INDIAN INSTITUTE OF TECHNOLOGY MADRAS

Department of Chemical Engineering

CH5115 : Parameter and State Estimation (Jul-Nov 2020)
Solutions to Assignment 3

Marks Distribution

	Question 1	Question 2	Question 3	Question 4	Question 5
(a)	5	7	10	10	8
(b)	5	5	10	10	10
(c)	5	5	–	–	–
(d)	–	10	–	–	–

Question 1

(a)

Calculating CDF of Y:

$$\begin{aligned} F_Y(y) &= \int_\theta^y f(y)dy = \int_\theta^y e^{-(y-\theta)}dy \\ &= 1 - e^{-(y-\theta)} \end{aligned}$$

$$F_Y(y) = \begin{cases} 1 - e^{-(y-\theta)}, & \text{where, y } > \theta \\ 0 & \text{otherwise} \end{cases}$$

Let the N observations be y_1, y_2, y_3, ... y_N such that $y_1 \leq y_2 \leq y_3 ..... \leq y_N$

We get:

$$T_N = 2\min(\boldsymbol{y_N}) = 2y_1$$

$$\begin{aligned}
F_{T_N} &= P(2y_1 \le y) \\
&= P(y_1 \le \frac{y}{2}) \\
&= 1 - P(y_1 > \frac{y}{2}) \\
&= 1 - P(y_1 > \frac{y}{2}, y_2 > \frac{y}{2}, ...y_N > \frac{y}{2}) \\
&= 1 - P(y_1 > \frac{y}{2})P(y_2 > \frac{y}{2})....P(y_N > \frac{y}{2}) \\
&= 1 - \left[P(y_1 > \frac{y}{2})\right]^N & = 1 - e^{-N(\frac{y}{2}-\theta)} \text{for y} > 2\theta
\end{aligned}$$

Therefore,

$$F_{T_N}(t) = \begin{cases} 1 - e^{-N(\frac{t}{2}-\theta)} & \text{for t} > 2\theta \\ 0 & \text{otherwise} \end{cases}$$

Upon differentiating $F_{T_N}(t)$, we obtain the PDF:

$$f_{T_N}(t) = \begin{cases} \frac{N}{2}e^{-N(\frac{y}{2}-\theta)} & \text{for t} > 2\theta \\ 0 & \text{otherwise} \end{cases}$$

(b)

$$\begin{aligned}
E[T_N] &= \int_{2\theta}^{\infty} \frac{N}{2}e^{-N(\frac{y}{2}-\theta)}dt \\
&= 2\left(\theta + \frac{1}{N}\right) \\
&\neq \theta
\end{aligned}$$

Therefore, T_N is a biased estimator of θ. Applying a correction factor to get an unbiased estimator T_N', we get:

$$\begin{aligned}
T_N' &= \frac{T_N}{2} - \frac{1}{N} \\
&= \min(\boldsymbol{y_N}) - \frac{1}{N}
\end{aligned}$$

(c)

To prove that T'_N converges to θ in probability, we need to prove that:

$$\lim_{N\to\infty} P\left(|T'_N - \theta| \geq \epsilon\right) = 0, \quad \text{for all } \epsilon > 0$$

$$\begin{aligned}
\lim_{N\to\infty} P\left(|T'_N - \theta| \geq \epsilon\right) &= \lim_{N\to\infty} \left[P\left(T'_N \leq \theta - \epsilon\right) + P\left(T'_N \geq \theta + \epsilon\right)\right] \\
&= \lim_{N\to\infty} P\left(T'_N \leq \theta - \epsilon\right) + \lim_{N\to\infty} P\left(T'_N \geq \theta + \epsilon\right) \\
&= \lim_{N\to\infty} F_{T'_N}\left(\theta - \epsilon\right) + \lim_{N\to\infty} P\left(T'_N \geq \theta + \epsilon\right) \\
&= 0 + \lim_{N\to\infty} P\left(T'_N \geq \theta + \epsilon\right) \quad (\text{since } F_{T'_N}(t) = 0 \text{ for t } \leq \theta - \frac{1}{N}) \\
&\cong 1 - \lim_{N\to\infty} P\left(T'_N \leq \theta + \epsilon\right) \\
&= 1 - \lim_{N\to\infty} F_{T'_N}\left(\theta + \epsilon\right) \\
&= 1 - 1 \\
&= 0
\end{aligned}$$

Hence, proved.

