ISYE 6740, Fall 2024, Homework 3

100 points + 10 bonus points

Prof. Yao Xie

1. Conceptual questions. [20 points]

1. (5 points) Please compare the pros and cons of KDE over histogram, and give at least one advantage and disadvantage to each.

   Answer:
   Histogram: it is more efficient than KDE but when the sample size is small, the estimates is noisy because the outputs depend on how you put your bins.
   KDE: it is more smoother than histogram but KDE can be computationally more expensive and memory inefficient compared with histograms

2. (5 points) Why you cannot use maximum likelihood estimation to directly estimate GMM? Then how to estimate the model of GMM?

   Answer:
   In GMM, each data point belongs to one of the Gaussian components, but this is hidden, which makes it hard to know which component a data point belongs to.
   We can use expectation maximization algorithm to estimate the parameters of GMM.

3. (5 points) For the EM algorithm for GMM, please show how to use the Bayes rule to drive τ_k^i in a closed-form expression.

   Answer:
   For each data point x^i, compute $p(z^i = k|x^i)$ for each k:

$$\tau_k^i = p(z^i = k|x^i, \theta^t) = \frac{P(z^i = k, x^i)}{P(x^i)} = \frac{p(x^i|z^i = k)p(z^i = k)}{\sum_{k'=1..K} p(z^i = k', x^i)}$$
$$= \frac{P(z^i = k)P(x^i|z^i = k)}{\sum_{k'=1..K} P(z^i = k')P(x^i|z^i = k')} = \frac{\pi_k \mathcal{N}(x^i|\mu_k, \Sigma_k)}{\sum_{k'=1..K} \pi_{k'} \mathcal{N}(x^i|\mu_{k'}, \Sigma_{k'})}$$

4. (5 points) Explain how to choose the kernel bandwidth for KDE?

   Answer:
   Silverman's rule of thumb method is robust but doesn't work well in complicated cases especially for data with multiple modes or skewed distributions.
   Cross-validation method maximizes the likelihood of the KDE fitting the data, it is more accurate but computationally expensive.

2. Density estimation: Psychological experiments. [40 points]

In *Kanai, R., Feilden, T., Firth, C. and Rees, G., 2011. Political orientations are correlated with brain structure in young adults. Current biology, 21(8), pp.677-680.*, data are collected to study whether or not the two brain regions are likely to be independent of each other and considering different types of political view **For this question; you can use third party histogram and KDE packages; no need to write your own.** The data set n90pol.csv contains information on 90 university students who participated in a psychological experiment designed to look for relationships between the size of different regions of the brain and political views. The variables amygdala and acc indicate the volume of two particular brain regions known to be involved in emotions and decision-making, the amygdala and the anterior cingulate cortex; more exactly, these are residuals from the predicted volume, after adjusting for height, sex, and similar body-type variables. The variable orientation gives the students' locations on a five-point scale from 1 (very conservative) to 5 (very liberal). Note that in the dataset, we only have observations for orientation from 2 to 5.

Recall in this case, the kernel density estimator (KDE) for a density is given by

$$p(x) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{h}K\left(\frac{x^i - x}{h}\right),$$

where x^i are two-dimensional vectors, $h > 0$ is the kernel bandwidth, based on the criterion we discussed in lecture. For one-dimensional KDE, use a one-dimensional Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

For two-dimensional KDE, use a two-dimensional Gaussian kernel: for

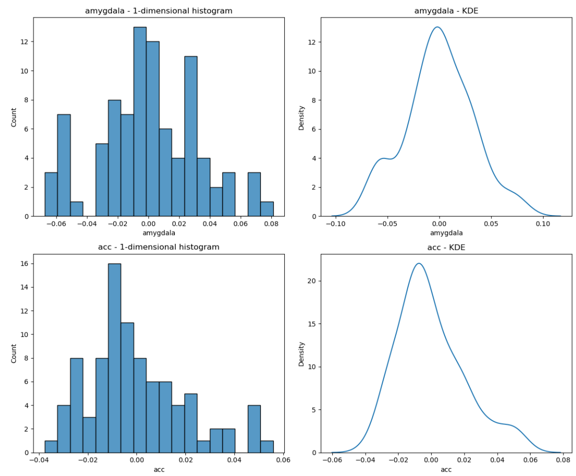
$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2,$$

where x_1 and x_2 are the two dimensions respectively

$$K(x) = \frac{1}{2\pi}e^{-\frac{(x_1)^2+(x_2)^2}{2}}.$$

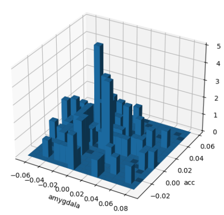
1. (5 points) Form the 1-dimensional histogram and KDE to estimate the distributions of amygdala and acc, respectively. For this question, you can ignore the variable orientation. Decide on a suitable number of bins so you can see the shape of the distribution clearly. Set an appropriate kernel bandwidth $h > 0$.

