Outlier Detection

Madhur Bansal (210572)

2024-03-15

The objective is so generate simulated data along with some outliers. Then we have to propose a method detect the outliers and estimate the proportion of outliers in the data.

Simulating Data:

We generate a random $X_{(t)}$ from $L^2_{[0,1]}$ space using the following:

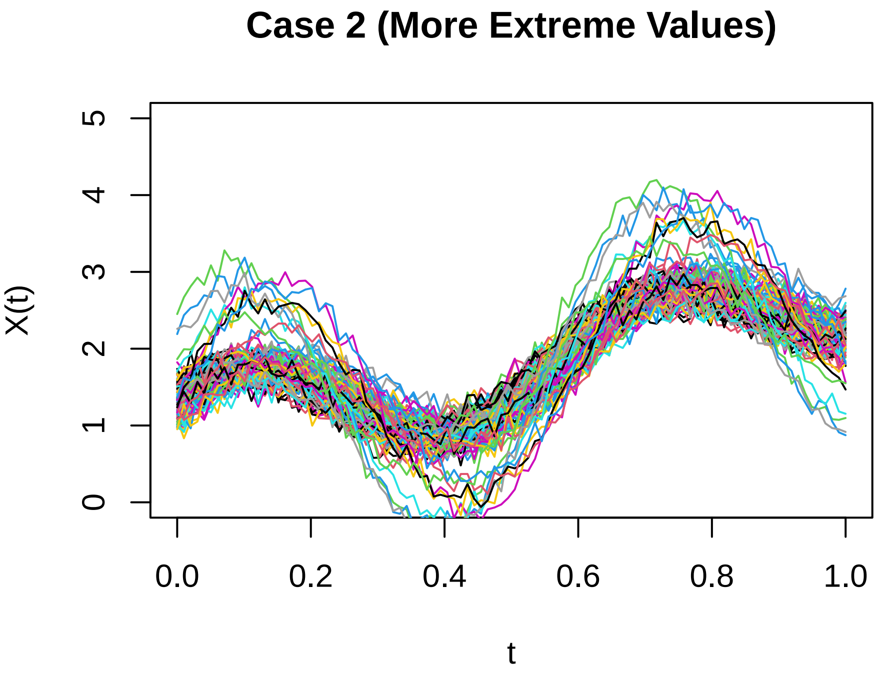
$$X_{(t)} = 0.5(c_1 sin(10.c_1.t) + c_2 sin(10.c_2.t) + c_3 sin(10.c_3.t)) + c_4$$

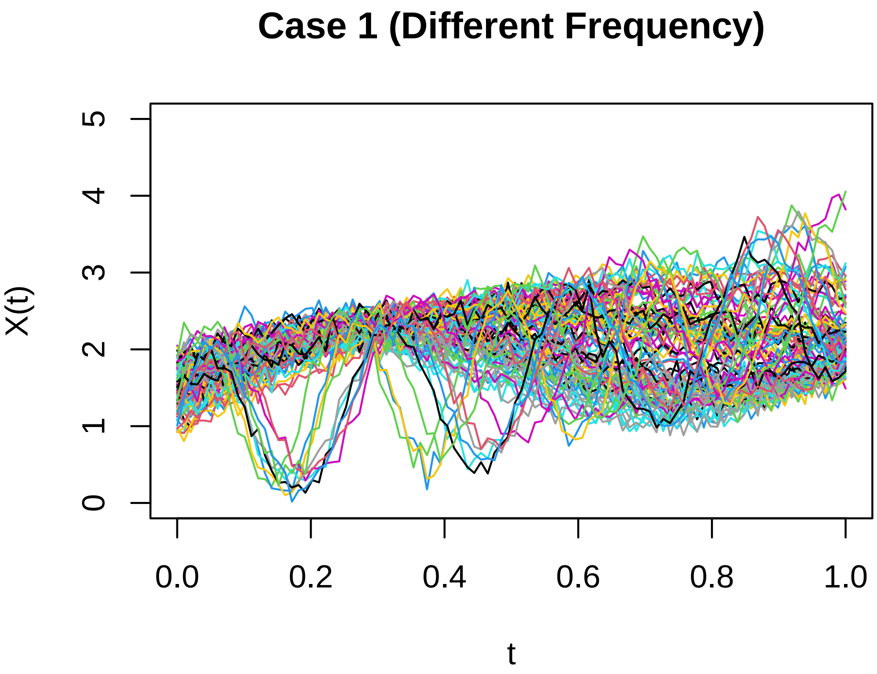
where c_i are random constants from $Unif(0,1)$.

