CS215: Data Analysis and Interpretation
Assignment 2

Anilesh Bansal, 22b0928
Arihant Vashista, 22b0958
Sanskar Shaurya, 22b0985

1 September 2023

Contents

1 Question 1

The cdf of the maximum Y_1 of the random variables $(X_1, X_2, ..., X_n)$ will be defined as $F_{Y_1}(y) = P(Y_1 \leq y)$, now as Y_1 is the max of all X_i, we have:

$$\begin{aligned} Y_1 &\geq X_i \ \forall i \in \{1, 2, ...n\} \\ \therefore Y_1 \leq y \implies X_i &\leq y \ \forall i \in \{1, 2, ...n\} \end{aligned}$$

This means that for a given y, the maximum Y_1 will be less than or equal to it if and only if all X_i are less than or equal to it. Hence, we have:

$$\begin{aligned} F_{Y_1}(y) &= P(Y_1 \leq y) \\ &= P(X_1 \leq y, X_2 \leq y, ..., X_n \leq y) \\ &= P(X_1 \leq y) \cdot P(X_2 \leq y) \cdots P(X_n \leq y) \text{ (As they are independent)} \\ &= [P(X_1 \leq y)]^n \text{ (As they are identically distributed)} \\ &= [F_X(y)]^n \end{aligned}$$

And the pdf of Y_1 will the derivative of the cdf above, which we get as:

$$\begin{aligned} f_{Y_1}(y) &= \frac{d}{dy}[F_X(y)]^n \\ &= n[F_X(y)]^{n-1} \cdot F'_X(y) \end{aligned}$$

Now the cdf of the minimum Y_2 will be defined as $F_{Y_2}(y) = P(Y_2 \leq y)$. If the minimum of a set of elements is less than or equal to a particular value, it means that at least one element exists in the set such that its value is less than or equal to that particular element. Otherwise, the minimum would be greater than that value. Therefore, $F_{Y_2}(y) =$ the probability that at least one of X_i is less than or equal to y. This is the complement of the probability that every element in X_i is greater than y. Hence, we have:

$$\begin{aligned} F_{Y_2}(y) &= 1 - P(X_1 > y, X_2 > y, ..., X_n > y) \\ &= 1 - P(X_1 > y) \cdot P(X_2 > y) \cdots P(x_n > y) \text{ (As they are independent)} \\ &= 1 - [P(X_1 > y)]^n \text{ (As they are identically distributed)} \\ &= 1 - [1 - F_X(y)]^n \end{aligned}$$

And the pdf of Y_2 will the derivative of the cdf above, which we get as:

$$\begin{aligned} f_{Y_2}(y) &= \frac{d}{dy}\{1 - [1 - F_X(y)]^n\} \\ &= n[1 - F_X(y)]^{n-1} \cdot F'_X(y) \end{aligned}$$

2 Question 2

We know that X is a mixture of n Gaussian variables, and every time we first choose the distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ with probability p_i and then take a sample form the distribution , we can say that the net expected value of the variable will be $E[X] = P(i^{th}$ distribution is chosen$)E[X_i]$ where summed over all i the individual Gaussian distributions.To do this more formally , let's define a new random variable I which is equal to i if the i^{th} distribution was chosen.Note that $P(I = i) = p_i$. Now,:

$$\begin{aligned} \mu = E[X] &= \sum_{i=1}^{k} E[X|I = i]P(I = i) \quad \text{From total probability law} \\ &= \sum_{i=1}^{k} p_i E[X_i] = \sum_{i=1}^{k} p_i \mu_i \end{aligned}$$

Using similar arguments above since we select the gaussian first with probability $p_i$the indivisual variance of the variable will be $E[(X_i - \mu)^2]$, where $\mu = E[X]$, the variance will be:

$$\begin{aligned}
Var(X) &= \sum_{i=1}^{k} p_i E[(X_i - \mu)^2] \\
&= \sum_{i=1}^{k} p_i E[X_i^2 - 2X_i\mu + \mu^2] \\
&= \sum_{i=1}^{k} p_i \left(E[X_i^2] - 2\mu E[X_i] + \mu^2\right) \\
&= \sum_{i=1}^{k} p_i E[X_i^2] - 2\mu \sum_{i=1}^{k} p_i E[X_i] + \mu^2 \sum_{i=1}^{k} p_i \\
&= \sum_{i=1}^{k} p_i(\sigma_i^2 + \mu_i^2) - 2\mu^2 + \mu^2 \\
&= \sum_{i=1}^{k} p_i(\sigma_i^2 + \mu_i^2) - \mu^2
\end{aligned}$$

