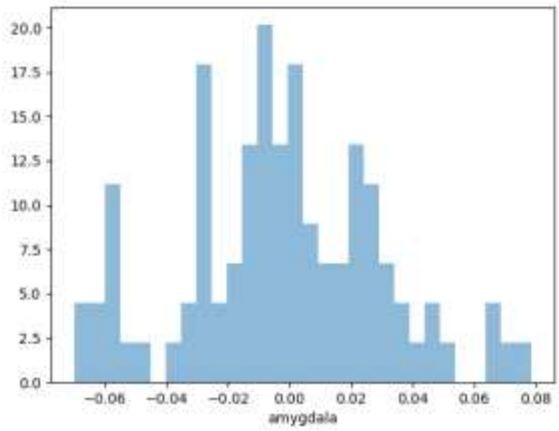
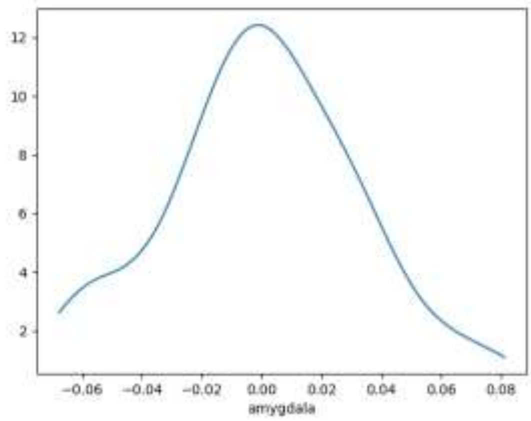
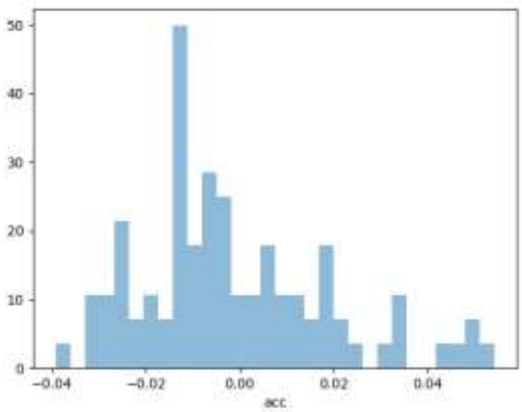
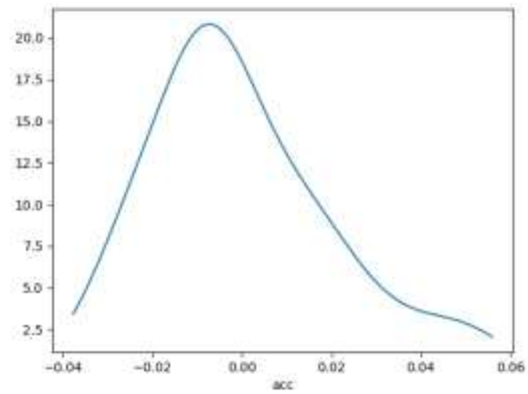
Name: Jianyuan Lu GT ID: 903633385

1.Density estimation

(a)

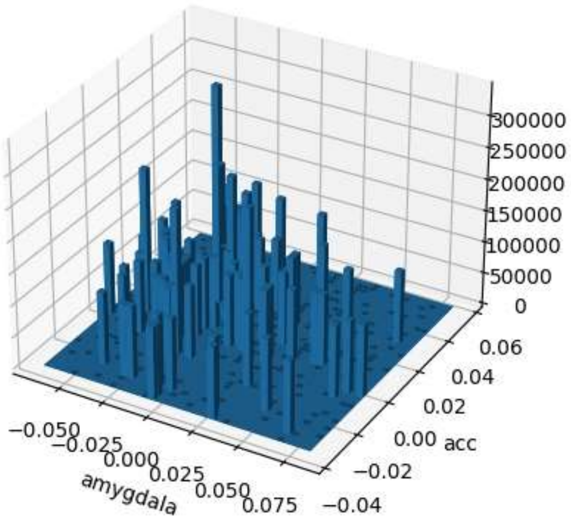
I use the rules of thumb to decide the bandwidth for KDE. The graph is shown as below:

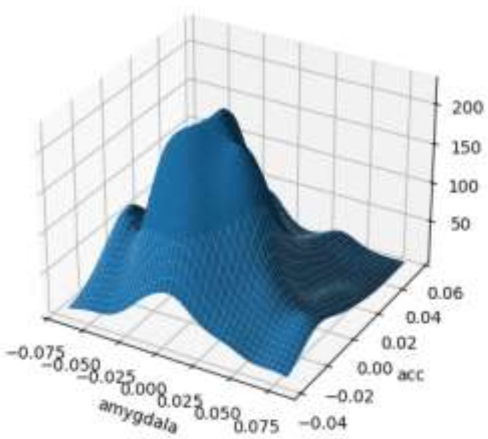
Amygdala	
	

Acc	
	

(b)

