CS215 Assignment 1 Report

Atharva Bendale (22B0901)
Vishal Bysani (22B1061)

August 2023

Contents

Introduction

In this report, we have written down our solutions for the problems in the Assigment 1 of CS 215 course on Data Analysis and Interpretation.

1 Problem 1

Let $X_1, X_2, ..., X_n$ be $n > 0$ independent identically distributed random variables with cdf $F_X(x)$ and pdf $f_X(x) = F'_X(x)$. Derive an expression for the cdf and pdf of $Y_1 = \max(X_1, X_2, ..., X_n)$ and $Y_2 = \min(X_1, X_2, ..., X_n)$ in terms of $F_X(x)$. [10 points]

Solution :

We are given that $F_{X_1}(x) = F_{X_2}(x) = \cdots = F_X(x)$ and $f_{X_1}(x) = f_{X_2}(x) = \cdots = f_X(x)$ since $X_1, X_2, ..., X_n$ are identically distributed random variables.

Now consider $Y_1 = \max(X_1, X_2, ..., X_n)$

$$F_{Y_1}(x) = P(Y_1 \leq x) \text{ Here, P denotes the probability measure}$$

$$\text{Since } Y_1 = \max(X_1, X_2, ..., X_n), \quad Y_1 \leq x \implies X_1 \leq x, X_2 \leq x, ..., X_n \leq x$$
$$\therefore P(Y_1 \leq x) = P(\textstyle\bigcap_{i=1}^{n} X_i \leq x)$$

$$\text{Since } X_1, X_2, ..., X_n \text{ are independent, } P(\textstyle\bigcap_{i=1}^{n} X_i \leq x) = \prod_{i=1}^{n} P(X_i \leq x)$$

$$\textstyle\prod_{i=1}^{n} P(X_i \leq x) = \prod_{i=1}^{n} F_{X_i}(x) = F_X^n(x)$$

$$\therefore F_{Y_1}(x) = F_X^n(x)$$

$$f_{Y_1}(x) = F'_{Y_1}(x) = \frac{\mathrm{d}F_X^n(x)}{\mathrm{d}x} = n \cdot F_X^{n-1}(x) \cdot F'_X(x) = n \cdot F_X^{n-1}(x) \cdot f_X(x)$$

$$\therefore \text{Cumulative distribution function of } Y_1 \text{ is } F_X^n(x)$$

$$\text{Probability density function of } Y_1 \text{ is } n \cdot F_X^{n-1}(x) \cdot f_X(x)$$

Now consider $Y_2 = \min(X_1, X_2, ..., X_n)$

$$F_{Y_2}(x) = P(Y_2 \leq x) = 1 - P(Y_2 > x)$$

$$\text{Since } Y_2 = \min(X_1, X_2, ..., X_n), \quad Y_2 > x \implies X_1 > x, X_2 > x, ..., X_n > x$$
$$\therefore P(Y_1 > x) = P(\textstyle\bigcap_{i=1}^{n} X_i > x)$$

$$\text{Since } X_1, X_2, ..., X_n \text{ are independent, } P(\textstyle\bigcap_{i=1}^{n} X_i > x) = \prod_{i=1}^{n} P(X_i > x)$$

$$\textstyle\prod_{i=1}^{n} P(X_i > x) = \prod_{i=1}^{n}(1 - F_{X_i}(x)) = (1 - F_X(x))^n \quad [\because P(X > x) = 1 - P(X \leq x) = 1 - F_X(x)]$$

$$\therefore F_{Y_2}(x) = 1 - (1 - F_X(x))^n$$

$$f_{Y_2}(x) = F'_{Y_2}(x) = \frac{\mathrm{d}(1 - (1 - F_X(x))^n)}{\mathrm{d}x} = n \cdot (1 - F_X(x))^{n-1} \cdot F'_x(x) = n \cdot (1 - F_X(x))^{n-1} \cdot f_x(x)$$

$$\therefore \text{Cumulative distribution function of } Y_2 \text{ is } 1 - (1 - F_X(x))^n$$

Probability density function of Y_2 is $n \cdot (1 - F_X(x))^{n-1} \cdot f_x(x)$

2 Problem 2

We say that a random variable X belongs to a Gaussian mixture model (GMM) if $X \sim \sum_{i=1}^{K} p_i \mathcal{N}(\mu_i, \sigma_i^2)$ where p_i is the 'mixing probability' for each of the K constituent Gaussians, with $\sum_{i=1}^{K} p_i = 1; \forall i, 0 \leq p_i \leq 1$. To draw a sample from a GMM, we do the following: (1) One of the K Gaussians is randomly chosen as per the PMF $\{p_1, p_2, ..., p_K\}$ (thus, a Gaussian with a higher mixing probability has a higher chance of being picked). (2) Let the index of the chosen Gaussian be (say) m. Then, you draw the value from $\mathcal{N}(\mu_m, \sigma_m^2)$. If X belongs to a GMM as defined here, obtain expressions for $E(X), \text{Var}(X)$ and the MGF of X.
Now consider a random variable of the form $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for each $i \in \{1, 2, ..., K\}$. Define another random variable $Z = \sum_{i=1}^{K} p_i X_i$ **where $\{X_i\}_{i=1}^{K}$ are independent random variables**. Derive an expression for $E(Z), \text{Var}(Z)$ and the PDF, MGF of Z. [2+2+2+2+2+2+3=15 points]

Solution:

Let $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for each $i \in \{1, 2, ..., K\}$

