Data Stream Anomaly Detection using Kalman Filter and Exponential Smoothing
By Vignesh Sundaram

1. Introduction:

Identifying anomalies is essential in many fields like finance, healthcare, manufacturing, and cybersecurity. Identifying abnormalities or variations from anticipated patterns is beneficial for promptly recognizing problems, preventing fraud, and maintaining system dependability. Exponential Smoothing and Kalman Filters are widely used methods for real-time anomaly detection because of their ability to forecast and reduce noise effectively, despite the presence of many other available techniques.

This report compares Exponential Smoothing and Kalman Filters for anomaly detection. Both approaches were used on a synthetic dataset that contained 1,000 data points with anomalies deliberately added. Each algorithm's performance was assessed using key metrics such as Precision, Recall, and F1 Score.

2. Methodologies:

In this project, I have discussed 2 main methodologies used in anomaly detection:

2.1. Exponential Smoothing:

Exponential Smoothing is a method for forecasting time series data that uses a weighted average of previous data points to estimate future outcomes. The technique current data points, allowing it to be responsive to changes in the trend and seasonality

2.1.1. Overview of the algorithm:

This implementation uses Holt-Winters Exponential Smoothing variant which is used for level, trend and seasonality components. These components are updated by the algorithm iteratively as and when we receive new data:

- Level (l_t) : Represents smoothed value of the series as time t.
- Trend (b_t) : Captures direction and magnitude of the trend.
- Seasonality (s_t) : Captures periodic fluctuations in the data

The forecast for the next point is derived as:

$$\hat{y}_{t+1}=l_t+b_t+s_{t-m}$$

Here m is the seasonality period.

2.1.2. Anomaly Detection in Exponential Smoothing:

We calculate residuals here to identify the anomalies. Residuals are the differences between actual values and forecasted values. A data point is an anomaly if the residual exceeds a threshold. Threshold is set as a multiple of rolling standard deviation of residuals.

2.2. Kalman Filter:

Kalman Filters are recursive algorithms used to determine state of a dynamic system from a series of incomplete and noisy measurements. Some areas where Kalman filters are used are in robotics, navigation and finance.

2.2.1. Overview of the algorithm:

With every new measurement, the Kalman Filter Anomaly Detector changes its estimation of the present state. The procedure entails:

- Prediction Step: Estimating the current state and its uncertainty which depends on the previous state.

- Update Step: Refining our calculations using the new measurement, balancing between the prediction and the measurement based on their uncertainties.

The residual is then computed as the difference between the actual measurement and the predicted estimate. This residual serves is used for anomaly detection.

2.2.2. Anomaly Detection in Kalman Filters:

Anomalies here are also detected using residuals. A point is classified as an anomaly if its residual exceeds a threshold, which is multiple of standard deviation of residuals.

3. Results:

We have used a simulator with 1000 data points with 5% anomaly injection rate. We have the following performance metrics:

3.1. Exponential Smoothing:

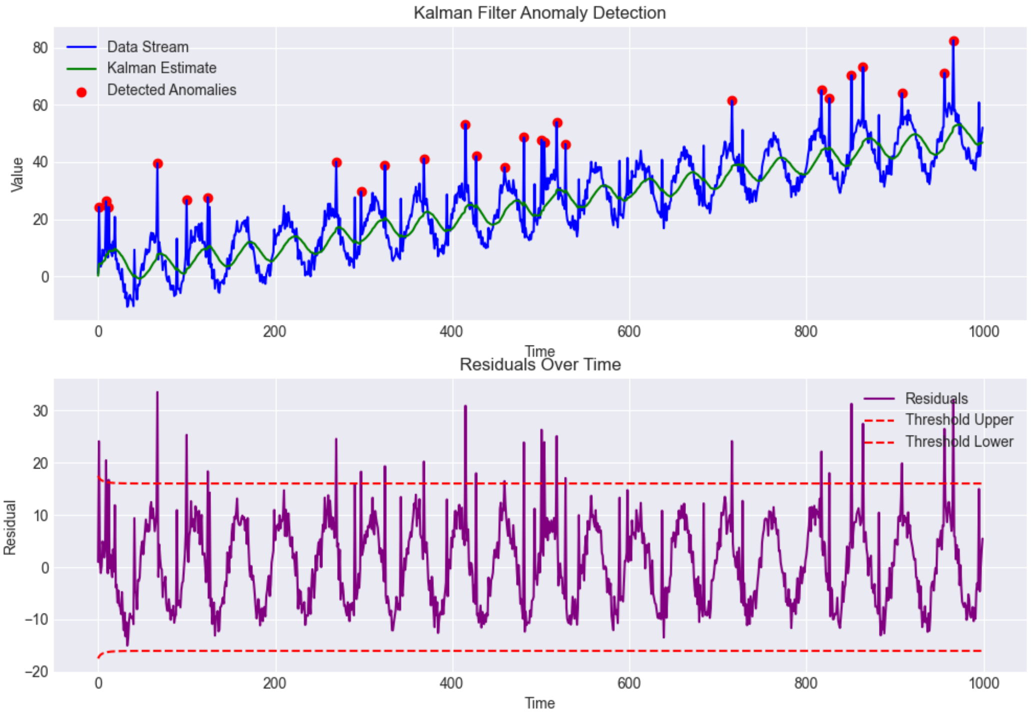
Total Data Points: 1000
True Positives (TP): 30
False Positives (FP): 1
True Negatives (TN): 941
False Negatives (FN): 28