Question 2

Given that

$$V[n] = \sum_{k=0}^{N-1} v[k]e^{-2\pi j f_n k}, \quad \text{where, } f_n = \frac{n}{N}, n = 0, 1, \ldots, N-1$$

$$\mathbb{P}n = \frac{|V[n]|^2}{N} = \frac{a_n^2 + b_n^2}{N}, \quad a_n = \Re(V[n])\text{and}b_n = \Im(V[n])$$

(a)

$$a_n = \sum_{k=0}^{N-1} v[k]\cos(2\pi f_n k)$$

$$b_n = -\sum_{k=0}^{N-1} v[k]\sin(2\pi f_n k)$$

both a_n and b_n are linear combination of GWN $v[k]$, therefore, DFT coefficients a_n and b_n are also Gaussian distributed.

$$E(a_n) = \mu_{a_n} = E(v[k]\cos(2\pi f_n k))$$
$$E(b_n) = \mu_{b_n} = E(v[k]\sin(2\pi f_n k))$$

Since, $\mu_v = 0, \implies, \quad \mu_{a_n} = \mu_{a_n} = 0$

Calculation of Variance

$$\begin{aligned}
E(a_n^2) &= E\bigg(\sum_{k=0}^{N-1} v[k]\cos(2\pi f_n k)\bigg)^2 + (\mu_{a_n})^2 \\
\sigma_{a_n}^2 &= E\bigg(\sum_{k=0}^{N-1} v[k]\cos(2\pi f_n k)\bigg)^2 \\
&= \sum_{k=0}^{N-1} \sigma_v^2 \cos^2(2\pi f_n k) \\
&= \frac{1}{2}\sum_{k=0}^{N-1}(1+\cos(4\pi f_n k)) \\
&= \frac{N}{2} + \frac{1}{2}\sum_{k=0}^{N-1}\cos(4\pi f_n k)
\end{aligned}$$

We know that $\sum_{k=0}^{N-1}\cos(\omega k)$ is:

$$\sum_{k=0}^{N-1}\cos(\omega k) = \begin{cases} N, & \sin(\omega/2) = 0 \\ \dfrac{\sin(N\omega/2)}{\sin(\omega/2)}\cos((N-1)\omega/2), & \text{otherwise} \end{cases}$$
$$\sum_{k=0}^{N-1}\cos(4\pi f_n k) = \begin{cases} N, & \sin(4\pi f_n/2) = 0 \\ \dfrac{\sin(2\pi f_n N)}{\sin(2\pi f_n)}\cos(2\pi(N-1)f_n), & \text{otherwise} \end{cases}$$

Given that $f_n = \frac{RN+1}{N}$, where RN is the last two digits of roll no, above equation can be written as:

$$\begin{aligned}
\sum_{k=0}^{N-1} \cos(4\pi f_n k) &= \begin{cases} N, & \sin(2\pi f_n) = 0 \\ \dfrac{\sin(2\pi RN)}{\sin\left(\frac{2\pi RN}{N}\right)} \cos(\frac{2\pi RN(N-1)}{N}), & \text{otherwise} \end{cases} \\
&= \begin{cases} N, & \sin(2\pi f_n) = 0 \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

because, $\sin(2\pi RN)$ will always be 0.
Therefore, $\sigma_{a_n}^2$ is

$$\begin{aligned}
\sigma_{a_n}^2 &= \begin{cases} \frac{N}{2} + \frac{1}{2}N, & \text{if} \sin(2\pi f_n) = 0 \\ \frac{N}{2} + 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} N, & \text{if} \sin(2\pi f_n) = 0 \\ \dfrac{N}{2}, & \text{otherwise} \end{cases}
\end{aligned}$$

Similarly, variance of b_n is given by

$$\begin{aligned}
\sigma_{b_n}^2 &= E\left(\sum_{k=0}^{N-1} -v[k] \sin(2\pi f_n k)^2\right) \\
&= \sum_{k=0}^{N-1} \sigma_v^2 \sin^2(2\pi f_n k)
\end{aligned}$$

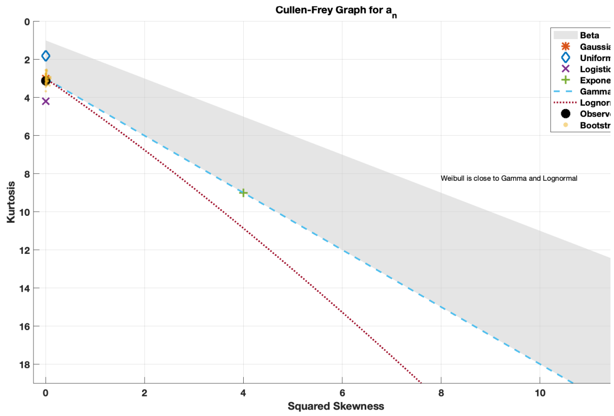
Using, $\cos(2\theta) = 1 - 2\sin^2\theta$

$$\begin{aligned}
\sigma_{b_n}^2 &= \frac{N}{2} - \frac{1}{2} \sum_{k=0}^{N-1} \cos(4\pi f_n k) \\
&= \begin{cases} 0, & \text{if} \sin(2\pi f_n) = 0 \\ \dfrac{N}{2}, & \text{otherwise} \end{cases}
\end{aligned}$$