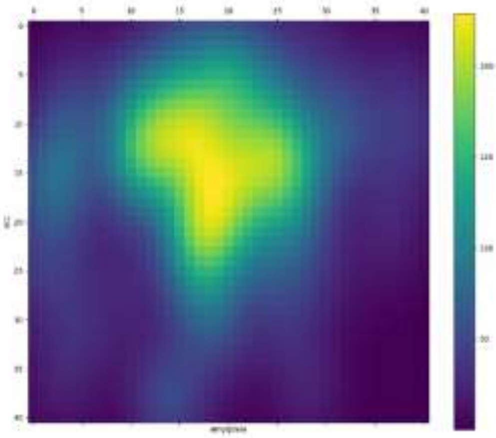
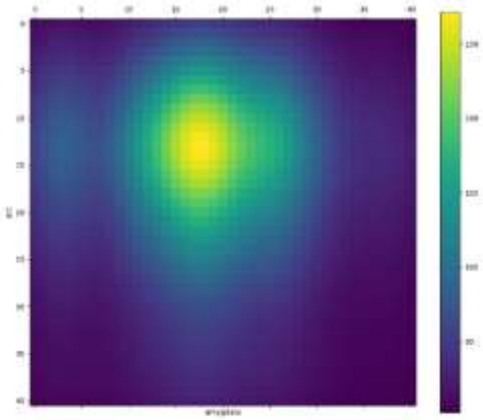
I set bandwidth=0.01 for KDE. The graph is shown as below:



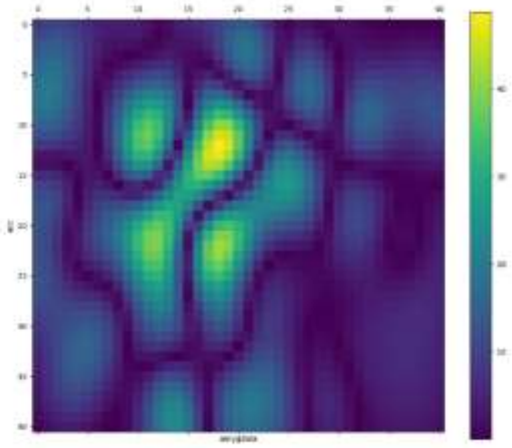


(c)

I set bandwidth=0.01 both for joint distribution and marginal distribution in KDE. The graph is shown as below:

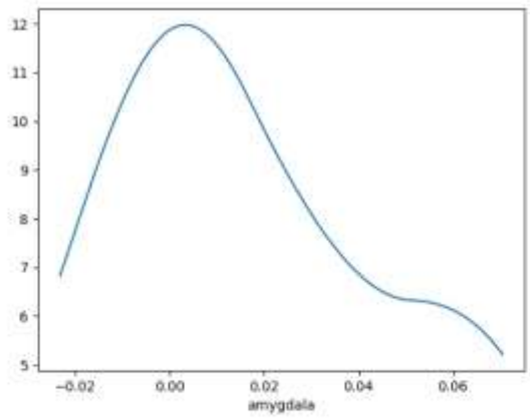
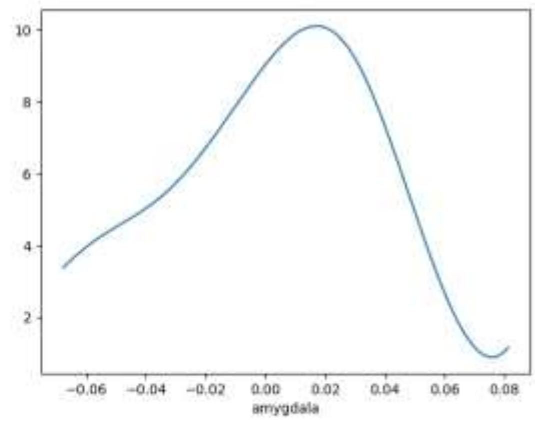
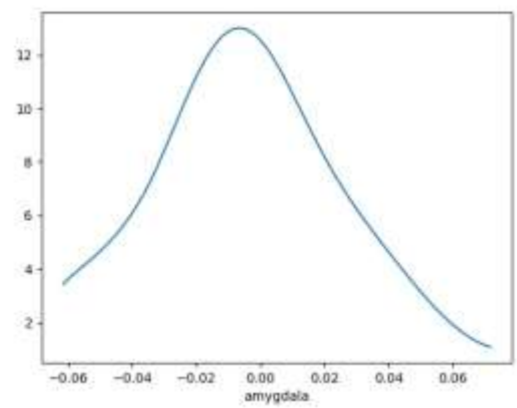
$p(amygdala, acc)$	$p(amygdala)p(acc)$
	

