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:

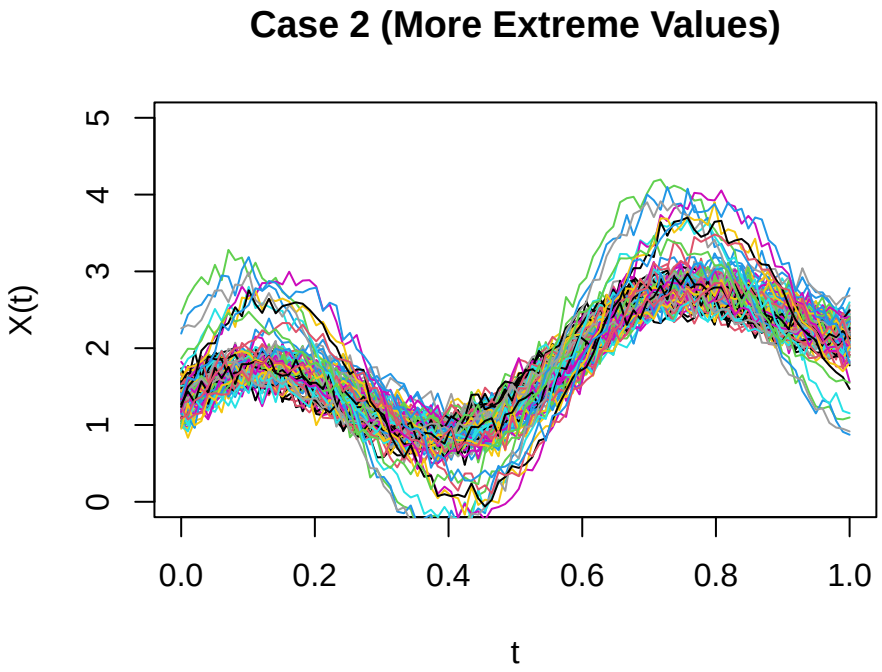
$$X_{(t)} = 0.5(c_1 sin(10.c_1.t) + c_2 sin(10.c_2.t) + c_3 sin(10.c_3.t)) + c_4$$

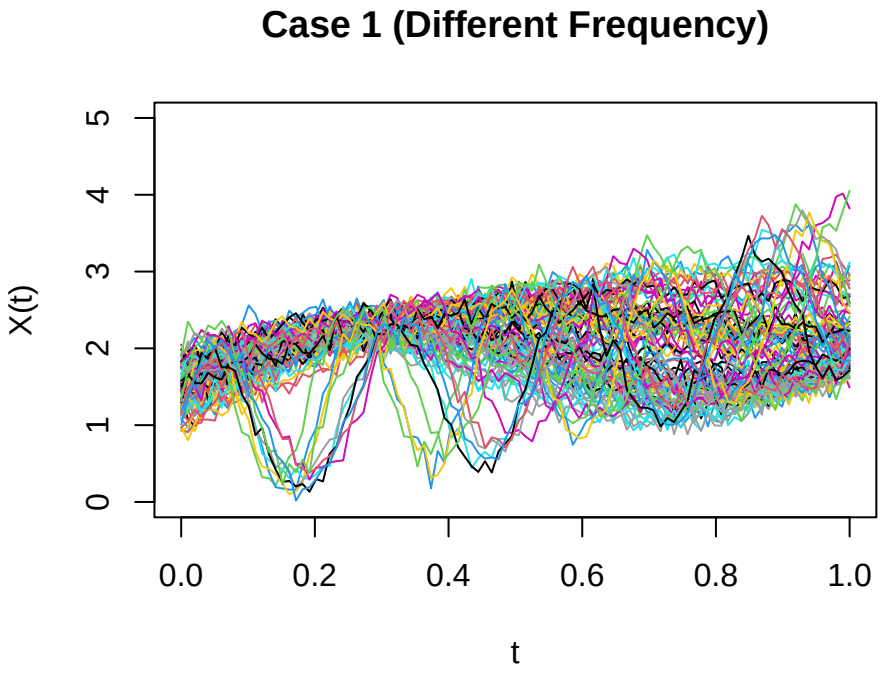
where c_i are random constants from $Unif(0, 1)$.

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)
X(t)
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
t


Case 1 (Different Frequency)
X(t)
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
t

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.

Then, I apply the following transformation to the data. This centers the data and allows to identify the outliers having different trend from rest of the data.