Therefore, $a_n \sim \mathcal{N}(0, \sigma_{a_n}^2)$ and $b_n \sim \mathcal{N}(0, \sigma_{b_n}^2)$.

```
%% Assignment 3
% December 26, 2020
% Question 2
% part (a)

N = 1000; R = 300;
RN = 24+1;
k = (0:1:N-1)';
fn = RN/N;
for i = 1:R
    vk(:,i) = randn(N,1);
    an(:,i) = sum(vk(:,i).*cos(2*pi*fn*k));
    bn(:,i) = sum(-vk(:,i).*sin(2*pi*fn*k));
end
stats_an = cfgraph(an);
title('Cullen-Frey Graph for a_n')
stats_bn = cfgraph(bn);
title('Cullen-Frey Graph for b_n')
figure;histogram(an);
figure;histogram(bn);

% Estimated variance
est_var_an = var(an);
est_var_bn = var(bn);

% Theoretical variance
var_an = [N N/2];
var_bn = [0 N/2];

est_var_an =

  501.9341

est_var_an =

  533.8614
```

C-F graphs for a_n and b_n shows that the DFT coefficients follows Gaussian distribution. It can also be observed that the estimated variance $\sigma^2_{a_n}$ and $\sigma^2_{b_n}$ are close to $N/2$.



0n(a)

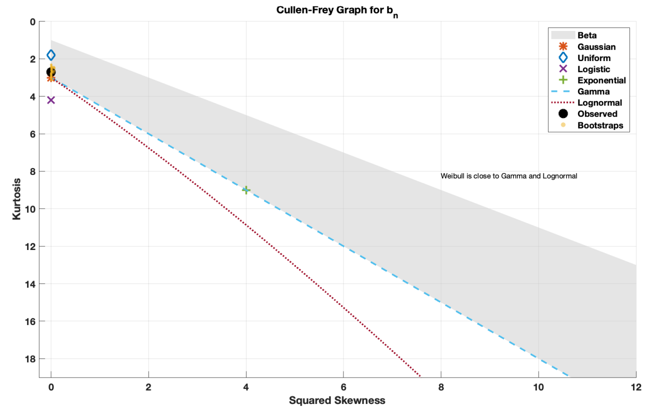


Figure 1: (b)

Figure 2: C-F graphs for estimated a_n and b_n

(b)

Since, DFT coefficients a_n and b_n are Gaussian distributed and are orthogonal to each other, a_n and b_n are uncorrelated.
Correlation at lags l

$$\begin{aligned}
\rho_{a_n b_n}[l] &= \frac{\sigma_{a_n b_n}[l]}{\sigma_{a_n}\sigma_{b_n}} \\
\sigma_{a_n b_n}[l] &= E\left(\sum_{k=0}^{N-1} v[k]\cos(2\pi f_n k)\sum_{k=0}^{N-1} -v[k-l]\sin(2\pi f_n(k-l))\right) \\
&= E\left(-\sum_{k=0}^{N-1} v[k]v[k-l]\cos(2\pi f_n k)\sin(2\pi f_n(k-l))\right)
\end{aligned}$$

Since, $v[k]$ ~GWN, $\implies E(v[k]v[k-l]) = 0$ at lags $l \neq 0$.
Correlation at lag $= 0$,

$$\begin{aligned}
\rho_{a_n b_n}[0] &= \frac{\sigma_{a_n b_n}[0]}{\sigma_{a_n}\sigma_{b_n}} \\
\sigma_{a_n b_n}[0] &= E\left(\sum_{k=0}^{N-1} v[k]\cos(2\pi f_n k)\sum_{k=0}^{N-1} -v[k]\sin(2\pi f_n k)\right)
\end{aligned}$$

$$
\begin{aligned}
&= E\Bigg(- \sum_{k=0}^{N-1} v^2[k] \cos(2\pi f_n k) \sin(2\pi f_n k) \Bigg) \\
&= -\frac{1}{2} \sum_{k=0}^{N-1} \sin(4\pi f_n k) \\
\implies \rho_{a_n b_n}[0] &= -\frac{1}{N} \sum_{k=0}^{N-1} \sin(4\pi f_n k)
\end{aligned}
$$

Therefore,

$$
\rho_{a_n b_n}[l] = \begin{cases} -\frac{1}{N} \sum_{k=0}^{N-1} \sin(4\pi f_n k), & l = 0 \\ 0, & \text{otherwise} \end{cases}
$$

(c)