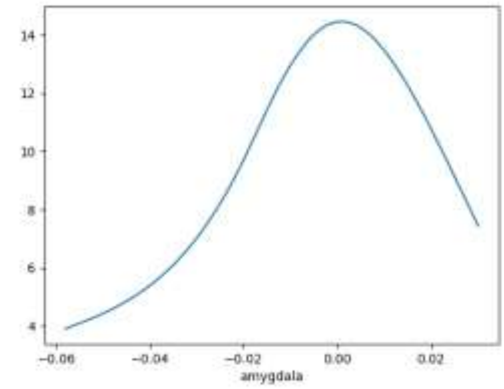
The difference for these two graphs is:

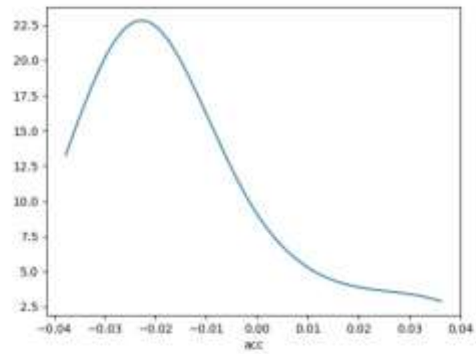
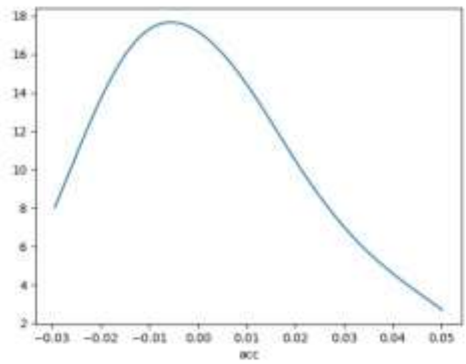
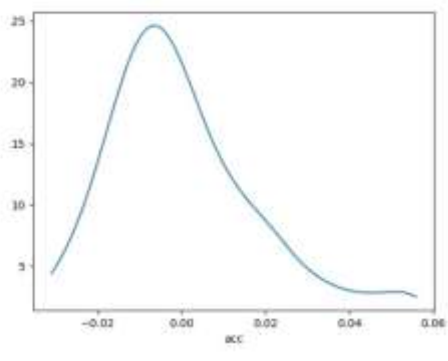
$\lvert p(amygdala, acc) - p(amygdala)p(acc) \rvert$


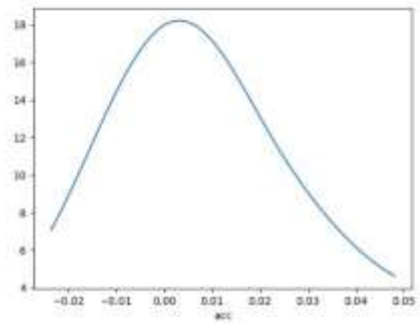
As you can see, the difference between $p(amygdala, acc)$ and $p(amygdala)p(acc)$ roughly doesn't equal to 0, which means that the two parts of brains are related with each other in this area..

(d)

Orientation	amygdala
c=2	
c=3	
c=4	

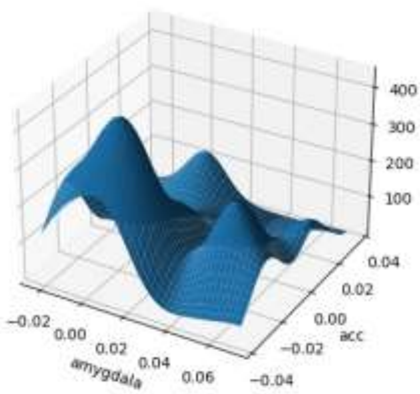
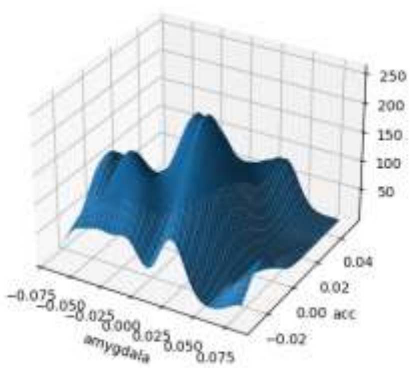
c=5	

