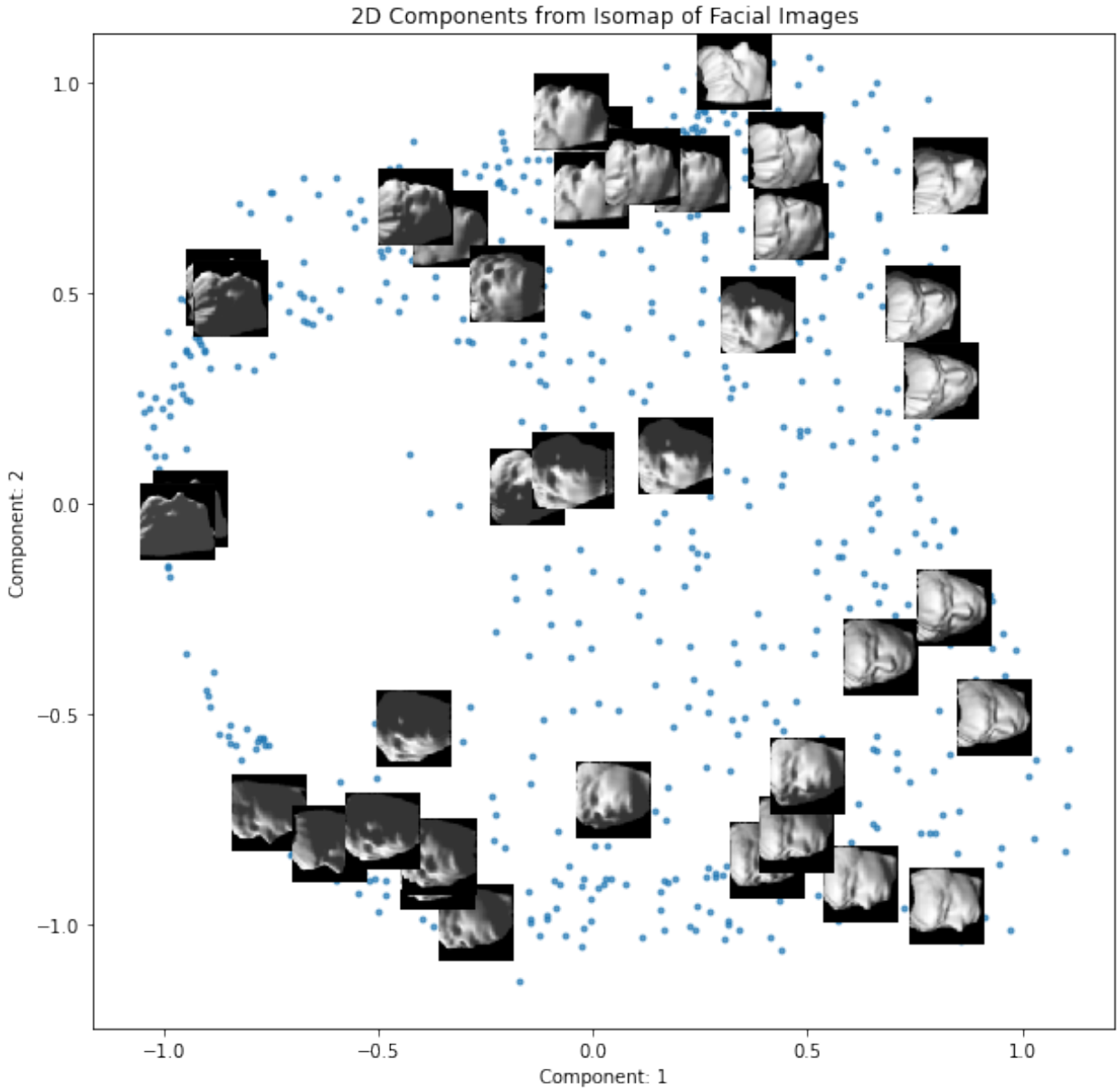

Figure 4: Isomap - euclidean distance

(c) (10 points) Now choose ℓ_1 distance (or Manhattan distance) between images (recall the definition from "Clustering" lecture)). Repeat the steps above. Use ϵ-ISOMAP to obtain a $k = 2$ dimensional embedding. Present a plot of this embedding. Do you see any difference by choosing a different similarity measure by comparing results in Part (b) and Part (c)?

Answer:
Observations:

(a) The resulting network graph is somewhat similar to euclidean graph.

(b) I see few faces grouped in the incorrect groups compared to euclidean distance.

The network graph I obtained for isomap with manhattan distance:

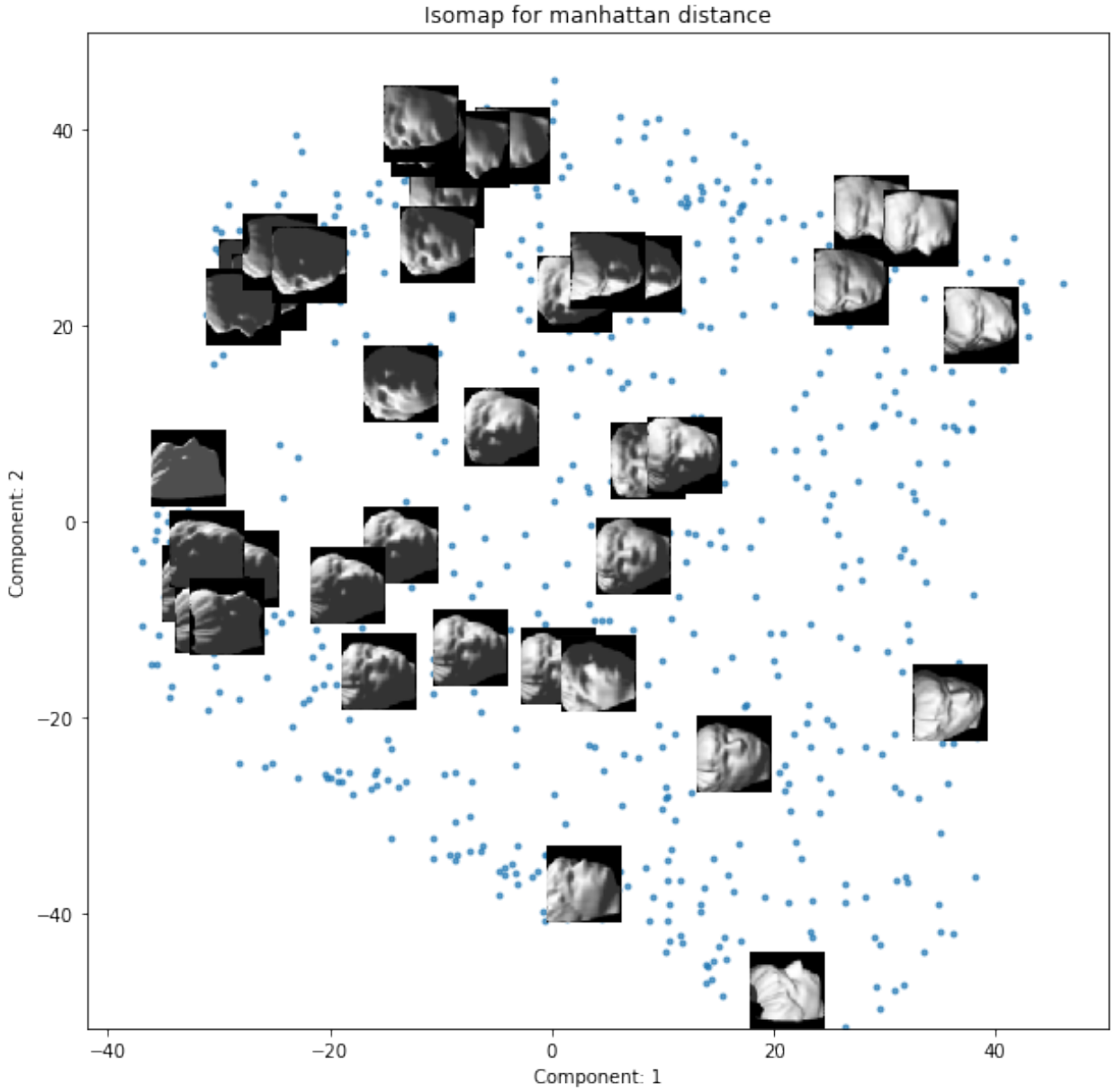

Figure 5: Isomap - manhattan distance

(d) (10 points) Perform PCA (you can now use your implementation written in Question 1) on the images and project them into the top 2 principal components. Again show them on a scatter plot. Explain whether or you see a more meaningful projection using

ISOMAP than PCA.

Answer:
I have run the PCA algorithm on the distances matrix calculated by useing cdist function for euclidean distance.

(a) PCA model, confuses quite a few faces into wrong categories compared to ISOMAP.

(b) The groupings can be logically identified as faces poinitng in certain direction, but there accuracy seems low.

(c) From the newtork graph, clearly ISOMAP has more meaningful projections compared to PCA.

Resulting network graph:

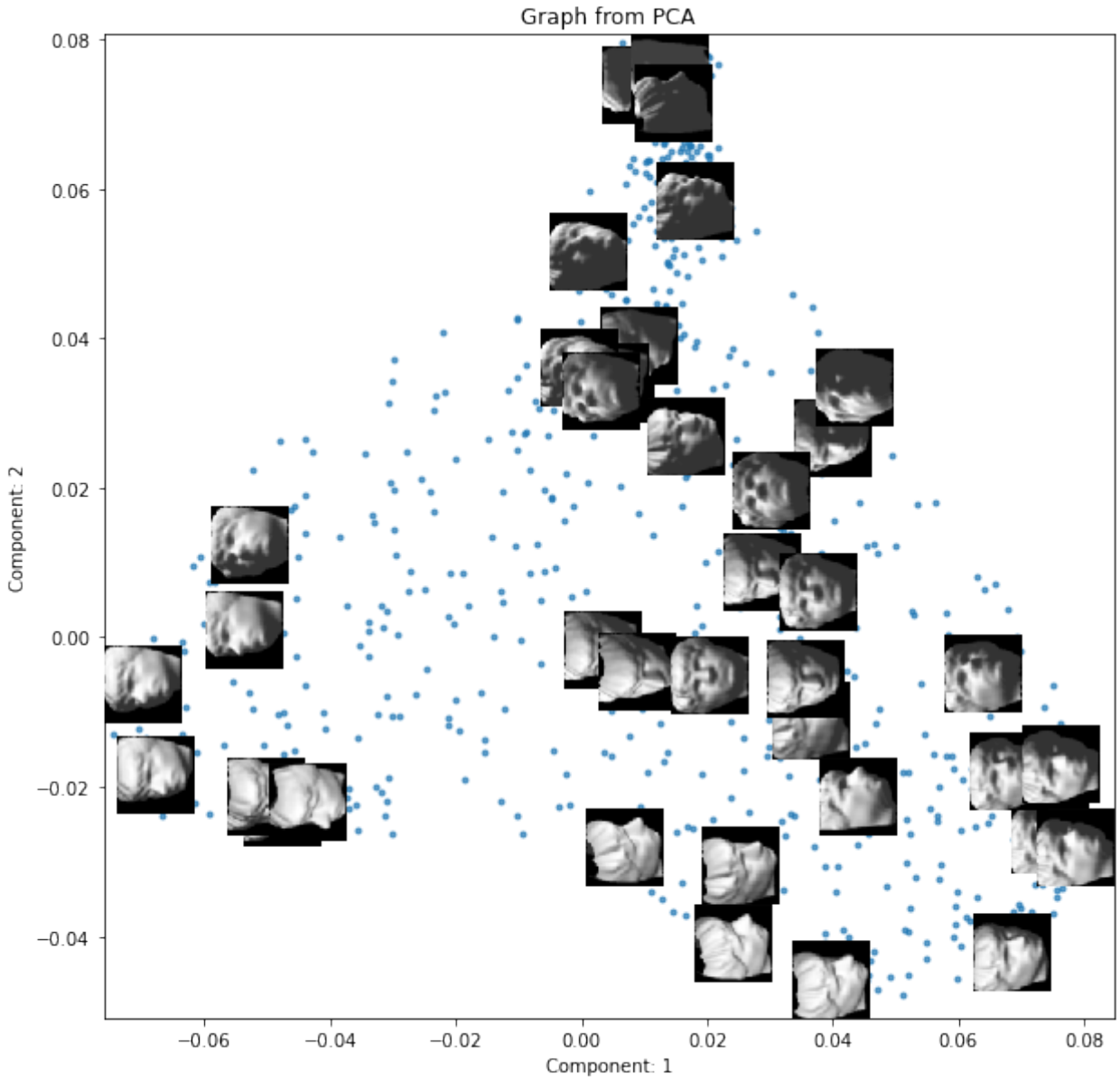

Figure 6: PCA - euclidean distance

2. Density estimation: Psychological experiments. (50 points)

(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. Decide on a suitable number of bins so you can see the shape of the distribution clearly.

Answer:

For Histogram: I ran the plot for various bin size. I have used seaborn package's hist-plot method to plot the histogram. This method takes in bin width instead of size. Calculation I used for bin width:

$$bin\_width = (maxValueInData - minValueInData)/number\_of\_bins$$

Looking at the plots, we can conclude the number of bins = 10 produces a better distribution plot.

For KDE plot, am using seaborn packages's kdeplot method. I calculated the factor 'h' value using the formula provided by prof:

$$h \approx 1.06\hat{\sigma}n^{-1/5},$$

But this was resulting in plots not consistent with histograms. When I dug deeper into implementation of the seaborn package, the bandwidth parameter accounts the std-deviation part of the above equation by itself. We have to provide the rest of the value. I verified this by calculating the default co-varaince factor that will be used if we dont provide any value.

band width factor calculated = (n ** (-1/5))
band width factor calualted = 90 ** (-1/5)= 0.4065851364889782

band width factor caluclated by gaussian kde library method = 0.4065851364889782

band width used by the method = bandwidth factor / std deviation of the data
band width used by the method = 0.4065851364889782 / 0.020321534068902875
band width used by the method = 0.008262433703070296

calculated by formula prof provided
h $\approx 1.06\hat{\sigma}n^{-1/5}$
$h \approx 1.06\hat{0}.001984299693119354 * 90^{-1/5}$,
$h \approx 0.00085519396$

Below are the histogram plots and kde plots for 1d:

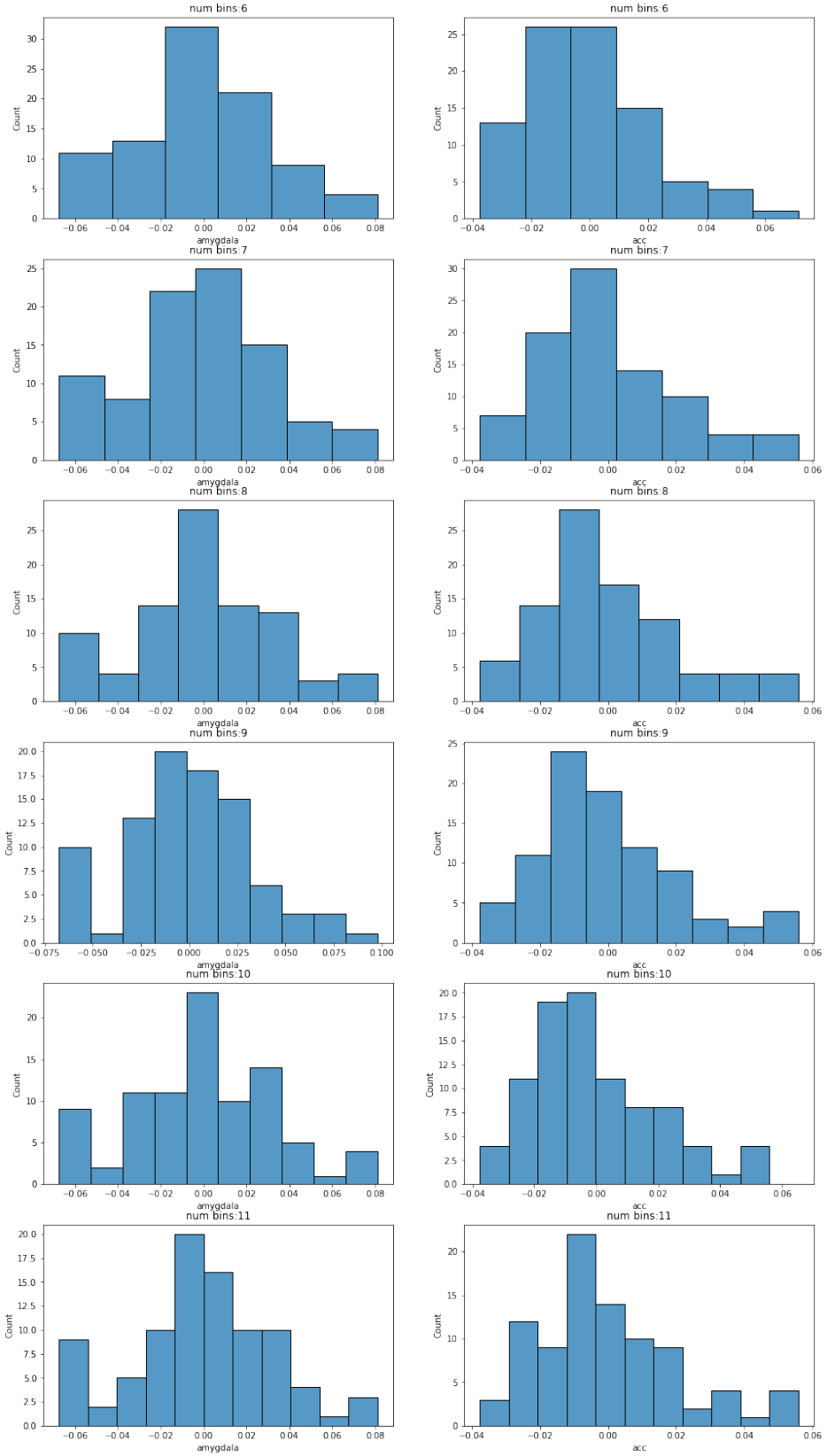

Figure 7: 1d Histogram

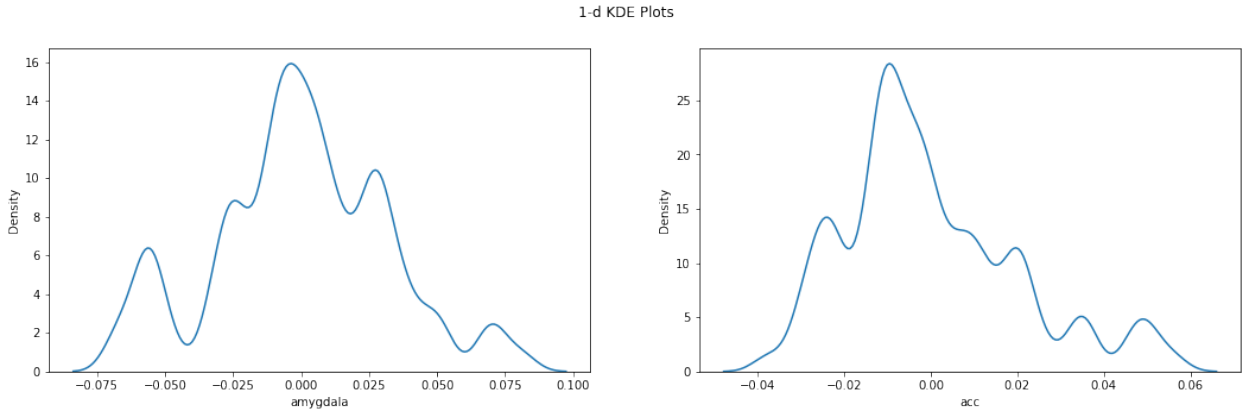

Figure 8: 1d KDE

(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.

Answer:
I have printed the heat map and 3d projection of 2 dimensional histogram. Based on the 3d projection of the bins, we can say bin size of 10 is appropriate value for this data.