Here we used the fact that $\sigma_i^2 = E[X_i^2] - E[X_i]^2$.
The MGF of a random variable is $\phi_X(t) = E[e^{Xt}]$. Again in this case we will choose a Gaussian with probabilty p_i and the indivisual MGF will be $E[e^{X_i t}] = e^{\mu_i t + \frac{t^2\sigma_i^2}{2}}$. Hence the net MGF is:

$$\begin{aligned}
\phi_X(t) = E[e^{Xt}] &= \sum_{i=1}^{k} p_i E[e^{tX_i}] \\
&= \sum_{i=1}^{k} p_i e^{\mu_i t + \frac{t^2\sigma_i^2}{2}}
\end{aligned}$$

Now the next variable we are given is $Z = \sum_{i=1}^{k} p_i X_i$, now this case is quite straight-forward, we just need to calculate the values directly by applying the formulas:-

$$E[Z] = E\left[\sum_{i=1}^{k} p_i X_i\right] = \sum_{i=1}^{k} p_i E[X_i] = \sum_{i=1}^{k} p_i \mu_i = \mu' \quad \text{Because E[.] is linear}$$

$$\begin{aligned}
Var(Z) &= Var\left(\sum_{i=1}^{k} p_i X_i\right) \\
&= \sum_{i=1}^{k} Var(p_i X_i) \quad \text{beacuse } X_i \text{ are } \textbf{independent random variables} \\
&= \sum_{i=1}^{k} p_i^2 \sigma_i^2 = \sigma'^2
\end{aligned}$$

$$
\begin{aligned}
\phi_Z(t) = E[e^{Zt}] &= E\left[e^{t\sum_{i=1}^{k} p_i X_i}\right] \\
&= E\left[\prod_{i=1}^{k} e^{tp_i X_i}\right] \\
&= \prod_{i=1}^{k} E\left[e^{tp_i X_i}\right] \quad \text{because } e^{X_i} \text{ are also } \textbf{independent variables} \\
&= \prod_{i=1}^{k} exp\left(\mu_i p_i t + \frac{\sigma_i^2 t^2 p_i^2}{2}\right) \\
&= exp\left(t\sum_{i=1}^{k} p_i\mu_i + \frac{t^2}{2}\sum_{i=1}^{k} p_i^2\sigma_i^2\right) = e^{\mu' t + \frac{t^2\sigma'^2}{2}}
\end{aligned}
$$

Note here we used the fact that e^{X_i} are also independent variables. This fact can be proven but since we have already done that in the tutorial we haven't included the proof here.

Further observe that the MGF(Z) is identical to the MGF of a gausssian variable with mean μ' and standard deviation σ', hence we can conclude that Z is also a Gaussian variable , so we can directly write the PDF as:

$$
f_Z(z) = \frac{1}{\sigma'\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-\mu'}{\sigma'}\right)^2}
$$

3 Question 3

Let us take a random variable $Y = X - \mu + t$, where t is some arbitrary positive number, then

$$
\begin{aligned}
Y^2 &\geq (\tau + t)^2 \quad \text{For some } \tau > 0 \\
\implies |Y| &\geq (\tau + t) \quad \text{As } \tau + t \geq 0 \\
\implies Y \geq \tau + t \quad &\text{OR} \quad -Y \leq \tau + t \\
\text{Thus, } P\{Y^2 \geq (\tau + t)^2\} &\geq P(Y \geq \tau + t) = P(X - \mu \geq \tau) \\
\text{As } Y \geq \tau + t &\iff X - \mu \geq \tau
\end{aligned}
$$