To find the PDF of X,

$$P(X = x) = \sum_{i=1}^{K} P(X = X_i) P(X_i = x | X = X_i) \qquad [P(X = X_i) = p_i]$$

$$= \sum_{i=1}^{K} p_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2}$$

For expectation value of X,

$$E(X) = \int_{-\infty}^{+\infty} x P(X = x) dx = \int_{-\infty}^{+\infty} x \sum_{i=1}^{K} p_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2} dx$$

$$= \sum_{i=1}^{K} p_i \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2} dx = \sum_{i=1}^{K} p_i E(X_i) = \sum_{i=1}^{K} p_i \mu_i$$

For variance of X,

$$\text{Var}(X) = E((X - E(X))^2 = E(X^2) - (E(X))^2$$

To find $E(X^2)$,

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 P(X = x) dx = \int_{-\infty}^{+\infty} x^2 \sum_{i=1}^{K} p_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2} dx$$

$$= \sum_{i=1}^{K} p_i \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2} dx = \sum_{i=1}^{K} p_i E(X_i^2)$$

In slides, it is given that $E(X) = \mu$ and $E((X - \mu)^2) = \sigma^2$ for a Gaussian random variable. We get $E((X - \mu)^2) = E(X^2) - (E(X))^2 = \sigma^2 \implies E(X^2) = \mu^2 + \sigma^2$

$$\therefore E(X^2) = \sum_{i=1}^{K} p_i (\mu_i^2 + \sigma_i^2) [\because E(X_i^2) = \mu_i^2 + \sigma_i^2]$$

$$\therefore \text{Var}(X) = \sum_{i=1}^{K} p_i (\mu_i^2 + \sigma_i^2) - (\sum_{i=1}^{K} p_i \mu_i)^2$$

For MGF of X,

$$\text{MGF of X} = \phi_X(t) = E(e^{tX}) = \int_{-\infty}^{+\infty} e^{tx} P(X = x)dx = \int_{-\infty}^{+\infty} e^{tx} \sum_{i=1}^{K} p_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2}$$
$$\sum_{i=1}^{K} p_i \int_{-\infty}^{+\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2} = \sum_{i=1}^{K} p_i \phi_{X_i}(t) = \sum_{i=1}^{n} p_i e^{\mu_i t + \sigma_i^2 t^2/2}$$

Now consider $Z = \sum_{i=1}^{K} p_i X_i$ where $\{X_i\}_{i=1}^{K}$ are independent random variables, then,

$$E(Z) = \sum_{i=1}^{K} p_i \mu_i$$

For variance of Z,

$$\text{Var}(Z) = E[(X - E(X))^2] = E[(\sum_{i=1}^{K} p_i X_i - \sum_{i=1}^{K} p_i \mu_i)^2] = E[(\sum_{i=1}^{K} p_i (X_i - \mu_i))^2]$$

$$\therefore \text{Var}(X) = E[\sum_{i=1}^{K} p_i^2 (X_i - \mu_i)^2 + \sum_{\substack{i \\ i \neq j}} \sum_j p_i p_j (X_i - \mu_i)(X_j - \mu_j)]$$
$$= \sum_{i=1}^{K} E[p_i^2 (X_i - \mu_i)^2] + \sum_{\substack{i \\ i \neq j}} \sum_j E[p_i p_j (X_i - \mu_i)(X_j - \mu_j)]$$
$$= \sum_{i=1}^{K} p_i^2 E[(X_i - \mu_i)^2] + \sum_{\substack{i \\ i \neq j}} \sum_j p_i p_j E[(X_i - \mu_i)(X_j - \mu_j)]$$
$$= \sum_{i=1}^{K} p_i^2 \sigma_i^2 + \sum_{\substack{i \\ i \neq j}} \sum_j p_i p_j E[(X_i - \mu_i)(X_j - \mu_j)]$$
$$= \sum_{i=1}^{K} p_i^2 \sigma_i^2 + \sum_{\substack{i \\ i \neq j}} \sum_j p_i p_j (E(X_i X_j) - E(X_i)E(X_j))$$
$$= \sum_{i=1}^{K} p_i^2 \sigma_i^2 \quad [\because E(XY) = E(X)E(Y) \text{ if } X, Y \text{ are independent random variables}]$$

Since Z is a linear combination of independent normal random variables, $Z \sim \mathcal{N}(E(Z), \text{Var}(Z))$
This can be proved by showing that the MGF of Z is that of a Gaussian with parameters $E(Z)$ and $\text{Var}(Z)$

$$\text{MGF of Z} = E(e^{tZ}) = E(e^{\sum_{i=1}^{K} t p_i X_i}) = \sum_{x_1, x_2, \ldots, x_n} e^{\sum_{i=1}^{K} t p_i x_i} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$
$$= \sum_{x_1, x_2, \ldots, x_n} e^{\sum i=1^K t p_i x_i} P(X_1 = x_1) P(X_2 = x_2) \ldots P(X_n = x_n)$$
$$= \sum_{x_1} e^{t p_1 x_1} P(X_1 = x_1) \sum_{x_2} e^{t p_2 x_2} P(X_2 = x_2) \ldots \sum_{x_n} e^{t p_n x_n} P(X_n = x_n)$$
$$= \prod_{i=1}^{K} E(e^{t p_i X_i}) = \prod_{i=1}^{K} \phi_{X_i}(t) = \prod_{i=1}^{K} e^{p_i \mu_i t + p_i^2 \sigma_i^2 t^2/2} = e^{tE(Z) + t^2 Var(Z)/2}$$

Therefore, since MGF of X is that of a Gaussian random variables, and MGF is unique (in slides), X is also a Gaussian random variable.

