CS 215 - Data Analysis and Interpretation

Assignment 3

Saksham Rathi - 22B1003
Tanmay Gejapati - 22B0969
Aditya Singh - 22B1844

Contents

Question 1

Part (a)

X_1 denotes the number of times we have to pick a book such that we move from having picked books of 0 $(i-1)$ distinct colors to $1(i)$ distinct colors. Clearly if we pick any one book out of the n books, we will achieve our goal. So, the value of X_1 is 1. (This means that after we pick a single book, we will achieve our goal of having one distinct book.)

In the next part of the question, we are given $i-1$ distinct books. We need to calculate the probability of picking a book with different color (i.e. different from the previous i-1 colors). The number of books which are of a color different from the given $i-1$ colors is n(Total number of books of distinct colors) $-(i-1)=(n-i+1)$. The probability of choosing these books out of the n books is:

$$P = \frac{\text{Number of books which we can choose}}{\text{Total number of choices available}} = \frac{n-i+1}{n}$$

(The total number of choices always remain n because we are replacing every book after picking.)

Part (b)

According to the definition of geometric random variable and it's parameter given in the question statement, we have:

$$P(X_i = k) = (1-p)^{k-1}p$$

We need to find this p. (We can write this expression because X_i is already given to be a geometric random variable due to independence.)

$X_i = k$ means that we will have to pick k books for moving from having picked $i-1$ distinct colors to i distinct colors. This in turn means, that out of the k books, we need to pick, the first $k-1$ books have colors common with the $i-1$ distinct colors given and the k^{th} book has a color different from the $i-1$ colors given. Out of the $k-1$ books to be chosen, each will have $i-1$ options out of total n options. (They will have colors overlapping with the $i-1$ colors given initially.) So, each of these $k-1$ books have a probability of $\frac{i-1}{n}$ to have a color overlapping with the the $i-1$ colors. By product rule of counting, the probability that all the $k-1$ books have colors same as that of one of $i-1$ colors is $\prod_{j=1}^{k-1} \frac{i-1}{n} = (\frac{i-1}{n})^{(k-1)}$. The k^{th} should have color different from the $i-1$ colors given initially. The total number of options are n. Hence the k^{th} book will have $n-(i-1) = n-i+1$ options. The probability of doing this is $\frac{n-i+1}{n}$. Multiplying this with the initial probability product, we get $(\frac{i-1}{n})^{(k-1)} \times \frac{n-i+1}{n}$. Comparing this with the initial expression of definition of parameter p, we get: The parameter of geometric random variable X_i is $\frac{n-i+1}{n}$.

Part (c)

The expectation of a random variable is :

$$E(X) = \sum_i x_i P(X = x_i) \tag{1}$$

In our case:

$$P(Z = k) = (1-p)^{(k-1)}p$$

Combining these two equations, we get: (The values of k will range from 1 to ∞, for the total sum of probability to be one.)

$$E(Z) = \sum_{k=1}^{\infty} kP(Z = k) = \sum_{k=1}^{\infty} k((1-p)^{(k-1)}p)$$

$$\begin{aligned}
E(Z) &= \sum_{k=1}^{\infty} k((1-p)^{(k-1)}p) \\
&= p\sum_{k=1}^{\infty} k(1-p)^{(k-1)} \\
E(Z)\times(1-p) &= p\sum_{k=1}^{\infty} k(1-p)^{k} \\
E(Z)\times(1-(1-p)) &= p[(1+2(1-p)+3(1-p)^2+\dots+\dots) \\
&\quad -(1(1-p)+2(1-p)^2+\dots+\dots] \\
&\quad \text{(Subtracting above two eqs)} \\
E(Z)\times p &= p[1+(2-1)(1-p)+(3-2)(1-p)^2+\dots+\dots] \\
&\quad \text{(Combining terms with similar powers of (1-p))} \\
&= p[1+(1-p)+(1-p)^2+\dots+\dots] \\
&= p[\frac{1}{1-(1-p)}] \\
&\quad \text{(Applying the infinite geometric series formula} = \frac{a}{1-r}) \\
&= 1 \\
E(Z) &= 1/p
\end{aligned}$$

Hence, we have proved that the expected value of a geometric random variable with parameter p is 1/p.

Let us now calculate the variance of a geometric random variable.

$$Var(X) = E(X^2) - [E(X)]^2 \tag{2}$$

Let us calculate $E(X^2)$ (We will use the Law Of The Unconscious Statistician)

$$
\begin{aligned}
E(Z^2) &= \sum_{k=1}^{\infty} k^2((1-p)^{(k-1)}p) \\
&= p\sum_{k=1}^{\infty} k^2(1-p)^{(k-1)} \\
&= p[1+\sum_{k=1}^{\infty}(k+1)^2(1-p)^k](\text{Shifting of variable k}) \\
E(Z^2)(1-p) &= p\sum_{k=1}^{\infty} k^2(1-p)^k \\
E(Z^2)(1-p)^2 &= p\sum_{k=1}^{\infty} k^2(1-p)^{k+1} \\
&= p[\sum_{k=1}^{\infty}(k-1)^2(1-p)^{k+1}]
\end{aligned}
$$

(Manipulated, the initial term is 0)

$$
E(Z^2)+E(Z^2)(1-p)^2-2\times E(Z^2)(1-p) = p[1+\sum_{k=1}^{\infty}(1-p)^k((k+1)^2+(k-1)^2-2k^2]
$$

(Just adding the above three equations)

$$
\begin{aligned}
E(Z^2)(1+(1-p)^2-2(1-p)) &= p[1+\sum_{k=1}^{\infty}(1-p)^k(k^2+2k+1+k^2-2k+1-2k^2)] \\
E(Z^2)(1+1-2p+p^2-2+2p) &= p[1+\sum_{k=1}^{\infty}(1-p)^k(2)] \\
E(Z^2)(p^2) &= p[1+2\frac{1-p}{1-(1-p)}]
\end{aligned}
$$

(Applying the infinite geometric series formula $= \frac{a}{1-r}$)

$$
\begin{aligned}
E(Z^2) &= \frac{1}{p}[1+2\frac{1-p}{p}] \\
E(Z^2) &= \frac{2-p}{p^2}
\end{aligned}
$$

Using the obtained results:

$$
E(Z^2)-E(Z)^2 = \frac{2-p}{p^2}-(\frac{1}{p})^2 = \frac{1-p}{p^2}
$$

Therefore, the variance of a geometric random variable is $\frac{1-p}{p^2}$.

Part (d)

If A and B are two random variables, then we have:

$$
E(A+B) = E(A)+E(B)
$$

(Means just add up) In our case we have:

$$
X^{(n)} = X_1+X_2+\cdots+X_n
$$

$$E(X^{(n)}) = E(X_1 + X_2 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$$

As, we have proved in part(c) 1.3, the expectation value of a geometric random variable is $\frac{1}{p}$, where p is it's parameter.