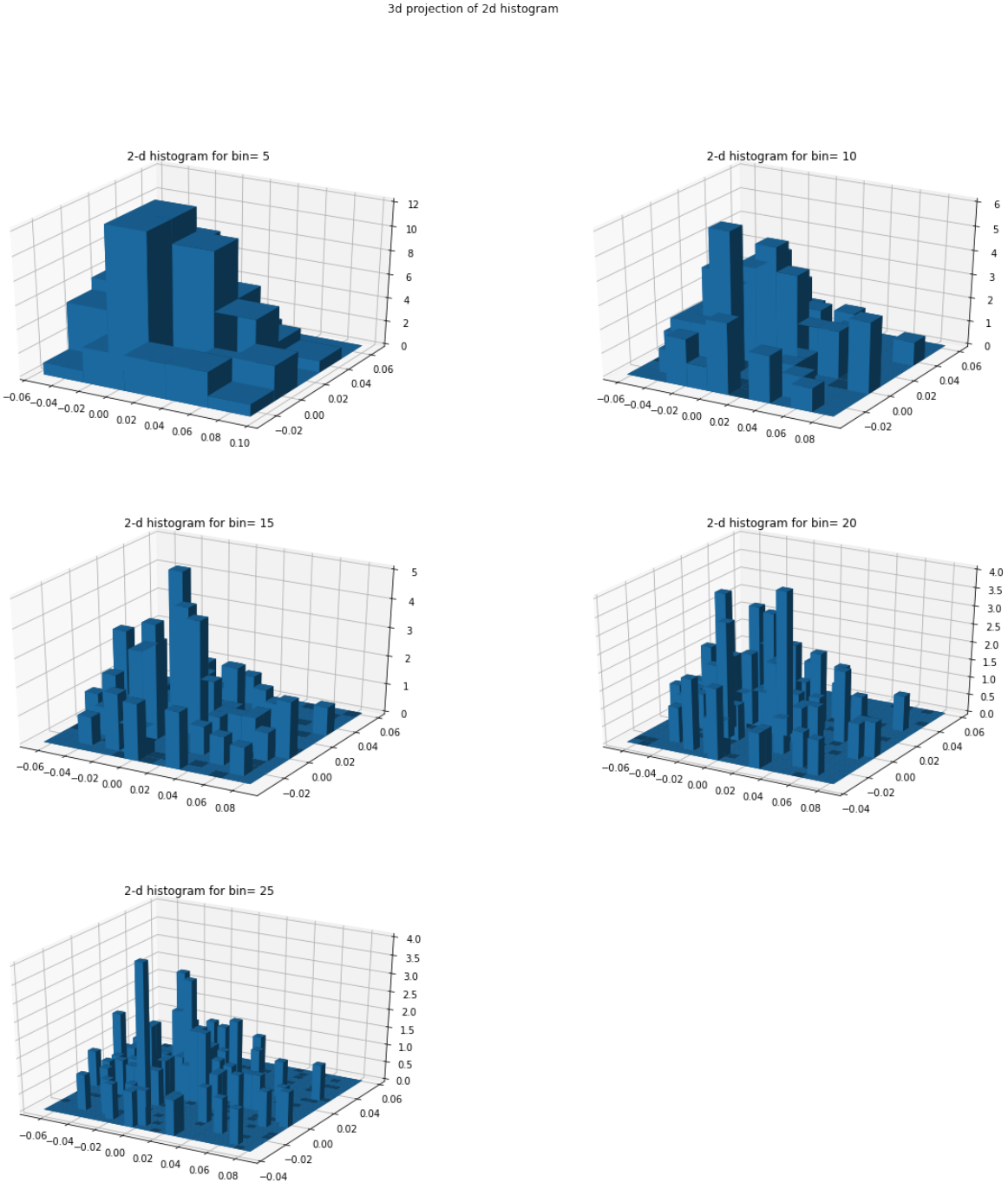

Figure 9: 3d projection of 2d histogram

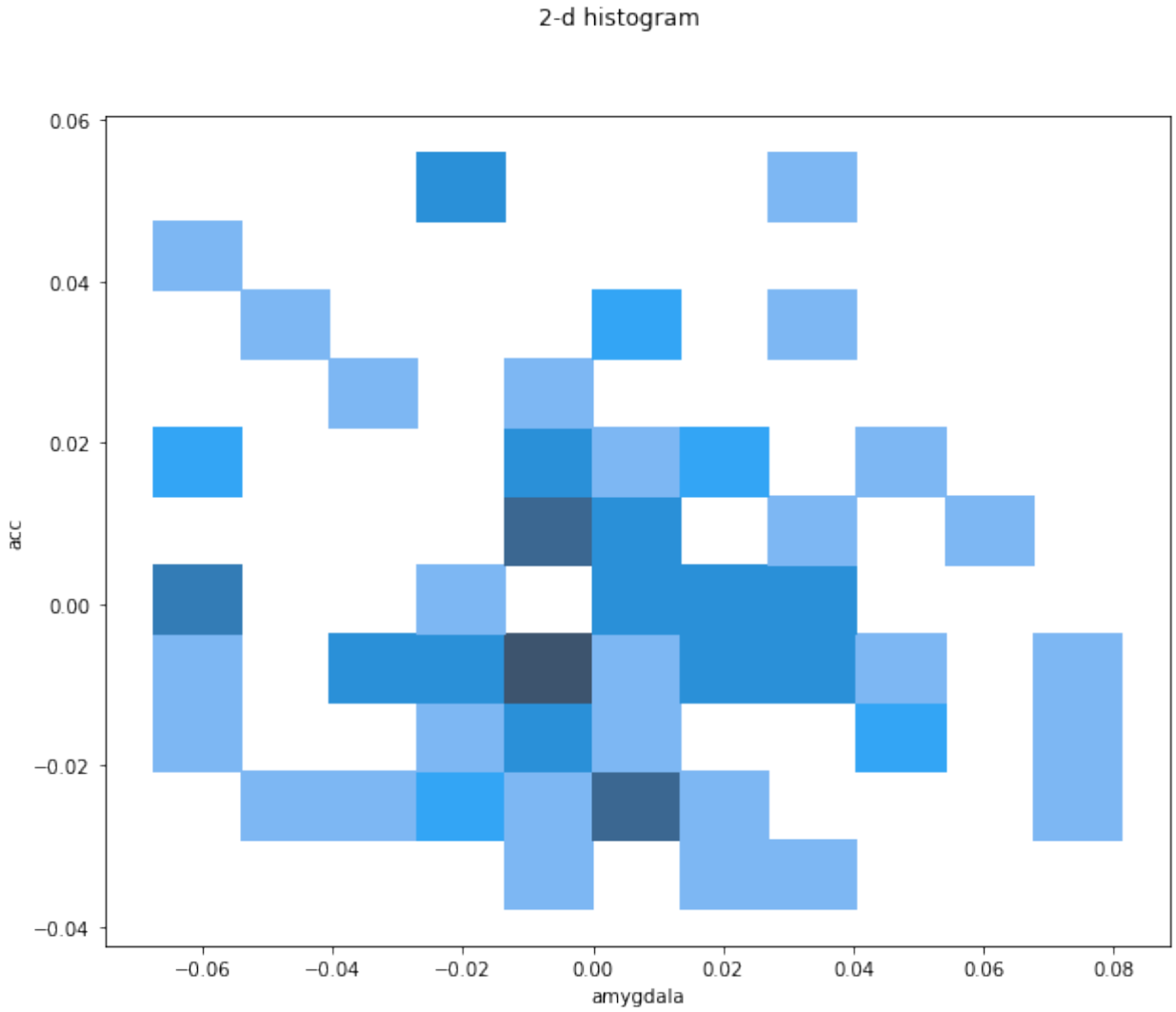

Figure 10: Heat Map of 2d histogram

(c) (10 points) 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 $h > 0$.