$$\text{PDF of } Z = f_Z(x) = \frac{1}{\sqrt{2\pi \text{Var}(Z)}} e^{\frac{-(x-E(Z))^2}{2\text{Var}(Z)}}$$
$$= \frac{1}{\sqrt{2\pi \sum_{i=1}^{K} p_i^2 \sigma_i^2}} e^{\frac{-(x - \sum_{i=1}^{K} p_i \mu_i)^2}{2 \sum_{i=1}^{K} p_i^2 \sigma_i^2}}$$

For X:

$E(X) = \sum_{i=1}^{K} p_i \mu_i$

$\text{Var}(X) = \sum_{i=1}^{K} p_i(\mu_i^2 + \sigma_i^2) - (\sum_{i=1}^{K} p_i \mu_i)^2$

MGF of $X = \sum_{i=1}^{n} p_i e^{\mu_i t + \sigma_i^2 t^2/2}$

For Z:

$E(Z) = \sum_{i=1}^{K} p_i \mu_i$

$\text{Var}(Z) = \sum_{i=1}^{K} p_i^2 \sigma_i^2$

PDF of $Z = \dfrac{1}{\sqrt{2\pi \sum_{i=1}^{K} p_i^2 \sigma_i^2}} e^{\frac{-(x - \sum_{i=1}^{K} p_i \mu_i)^2}{2 \sum_{i=1}^{K} p_i^2 \sigma_i^2}}$

MGF of $Z = \prod_{i=1}^{K} e^{p_i \mu_i t + p_i^2 \sigma_i^2 t^2/2}$

3 Problem 3

Using Markov's inequality, prove the following one-sided version of Chebyshev's inequality for random variable X with mean μ and variance σ^2: $P(X-\mu \geq \tau) \leq \frac{\sigma^2}{\sigma^2+\tau^2}$ if $\tau > 0$, and $P(X-\mu \geq \tau) \geq 1-\frac{\sigma^2}{\sigma^2+\tau^2}$ if $\tau < 0$. [15 points]

Solution :

To prove :

$$P(X-\mu \geq \tau) \leq \frac{\sigma^2}{\sigma^2+\tau^2} \qquad if \quad \tau > 0$$

$$P(X-\mu \geq \tau) \geq 1-\frac{\sigma^2}{\sigma^2+\tau^2} \qquad if \quad \tau < 0$$

Markov's inequality for any random variable X is :

$$P(X \geq a) \leq \frac{E(X)}{a} \qquad for \quad a > 0$$

We can also show that for any random variable X:

$$P(X \geq a) \leq P(X^2 \geq a^2) \leq \frac{E(X^2)}{a^2}$$

As :

$$P(X^2 \geq a^2) = P(X \geq a) + P(X \leq -a)$$
$$P(X^2 \geq a^2) \geq P(X \geq a)$$

As X is a random variable, X^2 is also a random variable, so by Markov's inequality :

$$P(X^2 \geq a^2) \leq \frac{E(X^2)}{a^2} \qquad for \quad a > 0$$

For some $t \in \mathbb{R}$, $t > 0$ and $\tau > 0$, if X is a random variable, then $(X-\mu+t)$ is also a random variable. We can show that :

$$P(X-\mu \geq \tau) = P(X-\mu+t \geq \tau+t)$$

$$P(X-\mu+t \geq \tau+t) \leq \frac{E\left[(X-\mu+t)^2\right]}{(\tau+t)^2}$$

$$= \frac{E\left[(X-\mu)^2\right] + E\left[2\cdot t\cdot(X-\mu)\right] + E\left[t^2\right]}{(\tau+t)^2}$$

$$= \frac{\sigma^2+t^2}{(\tau+t)^2}$$

We now find the minimum of the above expression with respect to t for the tightest bound (similar to what we did in proving Chebyshev's inequality).

$$\frac{d\left(\frac{\sigma^2+t^2}{(\tau+t)^2}\right)}{dt} = 0$$

$$\frac{2\cdot(\tau\cdot t-\sigma^2)}{(\tau+t)^3} = 0$$

$$t = \frac{\sigma^2}{\tau}$$

After using the above value of t, the tightest bound is :

$$P(X-\mu \geq \tau) \leq \frac{\sigma^2}{\sigma^2+\tau^2}$$

The above inequality is proved for $\tau > 0$. If $\tau < 0$,

$$P(X-\mu \geq \tau) = 1 - P(X-\mu \leq \tau) = 1 - P(\mu - X \geq -\tau) = 1 - P(\mu - X + t \geq t - \tau)$$

For calculating bounds on $P(\mu - X \geq -\tau)$ we proceed similar as in the first part. For some $t \in \mathbb{R}$, $t > 0$ and $-\tau > 0$, if X is a random variable, then $(\mu - X + t)$ is also a random variable. We can show that :

$$P(\mu - X + t \geq t - \tau) \leq \frac{E\left[(\mu - X + t)^2\right]}{(t-\tau)^2}$$

$$= \frac{E\left[(\mu - X)^2\right] + E\left[2 \cdot t \cdot (\mu - X)\right] + E\left[t^2\right]}{(t-\tau)^2}$$

$$= \frac{\sigma^2 + t^2}{(t-\tau)^2}$$

We minimize this expression with respect to t for tightest bound :

$$\frac{d\left(\frac{\sigma^2+t^2}{(t-\tau)^2}\right)}{dt} = 0$$

$$\frac{2 \cdot (t \cdot \tau + \sigma^2)}{(t-\tau)^2} = 0$$

$$t = -\frac{\sigma^2}{\tau}$$

After using the above value of t, the tightest bound is :

$$P(\mu - X + t \geq t - \tau) \leq \frac{\sigma^2}{\sigma^2+\tau^2}$$

$$1 - P(\mu - X + t \geq t - \tau) \geq 1 - \frac{\sigma^2}{\sigma^2+\tau^2}$$

$$\therefore P(X-\mu \geq \tau) \geq 1 - \frac{\sigma^2}{\sigma^2+\tau^2} \qquad for \quad \tau < 0$$

Hence proved.