From part(b) 1.2, we have that the parameter of X_i is $\frac{n-i+1}{n}$. Therefore:

$$E(X_i) = \frac{1}{p_{X_i}} = \frac{1}{\frac{n-i+1}{n}} = \frac{n}{n-i+1}$$

$$E(X^{(n)}) = E(X_1) + \cdots + E(X_n) = \sum_{i=1}^{n} \frac{n}{n-i+1} = n[\frac{1}{1} + \cdots + \frac{1}{n}]$$

This expression cannot be simplified further. Hence we have derived the expression for $E^{(n)}$.

Part (e)

Let us first start with writing the expression of $Var(X^{(n)})$.

$$
\begin{aligned}
Var(X^{(n)}) &= E((X^{(n)})^2) - (E(X^{(n)}))^2 \\
&= E((X_1 + X_2 + X_3 + \dots X_n)^2) - (E(X_1 + X_2 + \dots + X_n))^2 \\
&= E(\sum_{i=1}^{n} X_i^2 + 2\sum_{i}\sum_{i<j} X_i X_j) - (E(X_1) + E(X_2) + \dots + E(X_n))^2 \\
&(E(A+B) = E(A) + E(B), \text{for random variables } A \text{ and } B) \\
&= \sum_{i=1}^{n} E(X_i^2) + 2\sum_{i}\sum_{i<j} E(X_i X_j) - \sum_{i=1}^{n}(E(X_i))^2 - 2\sum_{i}\sum_{i<j} E(X_i)E(X_j) \\
&= \sum_{i=1}^{n}(E(X_i^2) - (E(X_i))^2) + 2\sum_{i}\sum_{i<j}(E(X_i X_j) - E(X_i)E(X_j)) \\
&= \sum_{i=1}^{n}(Var(X_i)) + 2\sum_{i}\sum_{i<j}(Cov(X_i, X_j)) \\
&(X_i \text{ is a geometric variable, so it's variance is } \frac{1-p}{p^2}, \text{ where } p \text{ is it's parameter.}) \\
&= \sum_{i=1}^{n}(\frac{1-p_i}{p_i^2}) + 2\sum_{i}\sum_{i<j}(Cov(X_i, X_j))
\end{aligned}
$$

(Since, X_i and X_j are geometric random variables with two different parameters p_i and p_j, they are independent. Moreover, even if we see by the definition of X_i, this fact is clear.)

$$
\begin{aligned}
&= \sum_{i=1}^{n}(\frac{1-p_i}{p_i^2}) \\
&\leq \sum_{i=1}^{n}\frac{1}{p_i^2} \\
&= \sum_{i=1}^{n}\frac{n}{(n-i+1)^2}
\end{aligned}
$$

$p_i = \dfrac{n-i+1}{n}$ from part (b) 1.2

$$
\begin{aligned}
&= n^2\sum_{i=1}^{n}\frac{1}{(n-i+1)^2} \\
&= n^2[\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \dots + \frac{1}{n^2}] \\
&\leq n^2[\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \dots + \frac{1}{n^2} + \frac{1}{(n+1)^2} + \dots] \\
&\leq n^2[\frac{\pi^2}{6}]
\end{aligned}
$$

From the information given in the question, we can upper-bound the above series

Hence, we have found an upper bound on the variance of $X^{(n)}$ which is $\frac{n^2\pi^2}{6}$.

Part (f)

The code for this part is present in the file $a3\_q1\_f.m$.

The following are the graphs of $E(X^{(n)})$ versus n for different values of n. The graphs are

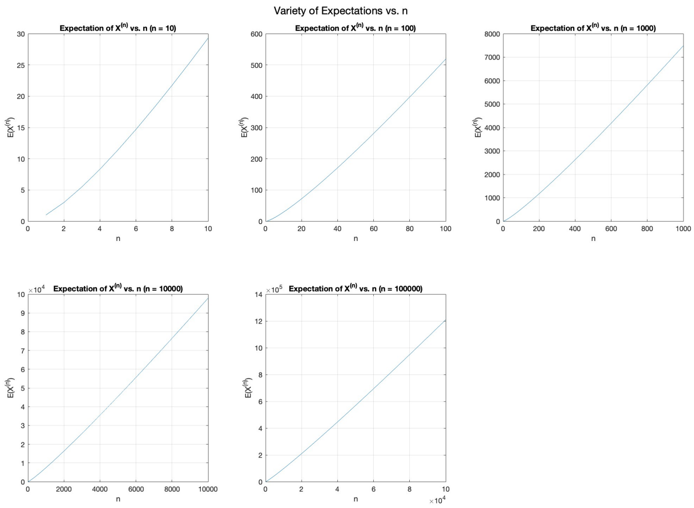


Figure 1: Graphs of $E(X^{(n)})$ versus n for different n

clearly monotonically increasing for all n.

In the next part of the question we need to find $f(n)$ such that $E(X^{(n)}) = \Theta(f(n))$. According to the definition of $\Theta(f(n))$, if $g(n) = \Theta(f(n))$, we must have:

$$c_1 f(n) \leq g(n) \leq c_2 f(n)$$

for some positive real constants c_1, c_2 and for all $n > n_0$, where $n_0 \in \mathbb{N}$.

Let us find the upper bound of $E(X^{(n)})$. For simplicity, let us consider $n = 2^m$, for some $m \geq 0$. From part (d) 1.4, we have:

$$E(X^{(n)}) = n[\frac{1}{1} + \cdots + \frac{1}{n}]$$

Since we are taking $n = 2^m$:

$$E(X^{(n)}) = 2^m[\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \cdots + \frac{1}{2^m - 1} + \frac{1}{2^m}]$$

$$\leq 2^m[\frac{1}{1} + \frac{1}{2} \times 2 + \frac{1}{4} \times 4 + \frac{1}{8} \times 8 + \cdots + \frac{1}{2^{m-1}} \times 2^{m-1} + \frac{1}{2^m}]$$

(Grouping terms based on the powers of two, and taking the largest element out of each group)

$$= 2^m[1 + 1 + \cdots + 1 + \frac{1}{2^m}]$$

(1 will occur m times, from 0 to m-1)

$$\leq 2^m[1 + 1 + \cdots + 1 + 1]$$

$$= n[m+1]$$

$$\leq n[2m]$$

$$= 2n[log(n)]$$

(Since, $n = 2^m, log(n) = m$)Thus we have:

$E(X^{(n)}) \leq 2nlog(n)$

Following a similar procedure, let us find the lower bound of $E(X^{(n)})$.