   Answer:
   The number of bins is 18, which visually arrive at a balance between noisy overfitting and over-smoothing.



2. (5 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.

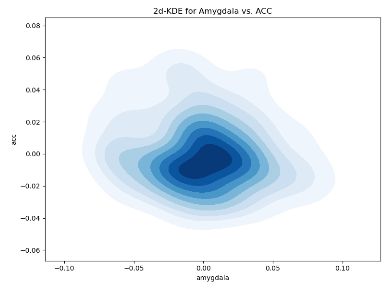
   Answer:
   I used the same number of bins as the first question. As it shows a clear shape of the distribution. I also tried several different bins, when the number is small, it doesn't show a clear plot. When the number is large, it may casue oevrfitting.



3. (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 $h > 0$.

   Please show the two-dimensional KDE (e.g., two-dimensional heat-map, two-dimensional contour plot, etc.)

   Answer:



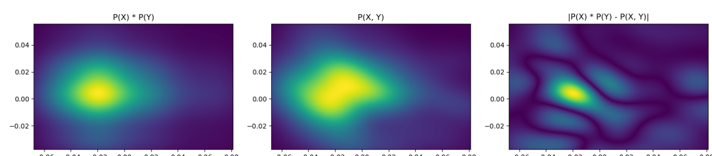
Please explain what you have observed: is the distribution unimodal or bi-modal? Are there any outliers?

Answer:
It is a bi-modal, as there are two peaks in the the center of the plot. Also, there are some outliers which make the plot not very regular and centrosymmetric.

Are the two variables (amygdala, acc) likely to be independent or not? Please support your argument with reasonable investigations.

Answer:

$$X \perp Y \quad \text{iff} \quad P(X = x, Y = y) = P(X = x)P(Y = y), \quad \forall x, y$$



I used scipy.stats gaussian_kde to get the kernel density estimation. The first plot is for the $P(x)p(y)$, the second plot is for $P(x, y)$ and the last plot is for the absolute difference between them. In the last plot, we can see that the different between the first two is obvious, which means the two variables are not independent.

4. (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.)

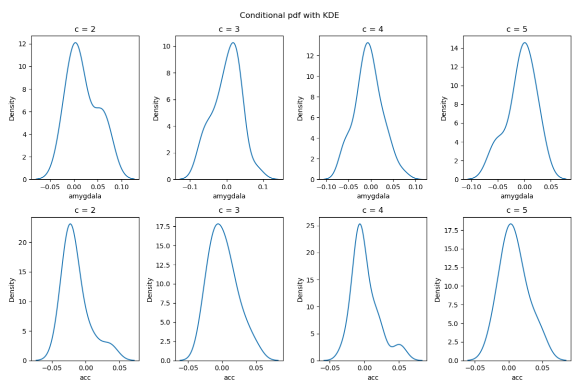
   Now please explain based on the results, can you infer that the conditional distribution of amygdala and acc, respectively, are different from $c = 2, \ldots, 5$? This is a type of scientific question one could infer from the data: Whether or not there is a difference between brain structure and political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

Now please also fill out the *conditional sample mean* for the two variables:

	$c = 2$	$c = 3$	$c = 4$	$c = 5$
amygdala	0.019088	0.000588	-0.00472	-0.005692
acc	-0.014769	0.001671	0.00131	0.008142

Remark: As you can see this exercise, you can extract so much more information from density estimation than simple summary statistics (e.g., the sample mean) in terms of explorable data analysis.

Answer:

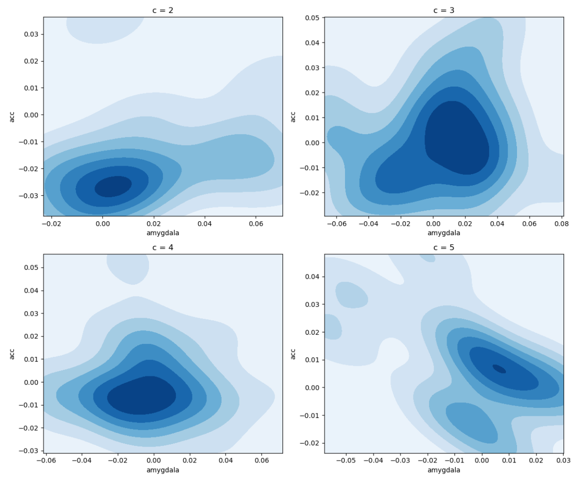


I choose bandwidth as 1 because they show me smoothing plots and do not focus more on detail. Based on the conditional distribution of amygdala and acc, which shows some differences across the different political orientations. The shape and peak of the distributions suggest that political orientation might have an influence on the structure of the amygdala and acc.

