Assignment 1: CS 215

Name:Challa Siva Ramya
Roll No:24B0941

Name:Gunda Joshmitha
Roll No:24B1098

Question 1

The following shows the comparison of moving filters (median, mean, and quartile)
The relative mean squared errors for different filters are:
for corrupted data with $f = 30\%$

- Median relative MSE: 33.527
- Mean relative MSE: 57.816
- Quartile relative MSE: 0.014

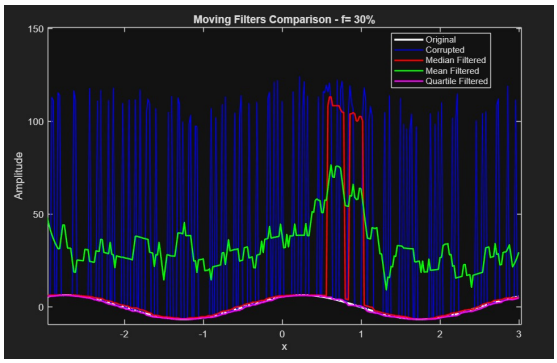

Figure 1: f=30%

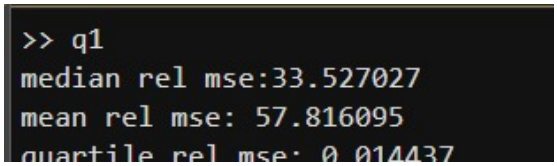

Figure 2: f=30%

for corrupted data with $f = 60\%$

- Median relative MSE: 442.085
- Mean relative MSE: 210.825
- Quartile relative MSE: 25.107

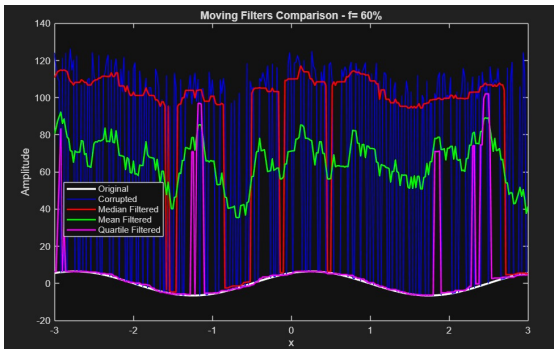

Figure 3: f=60%

```
median rel mse:442.084648
mean rel mse: 210.825062
quartile rel mse: 25.106715
```

Figure 4: f=60%

The quartile filter gives much lower relative MSE than both the median and mean because it considers only the first quarter portion of the data,which is not corrupted(until c is more than 75%) and ignores extreme values at the end(which are corrupted), effectively removing most of the corruption.
The median performs somewhat better when corruption is moderate (30%),but when more than half the values are corrupted, it fails,leading to higher MSE.When more than half of the values get corrupted the median of most of the $N(i)$ also gets corrupted leading to higher MSE.
The mean is sensitive to all values, so its MSE increases as the level of corruption rises.
So the quartile gives the least mean squared error.

2

Computing Mean

Given OldMean,NewMean,NewDataValue,

$$OldMean = \frac{(\sum_{i=1}^{n} x_i)}{n}$$

where x_i is element of the array A.

$$OldMean * n = \sum_{i=1}^{n} x_i \tag{1}$$

$$NewMean = \frac{(\sum_{i=1}^{n+1} x_i)}{n+1}$$

$x_{n+1} = NewDataValue$

$$NewMean = \frac{(\sum_{i=1}^{n} x_i) + NewDataValue}{n+1}$$

$$NewMean = \frac{OldMean \cdot n + NewDataValue}{n+1}$$

Computing Median

Since the array A is already sorted,case analysis when n is even and when n is odd is done.

When n is even,if the NewDataValue is less than $A[n/2]$,then it gets inserted to the left of $A[n/2]$ shifting the values greater than it to the right and then $A[n/2]$ becomes the mid value of the new array of size $n+1$.