Given that,

$$
\zeta_n = \frac{2\mathbb{P}(f_n)}{\gamma_{vv}(f_n)} = \frac{2(a_n^2 + b_n^2)}{N\gamma_{vv}(f_n)}
$$

$$
\gamma_{vv}(f_n) = \sum_{l=-\infty}^{\infty} \sigma_{vv}[l] e^{-j2\pi f l}
$$

It is given that $v[k]$ is a Gaussian white noise,

$$
\implies \sigma_{vv}[l] = \begin{cases} 1, & l = 0 \\ 0, & l \neq 0 \end{cases}
$$

$$
\zeta_n = \frac{2(a_n^2 + b_n^2)}{N}
$$

It is proved in the above segment of question that a_n and b_n are Gaussian distributed and uncorrelated to each other, it implies that a_n and b_n are *independent* variables. Also, ζ_n is a linear combination of two Gaussian distributed and independent variables, therefore, ζ_n follows χ^2 distribution with degree of freedom, $\nu = 2$. We know that mean and variance of χ^2 distributed variable is ν and 2ν respectively. Thus, ζ_n follows χ^2 distribution with mean 2 and variance 4.

(d)

Consistency in terms of mean square sense

$$
\begin{aligned}
\lim_{N\to\infty} E\Big(\mathbb{P}(f_n) - \gamma_{vv}(f_n)\Big)^2 &= \lim_{N\to\infty} E\Big(\frac{2(a_n^2+b_n^2)}{N} - \gamma_{vv}(f_n)\Big)^2 \\
&= \lim_{N\to\infty} E\Big(\frac{2(a_n^2+b_n^2)}{N} - 1\Big)^2 \\
&= \lim_{N\to\infty} E\Big(\frac{a^4+b^4+2a^2b^2}{N^2} - 2\frac{a^2+b^2}{N} + 1\Big) \\
&= \lim_{N\to\infty} E\Big(\frac{a^2+b^2}{N}\Big)^2 - E\Big(\frac{2(a^2+b^2)}{N}\Big) + 1 \\
&= \lim_{N\to\infty} E\Big(\frac{\zeta_n}{2}\Big)^2 - E(\zeta_n) + 1 \\
&= \lim_{N\to\infty} (2 - 2 + 1) = 1
\end{aligned}
$$

As $N \to \infty$, $\lim_{N\to\infty} E(\mathbb{P}(f_n) - \gamma_{vv}(f_n)) = 1$. Therefore, $\mathcal{P}(f_n)$ is **not a consistent** estimate of $\gamma_{vv}(f_n)$ in mean-square sense.

MATLAB code for checking consistency

```matlab
N = 1000;
RN = 24+1;
fn = RN/N;
vk = randn(N,1);
for i = 25:N
    an(i) = sum(vk(1:i).*(cos(2*pi*fn*i))');
    bn(i) = sum(-vk(1:i).*(sin(2*pi*fn*i))');
    P(i) = (an(i)^2+ bn(i)^2)/i;
    acf_vk(:,i) = autocorr(vk(1:i));
    fft_acf(:,i) = fft(acf_vk(:,i));
    gamma_vk(:,i) = fft_acf(:,i).*conj(fft_acf(:,i));
    mse_P(i-24) = var(gamma_vk(:,i)) + (P(i) - mean(gamma_vk(:,i)))^2;
end
figure; plot(25:N,mse_P,'LineWidth',2);
xlabel('Sample size (N)')
ylabel('MSE'); title('Consistency in mean-square sense');
```

As observed from Fig 3, with increasing sample size, mean-square error tends to 1 instead of 0. Therefore, $\mathbb{P}(f_n)$ is not a consistent estimate of $\gamma_{vv}(f_n)$