$$T(X_i(t)) = X_i(t) - \frac{1}{T}\sum_{j=1}^{T} X_i(j) \forall i = 1(1)n$$

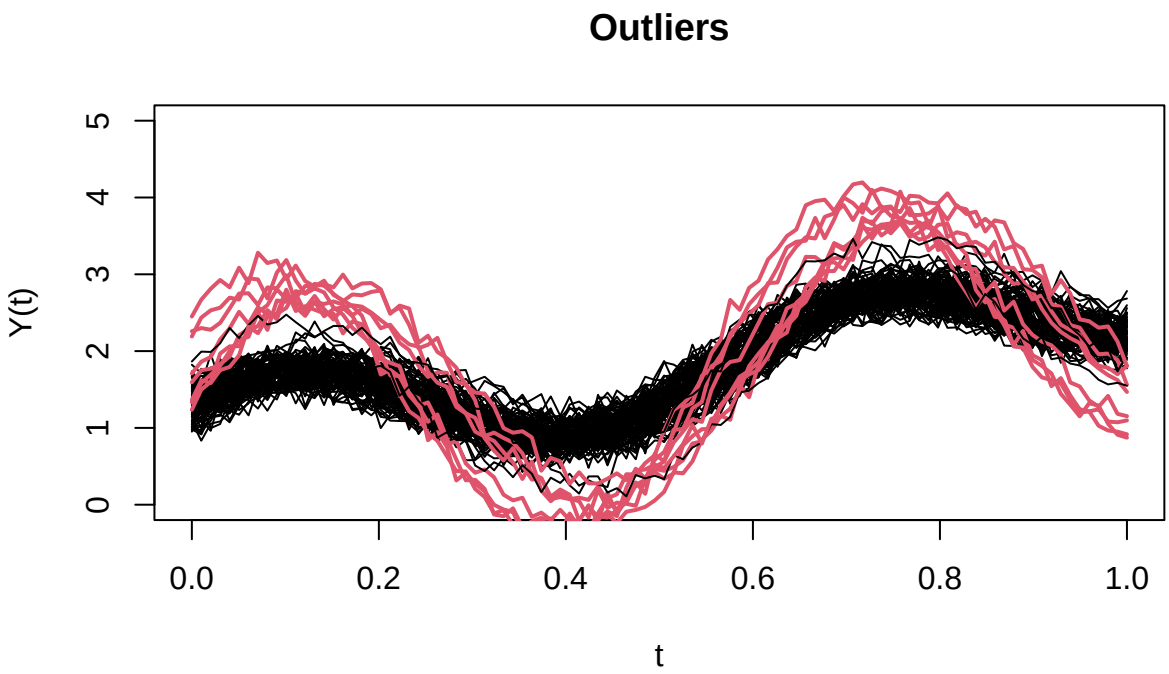
After applying the transformation, I again find the outliers using 95% region (decribed above). Then, finally I apply a transformation to normalize the centred data. This helps in detecting the outliers which may have different shape than the data.

$$T(X_i(t)) = \frac{X_i(t)}{||X_i(t)||_{L_2}} \forall i = 1(1)n$$

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. I also referenced this blog: https://www.lancaster.ac.uk/stor-i-student-sites/harini-jayaraman/anomaly-detection-in-functional-data in order to improve my implementation.

Case 1

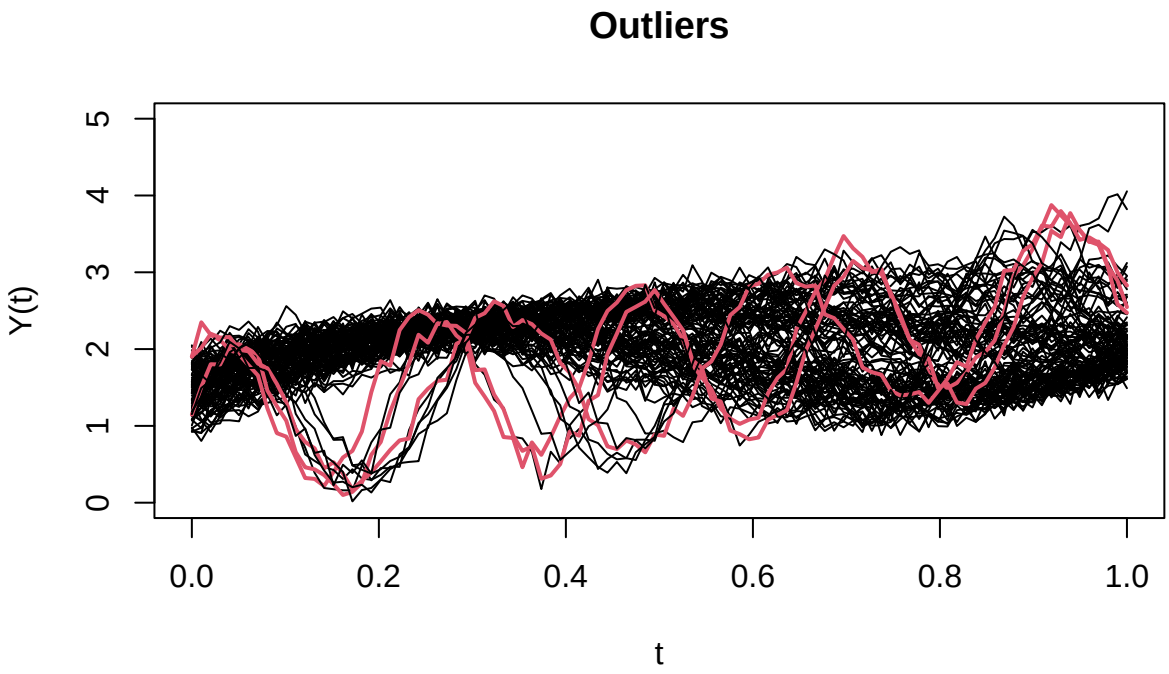
The method works well in this case. It is able to identify most of the outliers correctly.



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

Case 2

In this case also, the method was able to identify most of the outliers.




```
##            predicted_outlier
## is_outlier  0  1
##          0 90  0
##          1  7  3
```

Conclusion

The method successfully identified most of the outliers in two scenarios: functions with more extreme values and those with higher frequency patterns. This shows that the approach is effective in detecting different types of anomalies in the data.