Thus $P\{Y^2 \geq (\tau + t)^2\} \geq P(X - \mu \geq \tau)$. Now, from Markov's inequality, we get:

$$
P\{Y^2 \geq (\tau + t)^2\} \leq \frac{E[Y^2]}{(\tau + t)^2}
$$

Now,

$$
\begin{aligned}
E[Y^2] &= E[(X - \mu + t)^2] \\
&= E[(X - \mu)^2] + 2 \cdot t \cdot E[X - \mu] + t^2 \\
&= \sigma^2 + 0 + t^2
\end{aligned}
$$

Hence,

$$
\begin{aligned}
P(X - \mu \geq \tau) &\leq P\{Y^2 \geq (\tau + t)^2\} \\
&\leq \frac{\sigma^2 + t^2}{(\tau + t)^2}
\end{aligned}
$$

The RHS of this inequality is a function of t and it attains its minimum when $t = \sigma^2/\tau$. We get this by analyzing the function for $t > 0$. Now, putting this value in the inequality, we get:

$$\begin{aligned} P(X-\mu \geq \tau) &\leq \frac{\sigma^2 + \frac{\sigma^4}{\tau^2}}{(\frac{\sigma^2}{\tau} + \tau)^2} \\ \implies P(X-\mu \geq \tau) &\leq \frac{\sigma^2}{\sigma^2+\tau^2} \quad \text{if } \tau > 0 \end{aligned}$$

Now, let us consider the case where $\tau < 0$. Consider a random variable $Y = \mu - X$, then:

$$\begin{aligned} X - \mu &\leq \tau \\ \implies Y &\geq -\tau \end{aligned}$$

Thus, $P(X-\mu \leq \tau) = P(Y \geq -\tau)$. Now $E[Y] = \mu - E[X] = 0$ and $\sigma_y = E[(Y-\mu_y)^2] = E[Y^2] = E[(\mu - X)^2] = \sigma$. As $-\tau$ is a positive number, we can apply the above inequality we got for the random variable Y and get:

$$\begin{aligned} P(Y - \mu_y \geq -\tau) &\leq \frac{\sigma_y^2}{\sigma_y^2 + (-\tau)^2} \\ \implies P(Y \geq -\tau) &\leq \frac{\sigma^2}{\sigma^2+\tau^2} \\ \implies P(X-\mu \leq \tau) &\leq \frac{\sigma^2}{\sigma^2+\tau^2} \end{aligned}$$

Now $P(X-\mu \geq \tau) = P(X-\mu > \tau) + P(X - \mu = \tau)$ and $P(X-\mu > \tau) = 1 - P(X-\mu \leq \tau)$ hence

$$\begin{aligned} P(X-\mu \geq \tau) &\geq P(X-\mu > \tau) \\ &\geq 1 - \frac{\sigma^2}{\sigma^2+\tau^2} \end{aligned}$$

4 Question 4

Consider a non-negative increasing function $\Psi(x) : \mathbb{R} \to \mathbb{R}$.

$$\begin{aligned} E[\Psi(X)] = \int_{-\infty}^{\infty} \Psi(x) f_X(x) dx &\geq \int_a^{\infty} \Psi(x) f_X(x) dx \\ &\geq \int_a^{\infty} \Psi(a) f_X(x) dx \quad \text{since } x \geq a \implies \Psi(x) \geq \Psi(a) \\ &\geq \Psi(a) P(X \geq a) \end{aligned} \tag{1}$$

Similarly for a non-negative decreasing function $\psi(x) : \mathbb{R} \to \mathbb{R}$, we can say

$$\begin{aligned} E[\psi(X)] = \int_{-\infty}^{\infty} \psi(x) f_X(x) dx &\geq \int_{-\infty}^{a} \psi(x) f_X(x) dx \\ &\geq \int_{\infty}^{a} \psi(a) f_X(x) dx \quad \text{since } x \leq a \implies \psi(x) \geq \psi(a) \\ &\geq \psi(a) P(X \leq a) \end{aligned} \tag{2}$$

Now we can simply substitute $\Psi(x) = e^{tx}$ for $t > 0$ as it would be an increasing function and $\psi(x) = e^{tx}$ for $t < 0$ as it would be and decreasing function to get:

$$P(X \geq a) \leq \frac{E[e^{tX}]}{e^{tx}} = e^{-tx}\phi_X(t) \quad \text{for } t > 0 \text{ and}$$
$$P(X \leq a) \leq \frac{E[e^{tX}]}{e^{tx}} = e^{-tx}\phi_X(t) \quad \text{for } t < 0$$