Similarly, if n is greater than $A[\frac{n}{2}+1]$,the NewDataValue gets inserted to the right of $A[\frac{n}{2}+1]$ and the mid value of the array of size $n+1$ becomes $A[\frac{n}{2}+1]$. If the NewDataValueis less than $A[\frac{n}{2}+1]$ and greater than $A[\frac{n}{2}]$ than the NewDataValuebecomes the midvalue of the new array.

When n is odd, the median of the original array is the middle element $A\left(\frac{n+1}{2}\right)$. After inserting the new value, array size becomes even, so the new median will be the average of two middle elements.
If the new data value is less than $A\left(\frac{n-1}{2}\right)$, then it gets inserted to the left of $A\left(\frac{n-1}{2}\right)$, and the two middle values of the new array become $A\left(\frac{n-1}{2}\right)$ and $A\left(\frac{n+1}{2}\right)$. Hence, the new median is their average.

If NewDataValue is greater than $A\left(\frac{n+3}{2}\right)$, it gets inserted to the right of $A\left(\frac{n+3}{2}\right)$, making the two middle values $A\left(\frac{n+1}{2}\right)$ and $A\left(\frac{n+3}{2}\right)$. The new median is their average.

Otherwise, if NewDataValue lies between $A\left(\frac{n-1}{2}\right)$ and $A\left(\frac{n+3}{2}\right)$, then the new median is the average of $A\left(\frac{n+1}{2}\right)$ and the new value itself.

Computing Standard Deviation

Given OldMean,OldVariance,NewMean and NewDataValue.

$$OldVariance = \frac{\sum_{i=1}^{n}(x_i - OldMean)^2}{n-1}$$

$$OldVariance = \frac{\sum_{i=1}^{n}(x_i^2 + OldMean^2 - 2 \cdot x_i \cdot OldMean)}{n-1}$$

$$OldVariance = \frac{(\sum_{i=1}^{n} x_i^2) + (\sum_{i=1}^{n} OldMean^2) - (\sum_{i=1}^{n} 2 \cdot x_i \cdot OldMean)}{n-1}$$

$$OldVariance \cdot (n-1) = (\sum_{i=1}^{n} x_i^2) + n \cdot (OldMean)^2 - 2 \cdot OldMean \sum_{i=1}^{n} x_i$$

$$OldVariance \cdot (n-1) = (\sum_{i=1}^{n} x_i^2) + n \cdot (OldMean)^2 - 2 \cdot n \cdot OldMean \cdot OldMean$$

$$OldVariance \cdot (n-1) = (\sum_{i=1}^{n} x_i^2) - n \cdot (OldMean)^2$$

$$(\sum_{i=1}^{n} x_i^2) = OldVariance \cdot (n-1) + n \cdot (OldMean)^2$$

$$NewVariance = \frac{\sum_{i=1}^{n+1}(x_i - NewMean)^2}{n}$$

$$NewVariance = \frac{(\sum_{i=1}^{n+1} x_i^2) - (n+1) \cdot (NewMean)^2}{n}$$

$$NewVariance = \frac{(\sum_{i=1}^{n} x_i^2) + (NewDataValue)^2 - (n+1) \cdot (NewMean)^2}{n}$$

$$NewVariance = \frac{OldVariance \cdot (n-1) + n \cdot (OldMean)^2 + NewDataValue^2 - (n+1) \cdot (NewMean)^2}{n}$$

$$NewStd = \sqrt{(\frac{OldVariance \cdot (n-1) + n \cdot (OldMean)^2 + NewDataValue^2 - (n+1) \cdot (NewMean)^2}{n})}$$

To add the NewDataValue to the histogram,we loop through the bins in the histogram.If the NewDataValue is greater than the the lower bound and less than the upper bound of a particular bin,the frequency of that particular bin increases by 1.If the value is less than the lower bound of every bin or greater than the upper bound of every bin ,then a new bin is created with frequency 1. To run the code,initailize an array and give required inputs to the functions and run the code.