4 Problem 4

Given stuff you've learned in class, prove the following bounds: $P(X \geq x) \leq e^{-tx}\phi_X(t)$ for $t > 0$, and $P(X \leq x) \leq e^{-tx}\phi_X(t)$ for $t < 0$. Here $\phi_X(t)$ represents the MGF of random variable X for parameter t. Now consider that X denotes the sum of n independent Bernoulli random variables $X_1, X_2, ..., X_n$ where $E(X_i) = p_i$. Let $\mu = \sum_{i=1}^{n} p_i$. Then show that $P(X > (1+\delta)\mu) \leq \frac{e^{\mu(e^t-1)}}{e^{(1+\delta)t\mu}}$ for any $t \geq 0, \delta > 0$. You may use the inequality $1 + x \leq e^x$. Further show how to tighten this bound by choosing an optimal value of t. [15 points]

Solution :

We know that Markov's inequality states that:
$P(X \geq x) \leq \dfrac{E(X)}{x} \quad \forall x > 0$

Consider the case of $t > 0$,
$P(X \geq x) = P(e^{tX} \geq e^{tx})[\because e^{tx}$ is monotonically increasing function $]$

$\leq \dfrac{E(e^{tX})}{e^{tx}} \quad [\because e^{tx} > 0,$ we can use Markov's inequality$]$

$\therefore P(X \geq x) \leq e^{-tx}\phi_X(t)$ for $t > 0$

In case $t < 0$,
$P(X \leq x) = P(e^{tX} \geq e^{tx})[\because e^{tx}$ is monotonically decreasing function $]$

$\leq \dfrac{E(e^{tX})}{e^{tx}} \quad [\because e^{tx} > 0,$ we can use Markov's inequality$]$

$\therefore P(X \leq x) \leq e^{-tx}\phi_X(t)$ for $t < 0$

We need to find the upper bound for $P(X > (1+\delta)\mu)$

$$P(X > (1+\delta)\mu) = P(e^{tX} > e^{(1+\delta)\mu t}) \quad [\because t \geq 0]$$

From Markov's inequality, we know that

$$P(e^{tX} > e^{(1+\delta)\mu t}) \leq \frac{E(e^{tX})}{e^{(1+\delta)\mu t}} \quad [\because e^{(1+\delta)\mu t} > 0]$$

In order to find $E(e^{tX})$,

$$E(e^{tX}) = E(e^{\sum_{i=1}^{n} tX_i}) = \sum_{x_1}\sum_{x_2}...\sum_{x_n} e^{tx_1}e^{tx_2}...e^{tx_n}P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

Since $X_1, X_2, ..., X_n$ are independent Bernoulli variables,

$$= \sum_{x_1} e^{tx_1}P(X_1 = x_1) \cdot \sum_{x_2} e^{tx_2}P(X_2 = x_2)... \sum_{x_n} e^{tx_n}P(X_n = x_n)$$

$$\implies \prod_{i=1}^{n}(1 - p_i + p_i e^t)$$

Now for any $i \in \{1, 2, ...n\}$,

$$(1 - p_i + p_i e^t) = 1 + p_i(e^t - 1) \leq e^{p_i(e^t-1)} \quad [\because 1 + x \leq e^x]$$

$$\prod_{i=1}^{n}(1 - p_i + p_i e^t) \leq \prod_{i=1}^{n} e^{p_i(e^t-1)} = e^{\sum_{i=1}^{n} p_i(e^t-1)} = e^{\mu(e^t-1)}$$

$\therefore E(e^{tX}) \leq e^{\mu(e^t-1)}$

Finally, $P(X > (1+\delta)\mu) = P(e^{tX} > e^{(1+\delta)\mu t}) \leq \dfrac{e^{\mu(e^t-1)}}{e^{(1+\delta)\mu t}}$

Hence Proved.

To tighen the bound, we must choose a value of t such that the right hand side of the inequality $P(X > (1+\delta)\mu) \leq \dfrac{e^{\mu(e^t-1)}}{e^{(1+\delta)t\mu}}$ is minimized.
Consider function $g(t) = \dfrac{e^{\mu(e^t-1)}}{e^{(1+\delta)t\mu}}$

To minimize $g(t)$,

$$g'(t) = 0$$

$$\implies \mu e^t e^{\mu(e^t-1)} e^{-(1+\delta)t\mu} - \mu(1+\delta)e^{\mu(e^t-1)}e^{-(1+\delta)t\mu} = 0$$

$$\implies e^t = 1+\delta$$

$$\therefore t = \ln(1+\delta)$$

To verify that this value of t, minimizes $g(t)$, let's compute $g''(t)$

$$g''(t) = \mu e^t e^{\mu(e^t-1)}e^{-(1+\delta)t\mu} + \mu^2 e^{2t}e^{\mu(e^t-1)}e^{-(1+\delta)t\mu} - 2\mu^2(1+\delta)e^t e^{\mu(e^t-1)}e^{-(1+\delta)t\mu} + \mu^2(1+\delta)^2 e^{\mu(e^t-1)}e^{-(1+\delta)t\mu}$$

$$g''(\ln(1+\delta)) = e^{\mu\delta}(1+\delta)^{-(1+\delta)\mu}(\mu(1+\delta) + \mu^2(1+\delta)^2 - 2\mu^2(1+\delta)^2 + \mu^2(1+\delta)^2) = e^{\mu\delta}(1+\delta)^{-(1+\delta)\mu}(\mu(1+\delta))$$

$$\therefore g''(\ln(1+\delta)) > 0$$

Therefore $t = \ln(1+\delta)$ is the optimal value to tighten the bound of the inequality.