Now we are given n indepedent **Bernoulli** Variables, X_i. We know for a bernoulli variable

$$E[e^{tX_i}] = 1 + p_i(e^t - 1) \leq e^{p_i(e^t-1)} \tag{3}$$

Also, we are given the random variable $X = \sum_{i=1}^{n} X_i$, consider:

$$\begin{aligned}
\phi_X(t) = E[e^{Xt}] &= E\left[e^{t\sum_{i=1}^{n} X_i}\right] \\
&= E\left[\prod_{i=1}^{n} e^{tX_i}\right] \\
&= \prod_{i=1}^{n} E\left[e^{tX_i}\right] \quad \text{because } e^{X_i} \text{ are also } \textbf{independent variables} \\
&\leq \prod_{i=1}^{n} e^{p_i(e^t-1)} \quad \text{using (3)} \\
&= e^{(e^t-1)\sum_{i=1}^{n} p_i} = e^{\mu(e^t-1)} \qquad (4)
\end{aligned}$$

Using (1) and (4), we can conclude:

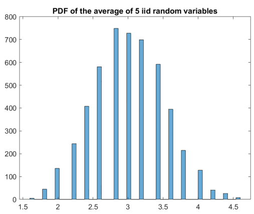
$$P(X > (1+\delta)\mu) \leq \frac{e^{\mu(e^t-1)}}{e^{(1+\delta)t\mu}} \quad \text{for } t \geq 0, \delta > 0 \tag{5}$$

To find the tightest bound we can different the expression in (5) and minimize w.r.t t. We get get $t = ln(1+\delta)$, substituting this value we get:

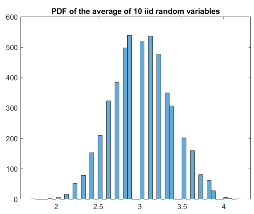
$$P(X > (1+\delta)\mu) \leq \frac{e^{\mu\delta}}{(1+\delta)^{(1+\delta)\mu}}$$

5 Question 5

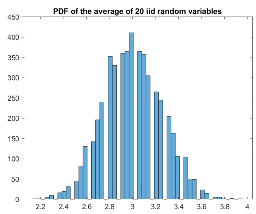
(a)