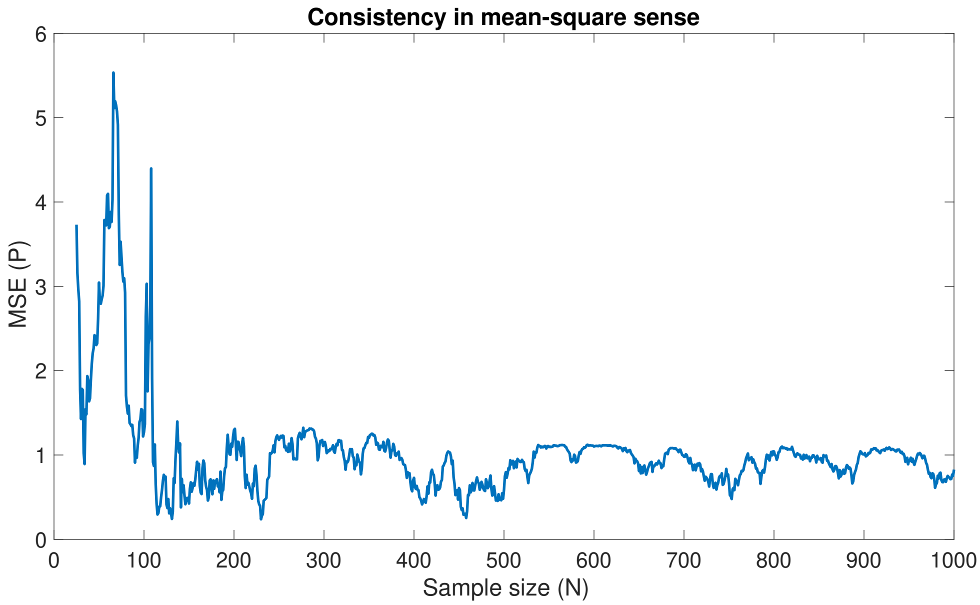


Figure 3: Consistency of $\mathbb{P}$ in mean-square sense

Question 3

(a)

Given that:

$$y[k] = A\sin\left(2\pi f_0 k\right) + e[k] \qquad e[k] \sim \mathcal{N}\left(0, \sigma_e^2\right)$$

The joint PDF for N observations is given by:

$$f(\mathbf{y};\theta) = \frac{1}{\left(2\pi\sigma_e^2\right)^{N/2}} \exp\left(-\frac{\sum_{k=0}^{N-1}\left(y[k] - A\sin\left(2\pi f_0 k\right)\right)^2}{2\sigma_e^2}\right)$$

Case 1 - if A and f_0 are unknown:

$$L(\theta;\mathbf{y}) = -\frac{N}{2}\ln\left(2\pi\sigma_e^2\right) - \frac{\sum_{k=0}^{N-1}\left(y[k]-A\sin\left(2\pi f_0 k\right)\right)^2}{2\sigma_e^2} \quad \text{where } \theta=[A,f_0]$$

$$\frac{\partial L}{\partial f_0} = \frac{2\pi A}{\sigma_e^2}\sum_{k=0}^{N-1}\left(y[k]-A\sin\left(2\pi f_0 k\right)\right)\left(k\cos\left(2\pi f_0 k\right)\right)$$

$$\frac{\partial L}{\partial A} = \frac{1}{\sigma_e^2}\sum_{k=0}^{N-1}\left(y[k]-A\sin\left(2\pi f_0 k\right)\right)\left(\sin\left(2\pi f_0 k\right)\right)$$

$$\frac{\partial^2 L}{\partial f_0^2} = \frac{4\pi^2 A}{\sigma_e^2}\sum_{k=0}^{N-1}k^2[-y[k]\sin\left(2\pi f_0 k\right)+A\sin^2\left(2\pi f_0 k\right)-A\cos^2\left(2\pi f_0 k\right)]$$

$$\frac{\partial^2 L}{\partial A^2} = \frac{-1}{\sigma_e^2}\sum_{k=0}^{N-1}\sin^2\left(2\pi f_0 k\right)$$

$$\frac{\partial^2 L}{\partial f_0 A} = \frac{2\pi}{\sigma_e^2}\sum_{k=0}^{N-1}k\left[y[k]\cos\left(2\pi f_0 k\right)-A\sin\left(4\pi f_0 k\right)\right]$$

$$I(\theta) = \begin{bmatrix} -E[\frac{\partial^2 L}{\partial f_0^2}] & -E[\frac{\partial^2 L}{\partial f_0 A}] \\ -E[\frac{\partial^2 L}{\partial f_0 A}] & -E[\frac{\partial^2 L}{\partial A^2}] \end{bmatrix}$$

$$I(\theta) = \begin{bmatrix} \frac{4\pi^2 A^2}{\sigma_e^2}\sum_{k=0}^{N-1}k^2\cos^2\left(2\pi f_0 k\right) & \frac{2\pi A}{\sigma_e^2}\sum_{k=0}^{N-1}k\sin\left(2\pi f_0 k\right)\cos\left(2\pi f_0 k\right) \\ \frac{2\pi A}{\sigma_e^2}\sum_{k=0}^{N-1}k\sin\left(2\pi f_0 k\right)\cos\left(2\pi f_0 k\right) & \frac{1}{\sigma_e^2}\sum_{k=0}^{N-1}\sin^2\left(2\pi f_0 k\right) \end{bmatrix}$$

$$|I(\theta)| = \frac{4\pi^2 A^2}{\sigma_e^4}\sum_{k=0}^{N-1}k^2\cos^2\left(2\pi f_0 k\right)\sum_{k=0}^{N-1}\sin^2\left(2\pi f_0 k\right) - (\sum_{k=0}^{N-1}k\sin\left(2\pi f_0 k\right)\cos\left(2\pi f_0 k\right))^2$$