5. (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.).

   Please explain based on the results, can you infer that the conditional distribution of two variables (amygdala, acc) are different from $c = 2, \ldots, 5$? 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.

Answer:
I tried several bandwidths and choose a small one instead of the default 1 which can capture details of the joint distribution between amygdala and acc for each political orientation. Based on the plot I generate, the joint distributions of amygdala and acc appear to differ across different political orientations.



3. Implementing EM for MNIST dataset. [40 points]

Implement the EM algorithm for fitting a Gaussian mixture model for the MNIST handwritten digits dataset. For this question, we reduce the dataset to be only two cases, of digits "2" and "6" only. Thus, you will fit GMM with $C = 2$. Use the data file data.mat or data.dat. True label of the data are also provided in label.mat and label.dat.

The matrix images is of size 784-by-1990, i.e., there are 1990 images in total, and each column of the matrix corresponds to one image of size 28-by-28 pixels (the image is vectorized; the original image can be recovered by mapping the vector into a matrix).

First use PCA to reduce the dimensionality of the data before applying to EM. We will put all "6" and "2" digits together, to project the original data into 4-dimensional vectors.

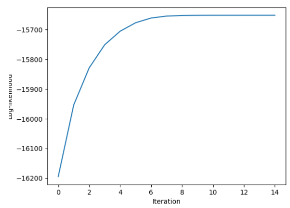
Now implement EM algorithm for the projected data (with 4-dimensions).
(In this question, we use the same set of data from the provided data files for training and testing)

1. (10 points) Implement EM algorithm yourself. Use the following initialization
   - initialization for mean: random Gaussian vector with zero mean
   - initialization for covariance: generate two Gaussian random matrix of size n-by-n: S_1 and S_2, and initialize the covariance matrix for the two components are

$\Sigma_1 = S_1 S_1^T + I_n$, and $\Sigma_2 = S_2 S_2^T + I_n$, where I_n is an identity matrix of size n-by-n.

Plot the log-likelihood function versus the number of iterations to show your algorithm is converging.

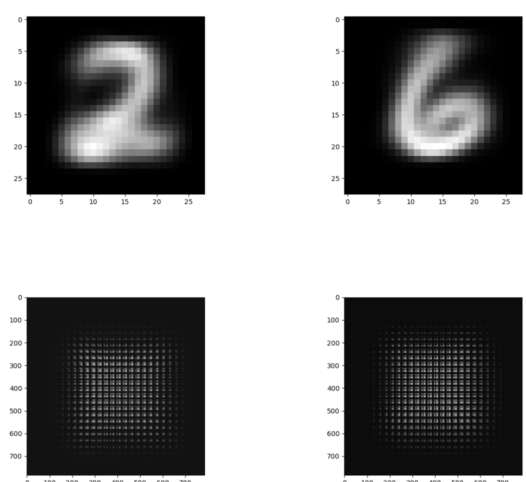
Answer:



2. (20 points) Report the fitted GMM model when EM has terminated in your algorithms as follows:

   - The numerical weights for each component
   - The mean of each component by mapping it back to the original space and reformatting the vectors into 28-by-28 matrices. These should be displayed as images, ideally corresponding to a kind of "average" of the images.
   - Two 4-by-4 covariance matrices by visualizing their intensities (i.e. a gray-scaled image or heatmap.)

   Answer:



3. (10 points) Use the τ_k^i to infer the labels of the images, and compare with the true labels. Report the mis-classification rate (1 - Accuracy) for digits "2" and "6" respectively. Perform K-means clustering with $K = 2$ (you may call a package or use the code from your previous homework). Find out the mis-classification rate for digits "2" and "6" respectively, and compare with GMM. Which one achieves the better performance?

   Answer:

   GMM mis-classification rate
   2: 7.41%
   6: 1.33%

   K-means mis-classification rate
   2: 6.2016%
   6: 7.9332%

   In conclusion, GMM has better performance.