$$
\begin{aligned}
E(X^{(n)}) &= 2^m[\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \cdots + \frac{1}{2^m - 1} + \frac{1}{2^m}] \\
&\geq 2^m[\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 4 + \frac{1}{16} \times 8 + \cdots + \frac{1}{2^m} \times 2^{m-1} + \frac{1}{2^m}] \\
&\text{(Grouping terms based on the powers of two, and taking the power of two just next of the group)} \\
&= 2^m[\frac{1}{2} + \frac{1}{2} + \cdots + \frac{1}{2} + \frac{1}{2^m}] \\
&(\frac{1}{2} \text{ will occur m times, from 0 to m-1}) \\
&\geq 2^m[\frac{1}{2} + \frac{1}{2} + \cdots + \frac{1}{2}] \\
&= 2^m[\frac{1}{2} \times (m-1)] \\
&\geq 2^m[\frac{1}{2} \times (\frac{m}{2})] \\
&\text{(For m} \geq 2) \\
&\geq 2^m[\frac{m}{4}] \\
&\text{(Since, } n = 2^m, log(n) = m)\text{Thus we have:}
\end{aligned}
$$

$$E(X^{(n)}) \geq n\frac{log(n)}{4}$$

From, the above two bounds, we have:

$$\frac{nlog(n)}{4} \leq E(X^{(n)}) \leq 2nlog(n)$$

Comparing the above equation with the definition of $\Theta(f(n))$, we get:

$$f(n) = nlog(n)$$

(Where the constants c_1 and c_2 are $\frac{1}{4}$ and 2 respectively).

(We did the above procedure for powers of two, just for simplicity. Even if n is not a power of two, we will be able to find constants c_1 and c_2, for which the function will be bounded by multipliers of f(n). Thus, for all n, f(n) will remain same.)

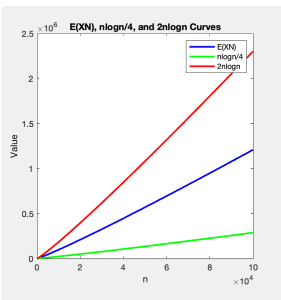


Figure 2: Comparison of $E(X^{(n)})$ versus nlogn bounds we got

Question 2

As F is invertible, then F must be strictly increasing function.

Part (a)

We need to prove that the random variable V which takes values $v_i, 1 \leq i \leq n$ follows the distribution F.
For a random variable U is drawn from Uniform(0, 1) where u_i are the drawn values, hence $v_i = F^{-1}(u_i)$
To prove that, it should be shown that $P(v_i \leq y) = F(y)$.
$v_i \leq y$ is equivalent to $F(v_i) \leq F(y)$ as F is strictly increasing function.

$$\begin{aligned} P(v_i \leq y) &= P(F(v_i) \leq F(y)) \\ &= P(u_i \leq F(y)) \\ &= F(y) \end{aligned}$$

The last equation is due to the uniformity of u_i.

$$\therefore P(v_i \leq y) = F(y)$$

Hence, proved.

Part (b)

Given,

$$F_e(x) = \frac{\sum_{i=1}^{n} \mathbf{1}(Y_i \leq x)}{n}$$

$$D = \max_x |F_e(x) - F(x)|$$

$$E = \max_{0 \leq y \leq 1} |\frac{\sum_{i=1}^{n} \mathbf{1}(U_i \leq y)}{n} - y|$$

$$D = \max_x |\frac{\sum_{i=1}^{n} \mathbf{1}(Y_i \leq x)}{n} - F(x)| = \max_{F(x)} |\frac{\sum_{i=1}^{n} \mathbf{1}(F(Y_i) \leq F(x))}{n} - F(x)|$$

Since F is strictly increasing.
Replacing $F(x)$ by y we can see that $D = E$ as $0 \leq F(x) \leq 1$ and $F(Y_i) = U_i$
Since $Y_i = F^{-1}(U_i)$ from part (a) because a distribution from F can be written in terms of a distribution from a **uniform** distribution.
Hence there distribution function will be same or $P(D \leq d) = P(E \leq d)$
Now $P(D \geq d) = 1 - P(D \leq d) = 1 - P(E \leq d) = P(E \geq d)$
Hence proved.

The interpretation of the above result can be as follows:
Any arbitrary empirical distribution will converge to the actual distribution to the same extent as the uniform distribution w.r.t. n, the number of samples.
Since this relationship is transitive in nature, we can extend this relation to the empirical distribution of any two random variables provided their distribution functions are continuous.
The most remarkable property of this result is that the random variable D is independent of the underlying function F, i.e. it is independent of the distribution of the original random variables.

Question 3

Part (a)

In this question, we are asked to perform maximum likelihood based plane fitting. The equation of the plane given is:

$$z = ax + by + c$$

We have access to accurate X and Y coordinates of some N points lying on the plane. We also have access to Z coordinates of these points, but those have been corrupted independently by noise from $\mathcal{N}(0, \sigma^2)$. Thus, we have:

$$z_i = ax_i + by_i + c + \epsilon_i$$

where ϵ_i is a random variable (noise) sampled from $\mathcal{N}(0, \sigma^2)$. i ranges from 1 to N Moreover, we also have ϵ_i and ϵ_j as independent $\forall i, j$.

The probability that the random variable from $\mathcal{N}(0, \sigma^2)$ takes value ϵ_i is:

$$f(\epsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\epsilon_i - \mu)}{2\sigma^2}}$$

In our case, we have: $\mu = 0$ and $\epsilon = z_i - (ax_i + by_i + c)$. Putting these into the expression we get:

$$f(\epsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(z_i - (ax_i + by_i + c))}{2\sigma^2}}$$

Since, ϵ_i and ϵ_j are independent, we have:

$$f(\epsilon_i, \epsilon_j) = f(\epsilon_i) \times f(\epsilon_j)$$

For all, ϵ_i, the log-likelihood function $\mathcal{L}$ will be:

$$\begin{aligned}
\mathcal{L} &= log(f(\epsilon_1, \epsilon_2, \ldots, \epsilon_N)) \\
&= log(f(\epsilon_1) f(\epsilon_2) \ldots f(\epsilon_N)) \\
&= log((\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(z_1 - (ax_1 + by_1 + c)^2)}{2\sigma^2}})(\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(z_2 - (ax_2 + by_2 + c))^2}{2\sigma^2}}) \ldots (\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(z_N - (ax_N + by_N + c))^2}{2\sigma^2}})) \\
&= log(\frac{1}{\sigma^N (2\pi)^{\frac{N}{2}}} e^{\frac{-\sum_{i=1}^{N}(z_i - (ax_i + by_i + c))^2}{2\sigma^2}})
\end{aligned}$$

$$\mathcal{L} = -N log(\sigma) - \frac{N}{2} log(2\pi) - \frac{\sum_{i=1}^{N}(z_i - (ax_i + by_i + c))^2}{2\sigma^2}$$

The above is the log-likelihood function $\mathcal{L}$ that needs to be maximized in order to determine a, b, c. Now, to determine the likelihood estimates of a, b, c, we will equate the derivatives of $\mathcal{L}$ with these variables to 0.