$$I(\theta)^{-1} = \frac{1}{|I(\theta)|}\begin{bmatrix} \frac{1}{\sigma_e^2}\sum_{k=0}^{N-1}\sin^2\left(2\pi f_0 k\right) & \frac{-2\pi A}{\sigma_e^2}\sum_{k=0}^{N-1}k\sin\left(2\pi f_0 k\right)\cos\left(2\pi f_0 k\right) \\ \frac{-2\pi A}{\sigma_e^2}\sum_{k=0}^{N-1}k\sin\left(2\pi f_0 k\right)\cos\left(2\pi f_0 k\right) & \frac{4\pi^2 A^2}{\sigma_e^2}\sum_{k=0}^{N-1}k^2\cos^2\left(2\pi f_0 k\right) \end{bmatrix}$$

There for the CRLB for f_0 and A are as follows:

$$\begin{aligned} var(\hat{f}_0) &\geq \left[I(\boldsymbol{\theta})^{-1}\right]_{11} \\ &= \frac{\sigma_e^2}{4\pi^2 A^2}\frac{\sum_{k=0}^{N-1}\sin^2\left(2\pi f_0 k\right)}{\sum_{k=0}^{N-1}k^2\cos^2\left(2\pi f_0 k\right)\sum_{k=0}^{N-1}\sin^2\left(2\pi f_0 k\right)-(\sum_{k=0}^{N-1}k\sin\left(2\pi f_0 k\right)\cos\left(2\pi f_0 k\right))^2} \end{aligned}$$

$$
\begin{aligned}
var(\hat{A}) &\geq \left[I(\boldsymbol{\theta})^{-1}\right]_{22} \\
&= \frac{\sigma_e^2 \sum_{k=0}^{N-1} k^2 \cos^2(2\pi f_0 k)}{\sum_{k=0}^{N-1} k^2 \cos^2(2\pi f_0 k) \sum_{k=0}^{N-1} \sin^2(2\pi f_0 k) - (\sum_{k=0}^{N-1} k \sin(2\pi f_0 k) \cos(2\pi f_0 k))^2}
\end{aligned}
$$

Case 2 - if A is known and f_0 is unknown:

$$
I(\theta) = \frac{4\pi^2 A^2}{\sigma_e^2} \sum_{k=0}^{N-1} k^2 \cos^2(2\pi f_0 k)
$$

Therefore the CRLB for f_0 is:

$$
\begin{aligned}
var(\hat{f}_0) &\geq I(\boldsymbol{\theta})^{-1} \\
&= \frac{\sigma_e^2}{4\pi^2 A^2 \sum_{k=0}^{N-1} k^2 \cos^2(2\pi f_0 k)}
\end{aligned}
$$

Case 3 - if A is unknown and f_0 is known:

$$
I(\theta) = \frac{1}{\sigma_e^2} \sum_{k=0}^{N-1} \sin^2(2\pi f_0 k)
$$

Therefore the CRLB for A is:

$$
\begin{aligned}
var(\hat{A}) &\geq I(\boldsymbol{\theta})^{-1} \\
&= \frac{\sigma_e^2}{\sum_{k=0}^{N-1} \sin^2(2\pi f_0 k)}
\end{aligned}
$$

Comparing the CRLB for case 1 with case 2 and 3:

$$
\begin{aligned}
\frac{CRLB_1}{CRLB_2} &= \frac{\sum_{k=0}^{N-1} k^2 \cos^2(2\pi f_0 k) \sum_{k=0}^{N-1} \sin^2(2\pi f_0 k)}{\sum_{k=0}^{N-1} k^2 \cos^2(2\pi f_0 k) \sum_{k=0}^{N-1} \sin^2(2\pi f_0 k) - (\sum_{k=0}^{N-1} k \sin(2\pi f_0 k) \cos(2\pi f_0 k))^2} \\
&= \frac{1}{1 - \frac{(\sum_{k=0}^{N-1} k \sin(2\pi f_0 k)\cos(2\pi f_0 k))^2}{\sum_{k=0}^{N-1} k^2 \cos^2(2\pi f_0 k) \sum_{k=0}^{N-1} \sin^2(2\pi f_0 k)}} \\
&\geq 1
\end{aligned}
$$

$$
\begin{aligned}
\frac{CRLB_1}{CRLB_3} &= \frac{\sum_{k=0}^{N-1} k^2 \cos^2(2\pi f_0 k) \sum_{k=0}^{N-1} \sin^2(2\pi f_0 k)}{\sum_{k=0}^{N-1} k^2 \cos^2(2\pi f_0 k) \sum_{k=0}^{N-1} \sin^2(2\pi f_0 k) - (\sum_{k=0}^{N-1} k \sin(2\pi f_0 k) \cos(2\pi f_0 k))^2} \\
&= \frac{1}{1 - \frac{(\sum_{k=0}^{N-1} k \sin(2\pi f_0 k)\cos(2\pi f_0 k))^2}{\sum_{k=0}^{N-1} k^2 \cos^2(2\pi f_0 k) \sum_{k=0}^{N-1} \sin^2(2\pi f_0 k)}} \\
&\geq 1
\end{aligned}
$$

We obtain a tighter bound lower bound when more information is available. Which is why CRLB is higher when both A and f_0 are unknown than when only one of the variables is unknown.