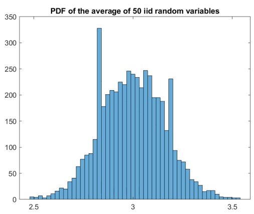
(a) n = 5



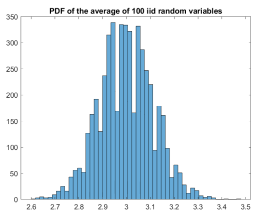
(b) n = 10



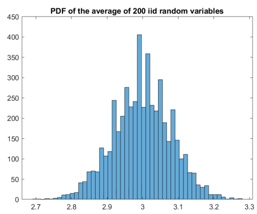
(c) n = 20



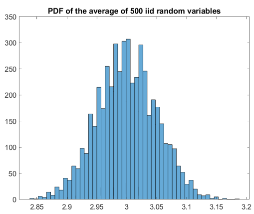
(d) n = 50



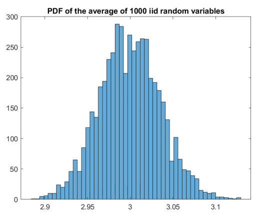
(e) n = 100



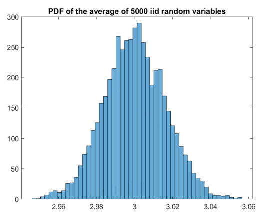
(f) n = 200



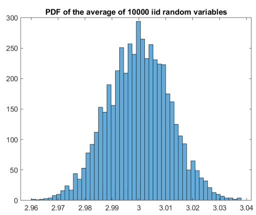
(g) n = 500



(h) n = 1000



(i) n = 5000

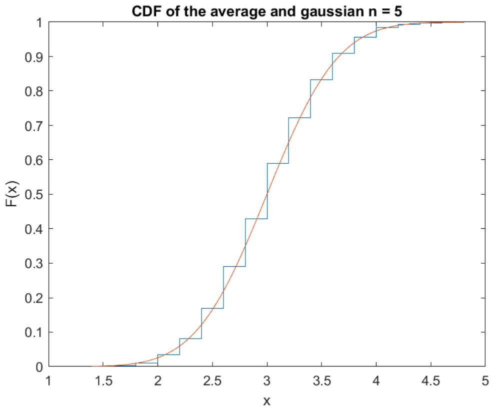


(j) n = 10000

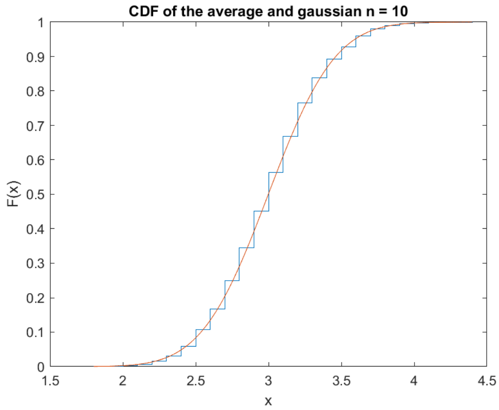
Figure 1: Histogram representing emperical average of n random variables

Note: The .m file used for this question is "Q5_a.m".

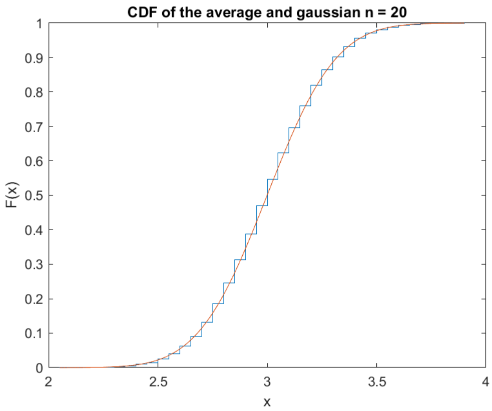
(b)