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial a} &= -[\frac{1}{2\sigma^2} \times 2(\sum_{i=1}^{N}(z_i - (ax_i + by_i + c)) \times -x_i)] \\
&= \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (ax_i + by_i + c)) \times x_i) \\
\frac{\partial \mathcal{L}}{\partial a} &= 0 \\
&\text{(For maximizing } a\text{)}
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (ax_i + by_i + c)) \times x_i) = 0 \\
\Rightarrow (\sum_{i=1}^{N}(z_i - (ax_i + by_i + c)) \times x_i) = 0
\end{aligned}
$$

Similarly for b:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b} &= -[\frac{1}{2\sigma^2} \times 2(\sum_{i=1}^{N}(z_i - (ax_i + by_i + c)) \times -y_i)] \\
&= \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (ax_i + by_i + c)) \times y_i) \\
\frac{\partial \mathcal{L}}{\partial b} &= 0 \\
&\text{(For maximizing } b\text{)}
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (ax_i + by_i + c)) \times y_i) = 0 \\
\Rightarrow (\sum_{i=1}^{N}(z_i - (ax_i + by_i + c)) \times y_i) = 0
\end{aligned}
$$

and for c:

$$\frac{\partial \mathcal{L}}{\partial c} = -[\frac{1}{2\sigma^2} \times 2(\sum_{i=1}^{N}(z_i - (ax_i + by_i + c)) \times -1)]$$
$$= \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (ax_i + by_i + c)) \times 1)$$
$$\frac{\partial \mathcal{L}}{\partial c} = 0$$
(For maximizing c)

$$\Rightarrow \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (ax_i + by_i + c))) = 0$$
$$\Rightarrow (\sum_{i=1}^{N}(z_i - (ax_i + by_i + c))) = 0$$

From the above partial-derivative expressions, we have got three equations:

$$a(\sum_{i=1}^{N} x_i^2) + b(\sum_{i=1}^{N} x_i y_i) + c(\sum_{i=1}^{N} x_i) = \sum_{i=1}^{N} z_i x_i \tag{3}$$
$$a(\sum_{i=1}^{N} x_i y_i) + b(\sum_{i=1}^{N} y_i^2) + c(\sum_{i=1}^{N} y_i) = \sum_{i=1}^{N} z_i y_i \tag{4}$$
$$a(\sum_{i=1}^{N} x_i) + b(\sum_{i=1}^{N} y_i) + c(N) = \sum_{i=1}^{N} z_i \tag{5}$$

Let us express these equations in terms of matrix and vector notation:

$$\begin{bmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} y_i^2 & \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} y_i & N \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} z_i x_i \\ \sum_{i=1}^{N} z_i y_i \\ \sum_{i=1}^{N} z_i \end{bmatrix} \tag{6}$$

The matrix equations involve the following coefficient matrix, they are:

$$\mathbf{P} = \begin{bmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} y_i^2 & \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} y_i & N \end{bmatrix} \tag{7}$$

The following variable vector:

$$\lambda = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \tag{8}$$

And the following vector on the right hand side of the equation:

$$\mathbf{q} = \begin{bmatrix} \sum_{i=1}^{N} z_i x_i \\ \sum_{i=1}^{N} z_i y_i \\ \sum_{i=1}^{N} z_i \end{bmatrix} \tag{9}$$

The overall equation to be solved is:

$$\mathbf{P}\lambda = \mathbf{q} \tag{10}$$

The solution to this equation provided $\mathbf{P}$ is an invertible matrix is :

$$\lambda = \mathbf{P^{-1}q} \tag{11}$$

This solution will provide the values of a, b, c, so as to maximize the log-likelihood function $\mathcal{L}$. Thus, we have used parameter estimation to achieve our goal.

Part (b)

The equation given is:

$$z = a_1x^2 + a_2y^2 + a_3xy + a_4x + a_5y + a_6$$

We have access to accurate X and Y coordinates of some N points lying on the plane. We also have access to Z coordinates of these points, but those have been corrupted independently by noise from $\mathcal{N}(0, \sigma^2)$. Thus, we have:

$$z = a_1x^2 + a_2y^2 + a_3xy + a_4x + a_5y + a_6 + \epsilon_i$$

where ϵ_i is a random variable (noise) sampled from $\mathcal{N}(0, \sigma^2)$. i ranges from 1 to N Moreover, we also have ϵ_i and ϵ_j as independent $\forall i, j$.

The probability that the random variable from $\mathcal{N}(0, \sigma^2)$ takes value ϵ_i is:

$$f(\epsilon_i) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(\epsilon_i-\mu)^2}{2\sigma^2}}$$

In our case, we have: $\mu = 0$ and $\epsilon = z_i - (a_1x^2 + a_2y^2 + a_3xy + a_4x + a_5y + a_6)$. Putting these into the expression we get:

$$f(\epsilon_i) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(z_i-(a_1x^2+a_2y^2+a_3xy+a_4x+a_5y+a_6))^2}{2\sigma^2}}$$

Since, ϵ_i and ϵ_j are independent, we have:

$$f(\epsilon_i, \epsilon_j) = f(\epsilon_i) \times f(\epsilon_j)$$

For all, ϵ_i, the log-likelihood function $\mathcal{L}$ will be:

$$\begin{aligned}
\mathcal{L} &= log(f(\epsilon_1, \epsilon_2, \dots, \epsilon_N)) \\
&= log(f(\epsilon_1)f(\epsilon_2)\dots f(\epsilon_N)) \\
&= log((\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(z_1-(a_1x_1^2+a_2y_1^2+a_3x_1y_1+a_4x_1+a_5y_1+a_6)^2)}{2\sigma^2}})(\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(z_2-(a_1x_2^2+a_2y_2^2+a_3x_2y_2+a_4x_2+a_5y_2+a_6))^2}{2\sigma^2}})\dots \\
&(\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(z_N-(a_1x_N^2+a_2y_N^2+a_3x_Ny_N+a_4x_N+a_5y_N+a_6))^2}{2\sigma^2}})) \\
&= log(\frac{1}{\sigma^N(2\pi)^{\frac{N}{2}}}e^{\frac{-\sum_{i=1}^N(z_i-(a_1x_i^2+a_2y_i^2+a_3x_iy_i+a_4x_i+a_5y_i+a_6))^2}{2\sigma^2}})
\end{aligned}$$

$$\mathcal{L} = -Nlog(\sigma) - \frac{N}{2}log(2\pi) - \frac{\sum_{i=1}^N(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6))^2}{2\sigma^2}$$

The above is the log-likelihood function $\mathcal{L}$ that needs to be maximized in order to determine $a_1, a_2, a_3, a_4, a_5, a_6$. Now, to determine the likelihood estimates of $a_1, a_2, a_3, a_4, a_5, a_6$, we will equate the derivatives of $\mathcal{L}$ with these variables to 0.