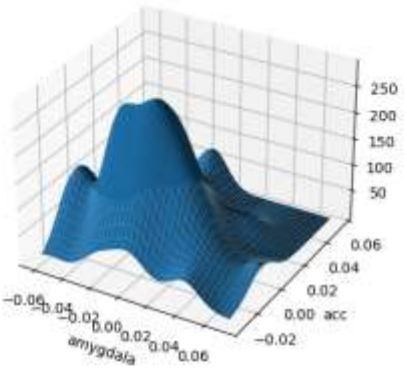
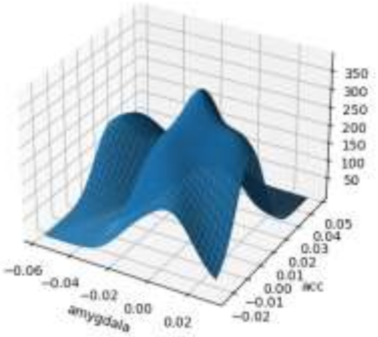
Orientation	acc
c=2	
c=3	
c=4	

c=5	

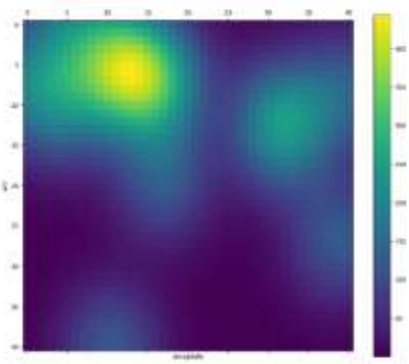
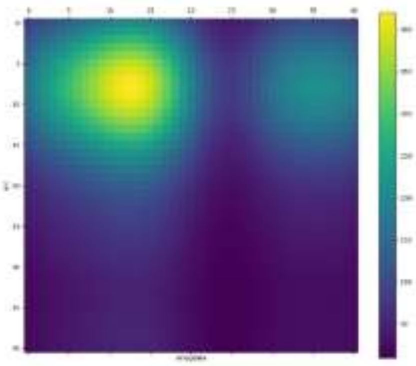
(e)

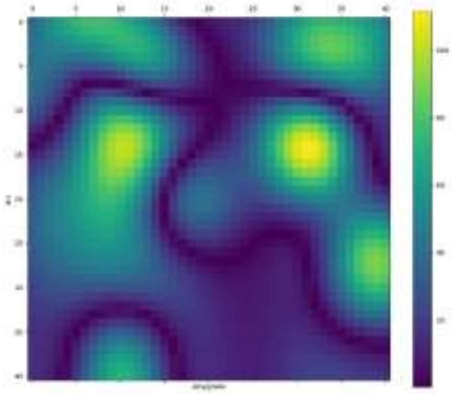
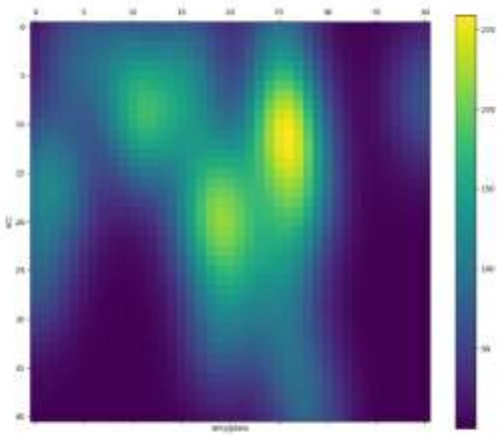
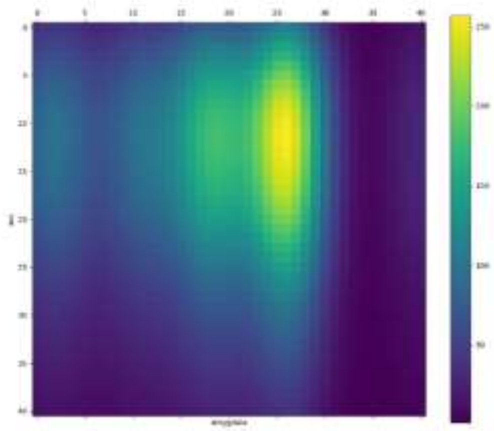
The joint distribution is shown as below:

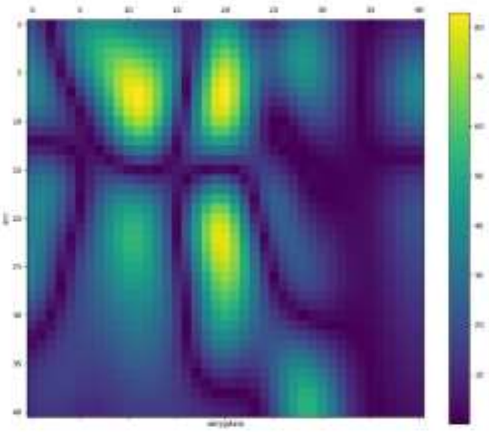
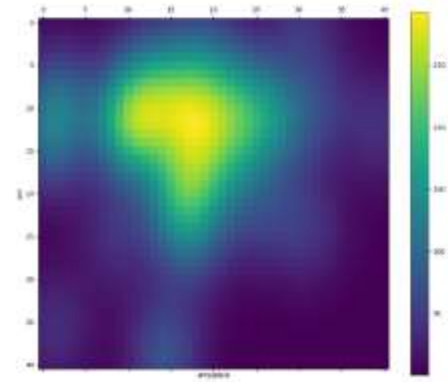
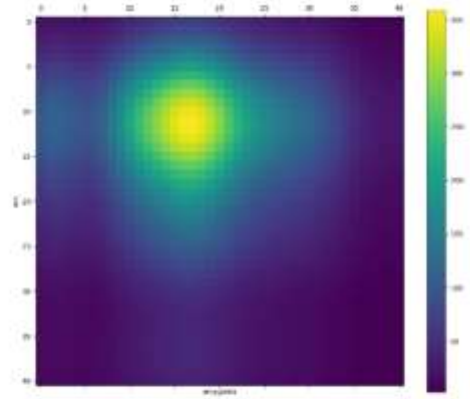
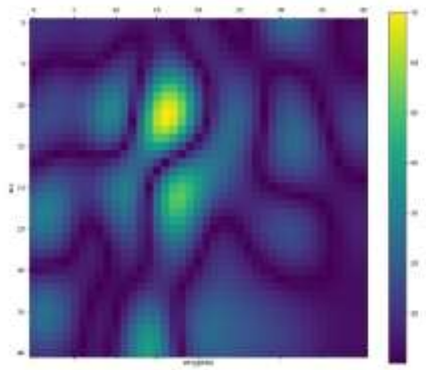
Orientation	joint distribution
c=2	
c=3	

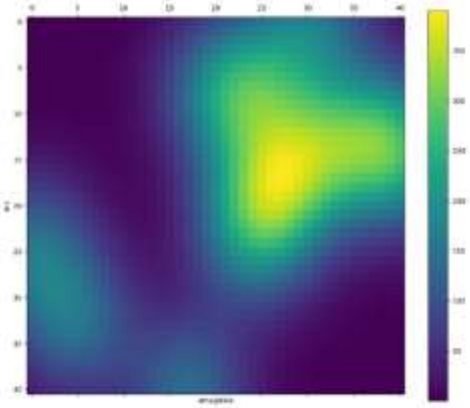
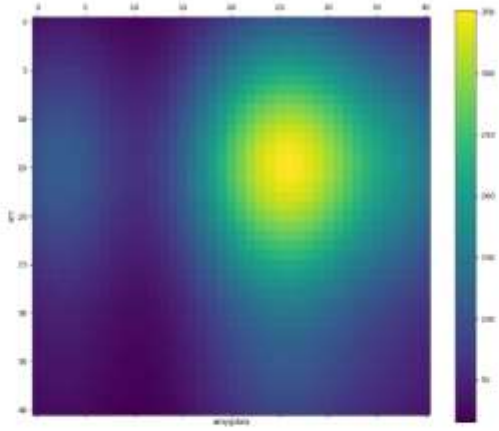
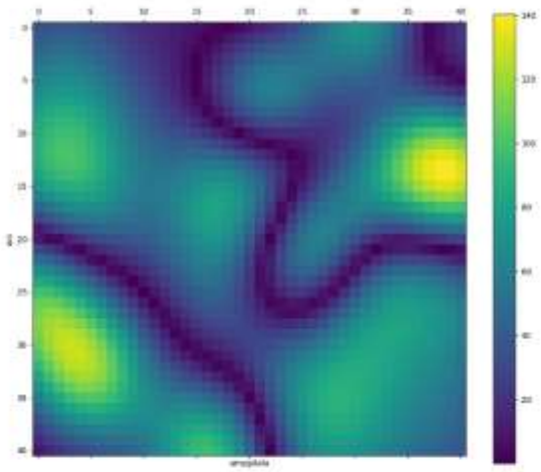
c=4	
c=5	

(f)