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

5 Problem 5

Consider N independent random variables $X_1, X_2, ..., X_N$, such that each variable X_i takes on the values $1, 2, 3, 4, 5$ with probability $0.05, 0.4, 0.15, 0.3, 0.1$ respectively. For different values of $N \in \{5, 10, 20, 50, 100, 200, 500, 1000, 5000, 10000\}$, do as follows:

1. Plot the (empirically determined) distribution of the average of these random variables ($X_{avg}^{(N)} = \sum_{i=1}^{N} X_i/N$) in the form of a histogram with 50 bins.

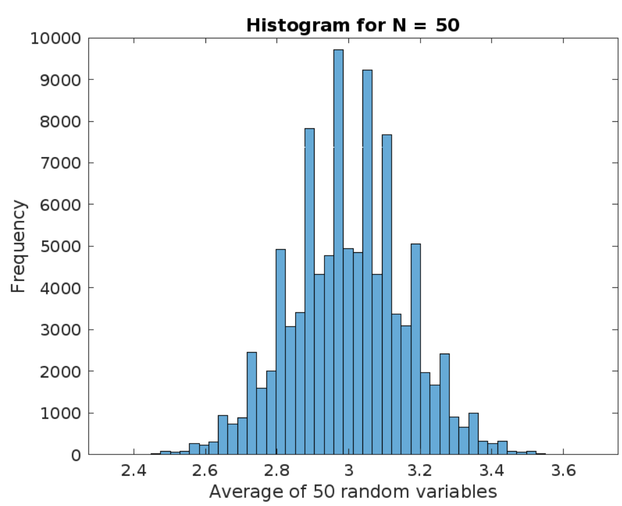
2. Empirically determine the CDF of $X_{avg}^{(N)}$ using the `ecdf` command of MATLAB (this is called the empirical CDF). On a separate figure, plot the empirical CDF. On this, overlay the CDF of a Gaussian having the same mean and variance as $X_{avg}^{(N)}$. To get the CDF of the Gaussian, use the `normcdf` function of MATLAB.

3. Let $E^{(N)}$ denote the empirical CDF and $\Phi^{(N)}$ denote the Gaussian CDF. Compute the maximum absolute difference (MAD) between $E^{(N)}(x)$ and $\Phi^{(N)}(x)$ numerically, at all values x returned by `ecdf`. For this, read the documentation of `ecdf` carefully. Plot a graph of MAD as a function of N. [3+3+4 = 10 points]

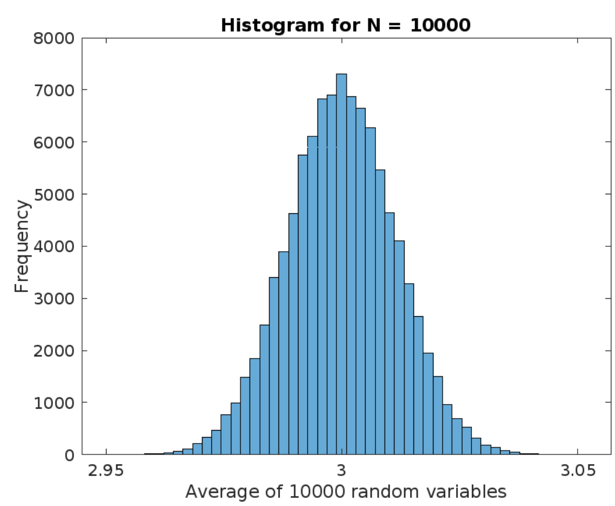
Solution :

The scripts for generating the histograms and plots are saved in the folder **Problem5** under the names `generate.m` and `ECDF.m` respectively. The histograms generated are stored inside the **Histograms** folder under **Problem5**. The plots of ECDF and NCDF are stored inside the **ECDF_NCDF** folder under **Problem5**. The plot of MAD as a function of N is stored as **MAD.png** inside the **Problem5** folder.

1. The histogram of the distribution of the average of these random variables is generated on running the command `run("generate.m")`. It generates 10 images of the histograms for each value of N, saved under the names **HistN.png** where $N \in \{5, 10, 20, 50, 100, 200, 500, 1000, 5000, 10000\}$.