For a_1:

$$\frac{\partial \mathcal{L}}{\partial a_1} = -[\frac{1}{2\sigma^2} \times 2(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times -x_i^2)]$$
$$= \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times x_i^2)$$
$$\frac{\partial \mathcal{L}}{\partial a_1} = 0$$

(For maximizing a_1)

$$\Rightarrow \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times x_i^2) = 0$$
$$\Rightarrow (\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times x_i^2) = 0$$

For a_2:

$$\frac{\partial \mathcal{L}}{\partial a_2} = -[\frac{1}{2\sigma^2} \times 2(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times -y_i^2)]$$
$$= \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times y_i^2)$$
$$\frac{\partial \mathcal{L}}{\partial a_2} = 0$$

(For maximizing a_2)

$$\Rightarrow \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times y_i^2) = 0$$
$$\Rightarrow (\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times y_i^2) = 0$$

For a_3:

$$\frac{\partial \mathcal{L}}{\partial a_3} = -[\frac{1}{2\sigma^2} \times 2(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times -x_iy_i)]$$
$$= \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times x_iy_i)$$
$$\frac{\partial \mathcal{L}}{\partial a_3} = 0$$

(For maximizing a_3)

$$\Rightarrow \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times x_iy_i) = 0$$
$$\Rightarrow (\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times x_iy_i) = 0$$

For a_4:

$$\frac{\partial \mathcal{L}}{\partial a_4} = -[\frac{1}{2\sigma^2} \times 2(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times -x_i)]$$
$$= \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times x_i)$$
$$\frac{\partial \mathcal{L}}{\partial a_4} = 0$$

(For maximizing a_4)

$$\Rightarrow \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times x_i) = 0$$
$$\Rightarrow (\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times x_i) = 0$$

For a_5:

$$\frac{\partial \mathcal{L}}{\partial a_5} = -[\frac{1}{2\sigma^2} \times 2(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times -y_i)]$$

$$= \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times y_i)$$

$$\frac{\partial \mathcal{L}}{\partial a_5} = 0$$

(For maximizing a_5)

$$\Rightarrow \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times y_i) = 0$$

$$\Rightarrow (\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times y_i) = 0$$

For a_6:

$$\frac{\partial \mathcal{L}}{\partial a_6} = -[\frac{1}{2\sigma^2} \times 2(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times -1)]$$

$$= \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)) \times 1)$$

$$\frac{\partial \mathcal{L}}{\partial a_6} = 0$$

(For maximizing a_6)

$$\Rightarrow \frac{1}{\sigma^2}(\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6))) = 0$$

$$\Rightarrow (\sum_{i=1}^{N}(z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6))) = 0$$

From the above partial-derivative expressions, we have got six equations:

$$a_1(\sum_{i=1}^{N} x_i^4) + a_2(\sum_{i=1}^{N} x_i^2y_i^2) + a_3(\sum_{i=1}^{N} x_i^3y_i) + a_4(\sum_{i=1}^{N} x_i^3) + a_5(\sum_{i=1}^{N} x_i^2y_i) + a_6(\sum_{i=1}^{N} x_i^2) = \sum_{i=1}^{N} z_ix_i^2 \quad (12)$$

$$a_1(\sum_{i=1}^{N} x_i^2y_i^2) + a_2(\sum_{i=1}^{N} y_i^4) + a_3(\sum_{i=1}^{N} x_iy_i^3) + a_4(\sum_{i=1}^{N} x_iy_i^2) + a_5(\sum_{i=1}^{N} y_i^3) + a_6(\sum_{i=1}^{N} y_i^2) = \sum_{i=1}^{N} z_iy_i^2 \quad (13)$$

$$a_1(\sum_{i=1}^{N} x_i^3y_i) + a_2(\sum_{i=1}^{N} x_iy_i^3) + a_3(\sum_{i=1}^{N} x_i^2y_i^2) + a_4(\sum_{i=1}^{N} x_i^2y_i) + a_5(\sum_{i=1}^{N} x_iy_i^2) + a_6(\sum_{i=1}^{N} x_iy_i) = \sum_{i=1}^{N} z_ix_iy_i \quad (14)$$

$$a_1(\sum_{i=1}^{N} x_i^3) + a_2(\sum_{i=1}^{N} x_iy_i^2) + a_3(\sum_{i=1}^{N} x_i^2y_i) + a_4(\sum_{i=1}^{N} x_i^2) + a_5(\sum_{i=1}^{N} x_iy_i) + a_6(\sum_{i=1}^{N} x_i) = \sum_{i=1}^{N} z_ix_i \quad (15)$$

$$a_1(\sum_{i=1}^{N} x_i^2 y_i) + a_2(\sum_{i=1}^{N} y_i^3) + a_3(\sum_{i=1}^{N} x_i y_i^2) + a_4(\sum_{i=1}^{N} x_i y_i) + a_5(\sum_{i=1}^{N} y_i^2) + a_6(\sum_{i=1}^{N} y_i) = \sum_{i=1}^{N} z_i y_i \quad (16)$$

$$a_1(\sum_{i=1}^{N} x_i^2) + a_2(\sum_{i=1}^{N} y_i^2) + a_3(\sum_{i=1}^{N} x_i y_i) + a_4(\sum_{i=1}^{N} x_i) + a_5(\sum_{i=1}^{N} y_i) + a_6(N) = \sum_{i=1}^{N} z_i \quad (17)$$

Let us express these equations in terms of matrix and vector notation:

$$\begin{bmatrix} \sum_{i=1}^N x_i^4 & \sum_{i=1}^N x_i^2 y_i^2 & \sum_{i=1}^N x_i^3 y_i & \sum_{i=1}^N x_i^3 & \sum_{i=1}^N x_i^2 y_i & \sum_{i=1}^N x_i^2 \\ \sum_{i=1}^N x_i^2 y_i^2 & \sum_{i=1}^N y_i^4 & \sum_{i=1}^N x_i y_i^3 & \sum_{i=1}^N x_i y_i^2 & \sum_{i=1}^N y_i^3 & \sum_{i=1}^N y_i^2 \\ \sum_{i=1}^N x_i^3 y_i & \sum_{i=1}^N x_i y_i^3 & \sum_{i=1}^N x_i^2 y_i^2 & \sum_{i=1}^N x_i^2 y_i & \sum_{i=1}^N x_i y_i^2 & \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^N x_i^3 & \sum_{i=1}^N x_i y_i^2 & \sum_{i=1}^N x_i^2 y_i & \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i y_i & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i^2 y_i & \sum_{i=1}^N y_i^3 & \sum_{i=1}^N x_i y_i^2 & \sum_{i=1}^N x_i y_i & \sum_{i=1}^N y_i^2 & \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i^2 & \sum_{i=1}^N y_i^2 & \sum_{i=1}^N x_i y_i & \sum_{i=1}^N x_i & \sum_{i=1}^N y_i & N \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N z_i x_i^2 \\ \sum_{i=1}^N z_i y_i^2 \\ \sum_{i=1}^N z_i x_i y_i \\ \sum_{i=1}^N z_i x_i \\ \sum_{i=1}^N z_i y_i \\ \sum_{i=1}^N z_i \end{bmatrix} \quad (18)$$