In this analysis, I have generated a total of 100 samples, with exactly 10 of them as outliers. I explore two different cases: Frequency Difference and Scale Difference

- In case 1, the outliers have similar shape to the data but attain more extreme values compared to the data. To introduce this difference, I multiply a random generated value uniform distribution $Unif(1,2)$ to the original function.
- In case 2, the outliers have a different frequency compared to the rest of the data. To achieve this, I multiply the frequency of sinusoidal function with 10*c where c ~ runif(1, 1.5)

Here are the plots of the datasets for the two cases:


Case 2 (More Extreme Values)


Case 1 (Different Frequency)

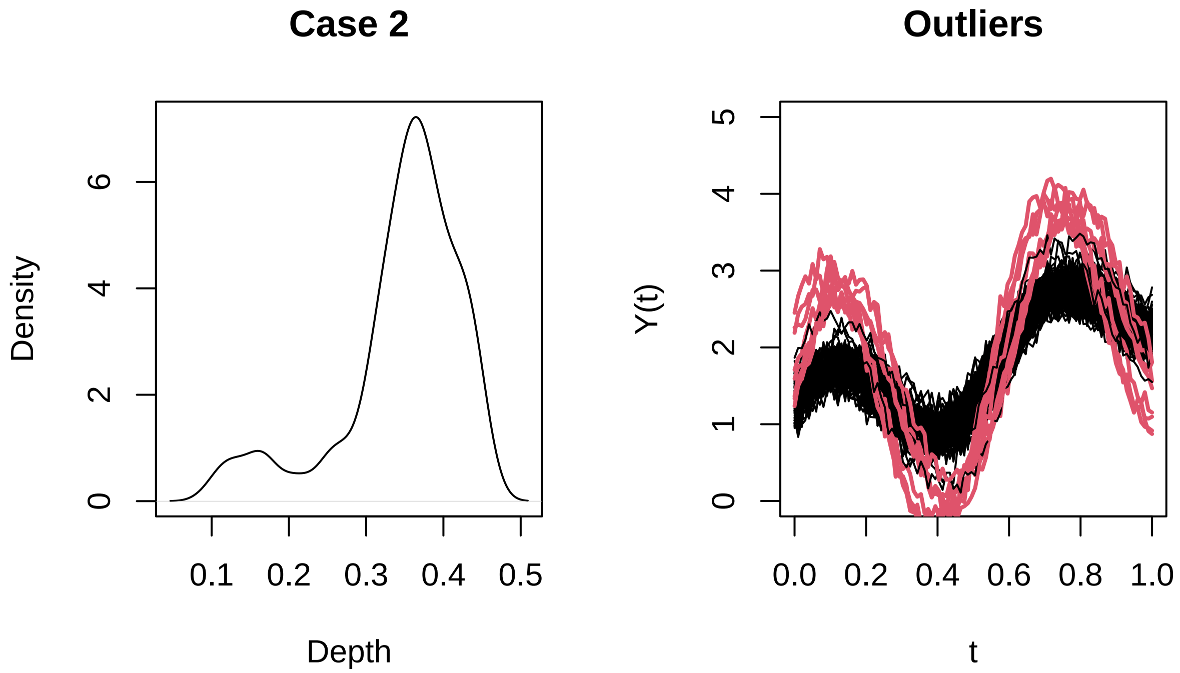
Proposed method:

To detect the outliers in the data, I calculate the depth of the $X_{(t)}$'s in the dataset. Subsequently, then assuming the depth comes from a normal distribution, a cut-off value based on the 95% region for $N(\mu, \sigma^2)$ where $\mu = mean(depth)$ and $\sigma^2 = Var(depth)$. Any samples that lie outside the 95% region, are labelled as outliers.