Orientation	$p(amygdala, acc)$	$p(amygdala)p(acc)$
c=2		

<table>
<tr><td rowspan="2"></td><td colspan="2">$|p(amygdala, acc) - p(amygdala)p(acc)|$</td></tr>
<tr><td colspan="2"></td></tr>
<tr><td rowspan="3">c=3</td><td>$p(amygdala, acc)$</td><td>$p(amygdala)p(acc)$</td></tr>
<tr><td></td><td></td></tr>
<tr><td colspan="2">$|p(amygdala, acc) - p(amygdala)p(acc)|$</td></tr>
</table>

<table>
<tr><td></td><td colspan="2"></td></tr>
<tr><td rowspan="4">c=4</td><td>$p(amygdala, acc)$</td><td>$p(amygdala)p(acc)$</td></tr>
<tr><td></td><td></td></tr>
<tr><td colspan="2">$|p(amygdala, acc) - p(amygdala)p(acc)|$</td></tr>
<tr><td colspan="2"></td></tr>
</table>

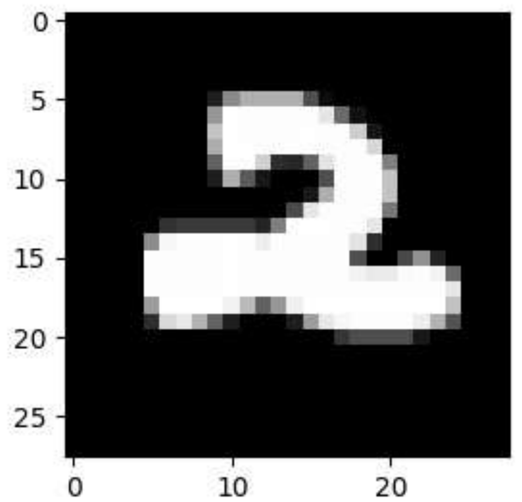
<table>
<tr><td rowspan="4">c=5</td><td>$p(amygdala, acc)$</td><td>$p(amygdala)p(acc)$</td></tr>
<tr><td></td><td></td></tr>
<tr><td colspan="2">$|p(amygdala, acc) - p(amygdala)p(acc)|$</td></tr>
<tr><td colspan="2"></td></tr>
</table>

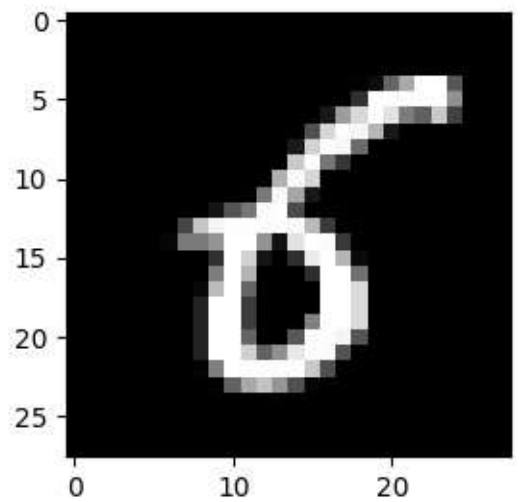
As you can see, for each Orientation, the difference $|p(amygdala, acc) - p(amygdala)p(acc)|$ does not equal to 0, which means the two parts of brain are related conditionally on the political orientation.

2 Implementing EM

(a).

The raw images for “2” and “6” are like the following:





(b).

For the E-step:

Since z^i is i.i.d and x^i is i.i.d, so we could get posterior distribution of $q(z^1, z^2, \ldots, z^m)$ like this:

$$q(z^1, z^2, \ldots, z^m) = \prod_{i=1}^{m} p(z^i | x^i, \theta^t)$$

To be detailed, I define:

$$\begin{aligned}
\tau_k^i = p\left(z^i = k \middle| x^i, \theta^t\right) &= \frac{p\left(x^i \middle| z^i = k, \theta^t\right) p(z^i = k)}{\sum_{j=1..K} p(z^i = k, x^i)} \\
&= \frac{\pi_k \mathrm{N}(x^i | \mu_k, \Sigma_k)}{\sum_{j=1..K} \pi_j \mathrm{N}(x^i | \mu_j, \Sigma_j)} \\
&= \frac{\pi_k \frac{1}{(2\pi)^{n/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}\left(x^i - \mu_k\right)^T {\Sigma_k}^{-1} (x^i - \mu_k)\right)}{\sum_{j=1..K} \pi_j \frac{1}{(2\pi)^{n/2} \left|\Sigma_j\right|^{1/2}} \exp\left(-\frac{1}{2}\left(x^i - \mu_j\right)^T {\Sigma_j}^{-1} (x^i - \mu_j)\right)}
\end{aligned}$$

$$= \frac{\pi_k \frac{1}{|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}\left(x^i-\mu_k\right)^T {\Sigma_k}^{-1}(x^i-\mu_k)\right)}{\sum_{j=1..K} \pi_j \frac{1}{|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}\left(x^i-\mu_j\right)^T {\Sigma_j}^{-1}(x^i-\mu_j)\right)}$$