4. De-bias review system using EM. [Bonus, 10 points]

In this question, we will develop an algorithm to remove individual reviewer's bias from their score. Consider the following problem. There are P papers submitted to a machine learning conference. Each of R reviewers reads each paper, and gives it a score indicating how good he/she thought that paper was. We let $x^{(pr)}$ denote the score that reviewer r gave to paper p. A high score means the reviewer liked the paper, and represents a recommendation from that reviewer that it be accepted for the conference. A low score means the reviewer did not like the paper.

We imagine that each paper has some "intrinsic" true value that we denote by μ_p, where a large value means it's a good paper. Each reviewer is trying to estimate, based on reading the paper, what μ_p is; the score reported $x^{(pr)}$ is then reviewer r's guess of μ_p.

However, some reviewers are just generally inclined to think all papers are good and tend to give all papers high scores; other reviewers may be particularly nasty and tend to give low scores to everything. (Similarly, different reviewers may have different amounts of variance in the way they review papers, making some reviewers more consistent/reliable than others.) We let ν_r denote the "bias" of reviewer r. A reviewer with bias ν_r is one whose scores generally tend to be ν_r higher than they should be.

All sorts of different random factors influence the reviewing process, and hence we will use a model that incorporates several sources of noise. Specifically, we assume that reviewers's scores are generated by a random process given as follows:

$$y^{(p)} \sim \mathcal{N}(\mu_p, \sigma_p^2)$$

$$z^{(r)} \sim \mathcal{N}(\nu_r, \tau_r^2)$$

$$x^{(pr)} | y^{(p)}, z^{(r)} \sim \mathcal{N}(y^{(p)} + z^{(r)}, \sigma^2).$$

The variables $y^{(p)}$ and $z^{(r)}$ are independent; the variables (x, y, z) for different paper-reviewer pairs are also jointly independent. Also, we only ever observe the $x^{(pr)}$s; thus, the $y^{(p)}$s and $z^{(r)}$s are all latent random variables.

We would like to estimate the parameters μ_p, σ_p^2, ν_r, τ_r^2. If we obtain good estimates of the papers "intrinsic values" μ_p, these can then be used to make acceptance/rejection decisions for the conference.

We will estimate the parameters by maximizing the marginal likelihood of the data $\{x^{(pr)}; p = 1, \ldots, P, r = 1, \ldots, R\}$. This problem has latent variables $y^{(p)}$s and $z^{(r)}$s, and the maximum likelihood problem cannot be solved in closed form. So, we will use EM.

Your task is to derive the EM update equations. For simplicity, you need to treat only $\{\mu_p, \sigma_p^2; p = 1 \ldots, P\}$ and $\{\nu_r, \tau_r^2; r = 1 \ldots R\}$ as parameters, i.e. treat σ^2 (the conditional variance of $x^{(pr)}$ given $y^{(p)}$ and $z^{(r)}$) as a fixed, known constant.

1. Derive the E-step (5 points)

   1.1. The joint distribution $p(y^{(p)}, z^{(r)}, x^{(pr)})$ has the form of a multivariate Gaussian density. Find its associated mean vector and covariance matrix in terms of the parameters μ_p, σ_p^2, ν_r, τ_r^2 and σ^2. [Hint: Recognize that $x^{(pr)}$ can be written as $x^{(pr)} = y^{(p)} + z^{(r)} + \epsilon^{(pr)}$, where $\epsilon^{(pr)} \sim \mathcal{N}(0, \sigma^2)$ is independent Gaussian noise.

   1.2. Derive an expression for $Q_{pr}(\theta'|\theta) = \mathbb{E}[\log p(y^{(p)}, z^{(r)}, x^{(pr)})|x^{(pr)}, \theta]$ using the conditional distribution $p(y^{(p)}, z^{(r)}|x^{(pr)})$ (E-step) (Hint, you may use the rules for conditioning on subsets of jointly Gaussian random variables.)

2. (5 points) Derive the M-step to update the parameters μ_p, σ_p^2, ν_r, and τ_r^2. [Hint: It may help to express an approximation to the likelihood in terms of an expectation with respect to $(y^{(p)}, z^{(r)})$ drawn from a distribution with density $Q_{pr}(y^{(p)}, z^{(r)})$.]

Remark: John Platt (whose SMO algorithm you've seen) implemented a method quite similar to this one to estimate the papers' true scores. (There, the problem was a bit more complicated because not all reviewers reviewed every paper, but the essential ideas are the same.) Because the model tried to estimate and correct for reviewers' biases, its estimates of the paper's value were significantly more useful for making accept/reject decisions than the reviewers' raw scores for a paper.