The matrix equations involve the following coefficient matrix, they are:

$$\mathbf{P} = \begin{bmatrix} \sum_{i=1}^N x_i^4 & \sum_{i=1}^N x_i^2 y_i^2 & \sum_{i=1}^N x_i^3 y_i & \sum_{i=1}^N x_i^3 & \sum_{i=1}^N x_i^2 y_i & \sum_{i=1}^N x_i^2 \\ \sum_{i=1}^N x_i^2 y_i^2 & \sum_{i=1}^N y_i^4 & \sum_{i=1}^N x_i y_i^3 & \sum_{i=1}^N x_i y_i^2 & \sum_{i=1}^N y_i^3 & \sum_{i=1}^N y_i^2 \\ \sum_{i=1}^N x_i^3 y_i & \sum_{i=1}^N x_i y_i^3 & \sum_{i=1}^N x_i^2 y_i^2 & \sum_{i=1}^N x_i^2 y_i & \sum_{i=1}^N x_i y_i^2 & \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^N x_i^3 & \sum_{i=1}^N x_i y_i^2 & \sum_{i=1}^N x_i^2 y_i & \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i y_i & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i^2 y_i & \sum_{i=1}^N y_i^3 & \sum_{i=1}^N x_i y_i^2 & \sum_{i=1}^N x_i y_i & \sum_{i=1}^N y_i^2 & \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i^2 & \sum_{i=1}^N y_i^2 & \sum_{i=1}^N x_i y_i & \sum_{i=1}^N x_i & \sum_{i=1}^N y_i & N \end{bmatrix} \quad (19)$$

The following variable vector:

$$\lambda = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} \quad (20)$$

And the following vector on the right hand side of the equation:

$$\mathbf{q} = \begin{bmatrix} \sum_{i=1}^N z_i x_i^2 \\ \sum_{i=1}^N z_i y_i^2 \\ \sum_{i=1}^N z_i x_i y_i \\ \sum_{i=1}^N z_i x_i \\ \sum_{i=1}^N z_i y_i \\ \sum_{i=1}^N z_i \end{bmatrix} \quad (21)$$

The overall equation to be solved is:

$$\mathbf{P}\lambda = \mathbf{q} \quad (22)$$

The solution to this equation provided $\mathbf{P}$ is an invertible matrix is :

$$\lambda = \mathbf{P^{-1}q} \quad (23)$$

This solution will provide the values of $a_1, a_2, a_3, a_4, a_5, a_6$, so as to maximize the log-likelihood function $\mathcal{L}$. Thus, we have used parameter estimation to achieve our goal.

Part (c)

The code for this part is present in the file: a3_q3_c.m

The data for this part was present in the file 'XYZ.txt'. This data was read by the matlab code using 'dlmread' function. The code prints the predicted equation of the plane and also the noise variance. The code also plots the plane, but this part has been commented out.

The plane will be of the form:

$$z = ax + by + c$$

But, the values of z have been corrupted by noise ϵ_i.

$$z_i = ax_i + by_i + c + \epsilon_i$$

Firstly, we assume that the noise ϵ_i is from a normal distribution with mean 0. So, under this assumption, we can use the equation which we derived in part(a) 3.1. The equation is:

$$\begin{bmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} y_i^2 & \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} y_i & N \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} z_i x_i \\ \sum_{i=1}^{N} z_i y_i \\ \sum_{i=1}^{N} z_i \end{bmatrix} \tag{24}$$

The matrix equations involve the following coefficient matrix, they are:

$$\mathbf{P} = \begin{bmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} y_i^2 & \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} y_i & N \end{bmatrix} \tag{25}$$

The following variable vector:

$$\lambda = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \tag{26}$$

And the following vector on the right hand side of the equation:

$$\mathbf{q} = \begin{bmatrix} \sum_{i=1}^{N} z_i x_i \\ \sum_{i=1}^{N} z_i y_i \\ \sum_{i=1}^{N} z_i \end{bmatrix} \tag{27}$$

The overall equation to be solved is:

$$\mathbf{P}\lambda = \mathbf{q} \tag{28}$$

The solution to this equation provided $\mathbf{P}$ is an invertible matrix is :

$$\lambda = \mathbf{P^{-1}q} \tag{29}$$

The calculations of the matrix P, vector Q have been done in the matlab code. Then the resulting values are stored in an array 'result'. The equation of the plane comes out to be:

$$z = 10.0022x + 19.998y + 29.9516$$

Then we proceed towards the calculation of the noise. The noise is the difference between the actual values of z and those we get from the plane. Thus noise is:

$$\epsilon_i = z_i - (ax_i + by_i + cz_i)$$

These values are stored in another array. The code also calculates the mean of the noise values. The mean value comes out to be $\mathbf{3.4147 \times 10^{-13}}$. Then, the code computes the variance of the noise values based on the mean calculated. The variance comes out to be **23.0685**.

The mean and the variance were calculated as:

$$\mu = \sum_{i=1}^{2000} (\epsilon_i)$$

$$\sigma^2 = \frac{\sum_{i=1}^{2000} (\epsilon_i - \mu)^2}{2000 - 1}$$

The following plot is obtained for the data given:

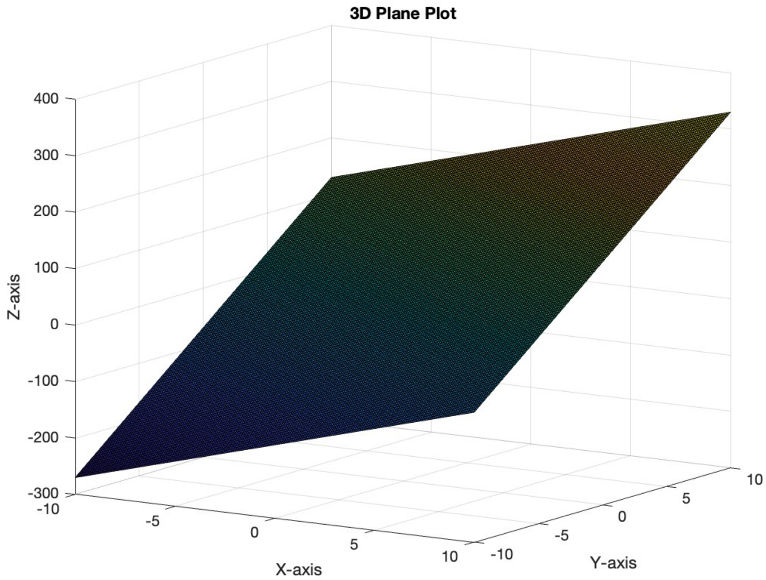


Figure 3: Predicted 2-D plane based on the data

Question 4

The code for this question is present in the file Q4_Final.m

The file will produce correct results when it is run on the online MATLAB website.