Precision: 0.97
Recall: 0.52
F1 Score: 0.67

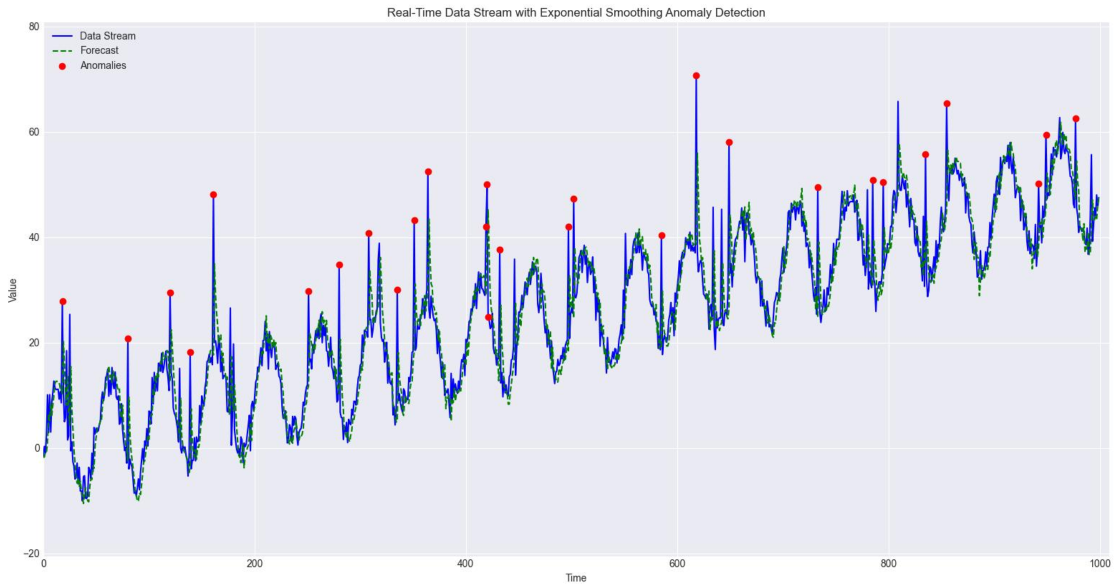
3.2. Kalman Filters:

Total Data Points: 1000
True Positives (TP): 25
False Positives (FP): 1
True Negatives (TN): 958
False Negatives (FN): 16
Precision: 0.96
Recall: 0.61
F1 Score: 0.75

Kalman Filter simulation:



Exponential smoothing simulation:



4. Discussion:

4.1. Algorithm Effectiveness:

4.1.1. Exponential smoothing:

- **Precision (0.97)**: Indicates that 97% of the detected anomalies were true anomalies. This high precision suggests that the algorithm correctly identifies anomalies in most of the cases.
- **Recall (0.52)**: Indicates that 52% of the actual anomalies were detected. The recall indicates that nearly half of the anomalies were missed.
- **F1 Score (0.67)**: Balances precision and recall, providing an overall measure of the algorithm's performance.

Effectiveness: This method is best in minimizing false positives and it can be used where false alarms are costly and disruptive. However, looking at the recall it seems that it may not recall all critical anomalies.

4.1.2. Kalman Filter:

- **Precision (0.96)**: Slightly lower than Exponential Smoothing but still very good, which indicates that most of the detected anomalies are true anomalies.

- **Recall (0.61)**: Higher than Exponential Smoothing, indicating that 61% of actual anomalies were identified.
- **F1 Score (0.75)**: Demonstrates a better balance between precision and recall compared to Exponential Smoothing.

Effectiveness: The Kalman Filter offers a better recall while maintaining high precision as compared to Exponential smoothing. This balance makes it more effective in scenarios where detecting a higher proportion of anomalies is crucial.

4.2. Correctness:

The high precision values confirms that both the detection mechanisms are accurately distinguishing between normal and anomalous data points based on the defined thresholds.

4.3. Comparison:

Metric	Exponential Smoothing	**Kalman Filter**
True Positives (TP)	30	25
False Positives (FP)	1	1
True Negatives (TN)	941	958
False Negatives (FN)	28	16
Precision	0.97	0.96
Recall	0.52	0.61
F1 Score	0.67	0.75

4.4. Reasons for performance difference:

The observed differences in performance can be due to the inherent characteristics of the two algorithms:

Exponential Smoothing:

- Simplicity: It is a straightforward forecasting method and it may not capture complex dynamics efficiently.
- Responsiveness: While it responds to trends and seasonality, it might lag in adapting to sudden changes

Kalman Filter:

- Dynamic Estimation: Continuously updates its estimates based on new measurements, allowing for better adaptation to changes.

- Noise Handling: Effectively reduces the impact of noise, improving anomaly detection accuracy.

5. Conclusion:

Both Kalman filters and exponential smoothing are useful methods for identifying anomalies, but they have different advantages and disadvantages.

Exponential smoothing is perfect for contexts where the cost of false alarms is significant since it provides great precision with few false positives. Its modest recall, however, suggests a tendency to overlook a sizable number of anomalies, which can be problematic in crucial applications.

Because Kalman Filters reduce false negatives without significantly raising false positives, they get a higher F1 Score by better balancing recall and precision. Because of this, even if there is a small loss of precision, the Kalman Filter is a more reliable option when detecting a higher percentage of abnormalities is crucial.

Appendix:

There are 2 python files: anomalyDetectionKalmanFilter.py and anomalyDetectionExponentialSmoothing.py, run these files separately to view the simulation and results.

Before running these files, run the requirements.txt using command “pip install -r requirements.txt”