Answer:
Ran the seaborn package's kdeplot on 2 dimensions with different bandwidth factors. The bandwidth with good contours were obtained at bw factor =0.5. When I checked the default implementation of method in the package, it uses scott method which is n**(-1/d+4) = n**(1/6) = 0.4723815372

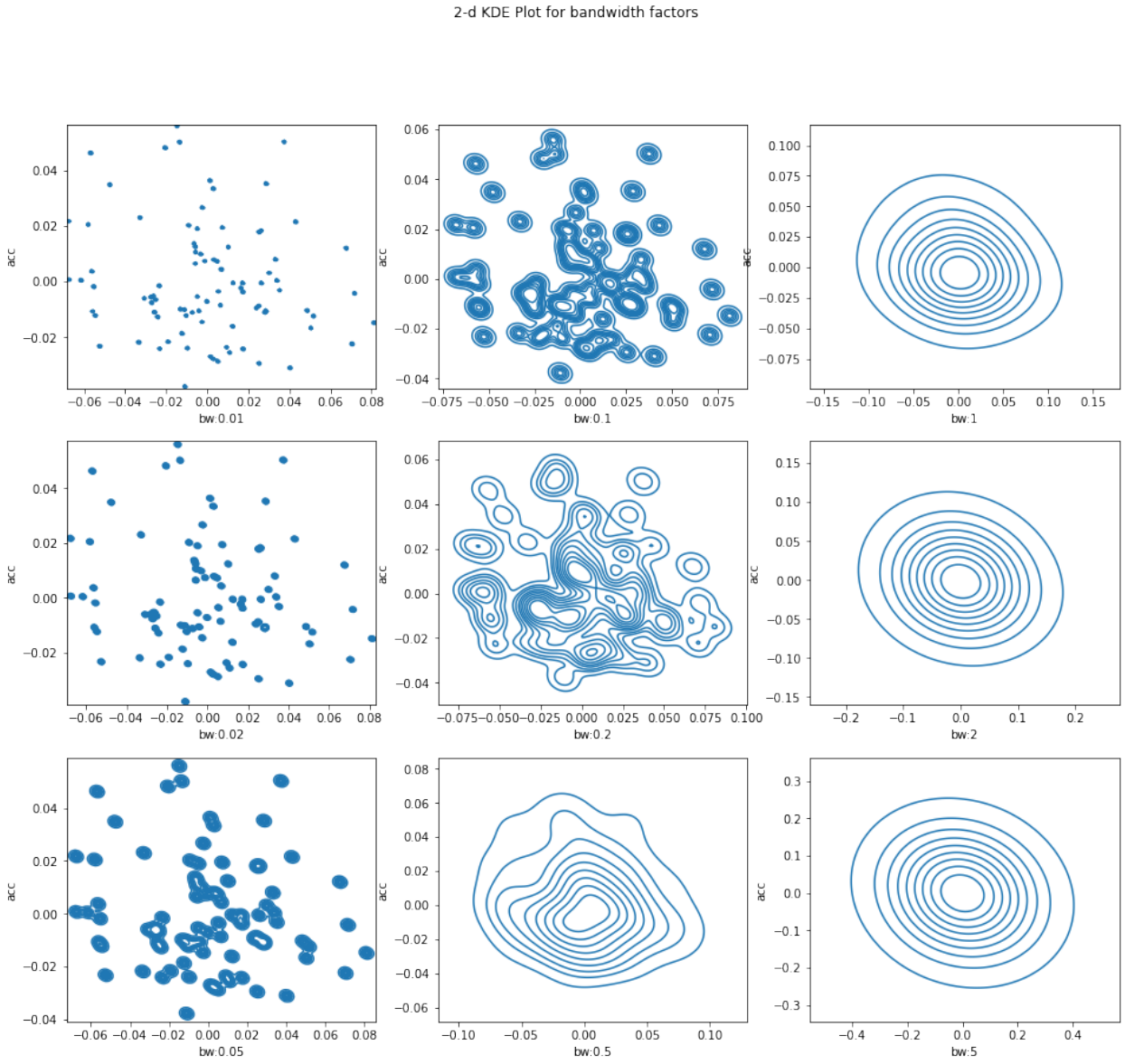

Figure 11: 2d kde

I have calculated p(amygdala,acc) - p(amygdala)*p(acc). If the variables are independent this factor will be remaining closer to zero. But we observe that this factor is fluctuating well above zero. Hence we can conclude that these 2 variables are not independent.