Part (b):

We estimate the Probability Distribution Function(PDF) of the distribution from the sample values stored in T using the formula of Kernel Density Estimation(KDE) given to us:

$$\hat{p}_n(x;\sigma) = \frac{\sum_{i=1}^{n} \exp\left(\frac{-(x-x_i)^2}{2\sigma^2}\right)}{n\sigma\sqrt{2\pi}}$$

Here x_i's are values obtained from the set T.

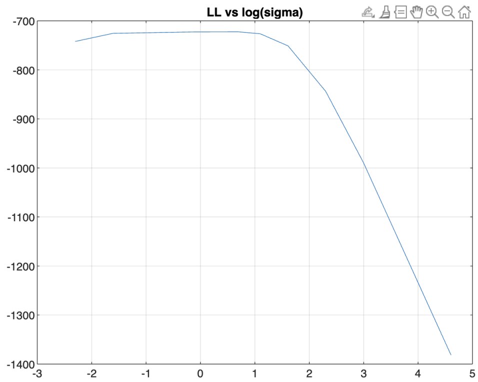
To find the joint likelihood of the samples in V, based on the estimate of the above PDF built from T with bandwidth parameter σ, we assume the individual samples to be independent of each other. Thus, the expression for the joint likelihood is,

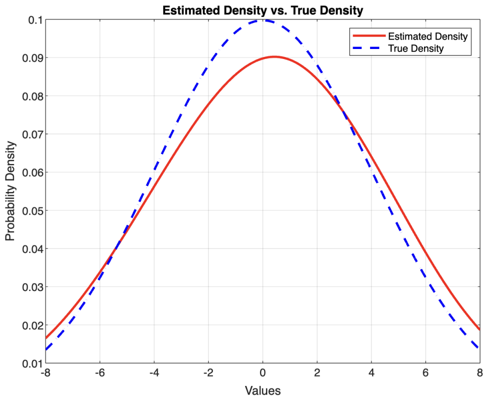
$$\hat{p}_n(x_1,\ldots,x_n;\sigma) = \prod_{j=1}^{n} \hat{p}_n(x_j;\sigma) = \prod_{j=1}^{n} \frac{\sum_{i=1}^{n} \exp\left(\frac{-(x_j-x_i)^2}{2\sigma^2}\right)}{n\sigma\sqrt{2\pi}}$$

All the relevant calculations of the log of LL etc. have been implemented in the MATLAB code submitted.

Part (c):

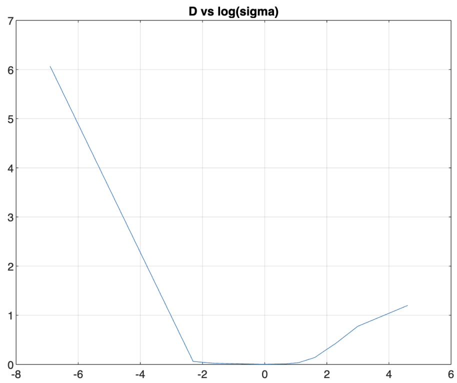
The plot of LL vs log σ has been attached below. The value of σ which gives the best LL value is 1. The MATLAB code prints this value. For this σ we draw the graph of $\hat{p}_n(x;\sigma)$ and over it we draw the graph of the true density which is N(0, 16).

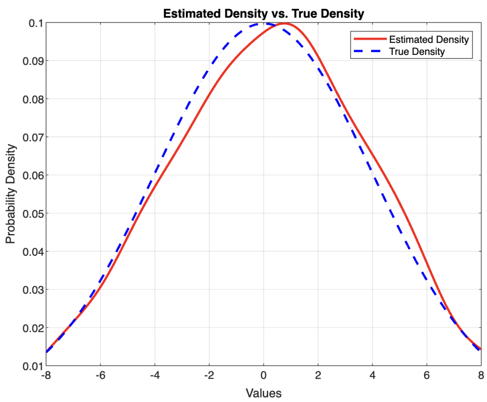




Part (d):

The graph of D vs logσ is attached below: The best D value was given when $\sigma = .9$, and the D value corresponding to this σ is 0.00357. For this σ the graph of $\hat{p}_n(x;\sigma)$ for the estimated and true density has been attached below.





Part (e):

In this question we divide our sample data into two disjoint parts T and V and then use T to create the PDF, i.e. to build our model, and V to test the performance of the model on unseen data. This is called cross-validation. The purpose of cross-validation is to assess how well a model or parameter generalizes to new, unseen data. When T and V are equal, we are not evaluating the

model's ability to generalize because it's already seen the validation data during training. So if T and V are equal then the whole process is less meaningful and doesn't provide useful information. In our case, we wanted to find the best value of σ, but if T and V are equal then we may find a value of σ that fits the data extremely well, but may not work for a new, unseen data set.
We may also have the following argument. If T = V, then we are find LL of T wrt. to T which doesn't have a maximum and approaches ∞ as $\sigma \to 0$. This means that $\sigma = 0$ is the parameter at which LL is maximised but we σ can't be 0. This is due to the fact that $\forall x \exists k < 750$, the value of $x - x_k = 0$ since V = T, which means for i = k, $\exp\left(\frac{-(x_j-x_i)^2}{2\sigma^2}\right) = 1 \forall \sigma$ and for i != k, $\exp\left(\frac{-(x_j-x_i)^2}{2\sigma^2}\right) = 0$ as $\sigma \to 0$.
Hence as $\sigma \to 0$, $\hat{p}_n(x;\sigma) = \lambda(\frac{1}{\sigma}) \to \infty \Rightarrow$ LL $\to \infty$. Hence no maximum likelihood for this LL.
Hence cross-validation procedure fails and gives wrong values for σ as the maximum value of LL occurs at a place where σ is not defined.