(b)

Given that:

$$y[k] = e[k] \sim \mathcal{N}\left(0, \sigma^2\right)$$

Let $\hat{\boldsymbol{\theta}}(\mathbf{y}) = \mathbf{A}\mathbf{y}$. For $\hat{\theta}$ to be the BLUE:

$$\mathbf{A}E(\mathbf{y}) = \theta_0 = \sigma^2$$

But $E(\mathbf{y}) = 0$. Hence BLUE for σ^2 can't be obtained without transforming the data. By squaring the data points, we get the following transformed data:

$$
\begin{aligned}
x[k] &= y[k]^2 \\
\mathbf{x} &= [y[1], y[2], y[3], ...y[N]]^T \quad \text{where } \mathbf{x} \text{ is a } n \text{ x } 1 \text{ vector}
\end{aligned}
$$

Applying the condition for unbaisedness:

$$
\begin{aligned}
\mathbf{A}E(\mathbf{x}) &= \sigma^2 \\
\mathbf{A}[\sigma^2, \sigma^2, ...]^T &= \sigma^2 \\
\mathbf{A}\left(\sigma^2 \mathbf{J}_{n,1}\right) &= \sigma^2 \quad \text{where } \mathbf{J}_{n,1} \text{ is an } n \text{ x } 1 \text{ matrix of all ones} \\
\mathbf{A}\mathbf{J}_{n,1} &= 1
\end{aligned}
$$

Applying condition for minimum variance:

$$\min_{\hat{\boldsymbol{\theta}}} E\left(\left\|\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\right\|_2^2\right)$$

From condition for unbiasedness it is known that $\mathbf{A}\mathbf{J}_{n,1} = 1$. Hence the following optimization problem is solved:

$$\begin{aligned}
&\min_{\mathbf{A}} E\left(\left\|\left(\mathbf{A}\mathbf{x} - \sigma^2\right)\right\|_2^2\right) - \lambda(\mathbf{A}\mathbf{J}_{n,1} - 1) \\
&\min_{\mathbf{A}} E\left((\mathbf{A}\mathbf{x})^2 + \sigma^4 - 2\sigma^2\mathbf{A}\mathbf{x}\right) - \lambda(\mathbf{A}\mathbf{J}_{n,1} - 1) \\
&\min_{\mathbf{A}} \sigma^4\mathbf{A}^{\mathbf{T}}\mathbf{A} + \sigma^4 - 2\sigma^4\mathbf{A}\mathbf{J}_{n,1} - \lambda\mathbf{J}_{n,1}
\end{aligned}$$

Differentiating w.r.t A and equating to zero:

$$\begin{aligned}
2\sigma^4 A^T - 2\sigma^4\mathbf{J}_{n,1} - \lambda\mathbf{J}_{n,1} &= 0 \\
A &= \mathbf{J}_{1,n}\left(\frac{2\sigma^4 + \lambda}{2\sigma^4}\right)
\end{aligned}$$

Multiplying both sides by $\mathbf{J}_{n,1}$:

$$\begin{aligned}
\mathbf{A}\mathbf{J}_{n,1} &= \mathbf{J}_{1,n}\mathbf{J}_{1,n}\left(\frac{2\sigma^4 + \lambda}{2\sigma^4}\right) \\
\mathbf{A}\mathbf{J}_{n,1} &= N\left(\frac{2\sigma^4 + \lambda}{2\sigma^4}\right) \\
\mathbf{A}\mathbf{J}_{n,1} &= 1 \quad \text{(from the unbiasedness condition)} \\
\frac{2\sigma^4 + \lambda}{2\sigma^4} &= \frac{1}{N}
\end{aligned}$$

Therefore we get:

$$\hat{\mathrm{A}}^{\star} = \frac{1}{N}\mathbf{J}_{1,n}$$

Question 4

(a)

Given that $\bar{x} = 14578$, $\sigma = 1845$ and $N = 100$. We know that,

$$\begin{aligned}
z_l \leq \frac{x - \bar{x}}{\frac{\sigma}{\sqrt{N}}} \leq z_u \\
z_l\frac{\sigma}{\sqrt{N}} + \bar{x} \leq x \leq z_u\frac{\sigma}{\sqrt{N}} + \bar{x}
\end{aligned}$$

Given that,

$$z_l \frac{\sigma}{\sqrt{N}} + \bar{x} = 12000$$
$$z_u \frac{\sigma}{\sqrt{N}} + \bar{x} = 16000$$

On solving the above two equations for z_l and z_u, we get, $z_l = -13.9$ and $z_u = 7.707$. It is observed from z table that $P(z_l \leq z \leq z_u) = 0.9999$. Therefore, we can assert with 99.99% confidence that the molecular weight of polymer is in between 12000 and 16000.