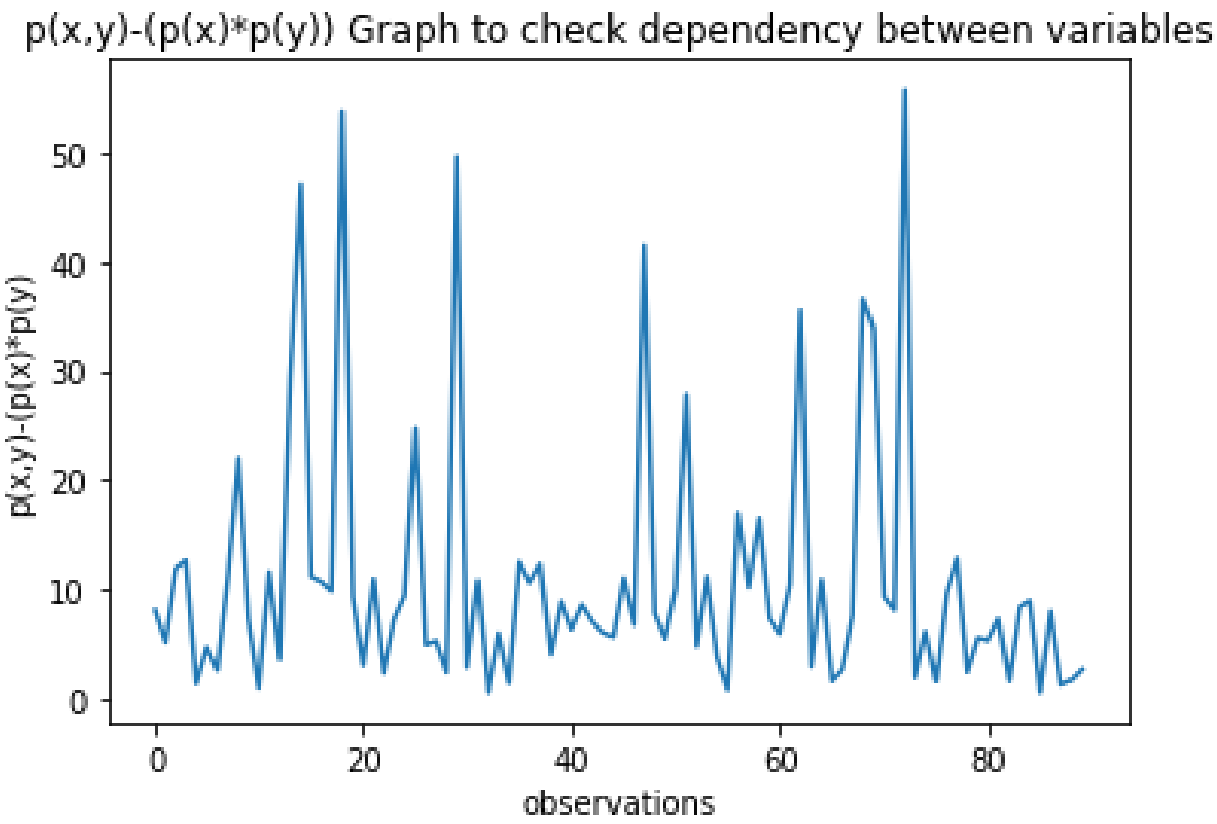

Figure 12: p(x,y)-p(x)*p(y)

(10 points) We will consider the variable orientation and consider conditional distributions. Please plot the estimated conditional distribution of amygdala conditioning on political orientation: $p(\text{amygdala}|\text{orientation} = c)$, $c = 2, \ldots, 5$, using KDE. Set an appropriate kernel bandwidth $h > 0$. Do the same for the volume of the acc: plot $p(\text{acc}|\text{orientation} = c)$, $c = 2, \ldots, 5$ using KDE. (Note that the conditional distribution can be understood as fitting a distribution for the data with the same orientation. Thus you should plot 8 one-dimensional distribution functions in total for this question.)

Answer:

The sample mean calculated for all the Cs:

C	2	3	4	5
Amygdala	0.01906153846153846	0.0005875	-0.004719512195121951	-0.0056916666666
ACC	-0.014769230769230769	0.0016708333333	0.0013097560975609756	0.0081416666666

Even though the means for these orientations are close to each other, it seems like the the data set is similar to for all orientations. But if we look at the conditional distributions below, we can see that distributions for these conditional orientations are different. This

tells that apart from the standard statistical measures like mean, distribution of the data gives full picture of the nature of the data.