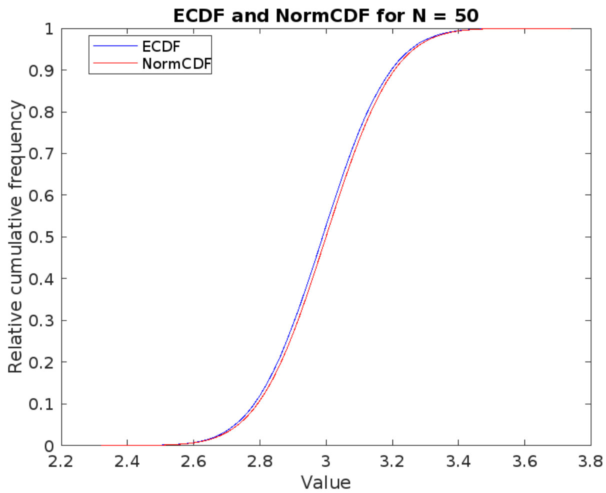
(a) Histogram for N=50



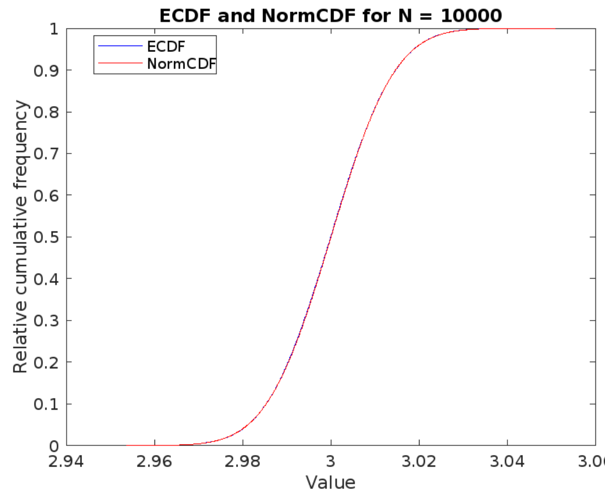
(b) Histogram for N=10000

From the above histograms, we can observe that as the value of N increases, the sample mean $X_{avg}^{(N)}$ assumes a Gaussian distribution. This is in accordance with the **Central Limit Theorem** done in class which says that the empirical mean calculated from a large number of samples is a random variable following a Gaussian distribution with mean equal to the true mean μ (of the distribution from which the samples were drawn). In this case the true mean is $E(X_i) = 0.05{\cdot}1{+}0.4{\cdot}2{+}0.15{\cdot}3{+}0.3{\cdot}4{+}0.1{\cdot}5 = 3$. We can see the concentration around 3 increases as N values increases.

2. The plots of empirical CDF and Gaussian cdf are generated by running the command `run("ECDF.m")`. On running this file, 10 images are generated, each containing the ECDF and NCDF for each value of N under the names **ecdf_normcdfN.png** where $N \in \{5, 10, 20, 50, 100, 200, 500, 1000, 5000, 10000\}$.



(a) ECDF and NCDF plot for N=50



(b) ECDF and NCDF plot for N=10000

From these plots we can again see that the empirical CDF of the sample mean $X_{avg}^{(N)}$ approaches the CDF of Gaussian variables with same mean and variance.

3. Along with the above 10 images, an 11th image of the Maximum Absolute Difference(MAD) vs N is also generated on running `ECDF.m`.

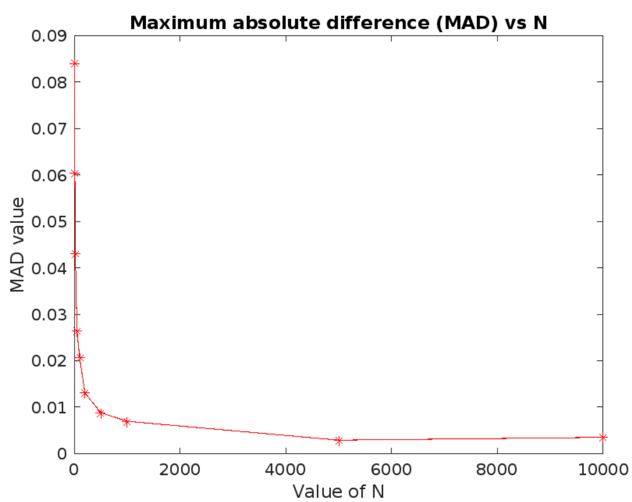


Figure 3: MAD vs N plot

Since the ECDF plot approaches NCDF plot as N increases, the maximum absolute difference between the two curves (MAD) decreases as N increases. As $N \longrightarrow \infty$, the two plots exactly coincide and MAD $\longrightarrow 0$.

6 Problem 6

Read in the images T1.jpg and T2.jpg from the homework folder using the MATLAB function `imread` and cast them as a double array using the code

```
im = double(imread('T1.jpg');
```

These are magnetic resonance images of a portion of the human brain, acquired with different settings of the MRI machine. They both represent the same anatomical structures and are perfectly aligned (i.e. any pixel at location (x, y) in both images represents the exact same physical entity). Consider random variables I_1, I_2 which denote the pixel intensities from the two images respectively. Write a piece of MATLAB code to shift the second image along the X direction by t_x pixels where t_x is an integer ranging from -10 to +10. While doing so, assign a value of 0 to unoccupied pixels. For each shift, compute the following measures of dependence between the first image and the *shifted version* of the second image:

- the correlation coefficient ρ,
- a measure of dependence called quadratic mutual information (QMI) defined as $\sum_{i_1}\sum_{i_2}(p_{I_1 I_2}(i_1, i_2) - p_{I_1}(i_1)p_{I_2}(i_2))^2$, where $p_{I_1 I_2}(i_1, i_2)$ represents the *normalized* joint histogram (*i.e.*, joint pmf) of I_1 and I_2 ('normalized' means that the entries sum up to one).

For computing the joint histogram, use a bin-width of 10 in both I_1 and I_2. For computing the marginal histogram, you need to integrate the joint histogram along one of the two directions respectively. You should write your own joint histogram routine in MATLAB - do not use any inbuilt functions for it. Plot a graph of the values of ρ versus t_x, and another graph of the values of QMI versus t_x.

Repeat exactly the same steps when the second image is a negative of the first image, i.e. $I_2 = 255 - I_1$.