Note: As a measure of depth, we are using Modified Band-Depth (MBD) presented by Sara Lopez-Pintado and Juan Romo. This is the link to the original paper: https://www.jstor.org/stable/40592217

Case 1

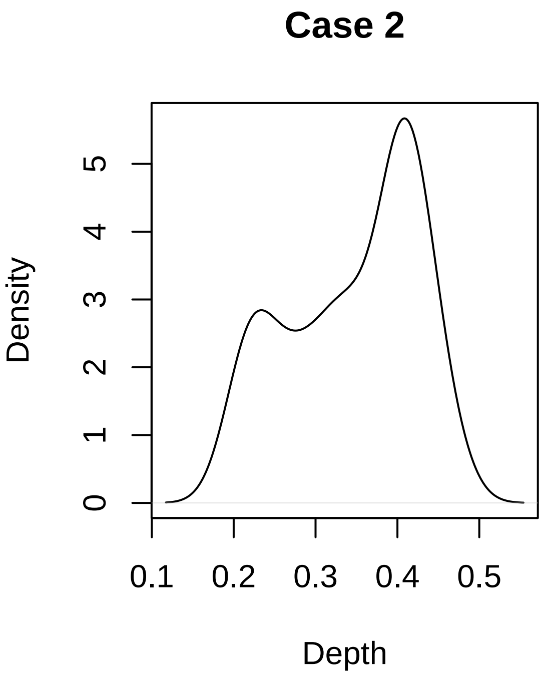
The method works well in this case. It is able to identify most of the outliers correctly.

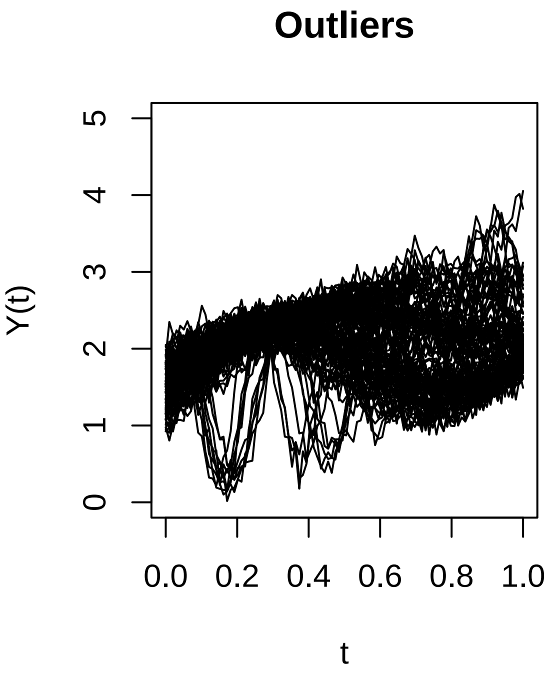


```
##            predicted_outlier
## is_outlier  0  1
##          0 90  0
##          1  2  8
```

Case 2

In this case, the method was unable to accurately predict the outliers. Due to higher frequency of the random function, the function spends most of the time inside the band. Consequently, these functions, which should be identified as outliers, are instead ranked with a greater depth than non-outliers





```
##            predicted_outlier
## is_outlier  0
##          0 90
##          1 10
```

Conclusion

The method was effective in detecting most outliers with more extreme values. However, it struggled to accurately identify outliers that had higher frequency but were on the same scale. To improve the performance, additional features that account for frequency variations can be included while detecting outliers.