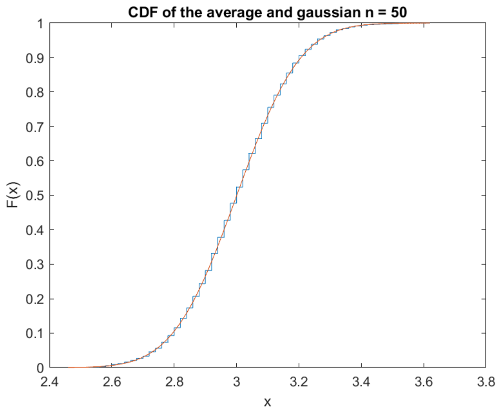
n = 5



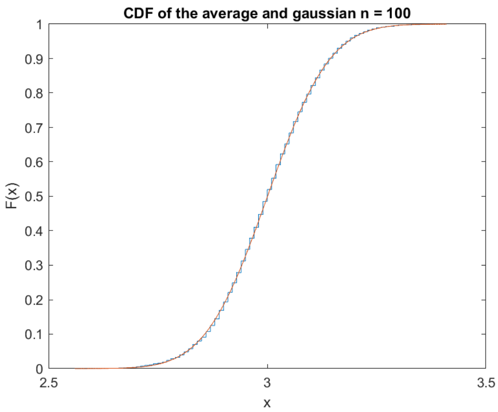
n = 10



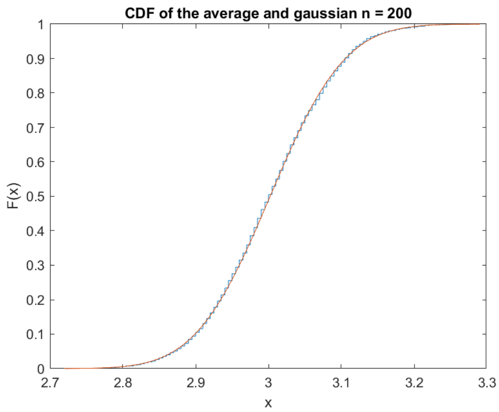
n = 20



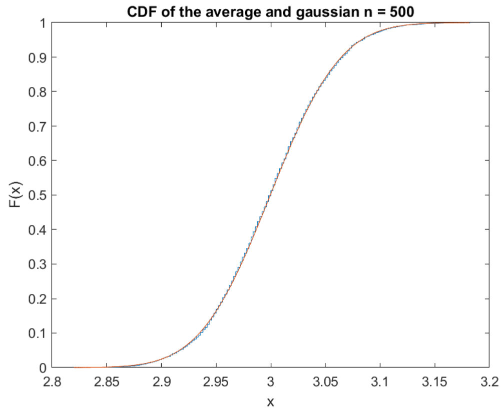
n = 50



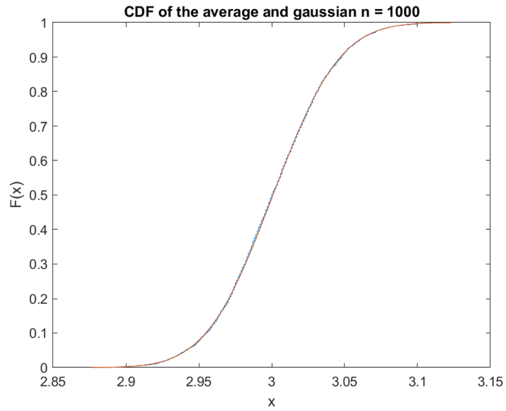
n = 100



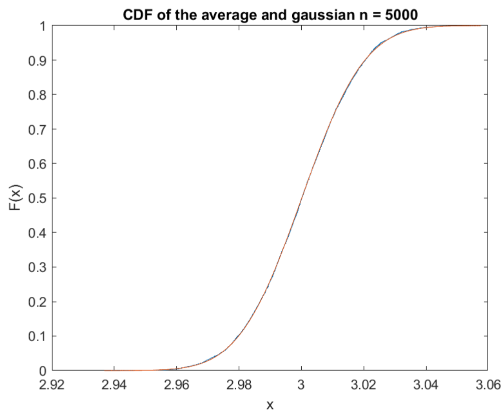
n = 200



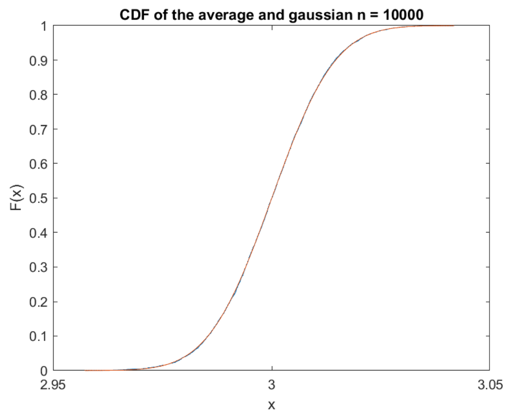
n = 500



n = 1000



n = 5000



n = 10000

Emperical CDF and the CDF of the predicted Gaussian

Note: The .m file used for this question is "Q5_b.m".

(c)

Here's the plot of the MAD as a function of N.

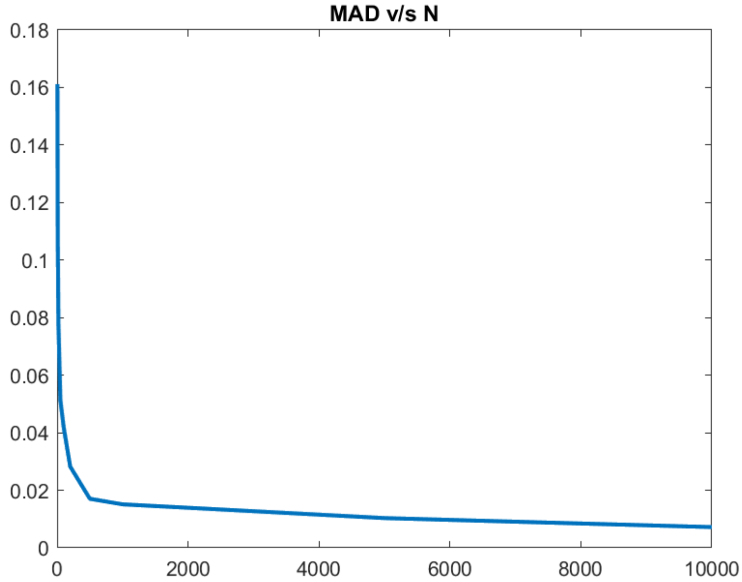


Figure 6: MAD v/s N

Note: The .m file used for this question is "Q5_c.m".