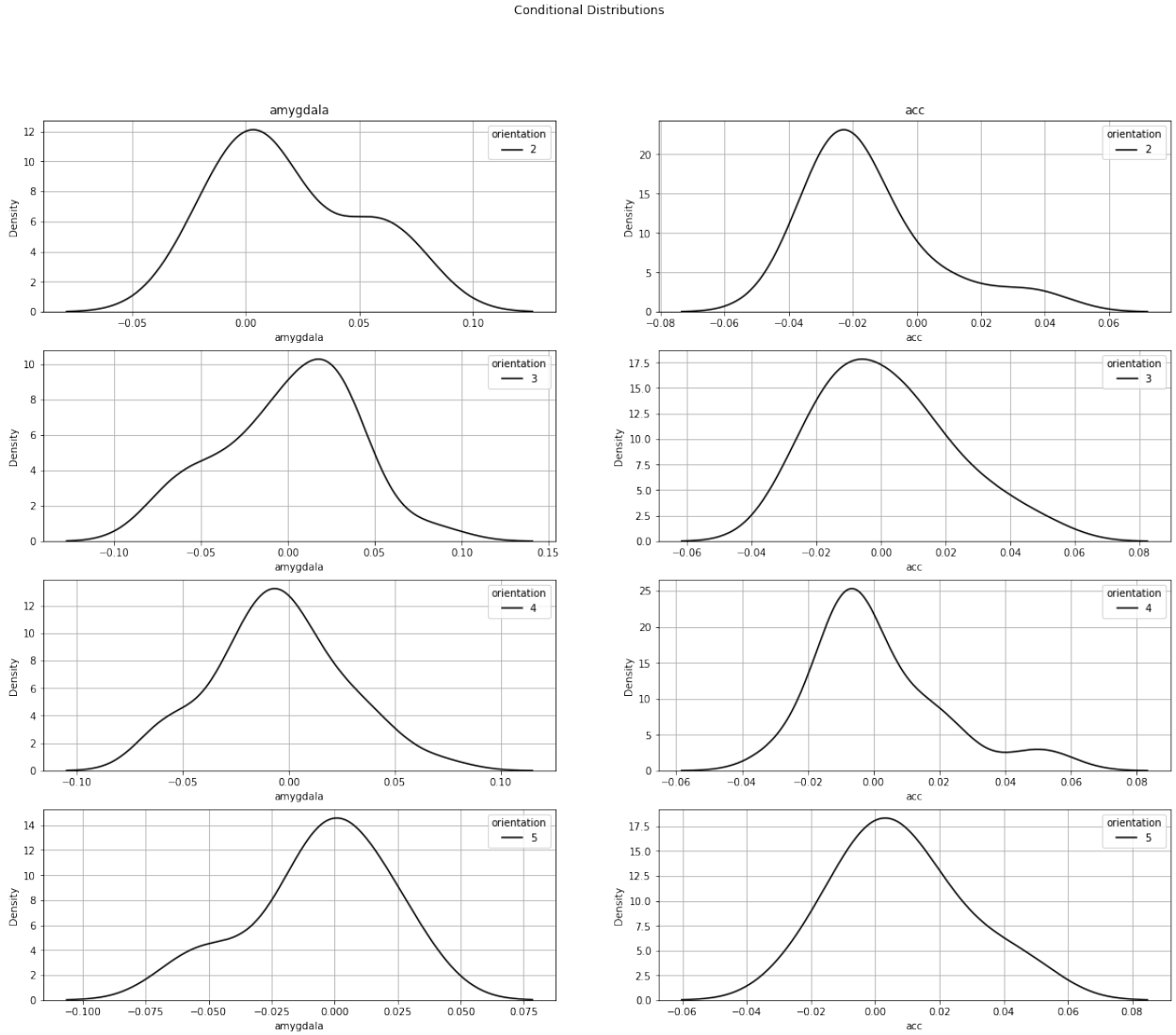

Figure 13: conditional distributions

(10 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: $p(\text{amygdala}, \text{acc}|\text{orientation} = c)$, $c = 2, \ldots, 5$. You will use two-dimensional KDE to achieve the goal; et an appropriate kernel bandwidth $h > 0$. Please show the two-dimensional KDE (e.g., two-dimensional heat-map, two-dimensional contour plot, etc.).

Answer: Based on the distribution plots we can conclude that:

1. Folks with orientation 2 have almost opposite amygdala and acc values compared to

orientation 5. This can be observed when we look at denser distribution point for o=2 is around (-0.02, -.025) and for o=5 is (0.01, 0.025).

2. The distribution of orientation 3 and 4 are somewhat similar at the denser parts. Both have concentrations around (0,0). The outliers can drag the distributions in the farther quartiles.

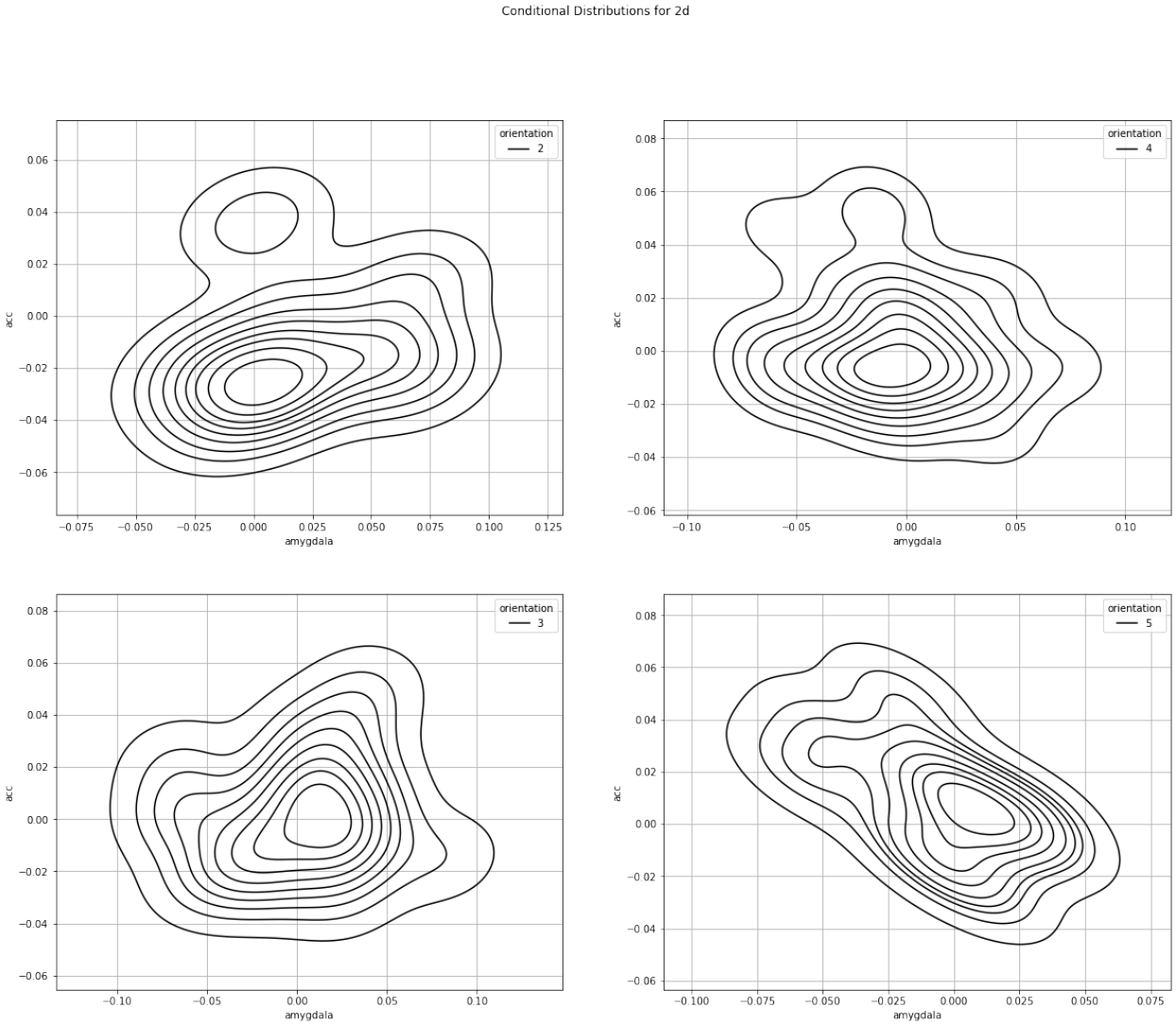

Figure 14: conditional distributions for 2d