3

Given that $P(A) \geq 1 - q_1 and P(B) \geq 1 - q_2$..
To prove $P(A, B) \geq 1 - (q_1 + q_2)$.
From Bonferroni's inequality,$P(A, B) \geq P(A) + P(B) - 1$

$$P(A, B) \geq 1 - q_1 + 1 - q_2 - 1$$

$$P(A, B) \geq 1 - (q_1 + q_2)$$

4

Let $P(R)$ be the probability that a bus is red in the town and $P(B)$ be the probability that a bus is blue in the town.
Then $P(R) = \frac{1}{100}$ and $P(B) = \frac{99}{100}$.
Also let $P(R_x)$ be the probability that the person XYZ has seen the bus to be red.
Since XYZ sees red objects as red 99% of time and blue objects as red 2%,

$$P(\frac{R_x}{R}) = \frac{99}{100}$$
$$P(\frac{R_x}{B}) = \frac{2}{100}$$

The probabilty that the bus was really red when XYZ saw it red is $P(\frac{R}{R_x})$

$$P(\frac{R}{R_x}) = \frac{P(R \cap R_x)}{P(R_x)} = \frac{P(R \cap R_x)}{P(R)P(\frac{R_x}{R}) + P(B)P(\frac{R_x}{B})}$$

$$P(\frac{R}{R_x}) = \frac{P(R)P(\frac{R_x}{R})}{P(R)P(\frac{R_x}{R}) + P(B)P(\frac{R_x}{B})}$$

$$P(\frac{R}{R_x}) = \frac{\frac{1}{100}\frac{99}{100}}{\frac{1}{100}\frac{99}{100} + \frac{99}{100}\frac{2}{100}}$$

$$P(\frac{R}{R_x}) = \frac{1}{3}$$

Question 5

(a)

Consider a village with 100 residents. Two candidates, A and B, are participating in an election. An exit poll is conducted on 3 residents with replacement.

The probability of a resident favoring A is: P(A) = 0.95, and favouring B is P(B) = 0.05

We want to calculate the probability that A is reported as winning according to the exit poll and A will win if:

- All 3 sampled residents vote for A, or
- 2 residents vote for A and 1 votes for B

Case I: 3 votes to A

$$P(A, A, A) = (0.95)(0.95)(0.95) = (0.95)^3$$

Case II: 2 votes to A, 1 vote to B

Here there are 3 possible arrangements:

$$\begin{aligned}\Pr(A, A, B) &= 0.95 \cdot 0.95 \cdot 0.05 \\ \Pr(A, B, A) &= (0.95) \cdot (0.05) \cdot (0.95) \\ \Pr(B, A, A) &= (0.05) \cdot (0.95) \cdot (0.95)\end{aligned}$$

Total probability:

$$P(A \text{ wins}) = 3 \cdot (0.95)^2(0.05)$$

Final Probability

$$P(A \text{ according to exit poll}) = (0.95)^3 + 3 \cdot (0.95)^2(0.05) = 0.99725$$

Thus, the probability that A is reported as the winner by the exit poll is approximately 0.99725.

(b)

if the village has 10,000 residents, and still P(A) is 0.95 and P(B) is 0.05 the probability that A is reported as the winner by the exit poll doesn't change as we are choosing random people with replacement and the probability of choosing a person voting for A and that of voting for B doesn't change during th 3 rounds.

Question 6

A village has m voters and the probability a person votes for A is $\frac{k}{m}$, and the probability a person votes for B is $1 - \frac{k}{m}$.

An exit poll asks randomly n voters with replacement (subset S). Define

$$X_i = \begin{cases} 1, & \text{if the } i\text{th voter voted for A} \\ 0, & \text{if the } i\text{th voter voted for B.} \end{cases}$$

Then,

$$q(S) = \frac{1}{n} \sum_{i \in S} x_i.$$

Since we are selecting people with replacement, the number of subsets S is m^n.