(b)

Given that

$$\text{institution A: } \bar{x}_1 = 85.2, \sigma_1 = 6.8, N_1 = 60$$
$$\text{institution B: } \bar{x}_2 = 87.2, \sigma_2 = 8.8, N_2 = 55$$

To investigate whether on average students from institution A performs better than institution B, let us formulate a hypothesis test as follows:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \Longrightarrow \text{Average performance from both the institutions is same}$$

$$H_1 : \mu_1 - \mu_2 \neq 0 \quad \Longrightarrow \text{Average performance from both the institutions is not same}$$

The variable of interest is $y = \bar{x}_1 - \bar{x}_2$ and since the sample size is small we will use $t-$statistic.

Test statistic for y is given by:

$$
\begin{aligned}
t &= \frac{y - \mu_y}{\sigma_y} \\
\mu_y &= 0 \text{ for } H_0 \\
\sigma_y &= \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_1^2}{N_1}} \\
t &= \frac{85.2 - 87.2}{\sqrt{\frac{6.8^2}{60} + \frac{8.8^2}{55}}} \\
&= -1.355
\end{aligned}
$$

The statistic value lies between the upper and lower bound for significance level of 0.05, therefore, Null hypothesis that the average performance of both the institutions is same can not be rejected in the favour of alternate hypothesis.

Question 5

(a)

An estimator is $\hat{\theta}(\mathbf{y_N})$ is said to be a sufficient statistic if the joint PDF of $\mathbf{y_N}$:

$$f(\mathbf{y_N};\theta) = g(\hat{\theta}, \theta)h(\mathbf{y_N})$$

Let Ω be a sample space and $E \subseteq \Omega$ be an event. The indicator function (or indicator random variable) of the event E, denoted by 1_E, is a random variable defined as follows:

$$1_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases}$$

From question 1:

$$
\begin{aligned}
f(\mathbf{y_N};\theta) &= \prod_{i=1}^{N} f(y_i) \\
&= \prod_{i=1}^{N} e^{-(y_i-\theta)} \quad \text{such that all } y_i > \theta \\
&= \prod_{i=1}^{N} e^{-(y_i-\theta)} 1\{y_i > \theta\} \\
&= 1\{\min(y_N) > \theta\}\left(\prod_{i=1}^{N} e^{-(y_i-\theta)}\right) \\
&= 1\left\{\frac{T_N}{2} > \theta\right\}\left(\prod_{i=1}^{N} e^{-(y_i-\theta)}\right)
\end{aligned}
$$

We have:

$$
\begin{aligned}
g(\hat{\theta},\theta) &= 1\left\{\frac{T_N}{2} > \theta\right\} \\
&= 1\left\{\frac{\hat{\theta}}{2} > \theta\right\}
\end{aligned}
$$

$$
h(\mathbf{y_N}) = \left(\prod_{i=1}^{N} e^{-(y_i-\theta)}\right)
$$

Therefore, T_N is a sufficient statistic for θ.

Since it is known that T_N is a sufficient statistic and T'_N is an unbiased estimator of θ, the Rao-Blackwell theorem can be applied to find the MVUE.

$$
\begin{aligned}
\hat{\theta_{\text{MVUE}}} &= E[T'_N|T_N] \\
&= E\left[\min(\mathbf{y_N}) - \frac{1}{N}\middle| 2\min(\mathbf{y_N})\right] \\
&= \min(\mathbf{y_N}) - \frac{1}{N}
\end{aligned}
$$

(b)

MATLAB code for comparing efficiencies

```
N = 1000;
```

```
R = 500;
theta = 3;
y = theta + exprnd(1,[N, R]);

theta_MVUE = min(y)-1/N;
theta_1 = mean(y) - 1;
theta_2 = median(y) - log(2);

efficiency_1 = var(theta_MVUE)/var(theta_1);
efficiency_2 = var(theta_MVUE)/var(theta_2);
```

After comparing the MVUE obtained in 5 a) with θ_1 and θ_2, we get the efficiencies (0.0013 and 0.0012 respectively) to be less than 1, as expected.