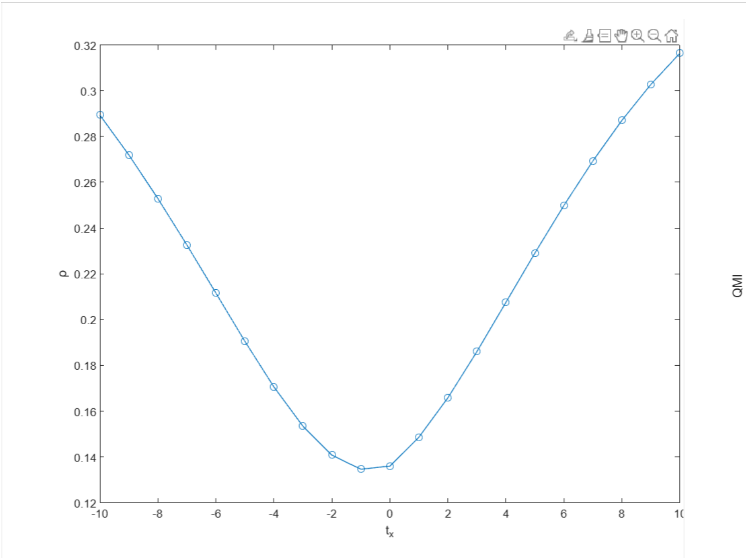
Comment on all the plots. In particular, what do you observe regarding the relationship between the dependence measures and the alignment between the two images? Your report should contain all four plots labelled properly, and the comments on them as mentioned before. [25 points]

Solution :

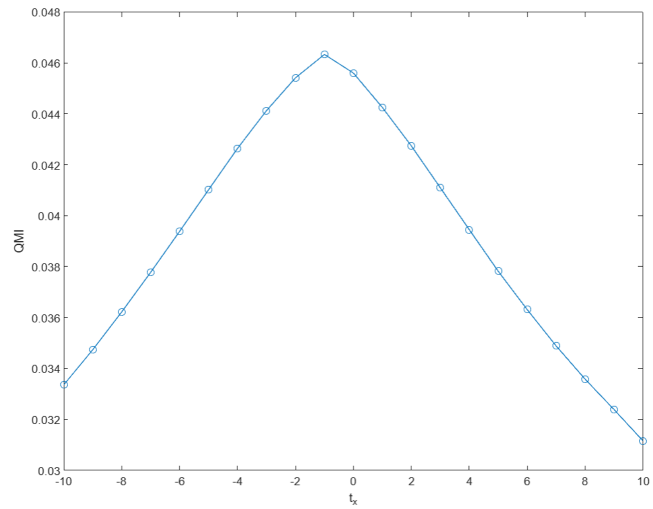
All the scripts are stored in the **Problem6** folder. The plots generates are also stored inside the same folder under the names `positive_correlation.png`, `negative_correlation.png`, `positive_QMI.png` and `negative_QMI.png`. Directions for code usage :

- Run the command `run("img_read.m")` if you want to calculate for $I_1,\ I_2$ or run command `run("negative_img_read.m")` if you want to calculate for $I_1,\ 255 - I_1$.
- Run command `run("cal_correlation.m")` for plotting ρ vs t_x graph.
- Run command `run("cal_QMI.m")` for plotting QMI vs t_x graph.

Plots for I_1, I_2



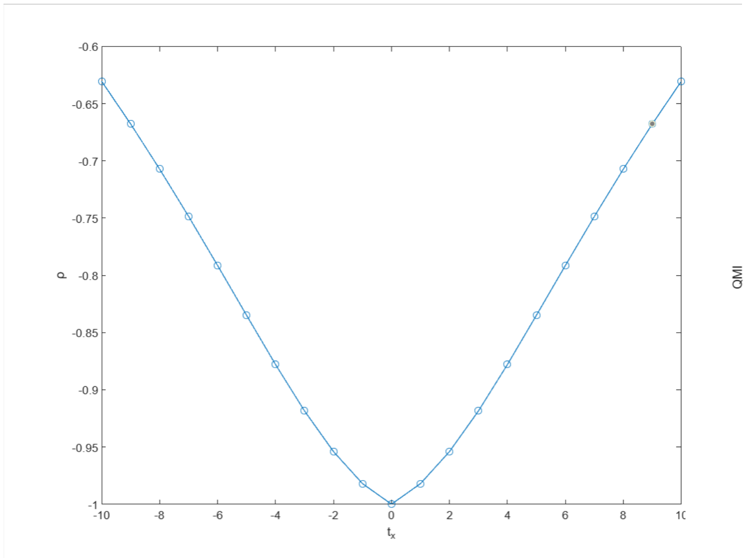
(a) Correlation coefficient (ρ) for I_1, I_2