Question 5

Objective: Put an (tight enough) upper bound on $P(S_n - E[S_n] > t)$.
We have,

$$S_n = \sum_{i=1}^{n} X_i$$

Since $a_i \leq X_i \leq b_i$ this implies:

$$\sum_{i=1}^{n} a_i \leq S_n \leq \sum_{i=1}^{n} b_i$$

For some $s > 0$, we can say:

$$P(S_n - E[S_n] > t) = P(s(S_n - E[S_n]) > st)$$

The above claim is true because the LHS and RHS scales up linearly, hence the sets in the respective probabilities have the same cardinality. Furthermore, if we raise both sides of the inequality to the power of e we get:

$$P(s(S_n - E[S_n]) > st) = P(e^{s(S_n - E[S_n])} > e^{st})$$

This is true because e^x is a monotonically increasing function.
Markov's inequality states that for any non-negative random variable and $a > 0$,

$$P(X > a) \leq \frac{E[X]}{a}$$
$$P(e^{s(S_n - E[S_n])} > e^{st}) \leq \frac{E[e^{s(S_n - E[S_n])}]}{e^{st}}$$

The given intermediate result (IR) is:

$$E[e^{s(X - E[X])}] \leq e^{s^2(b-a)^2/8}$$

Using the above inequality, we can write:

$$P(e^{s(S_n - E[S_n])} > e^{st}) \leq \frac{e^{s^2(\tilde{b}-\tilde{a})^2/8}}{e^{st}} = \exp\left(\frac{s^2(\tilde{b}-\tilde{a})^2}{8} - st\right)$$

Where $\tilde{a} = \sum_{i=1}^{n} a_i$ and $\tilde{b} = \sum_{i=1}^{n} b_i$ since these are the minimum and maximum values of S_n.
Define:

$$g(s) = \frac{s^2(\tilde{b}-\tilde{a})^2}{8} - st$$

To tighten the bound, we must find the minimum value of $g(s)$. This can be done by setting it's derivative equal to zero and substituting the optimal s to get the minimum value of g

$$g'(s) = \frac{s(\tilde{b}-\tilde{a})^2}{4} - t = 0$$
$$s = \frac{4t}{(\tilde{b}-\tilde{a})^2}$$
$$g_{min} = \frac{16t^2}{8(\tilde{b}-\tilde{a})^2} - \frac{4t^2}{(\tilde{b}-\tilde{a})^2}$$
$$= \frac{-2t^2}{(\tilde{b}-\tilde{a})^2}$$

Hence, an upper bound on the given probability is given by:

$$P(S_n - E[S_n] > t) = P(e^{s(S_n - E[S_n])} > e^{st}) \leq \exp\left(\frac{-2t^2}{(\tilde{b} - \tilde{a})^2}\right)$$

$$\therefore P(S_n - E[S_n] > t) \leq \exp\left(\frac{-2t^2}{(\sum_{i=1}^{n}(b_i - a_i))^2}\right)$$

Proving the intermediate result (IR)

Part (a)

The second derivative of e^x is e^x which is greater than 0 for all x, hence e^x is a convex function. By definition of convex function, every point on the line joining two points on the curve are greater than the value of the curve. Lets fix $(x_0, y_0) = (a, e^{sa})$ and $(x_1, y_1) = (b, e^{sb})$. (x, y) be any point on the line:

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0}$$

$$\frac{y - e^{sa}}{x - a} = \frac{e^{sb} - e^{sa}}{b - a}$$

$$y = e^{sa} + \frac{x - a}{b - a}(e^{sb} - e^{sa})$$

$$= \left(1 - \frac{x - a}{b - a}\right) e^{sa} + \frac{x - a}{b - a} e^{sb}$$

$$= \frac{b - x}{b - a} e^{sa} + \frac{x - a}{b - a} e^{sb}$$

And due to convexity:

$$e^{sx} \leq \frac{b - x}{b - a} e^{sa} + \frac{x - a}{b - a} e^{sb}, a \leq x \leq b$$

Part (b)

Taking $E[X] = 0$ we can take expectation on both sides of the inequality derived above:

$$E[e^{sx}] \leq \frac{E[b] - E[x]}{b - a} e^{sa} + \frac{E[x] - E[a]}{b - a} e^{sb}$$

$$\leq \frac{b \cdot e^{sa} - a \cdot e^{sb}}{b - a}$$

$$\frac{b \cdot e^{sa} - a \cdot e^{sb}}{b - a} = e^{sa} + \frac{a(e^{sa} - e^{sb})}{b - a}$$

$$= e^{sa}\left(1 + \frac{a(1 - e^{s(b-a)})}{b - a}\right)$$

$$= \exp\left(sa + \log\left(1 + \frac{a(1 - e^{s(b-a)})}{b - a}\right)\right)$$

Defining:

$$L(h) = \frac{ha}{b - a} + \log\left(1 + \frac{a(1 - e^{h})}{b - a}\right)$$

Which equates to the above equation for $h = s(b - a)$. Hence,

$$E[e^{sx}] \leq \exp[L(s(b - a))]$$

Part (c)

$$
\begin{aligned}
L(h) &= \frac{ha}{b-a} + \log\left(1 + \frac{a(1-e^h)}{b-a}\right) \\
L'(h) &= \frac{a}{b-a} + \frac{1}{1+\frac{a(1-e^h)}{b-a}} \frac{-a \cdot e^h}{b-a} \\
&= \frac{a}{b-a} + \frac{-a \cdot e^h}{b - a \cdot e^h} \\
L''(h) &= \frac{(-a \cdot e^h)(b - a \cdot e^h) - (-a \cdot e^h)(-a \cdot e^h)}{(b - a \cdot e^h)^2} \\
&= \frac{-ab \cdot e^h}{(b - a \cdot e^h)^2}
\end{aligned}
$$

We can apply the AM-GM inequality here as a is negative hence both the terms below are positives

$$
\begin{aligned}
\sqrt{(b)(-a \cdot e^h)} &\leq \frac{b - a \cdot e^h}{2} \\
-ab \cdot e^h &\leq \frac{(b - a \cdot e^h)^2}{4} \\
\frac{-ab \cdot e^h}{(b - a \cdot e^h)^2} &\leq \frac{1}{4} \\
\therefore L''(h) &\leq \frac{1}{4}
\end{aligned}
$$

Part (d)

Integrating both sides, we get:

$$\mathrm{E}[\mathrm{e}^{sx}] \leq \exp[L(s(b-a))]$$

Replacing x by $X - E[X]$ (considering the assumptions taken earlier), we get:

$$
\begin{aligned}
E[e^{s(X-E[X])}] &\leq \exp[L(s(b-a))], \\
\therefore E[e^{s(X-E[X])}] &\leq \exp\left(\frac{s^2(b-a)^2}{8}\right), \text{from previous results}
\end{aligned}
$$