If I take the expectation over $q(z^1, z^2, \dots, z^m)$ with respect to likelihood function $f(\theta)$, the expectation should be the lower bound of the maximum value of $f(\theta)$, since expectation is something like take the average. The lower bound of $f(\theta)$ is shown like the following:

$$\begin{aligned}
f(\theta) &= E_{q(z^1,z^2,\dots,z^m)}\left[log \prod_{i=1}^{m} p(z^i, x^i|\theta^t)\right] \\
&= E_{q(z^1,z^2,\dots,z^m)} \sum_{i=1}^{m} log\left[p(z^i, x^i|\theta^t)\right] \\
&= \sum_{i=1}^{m} E_{p(z^i|x^i,\theta^t)} log\left[p(z^i, x^i|\theta^t)\right] \\
&= \sum_{i=1}^{m}\sum_{k=1}^{K} \tau_k^i \left[log\left[p(z^i, x^i|\theta^t)\right]\right] \\
&= \sum_{i=1}^{m}\sum_{k=1}^{K} \tau_k^i \left[log\pi_k - \frac{1}{2}\left(x^i-\mu_k\right)^T {\Sigma_k}^{-1}\left(x^i-\mu_k\right) - \frac{1}{2} log|\Sigma_k| - \frac{n}{2}\log(2\pi)\right]
\end{aligned}$$

The M step is like this:

Now we want to maximize our lower bound $f(\theta)$, and now our unknown variables are π_k, μ_k, Σ_k. We want to find the expression of the unknown variables which can lead to the maximum of $f(\theta)$. We can notice that there is a constraint for one variable, that is $\sum \pi_k = 1$. So we could use Lagrange Multiplier Method:

$$L = \sum_{i=1}^{m}\sum_{k=1}^{K} \tau_k^i \left[log\pi_k - \frac{1}{2}\left(x^i-\mu_k\right)^T {\Sigma_k}^{-1}\left(x^i-\mu_k\right) - \frac{1}{2} log|\Sigma_k| - \frac{n}{2}\log(2\pi)\right] + \lambda(1 - \sum_{i=1}^{K} \pi_k)$$

Now I take the derivative with respect to each variable:

$$\begin{aligned}
\frac{\partial L}{\partial \pi_k} &= \sum_{i=1}^{m} \frac{\tau_k^i}{\pi_k} - \lambda = 0 \\
\Rightarrow \pi_k &= \frac{1}{\lambda}\sum_{i=1}^{m} \tau_k^i \\
\frac{\partial L}{\partial \mu_k} &= \sum_{i=1}^{m} \tau_k^i\left[{\Sigma_k}^{-1}\left(x^i-\mu_k\right)\right] = 0
\end{aligned}$$

$$= \sum_{i=1}^{m} \tau_k^i\left[(x^i - \mu_k)\right] = 0$$

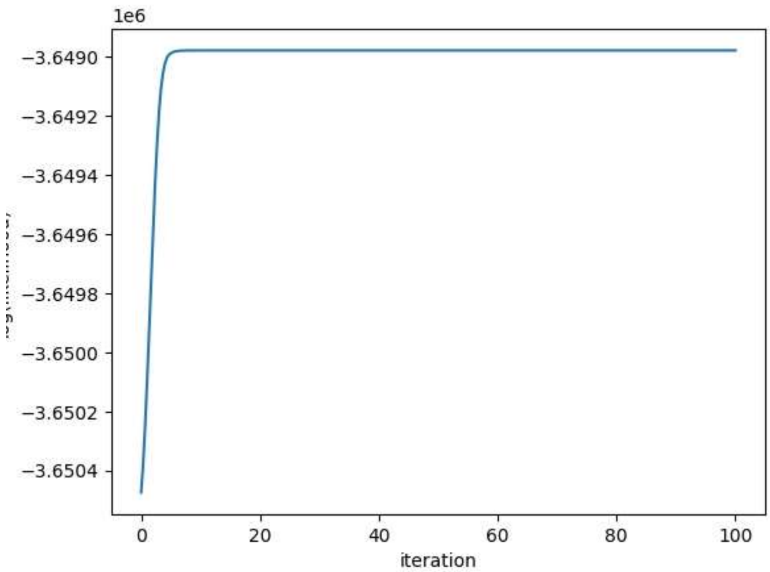
$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} \tau_k^i x^i}{\sum_{i=1}^{m} \tau_k^i}$$