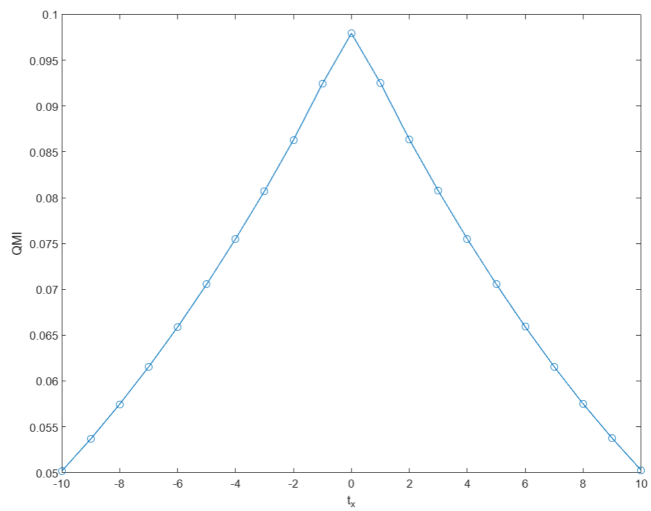
(b) QMI for I_1, I_2

Let's consider the case of images T_1 and T_2, these images are intuitively have a very strong correlation as they are the images of same brain structure taken with different settings. To capture this correlation quantitatively, we use two measures: the correlation coefficient and the QMI.
From the graph of Correlation coefficient we observe that the images are least correlated ($\rho \approx 0.13$)when $t_x = -1$. The correlation coefficient plot thus seems to show low correlation between the two images and implies the images are least correlated when the second image is offset by -1 pixel. But the graph of QMI attains a maxima at the same t_x.
As we know the two images are highly similar, we can say that QMI is a better metric for calculating similarity between images than correlation coefficient as it looks at the images more "intelligently" and can guage more complex relations. QMI correctly analyzes the two images and gives high correlation when the second image is offset by 1 pixel. This implies that the image 1 is perfectly aligned with image 2 moved to the left by 1 pixel.
QMI and correlation coefficient seem to give contradictory results because correlation coefficient looks only at the mean and variance, but QMI looks at the joint pdf, and hence gives correct results. The correlation coefficient fails here and associates a lower correlation value to actual high correlation value.

Plots for I_1, $255 - I_1$



(a) Correlation coefficient (ρ) for I_1, $255 - I_1$



(b) QMI for I_1, $255 - I_1$

In this part the correlation coefficient plot is symmetric, and shows complete negative correlation ($\rho = -1$) when $t_x = 0$ and it's higher for other values of t_x. This is expected as the other image is $255 - I_1$ (negative of I_1) and should have a high correlation with I_1 but in the negative direction and for other values of t_x the images deviate from exact opposite nature and that's why the correlation coefficient increases.
The QMI values are still positive as both the images are still similar, it also shows peak for $t_x = 0$ because the images are most similar in this case and it decreases due to offset as the images become less similar. Note that QMI is always non-negative and shows only magnitude of correlation, not the sign(positively or negatively correlated). Here since alignment was achieved at 0, both plots had extrema at 0.

7 Problem 7

Derive the covariance matrix of a multinomial distribution using moment generating functions. You are not allowed to use any other method. Since a covariance matrix $\boldsymbol{C}$ is square and symmetric, it is enough to derive expression for the diagonal elements C_{ii} and the off-diagonal elements $C_{ij}, i \neq j$. [10 points]

Solution :

Let there be a multinomial distribution of random variables $[X_1, X_2, \ldots, X_k]$ with $E[X_i] = p_i, n = \Sigma_{i=1}^{k} E[X_i]$. We have to derive the covariance matrix, $\mathbf{C}$ of all the random variables in the group using moment generating function. The size of this matrix will be $n \times n$, where :

$$C_{ij} = Cov(X_i, X_j)$$

Let $\mathbf{t} = [t_1, t_2, \ldots, t_n]$ be a vector of constants t_i.
Let X be another random variable such that :

$$X = t_1 \cdot X_1 + t_2 \cdot X_2 + \ldots t_n \cdot X_n$$

We know that the MGF of multinomial distribution is :

$$\Phi_X(\mathbf{t}) = (p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_{k-1} e^{t_{k-1}} + p_n e^{t_k})^n$$

We see that :

$$E[X_i] = \left[\frac{\partial \Phi_X(\mathbf{t})}{\partial t_i}\right]_{\mathbf{t}=0} = \left[n \cdot (p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_{k-1} e^{t_{k-1}} + p_n e^{t_k})^{n-1} \cdot p_i \cdot e^{t_i}\right]_{\mathbf{t}=0} = n \cdot p_i$$

$$E[X_i \cdot X_j] = \left[\frac{\partial^2 \Phi_X(\mathbf{t})}{\partial t_i \, \partial t_j}\right]_{\mathbf{t}=0} = \left[n \cdot (n-1)(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_{k-1} e^{t_{k-1}} + p_n e^{t_k})^{n-2} \cdot p_i \cdot p_j \cdot e^{t_i} \cdot e^{t_j}\right]_{\mathbf{t}=0}$$

$$= n \cdot (n-1) \cdot p_i \cdot p_j \qquad for\ i \neq j$$

$$E[X_i^2] = \left[\frac{\partial^2 \Phi_X(\mathbf{t})}{\partial t_i^2}\right]_{\mathbf{t}=0} = n \cdot (n-1) \cdot p_i^2 + n \cdot p_i$$

So for any general term $C_{ij} (i \neq j)$ of $\mathbf{C}$:

$$C_{ij} = E[X_i \cdot X_j] - E[X_i] \cdot E[X_j] \qquad for\ i \neq j$$

$$C_{ij} = n \cdot (n-1) \cdot p_i \cdot p_j - (n \cdot p_i) \cdot (n \cdot p_j) \qquad for\ i \neq j$$

$$C_{ij} = -n \cdot p_i \cdot p_j \qquad for\ i \neq j$$

For Any term C_{ii} :

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

$$C_{ii} = n \cdot (n-1) \cdot p_i^2 + n \cdot p_i \; - \; (n \cdot p_i) \cdot (n \cdot p_i)$$

$$C_{ii} = n \cdot p_i \cdot (1 - p_i)$$

So finally for Covariance matrix $\mathbf{C}$:

$$C_{ij} = \begin{cases} -n \cdot p_i \cdot p_j & if\ i \neq j, \\ n \cdot p_i \cdot (1 - p_i) & if\ i = j \end{cases}$$