6 Question 6

The plots of Correlation Coefficient vs Shift and QMI vs shift for both cases is given below

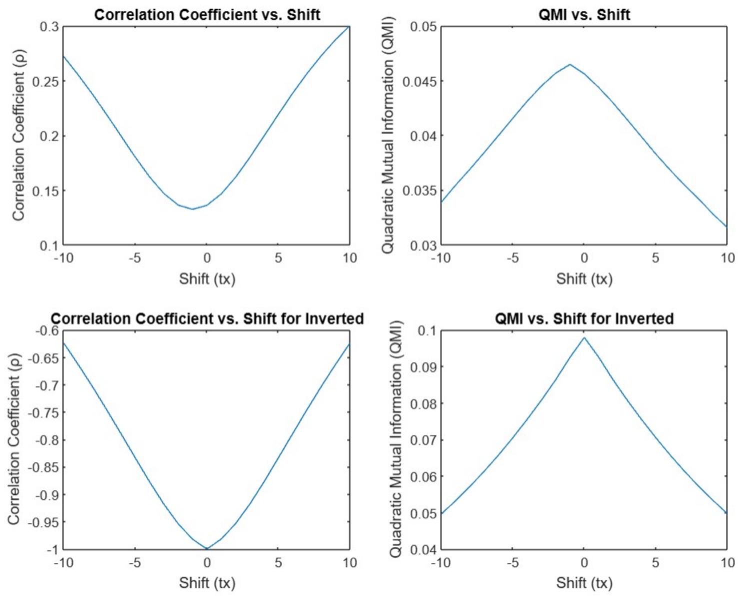
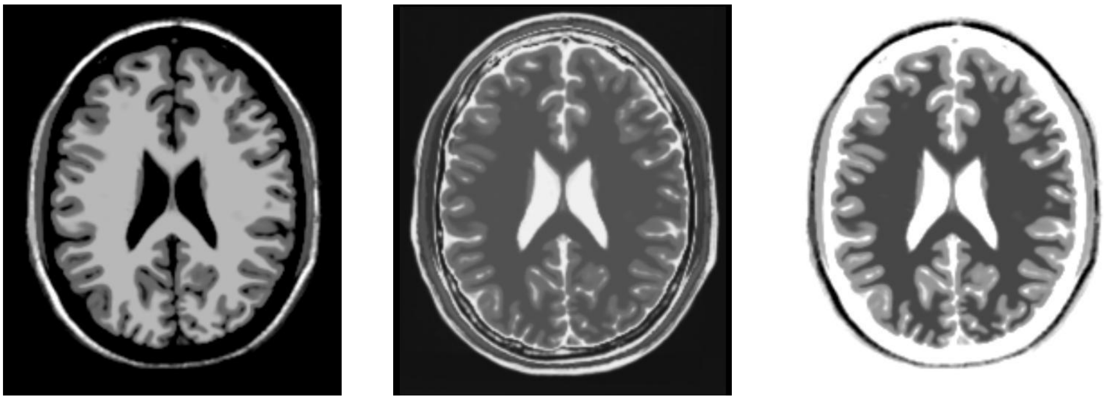


Figure 7: Q6

- For the case when Im2 is 'T2.jpg' the Correlation coefficient attains a minima for the value of shift $t_x = -1$. As the images T1 and T2 are approximately inverse of each other, when they match exactly we attain a more negative correlation and that's why there is a minima. IF there was no black background in the images, the minima would be close to -1 to show this negative correlation.

  However, due to the common black background in both the images, it skews the value of the correlation coefficient towards +1 and makes the absolute value of the correlation coefficient rather useless, only the difference in ρ as t_x changes makes sense.

- When image2 is inverse of image1, we see a perfect -1 correlation coefficient exactly when shift $t_x = 0$ and the correlation starts getting closer to 0 as we move away.