$$\frac{\partial L}{\partial \Sigma_k} = \sum_{i=1}^{m} \tau_k^i \left[-\frac{1}{2}\left(\Sigma_k{}^{-1}\right)^{-1}(x^i - \mu_k)(x^i - \mu_k)^T - \frac{1}{2}\Sigma_k{}^{-1}\right] = 0$$

$$\Rightarrow \Sigma_k = \frac{\sum_{i=1}^{m} \tau_k^i \left[-\frac{1}{2}(x^i - \mu_k)(x^i - \mu_k)^T\right]}{\sum_{i=1}^{m} \tau_k^i}$$

(c)

As what can be shown in the picture, the x Axis is the EM algorithm iteration times, and the y axis is the value of $\log(likelihood)$. The graph shows that with the iteration times increase, the value of $\log(likelihood)$ become steady and meet the local maximum value, which indicate that the EM algorithm converges.

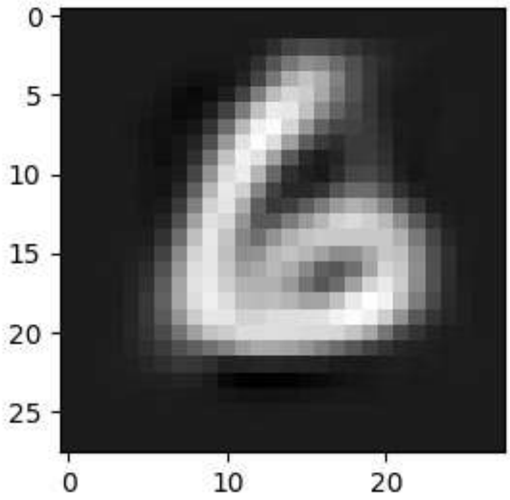
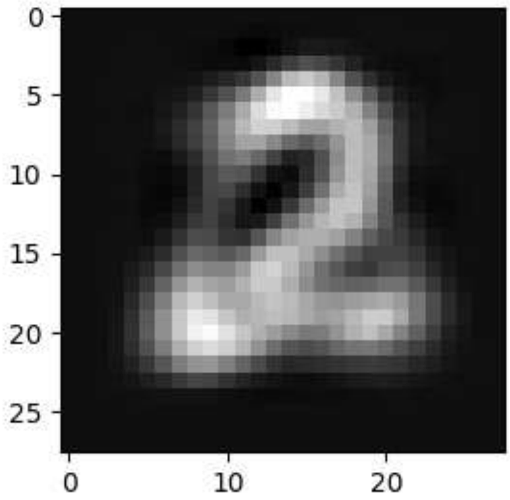


(d)

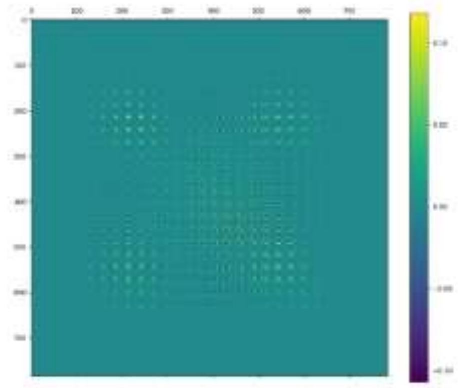
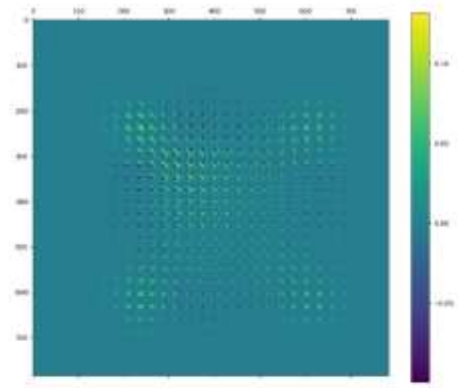
There are two components for the EM algorithm. The weights are shown as below:

	Component 1	Component 2
weight	0.5069	0.4931

The means for each component:

	Component 1	Component 2
mean		

The covariance for each component:

	Component 1	Component 2
covariance		

(e)

For the original data, I find that the 0~1031 sample is label “2”, which should belong to component 2. The 1032~1989 sample is label “6”, which should belong to component 1. The mismatch rates for Kmeans and EM are shown as below:

	Kmeans	EM
Mismatch Rate	6.3%	3.4%

So, we could conclude that the EM shows better behavior in clustering.