(a) Prove that $\sum_S \frac{q(S)}{m^n} = p$

We want the average of $q(S)$ over all subsets S of size n:

$$\sum_S \frac{q(S)}{m^n} = \frac{1}{m^n} \sum_S q(S).$$

$$= \frac{1}{m^n} \sum_S \left(\sum_{i \in I_S} \frac{x_i}{n} \right) = \frac{1}{m^n \cdot n} \sum_S \left(\sum_{i \in I_S} x_i \right).$$

Taking expectation:

$$= \frac{1}{m^n \cdot n} \sum_S \left(\sum_{i \in I_S} E(x_i) \right).$$

Since $E(x_i) = p$ from definition,

$$= \frac{1}{m^n \cdot n} \sum_S (n \cdot p) = \frac{1}{m^n \cdot n} (p \cdot n \cdot m^n) = p.$$

$$\therefore \quad \sum_S \frac{q(S)}{m^n} = p$$

(b) Prove that $\sum_S \frac{q^2(S)}{m^n} = \frac{p}{n} + \frac{p^2(n-1)}{n}$

$$\sum_S \frac{q^2(S)}{m^n} = \frac{1}{m^n} \sum_S q^2(S) = \frac{1}{m^n} \sum_S \left(\frac{1}{n} \sum_{i \in I_S} x_i \right)^2.$$

$$= \frac{1}{m^n \cdot n^2} \sum_S \left(\sum_{j=1}^{n} x_{ij}^2 + \sum_{j \neq k} x_{ij} x_{ik} \right).$$

Since $x^2 = x$ (as $x \in \{0, 1\}$),

$$= \frac{1}{m^n \cdot n^2} \sum_S \left(\sum_{j=1}^{n} p + \sum_{j \neq k} E(x_{ij} x_{ik}) \right).$$

Independence gives $E(x_{ij}x_{ik}) = E(x_{ij})E(x_{ik}) = p^2$.

$$= \frac{1}{m^n \cdot n^2} \sum_S \big(np + n(n-1)p^2\big).$$

Since there are m^n subsets S,

$$= \frac{1}{m^n \cdot n^2} \big(np + n(n-1)p^2\big) m^n.$$

$$= \frac{p}{n} + \frac{p^2(n-1)}{n}.$$

$$\therefore \quad \sum_S \frac{q^2(S)}{m^n} = \frac{p}{n} + \frac{p^2(n-1)}{n}.$$

(c) Prove that $\sum_S \frac{(q(S)-p)^2}{m^n} = \frac{p(1-p)}{n}$

$$\sum_S \frac{(q(S)-p)^2}{m^n} = \frac{1}{m^n} \sum_S (q(S)-p)^2.$$

$$= \frac{1}{m^n} \sum_S \big(q(S)^2 + p^2 - 2pq(S)\big).$$

Using results from (a) and (b): - From (b), $\sum_S q(S)^2 = m^n \left(\frac{p}{n} + \frac{p^2(n-1)}{n}\right)$. - From (a), $\sum_S q(S) = m^n p$.

Thus,

$$\sum_S \frac{(q(S)-p)^2}{m^n} = \frac{1}{m^n} \left(m^n \left(\frac{p}{n} + \frac{p^2(n-1)}{n} \right) + m^n p^2 - 2m^n p^2 \right).$$

$$= \frac{p}{n} + \frac{p^2(n-1)}{n} + p^2 - 2p^2 = \frac{p(1-p)}{n}.$$

$$\therefore \quad \sum_S \frac{(q(S)-p)^2}{m^n} = \frac{p(1-p)}{n}.$$

(d)

Chebyshev's inequality states that

$$S_k = \{x_i : |x_i - \bar{x}| \geq k\sigma\}, \qquad \frac{|S_k|}{N} \leq \frac{1}{k^2}.$$

From our question: $\bar{x} = p$ and $x_i = q(S)$

σ is the standard deviation which is equal to:

$$\sigma = \sqrt{\frac{\sum_S (q(S) - p)^2}{m^n - 1}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\frac{1}{k^2} = \frac{1}{\delta^2} \cdot \frac{p(1-p)}{n}$$

Larger samples make the pole more accurate so if the sample size is large enough then the exit poll will almost always reflect the real preference of the people in the village