(a) T1 (b) T2 (c) Inverted T1

- QMI compares the joint probability distribution of pixel intensities in both images to the product of their marginal probability distributions. In independent images, the joint distribution equals the product of the marginals, resulting in a QMI value of zero. When there are deviations from independence, QMI becomes nonzero, indicating dependence. From the graphs we can see there is a peak of QMI at shift $t_x = -1$ in first case and $t_x = 0$ in second case indicating that the images are most dependent at these shifts.

Note: Relevant Files for this question include *q6.m* containing the MATLAB code and *q6_combined.jpg* containing all the four plots.

7 Question 7

Let us consider a multinomial distribution with n trials and k possible outcomes with probability of success given by $(p_1, p_2, \cdots, p_k)$ respectively. The multinomial random variable is a vector $\mathbf{X} = \{X_1, X_2, \cdots, X_k\}$ with its i^{th} component corresponding to the number of successes of the i^{th} outcome.

The moment generating function for this multinomial distribution is given by

$$\phi_{\mathbf{X}}(\mathbf{t}) = (p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k})^n$$

- Off-diagonal Element C_{ij}

$$C_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)] = E(X_i X_j) - E(X_i)E(X_j)$$

Now, we know that

$$\begin{aligned}
E(X_i) &= \left.\frac{\partial \phi_{\mathbf{X}}(\mathbf{t})}{\partial t_i}\right|_{t_i=0} \\
&= \left. np_i e^{t_i}(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k})^{n-1}\right|_{t_i=0} \\
&= np_i
\end{aligned}$$

$$\begin{aligned}
E(X_i X_j) &= \left.\frac{\partial \phi_{\mathbf{X}}(\mathbf{t})}{\partial t_i \partial t_j}\right|_{\mathbf{t}=0} = \left.\frac{\partial}{\partial t_j}\left(\frac{\partial \phi_{\mathbf{X}}(\mathbf{t})}{\partial t_i}\right)\right|_{\mathbf{t}=0} \\
&= \frac{\partial\left(np_i e^{t_i}(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k})^{n-1}\right)}{\partial t_j} \\
&= np_i e^{t_i} \times \frac{\partial\left(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k}\right)^{n-1}}{\partial t_j} \\
&= \left. np_i e^{t_i} \times (n-1) p_j e^{t_j}\left(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k}\right)^{n-2}\right|_{\mathbf{t}=0} \\
&= n(n-1) p_i p_j
\end{aligned}$$

Substituting these expressions in the above formula for C_{ij}

$$\begin{aligned}
C_{ij} &= E[(X_i - \mu_i)(X_j - \mu_j)] = E(X_i X_j) - E(X_i)E(X_j) \\
&= n(n-1)p_i p_j - (np_i)(np_j) = -np_i p_j
\end{aligned}$$

as required.

- Diagonal element C_{ii}

$$C_{ii} = E[(X_i - \mu_i)^2] = E[X_i^2] - E[X_i]^2$$

Now, to find values of $E[X_i^2]$

$$\begin{aligned}
E(X_i^2) &= \left.\frac{\partial^2 \phi_{\mathbf{X}}(\mathbf{t})}{\partial t_i^2}\right|_{\mathbf{t}=0} \\
&= \frac{\partial\left(np_i e^{t_i}(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k})^{n-1}\right)}{\partial t_i} \\
&= np_i \times \frac{\partial e^{t_i}\left(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k}\right)^{n-1}}{\partial t_i} \\
&= np_i \times \left[e^{t_i}(n-1)p_i e^{t_i}\left(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k}\right)^{n-2} + \left(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k}\right)^{n-1} e^{t_i}\right] \\
&= np_i[(n-1)p_i + 1] \\
&= n(n-1)p_i^2 + np_i
\end{aligned}$$

Substituting in the above expression we get

$$\begin{aligned}
C_{ii} &= E[(X_i - \mu_i)^2] = E[X_i^2] - E[X_i]^2 \\
&= n(n-1)p_i^2 + np_i - (np_i)^2 \\
&= np_i - np_i^2 = np_i(1 - p_i)
\end{aligned}$$